The Fragility of Scientific Self-Correction

Felipe Romero
Philosophy-Neuroscience-Psychology Program
Washington University in St. Louis

Department of Logic and Philosophy of Science
University of California, Irvine
March 2015

Science as Self-Corrective

Self-Corrective Thesis (SCT): **Scientific method** will refute false theories and find closer approximations to the true theories in the **long run**.

> **C.S. Peirce**: "[Quantitative induction is] a method which, steadily persisted in, must lead to true knowledge in the long run of cases of its application", 1901

Reichenbach: Any legitimate scientific method should be reducible to quantitative induction, 1935.

Rescher, Levi, Hacking: Frequentist Statistics instantiates Peirce's Self-Corrective Thesis.

Mayo: "Peirce's self-corrective thesis provides a basis for justifying frequentist statistical methods in science", 2005

But is science really self-correcting?

"Researchers make unacknowledged decisions that may increase false positives".

Psychology Today, Nov 2011.

"Nobel laureate challenges psychologists to clean up their act".

Nature, Oct 2012.

"Scientists like to think of science as self-correcting. To an alarming degree, it is not".

The Economist, Oct 2013.

"Important findings haven't been replicated, and science may have to change its ways".

The Slate, Jul 2014.

Self-Corrective Thesis (SCT): **Scientific method** will refute false theories and find closer approximations to the true theories in the **long run**

Plan:

1. SCT*: SCT in terms of frequentist statistics.

2. Scientific Utopia: SCT* depends on idealized assumptions about the social structure of science.

3. Focus Shift: From methodology to social epistemology.

Bargh et. al, 1996 experiment

NEUTRAL WORDS

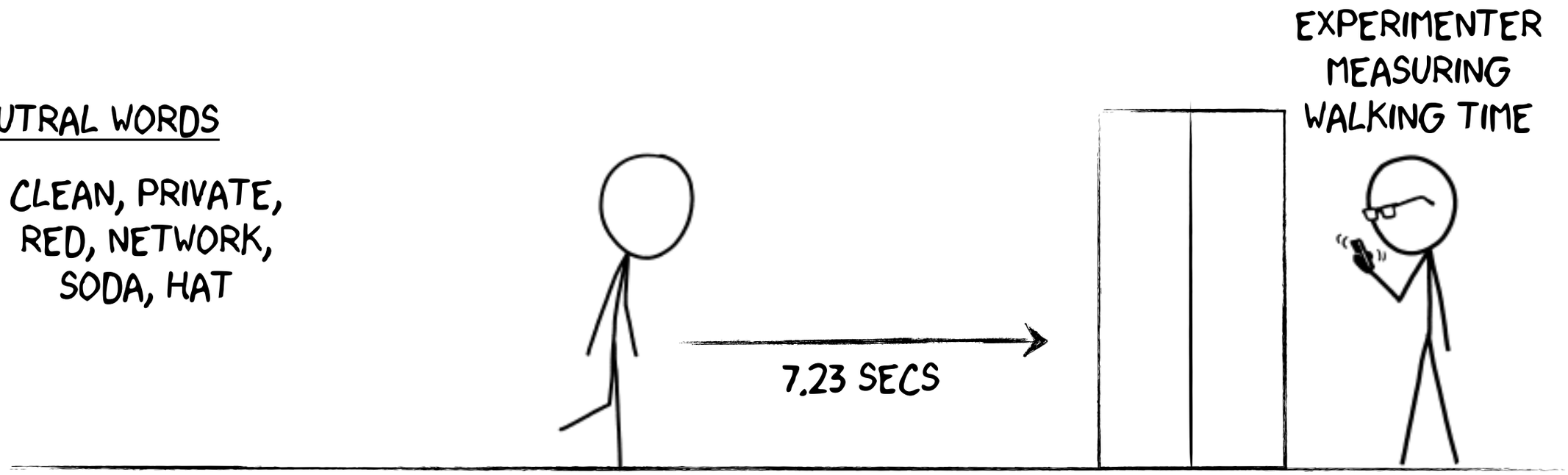
CLEAN, PRIVATE,
RED, NETWORK,
SODA, HAT

ELDERLY-RELATED WORDS

FLORIDA, LONELY,
GREY, WISE, BINGO,
FORGETFUL,
RETIRED

Modified cc images from xkcd.com

Bargh et. al, 1996 experiment



than did participants primed with polite-related stimuli. In Experiment 2, participants for whom an elderly stereotype was primed walked more slowly down the hallway when leaving the experiment than did control participants, consistent with the content of that stereotype. In Experiment 3, par-

Modified cc images from xkcd.com

Industry of "Social Priming" Research

Dozens of teams have been inspired by Bargh's work.

Citations to Bargh's et al. (mid 2014):

2491 in Google Scholar
1157 in PsycInfo (average in same journal is **128**)

The result has been incorporated in textbooks.

However, almost 20 years later...

Bargh's result does not replicate (Pashler 2008, Doyen 2012)

Other social priming findings are not replicating either.

“priming the stereotype of professors or the trait intelligent enhanced participants' performance on a scale measuring general knowledge. Also, priming the stereotype of soccer hooligans or the trait stupid reduced participants' performance on a general knowledge scale” (Dijksterhuis & van Knippenberg, 1998)

Failures to replicate reported in (Eder, Leipert, Musch, & Klauer, 2001) and (Shanks et al., 2013)

“participants who were exposed to honesty-related words admitted to having engaged in [excessive alcohol consumption] more than did participants who were exposed to neutral words” (Rasinski et al., 2005)

Failure to replicate reported by Pashler et al. (2013)

“reminders of money led to reduced helpfulness toward others […] participants primed with money preferred to play alone, work alone, and put more physical distance between themselves and a new acquaintance” (Vohs et al., 2006)

Failure to replicate reported by Grenier et al. (2012)

“participants who received a single exposure to an American flag exhibited a significant increase in Republican voting intentions, voting behavior, political beliefs, and implicit and explicit attitudes, with some effects lasting 8 months after the initial priming episode.” (Carter et al., 2011)

Failure to replicate reported by the “many labs” project (2013)

“priming the stereotype of professors or the trait intelligent enhanced participants' performance on a scale measuring general knowledge. Also, priming the stereotype of soccer hooligans or the trait stupid reduced participants' performance on a general knowledge scale” (Dijksterhuis & van Knippenberg, 1998)

Failures to replicate reported in (Eder, Leipert, Musch, & Klauer, 2001) and (Shanks et al., 2013)

“participants who to having
engaged in [exces ants who were
exposed to neutra

Failure to replicate

reputable scientists
reputable journals
standard methods
not fraud

“reminders of mor participants
primed with mone re physical
distance between , 2006)

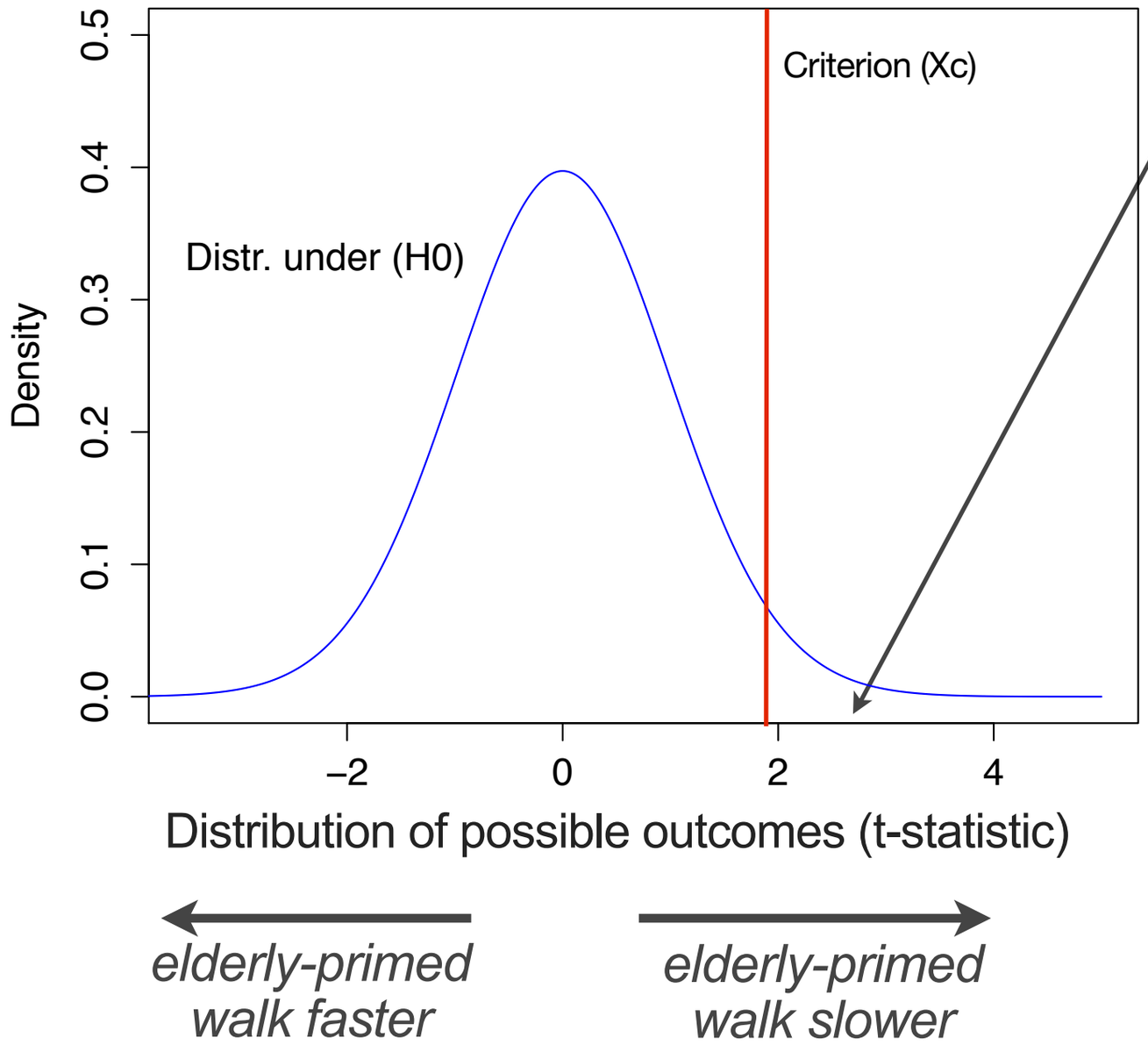
Failure to replicate

“participants who received a single exposure to an American flag exhibited a significant increase in Republican voting intentions, voting behavior, political beliefs, and implicit and explicit attitudes, with some effects lasting 8 months after the initial priming episode.” (Carter et al.,

Failure to replicate reported by the “many labs” project (2013)

Frequentist Inference - Null Hypothesis Significance Testing

Null **H0:** Elderly – Neutral = 0
Alternative **H1:** Elderly – Neutral > 0



If the observed data is unlikely under H0, then you reject H0 in favor of H1.

False Positive: Incorrect rejection of a null hypothesis.

Bargh's results:

p-value: 0.05
Effect size: Cohen's d = 0.81

d ≈ 0.2 (small)
d ≈ 0.5 (medium)
d > 0.8 (large)

Frequentist Inference - Other concepts

Statistical Power: Probability of detecting an effect under the assumption that there is a real effect.

Meta-analysis: Techniques to aggregate effect sizes from different experiments to get more robust estimates.

Replications

Direct Replication: The design mirrors the original experimental design in all causally relevant factors.

Conceptual Replication: A new design purported to find an effect that would be expected were the original effect true.

Few direct replications in psychology (Pashler & Harris, 2012)

SCT: **Scientific method** will refute false theories and find closer approximations to the true theories in the **long run**.

SCT*: Given a **series of replications** of an experiment, the **aggregation of their effect sizes** will approach the true effect size as the length of the series of replications increases.

Plan

1. **SCT*:** SCT in terms of frequentist statistics. ✓

2. **Scientific Utopia:** SCT* depends on idealized assumptions about the social structure of science.

   a. Assumptions of a scientific utopia.

   b. Simulations: In the utopia, SCT* works.

   c. In less utopian scenarios, SCT* doesn't work.

3. **Focus Shift:** From methodology to social epistemology.

A Scientific Utopia

A Scientific Utopia

Everything is published

Scientists publish all their results regardless of magnitude or direction.

Unlimited resources

Scientists have enough time and funds to run experiments and direct replications with large samples.

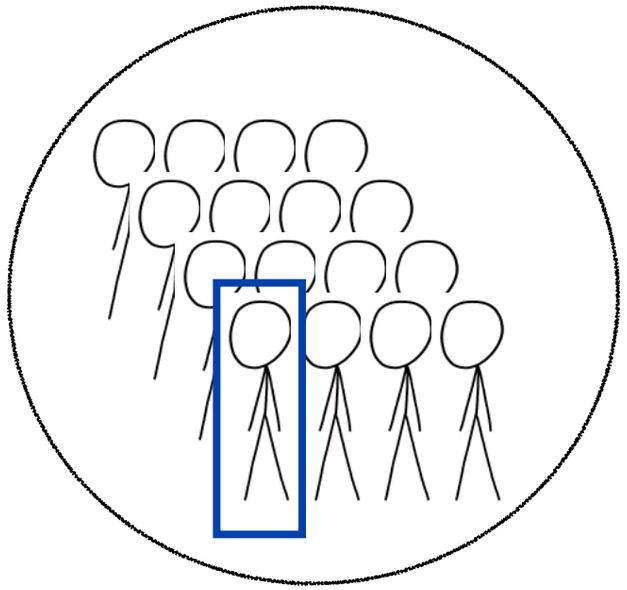
No direction bias

Scientists don’t adjust their experimental designs to meet prior expectations.

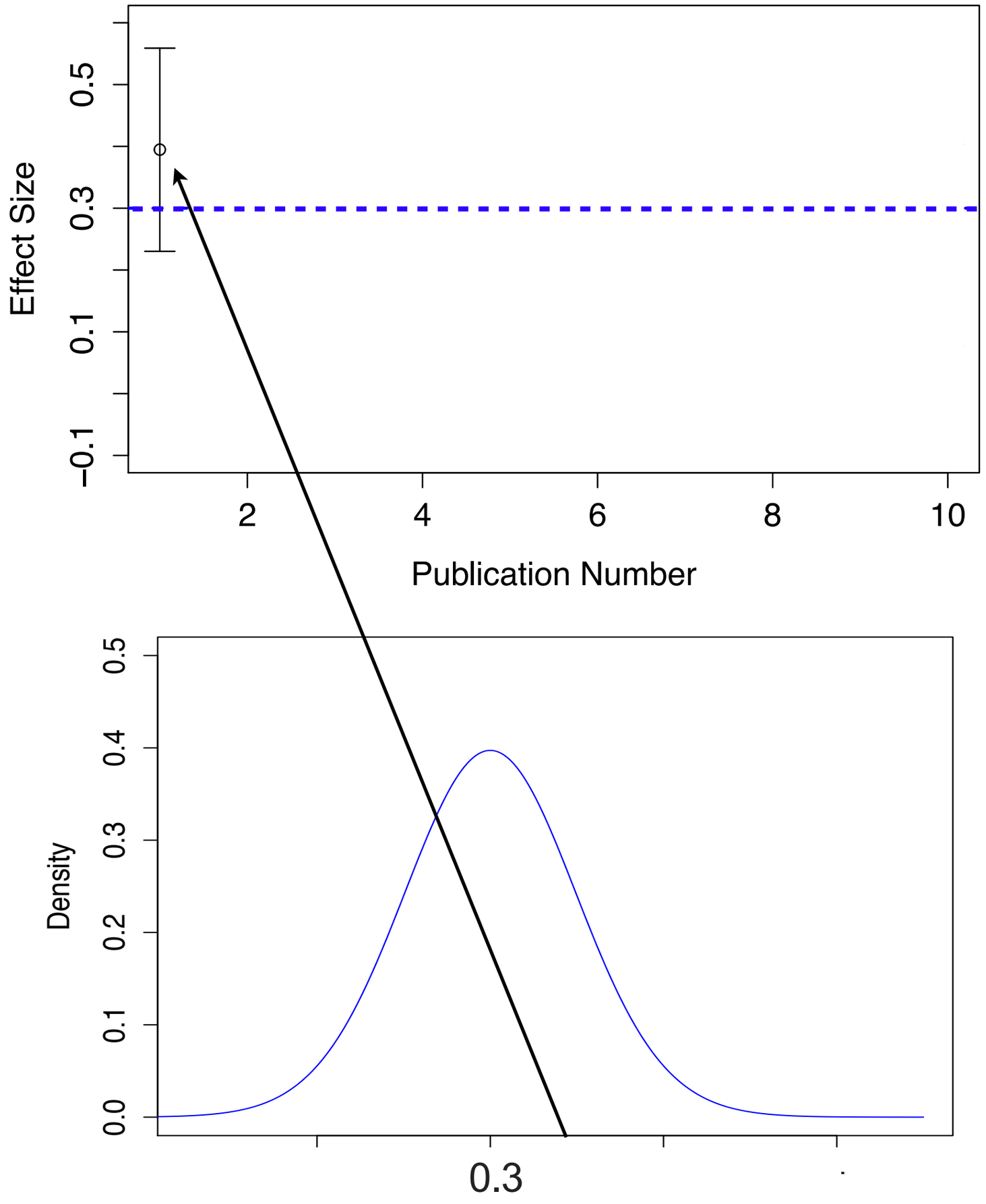
Total evidence

Scientists have access to (and use) all previous results regarding the phenomenon of interest.

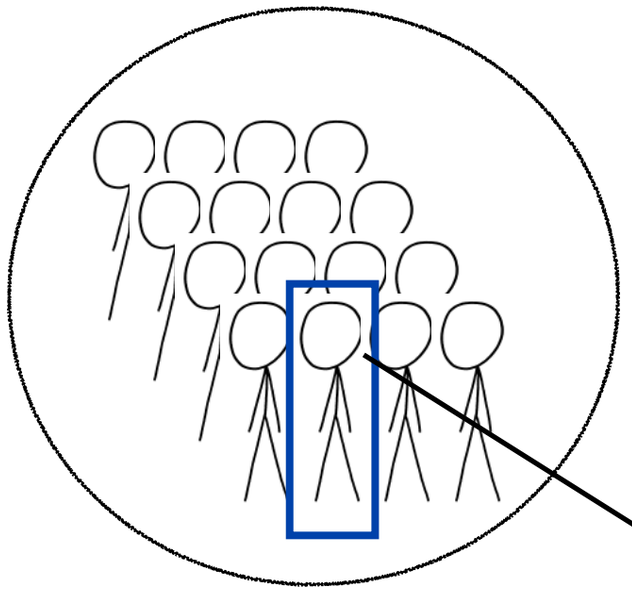
Simulating the Utopia



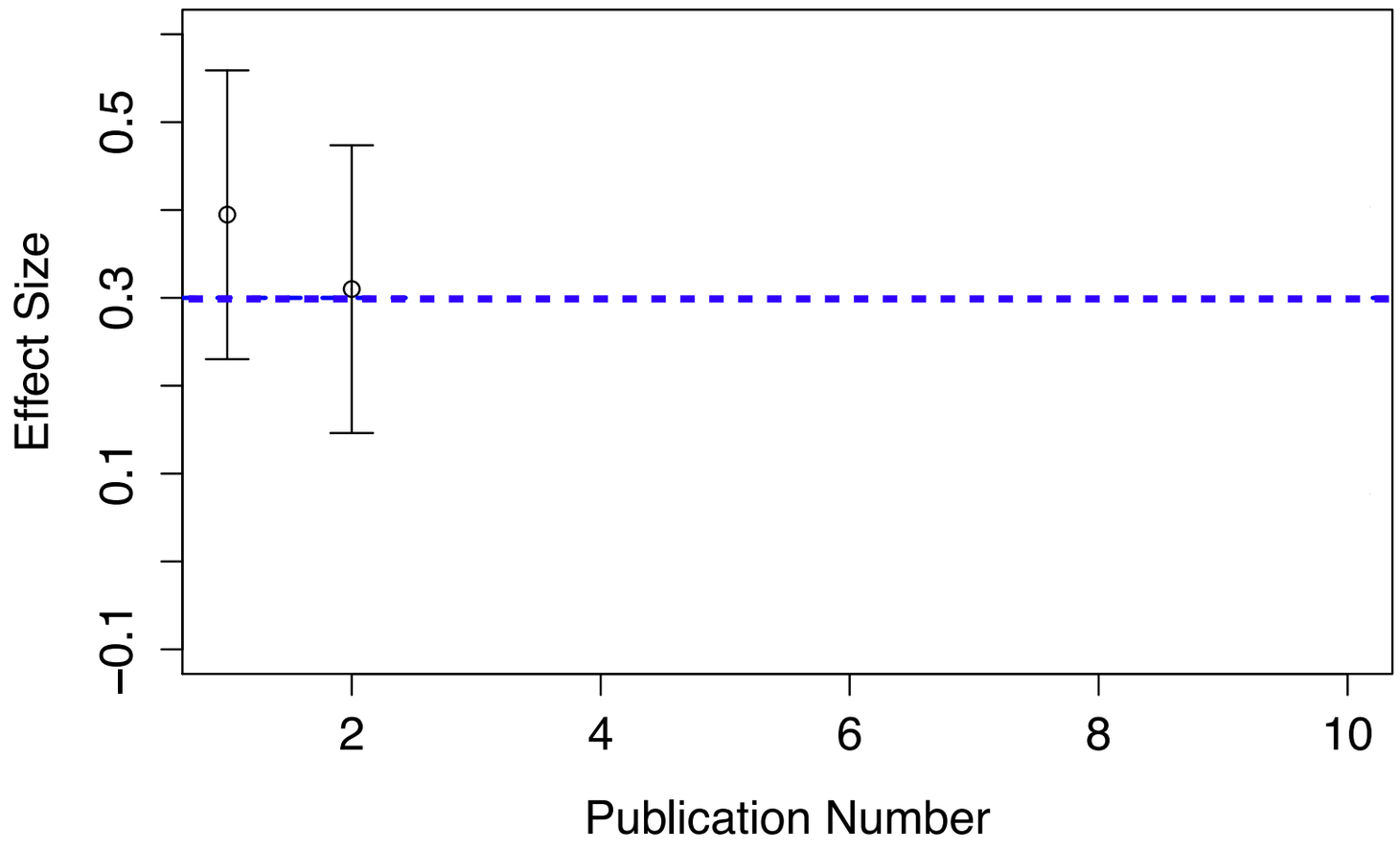
Society of Scientists

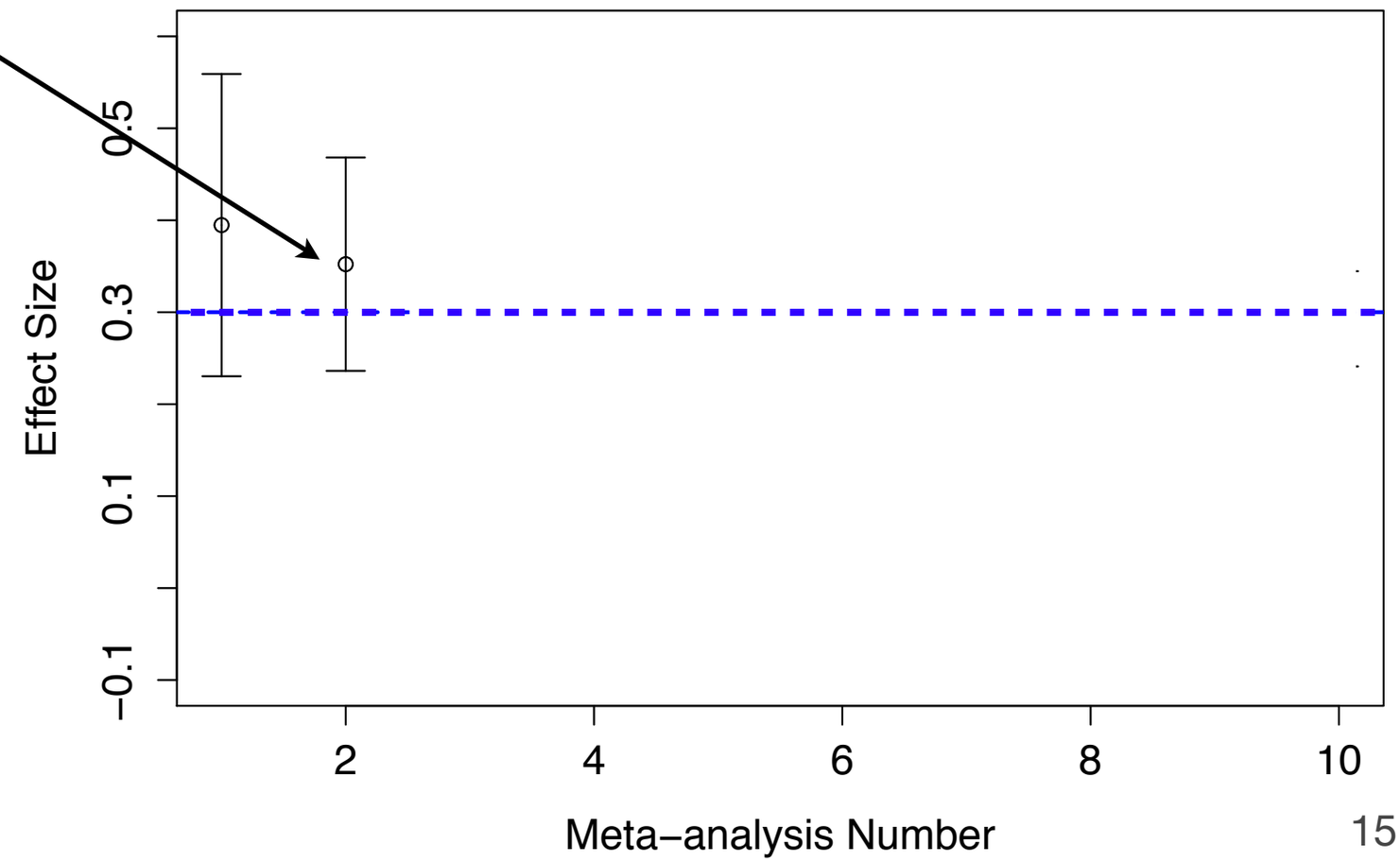


Simulating the Utopia

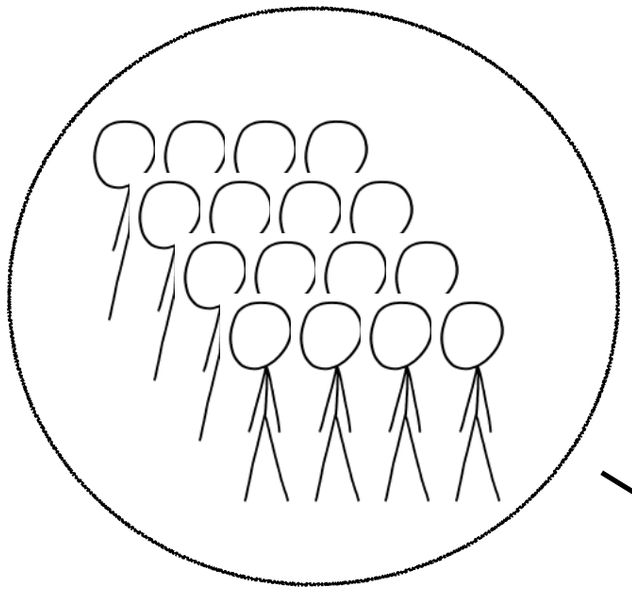


Society of Scientists

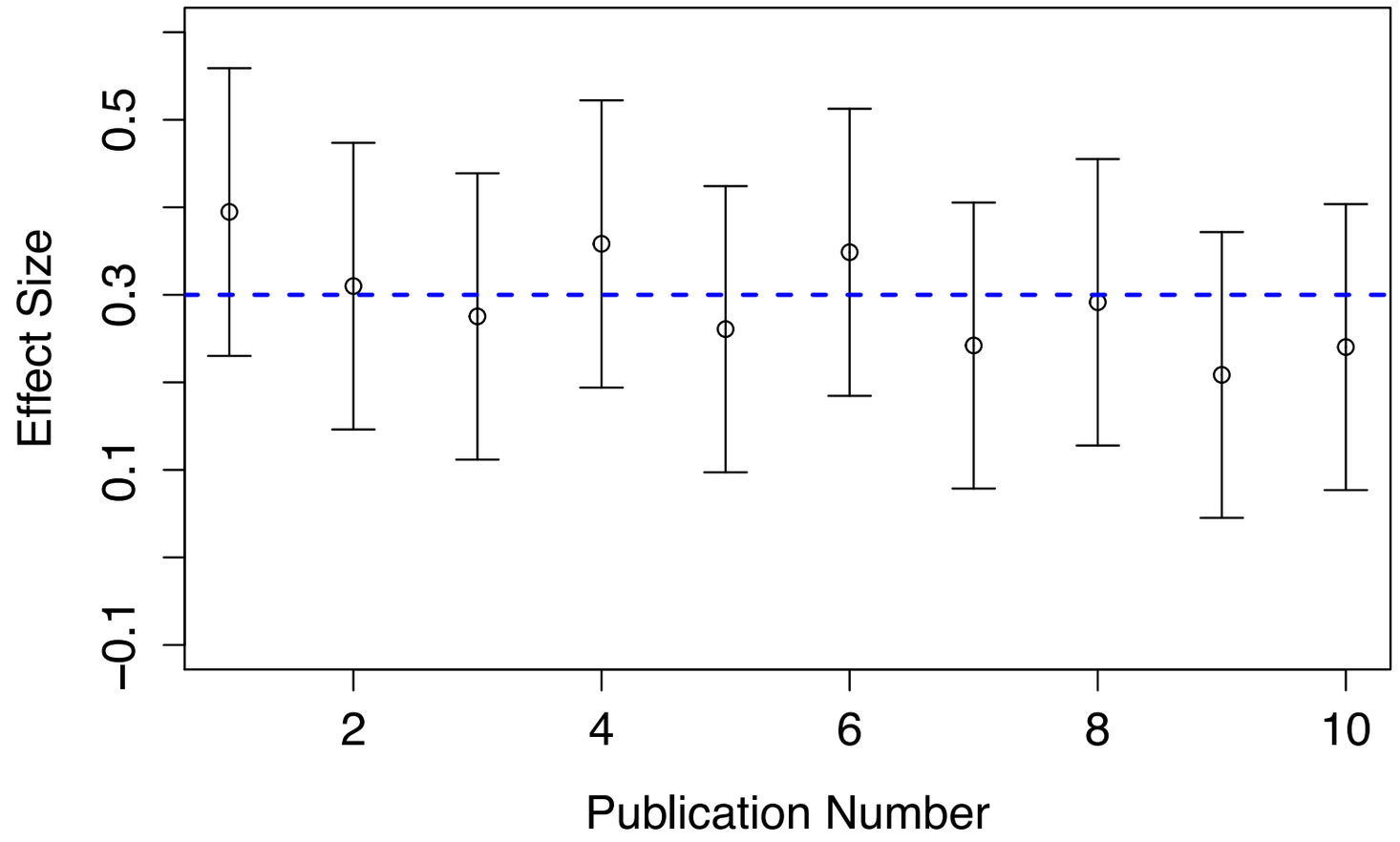


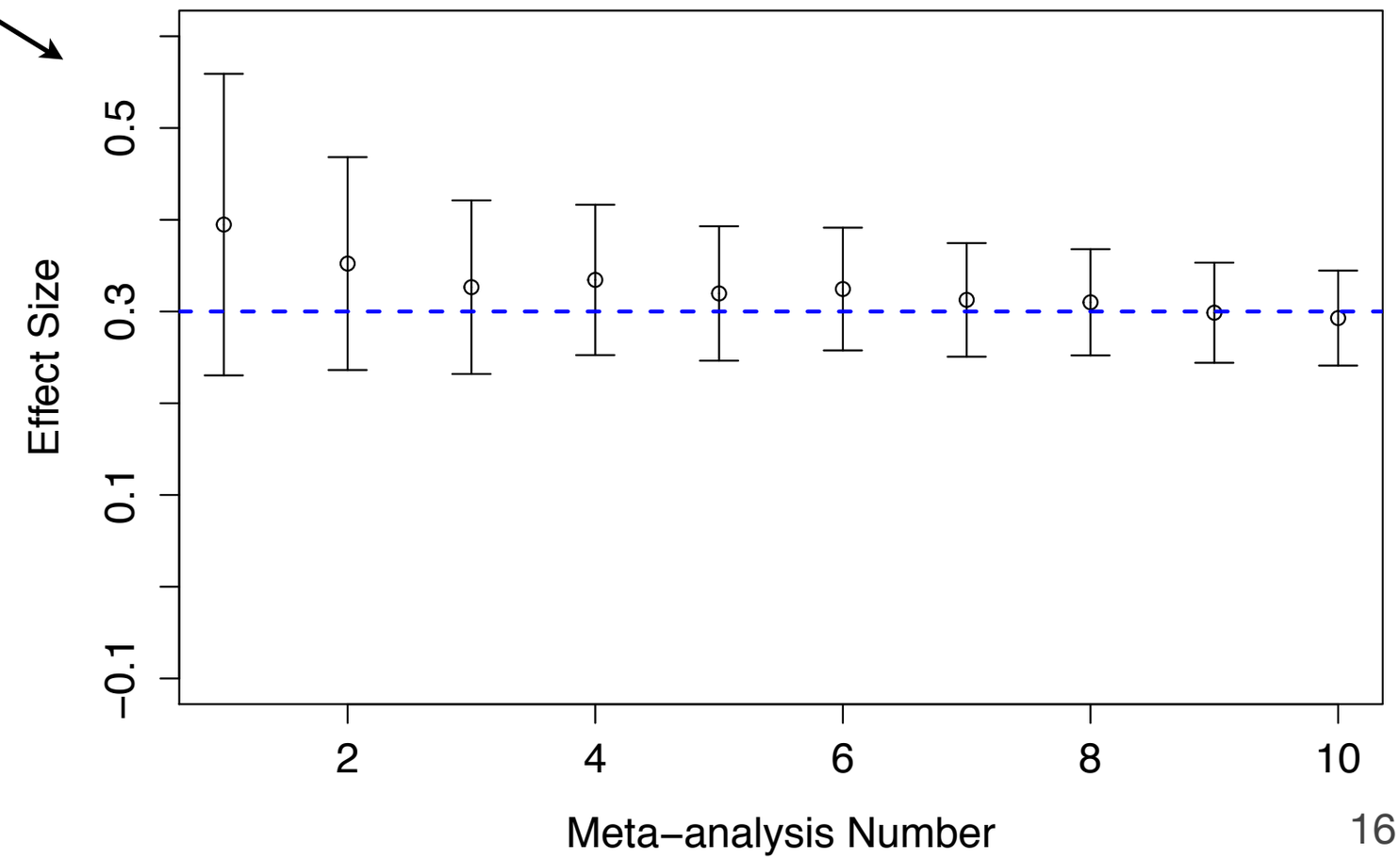


Simulating the Utopia

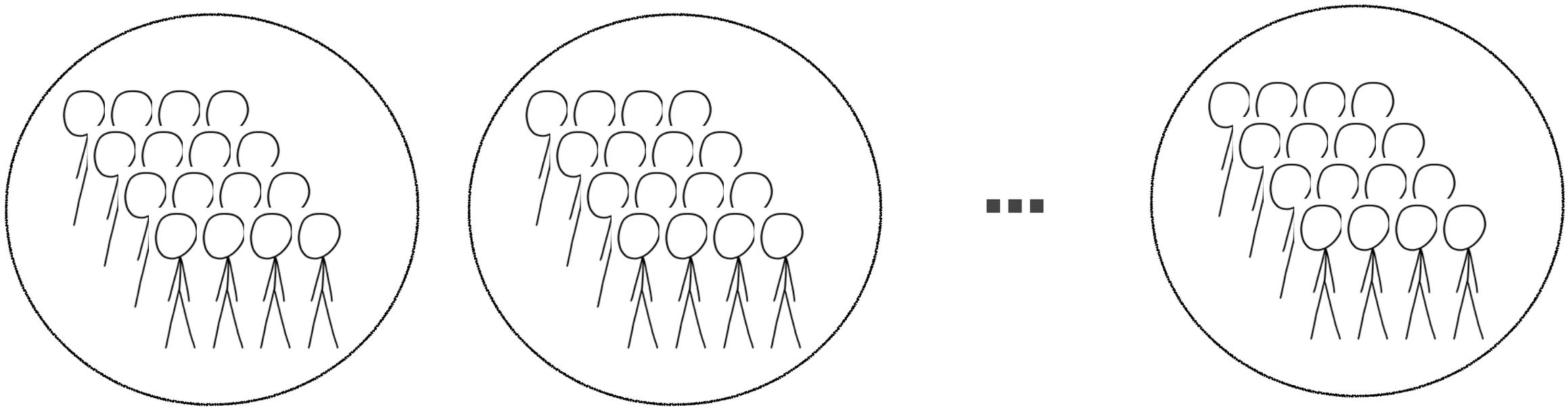


Society of Scientists





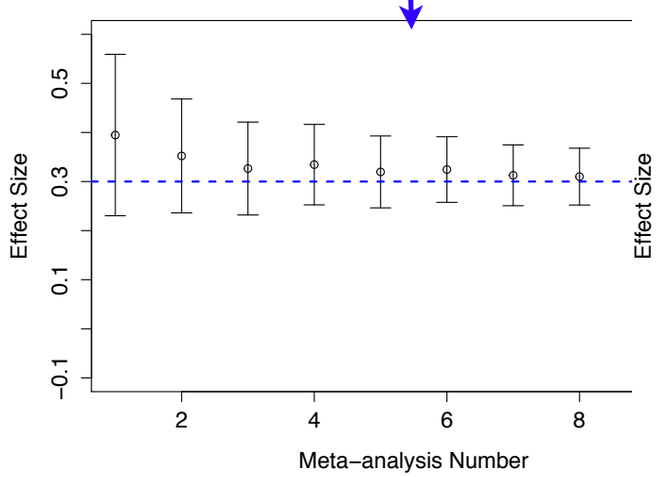
Simulating the Utopia

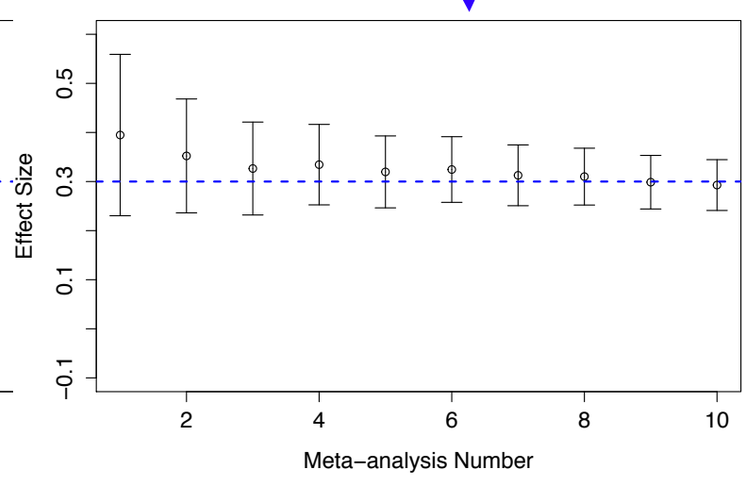


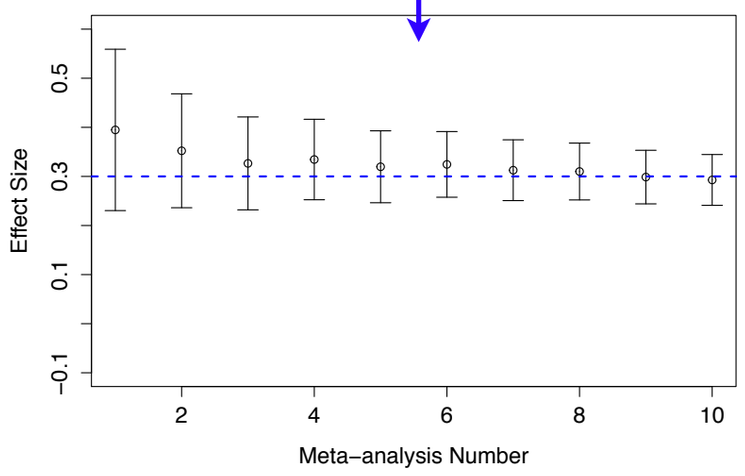
Society 1
(100 scientists)

Society 2

Society 1000







Aggregated meta-analysis

Let’s assume there is a real effect

Scenario 0 - Utopia

Everything is published

Unlimited resources

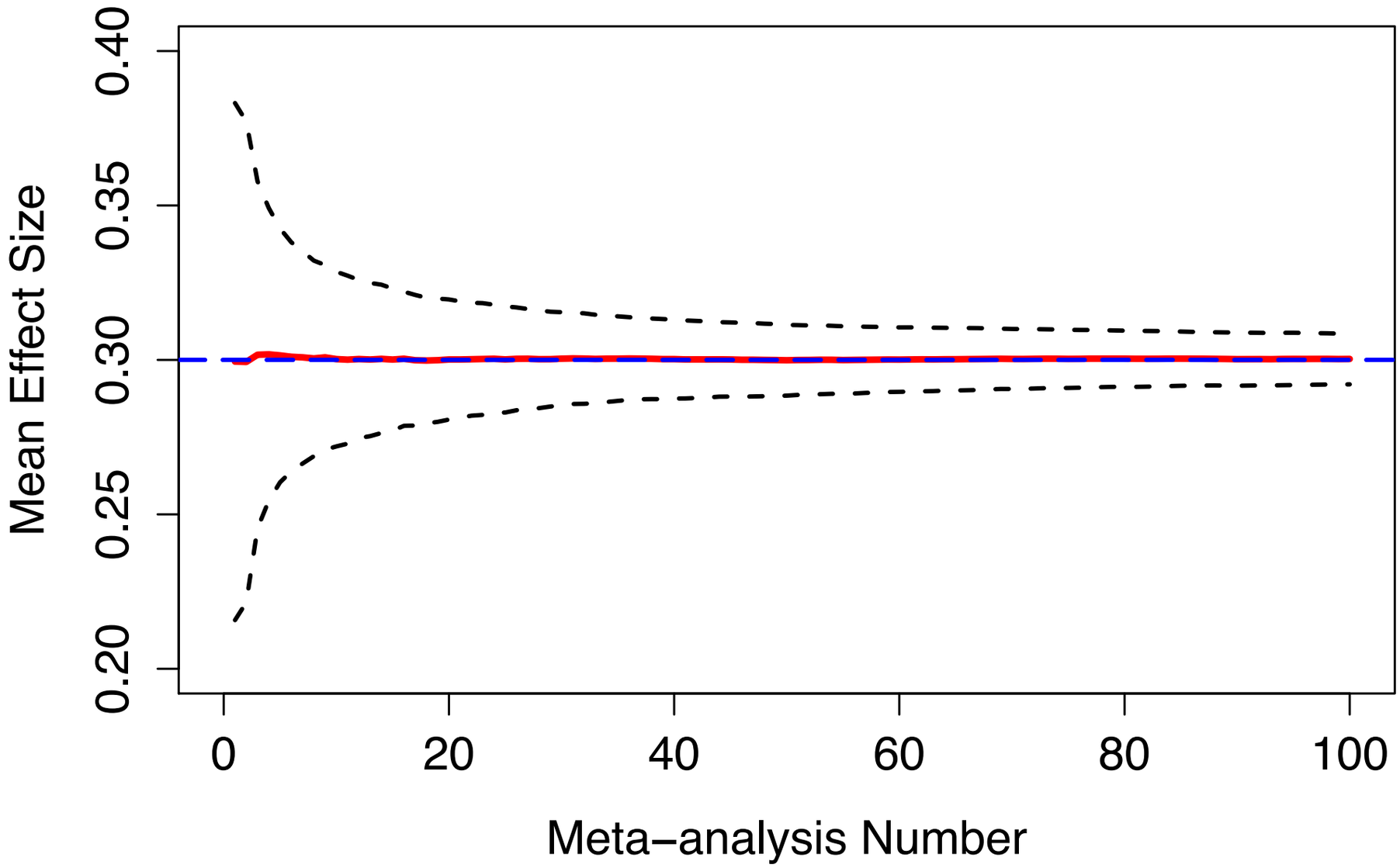
No direction bias

Total evidence

Real effect size: Cohen's *d* = 0.3

Sample size: 290 subjects

Statistical power: 0.95



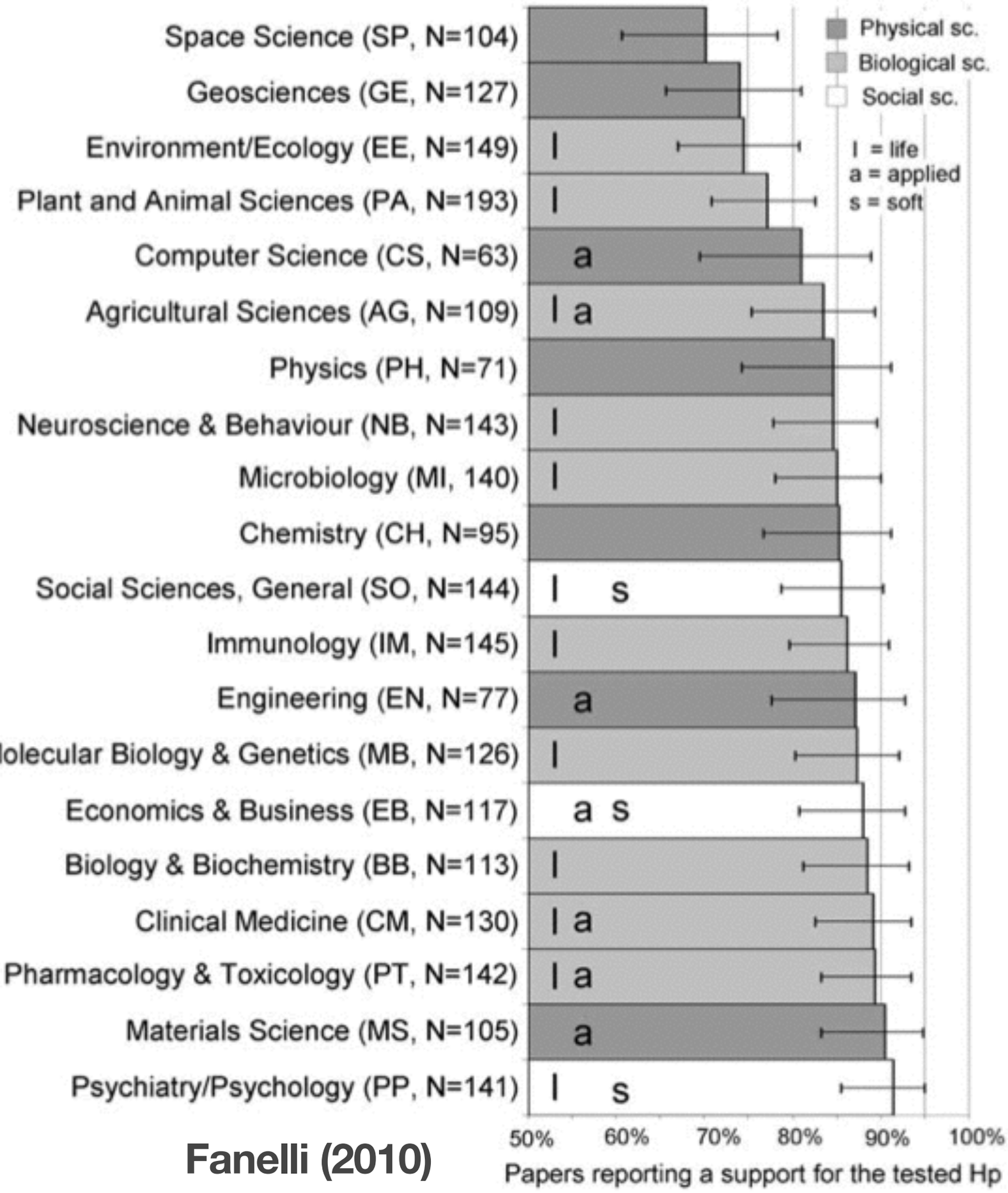
Scenario 1

~~Everything is published~~

Unlimited resources

No direction bias

Total evidence



Fanelli (2010)

Scenario 1

~~**Everything is published**~~

Unlimited resources

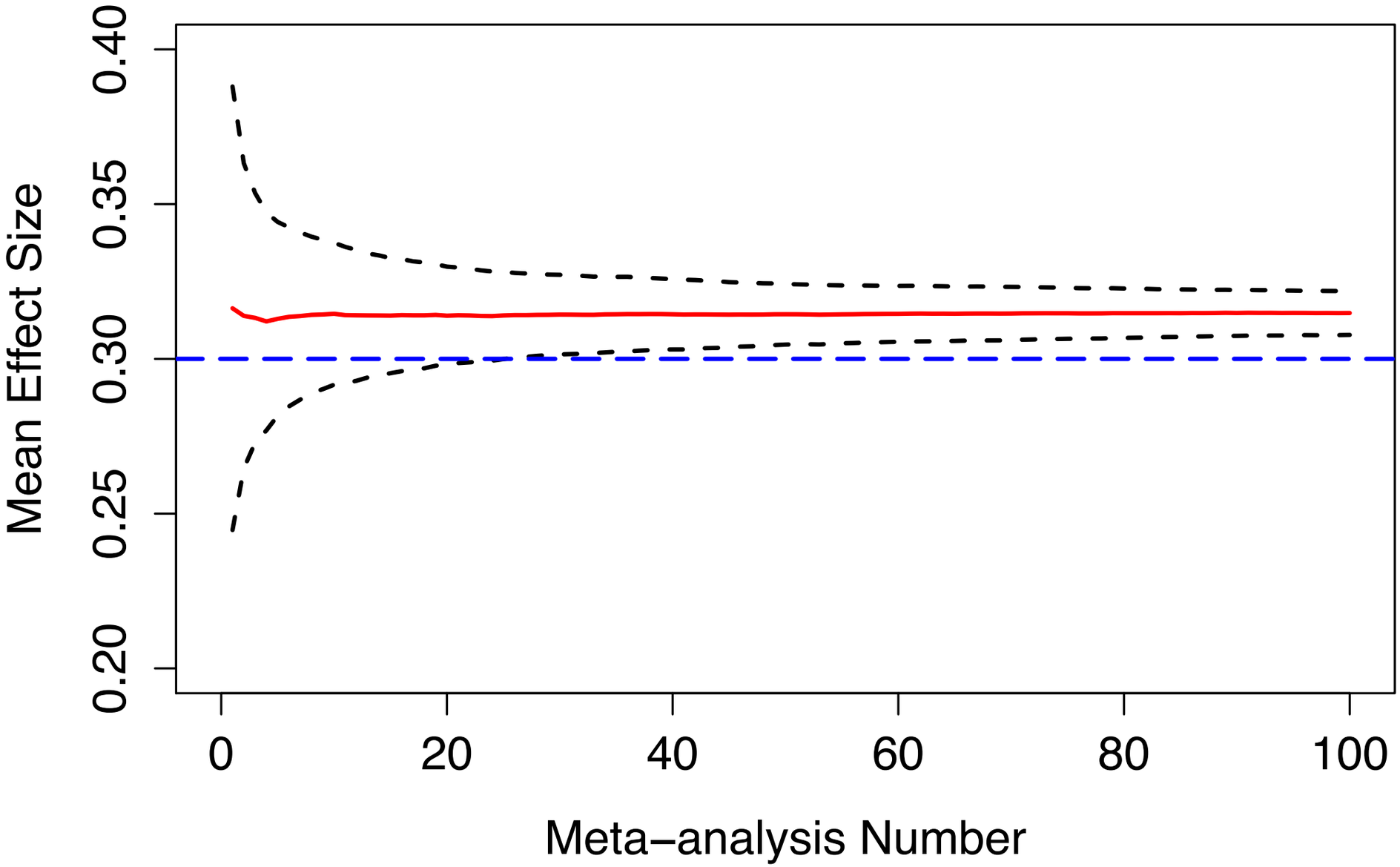
No direction bias

Total evidence

Real effect size: Cohen's *d* = 0.3

Sample size: 290 subjects

Statistical power: 0.95



} overestimation of the effect size

Scenario 2

~~**Everything is published**~~

~~**Unlimited resources**~~

No direction bias

Total evidence

Real effect size: Cohen's *d* = 0.3

Sample size: 87 subjects

Statistical power: 0.5

Limited resources -> more likely to have underpowered studies.

In psychology, statistical power has been historically below 0.5

- Jacob Cohen (1962). 70 articles, JAP. **Power = 0.46**
- Sedlmeier & Gigerenzer (1989). Same study, 24 years later. **Power = 0.44**
- Maxwell (2004). Little improvement in the 90s and 00s.
- Fraley & Vazire (2014). Still 0.5 in social-personality research.

Scenario 2

~~**Everything is published**~~

~~**Unlimited resources**~~

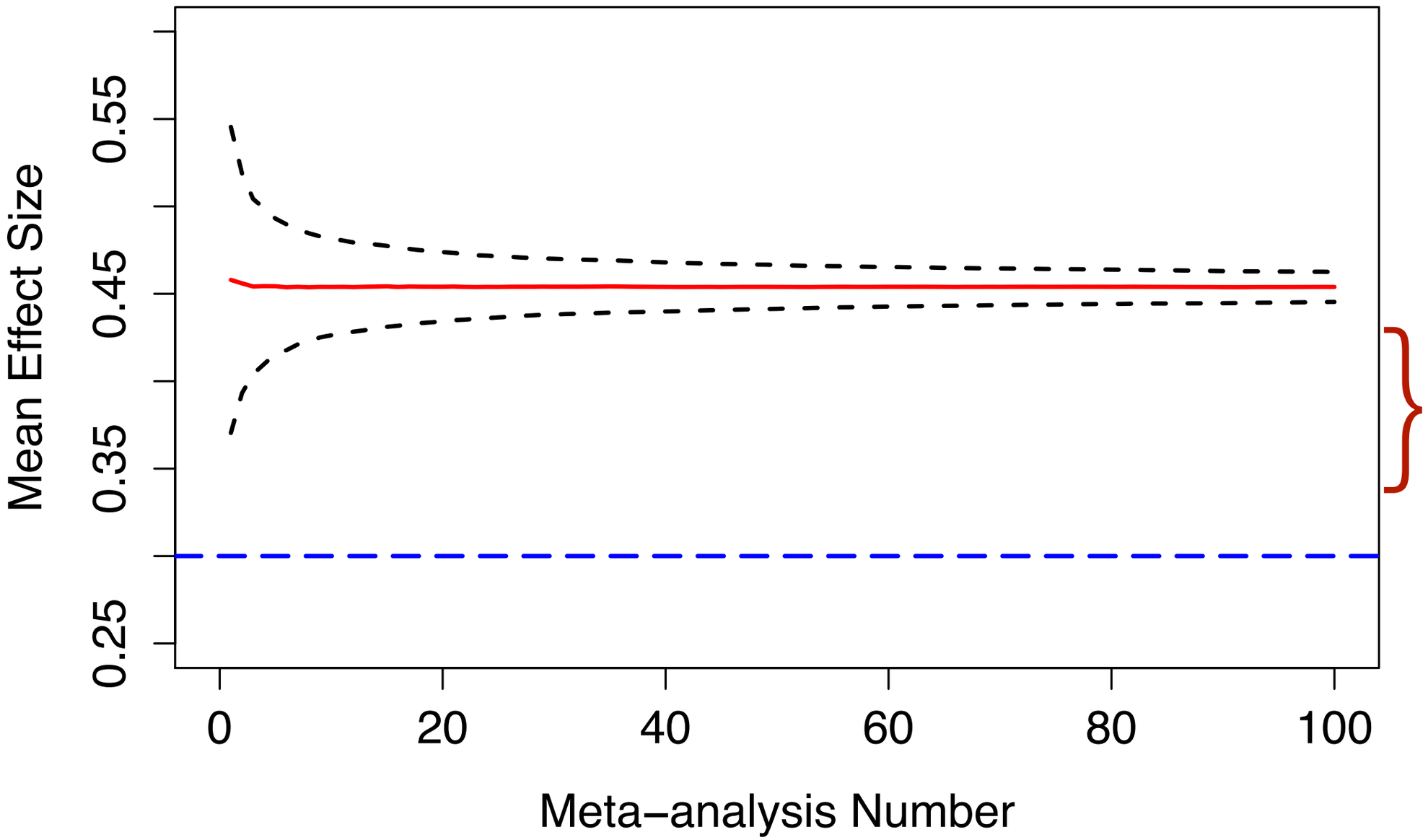
No direction bias

Total evidence

Real effect size: Cohen's *d* = 0.3

Sample size: 87 subjects

Statistical power: 0.5



Scenario 2

~~**Everything is published**~~

~~**Unlimited resources**~~

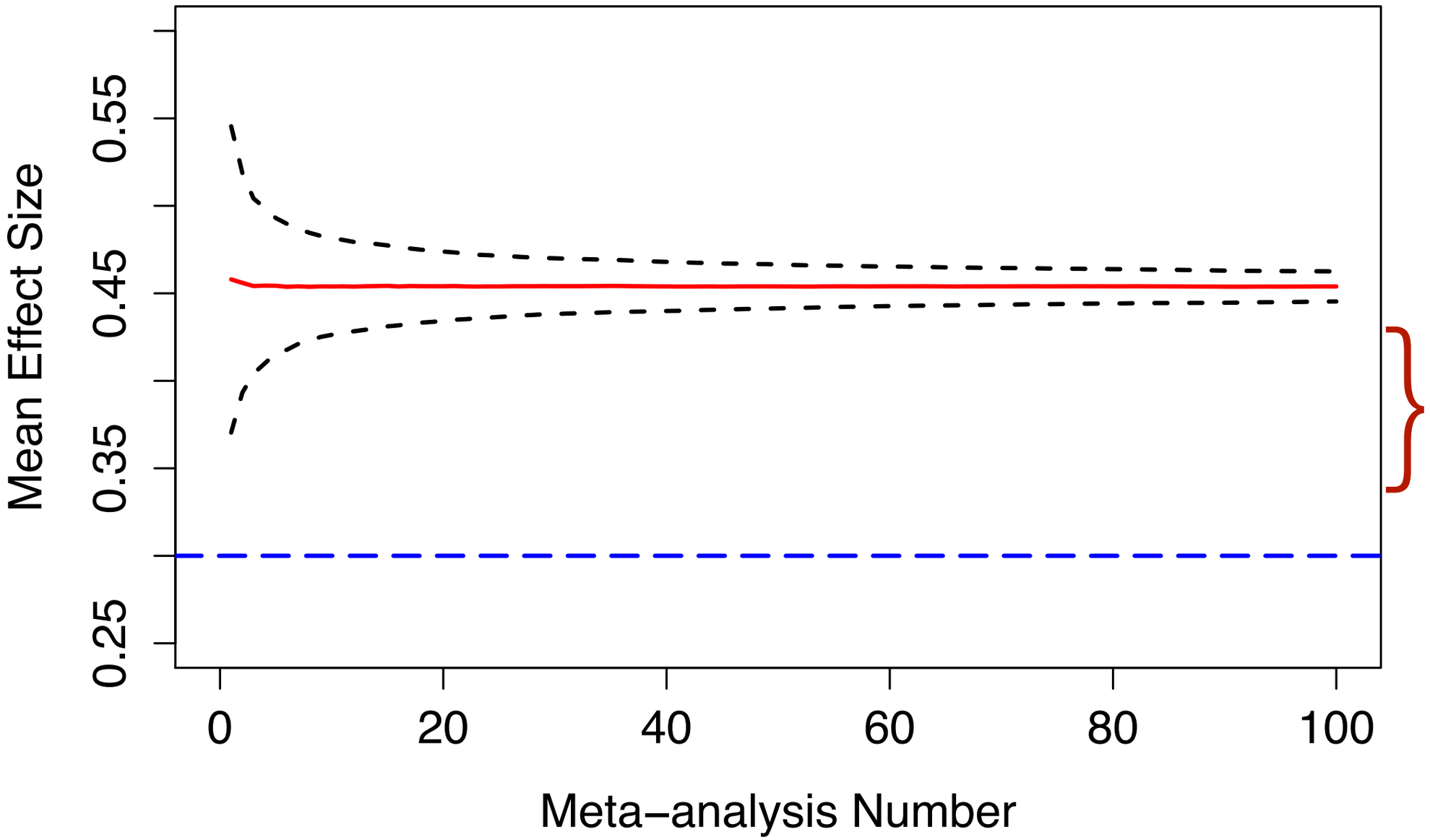
No direction bias

Total evidence

Real effect size: Cohen's *d* = 0.3

Sample size: 87 subjects

Statistical power: 0.5



large inflation of the effect size

See Ioannidis (2008), Button (2013)

What happens when there is not a real effect?

Scenario 3

Everything is published

Unlimited resources

No direction bias

Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 10395 subjects

Statistical power: 0.95

(required to detect a d = 0.05)

If you have enough statistical power to detect a very small effect and you fail, then you can accept an approximation to the Null. (See Machery 2011)

Scenario 3

Everything is published

Unlimited resources

No direction bias

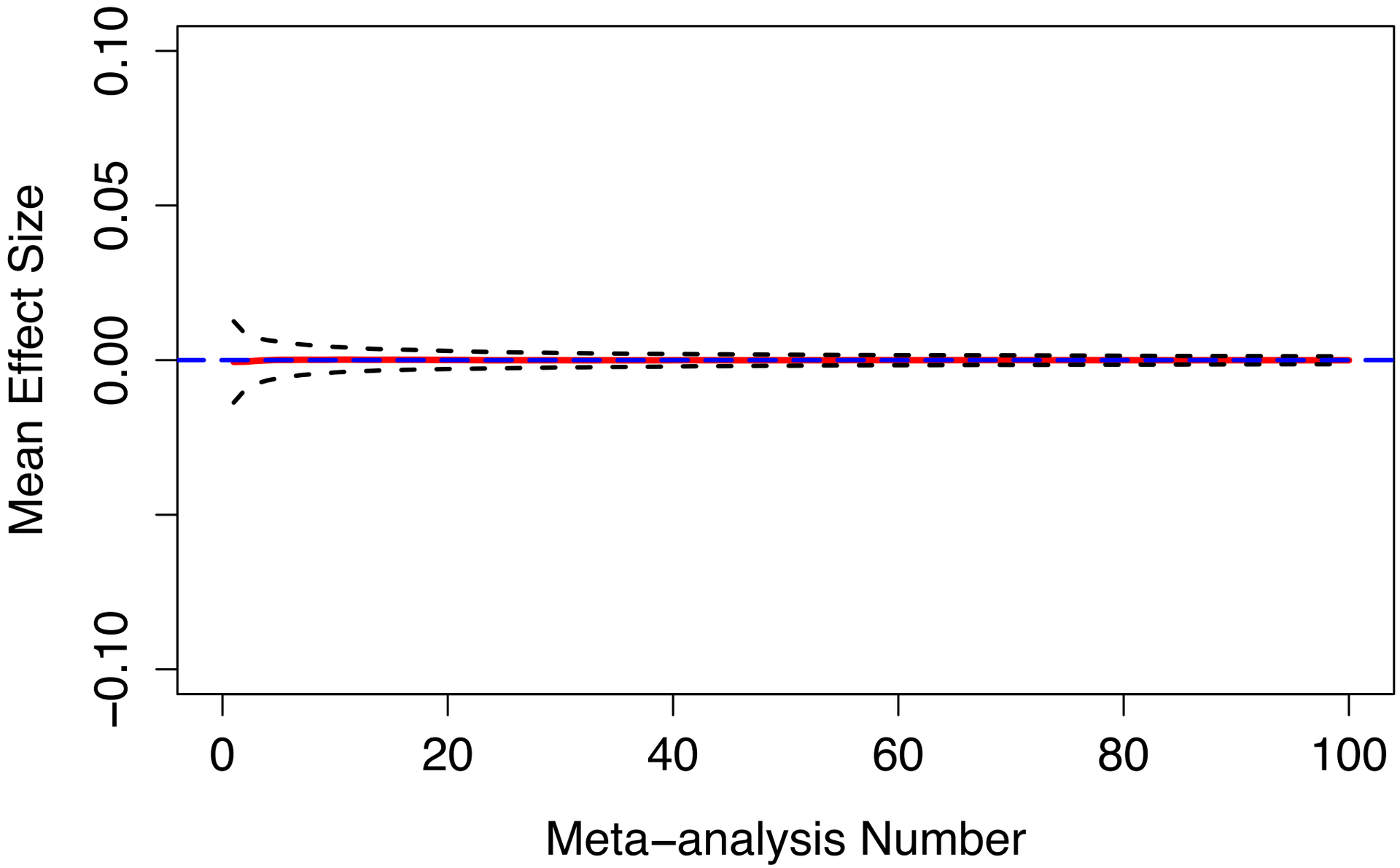
Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 10395 subjects

Statistical power: 0.95

(required to detect a d = 0.05)



Scenario 4

~~Everything is published~~

~~Unlimited resources~~

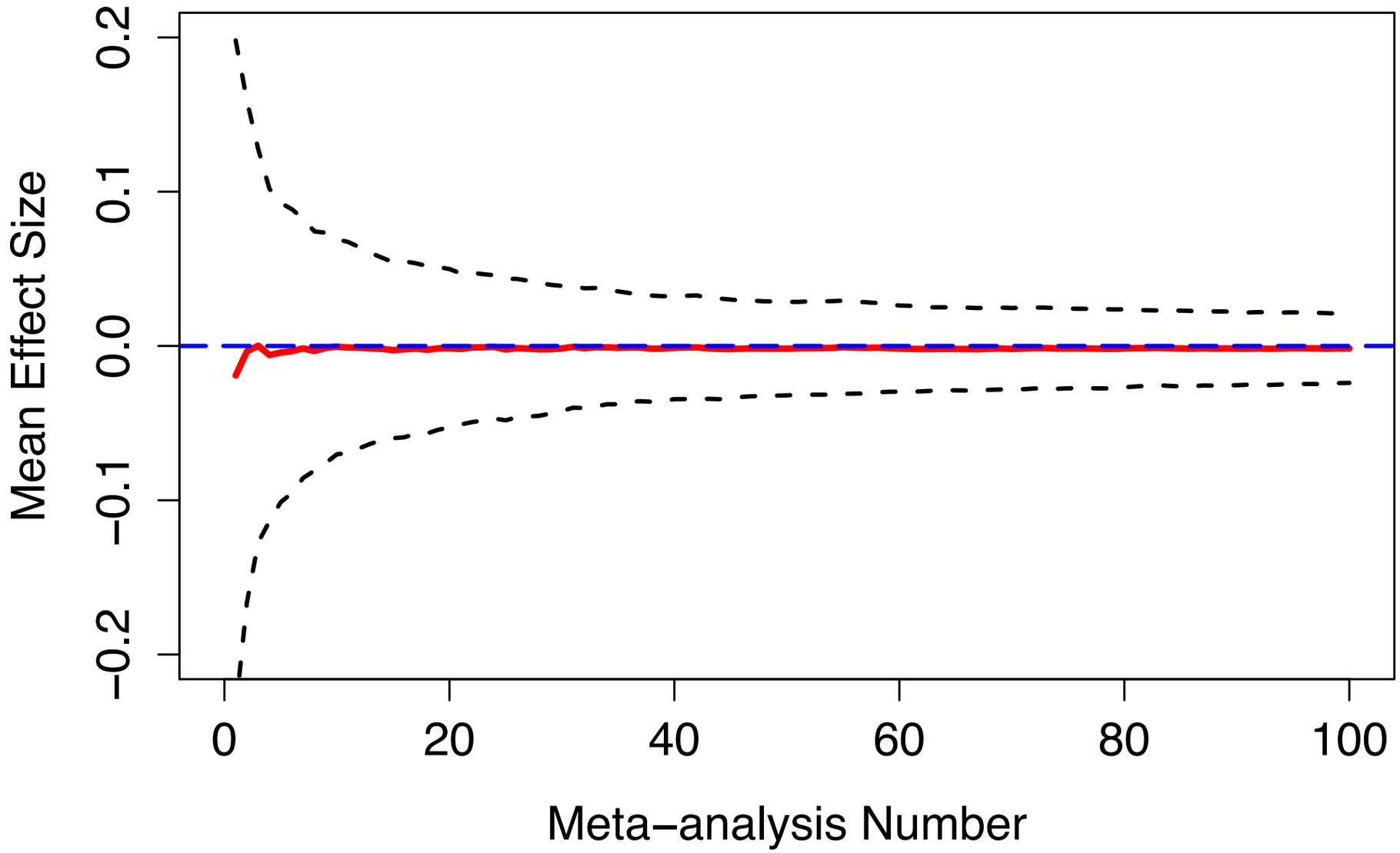
No direction bias

Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 290 subjects

Statistical power: 0.1 (for d = 0.05)



Scenario 4

~~Everything is published~~

~~Unlimited resources~~

No direction bias

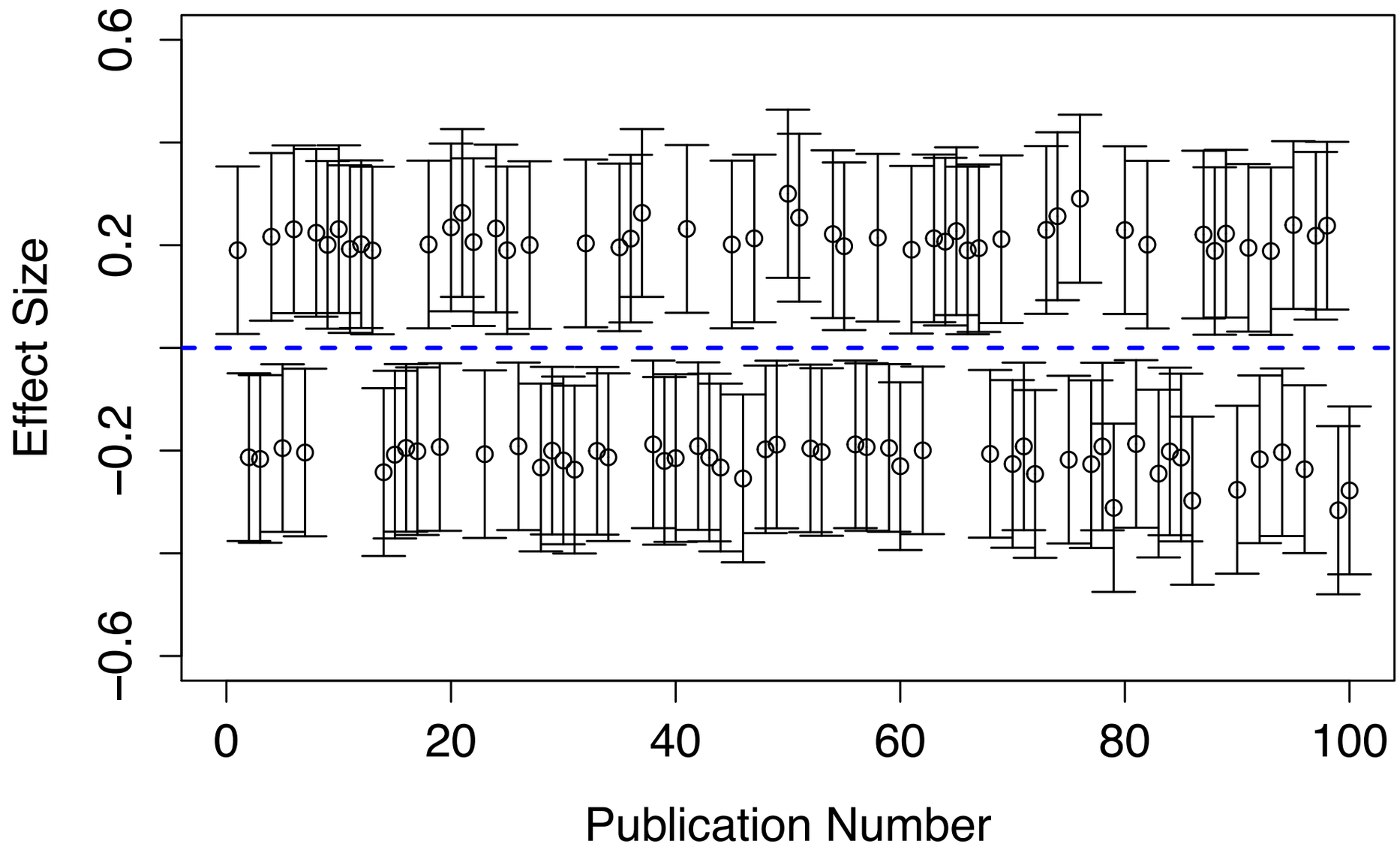
Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 290 subjects

Statistical power: 0.1 (for d = 0.05)

Slow convergence: 800 experiments for 1 publication



Scenario 5 - Direction Bias

~~Everything is published~~

~~Unlimited resources~~

~~No direction bias~~

Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 290 subjects

Statistical power: 0.1 (for d = 0.05)

Prior theoretical commitments

Motivated cognition

–> Selective reporting of effects in only one direction.

Scenario 5 - Direction Bias

~~Everything is published~~

~~Unlimited resources~~

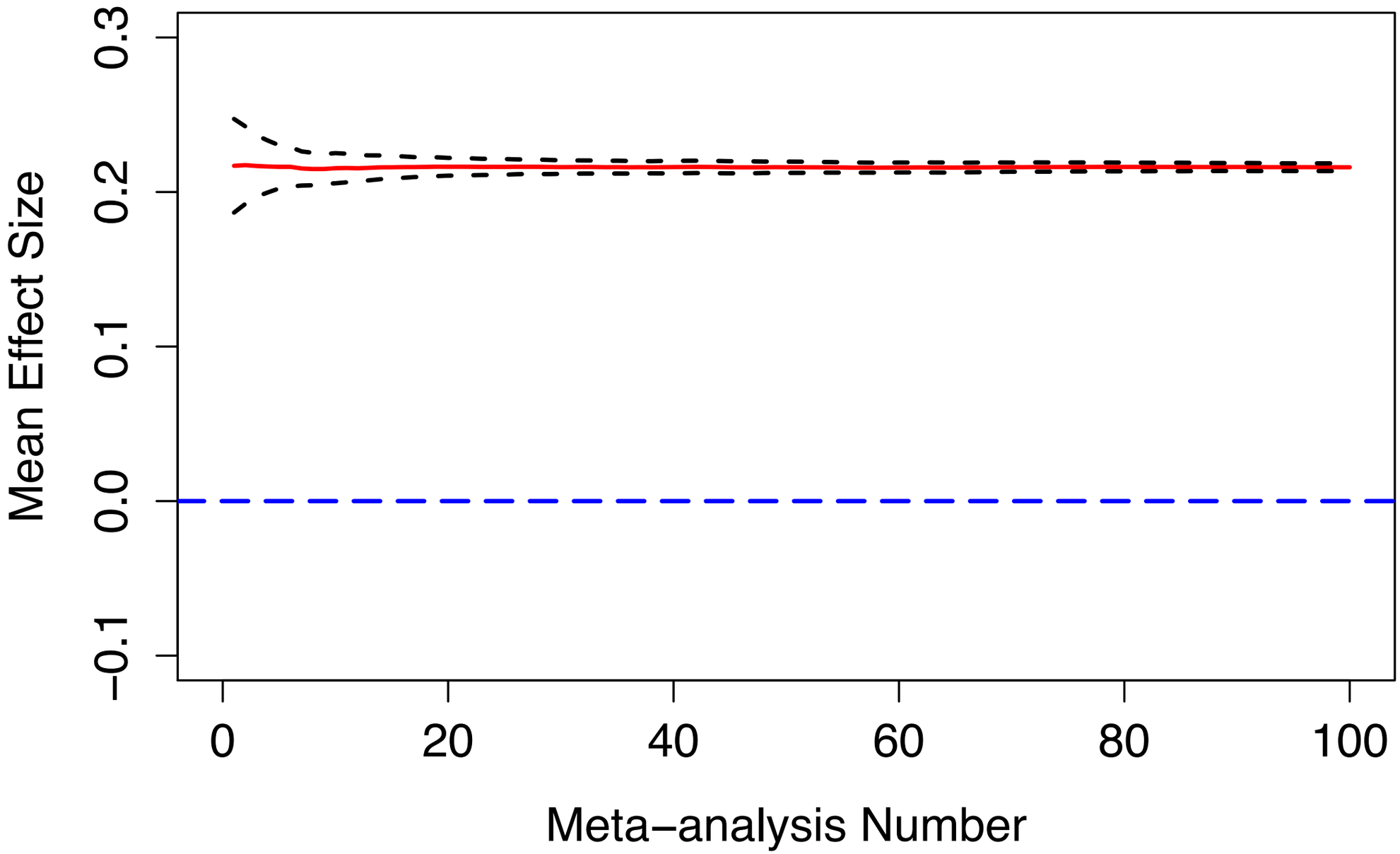
~~No direction bias~~

Total evidence

Real effect size: **Cohen's *d* = 0**

Sample size: 290 subjects

Statistical power: 0.1 (for d = 0.05)



Scenario 6 - Recency Bias

~~Everything is published~~

~~Unlimited resources~~

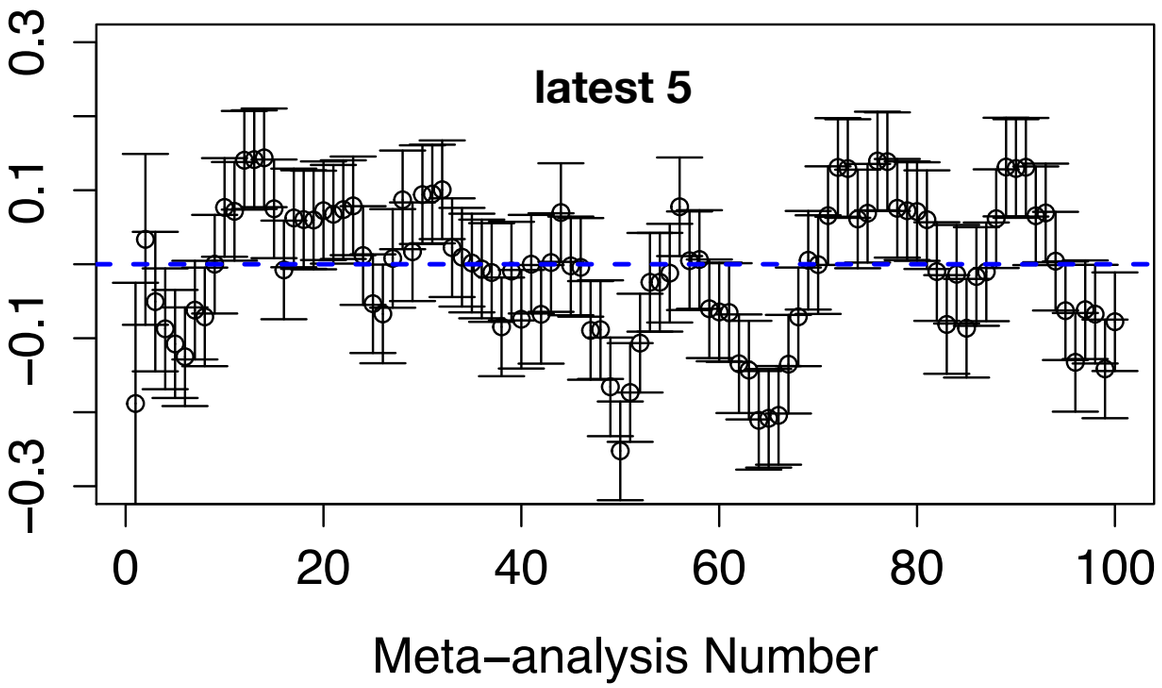
No direction bias

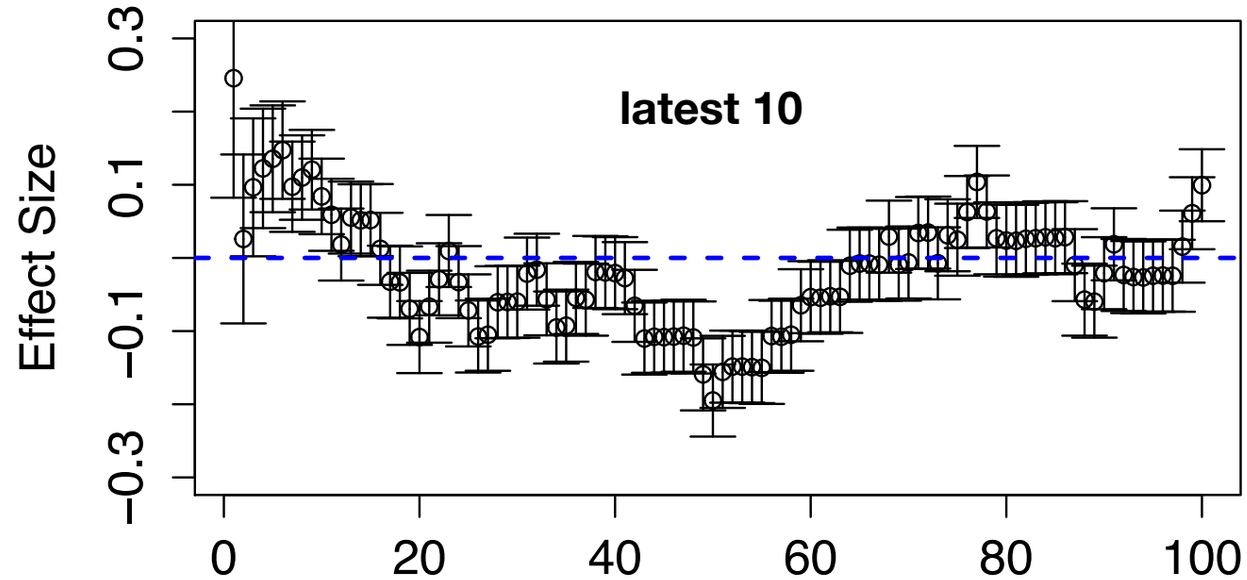
~~Total evidence~~

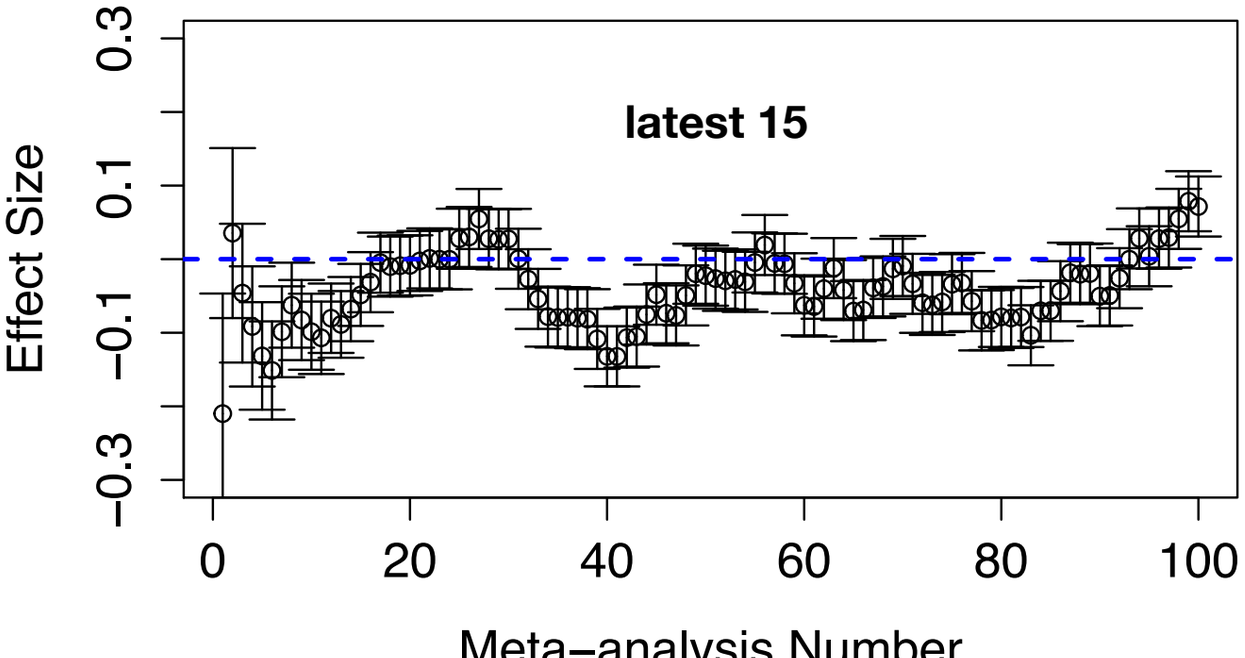
Real effect size: **Cohen's *d* = 0**

Sample size: 290 subjects

Statistical power: 0.1 (for d = 0.05)







Simulation Results

Scientific utopia self-corrects but self-correction is fragile.

When there is a real effect...

Publication of only significant results inflates effect size estimates (even more with **underpowered** studies).

When there is not a real effect...

Publication of only significant results does not inflate effect size estimates (but leads to convergence inefficiently)

Publication of only significant results and **direction bias** jointly inflate effect sizes.

Aggregation over recent publications produces effect size oscillations.

Notice: these results concern **direct** replications (contra Pashler & Harris, 2012; Makel et al, 2012)

Plan

1. **SCT*:** SCT in terms of frequentist statistics. ✓

2. **Scientific Utopia:** SCT* depends on idealized assumptions about the social structure of science.

   a. Assumptions of a scientific utopia. ✓

   b. Simulations: In the utopia, SCT* works.

   c. In less utopian scenarios, SCT* doesn't work.

3. **Focus Shift:** From methodology to social epistemology.

Scientific Self-Correction as an Interaction Effect

1. Inference Methods

Frequentist statistics (Peirce, 1901... Mayo, 2005)

Bayesian statistics

2. Social Structure of the Scientific Community

Incentive structures (Kitcher, 1990; Strevens, 2003)

Division of labor (Weisberg & Muldoon, 2009)

Topology of the community (Zollman, 2010)

Work in Progress - Infectious Falsehoods

Epistemic Trust is necessary for efficient division of cognitive labor (Hull 1988, Hardwig 1991)

Scientific Standing Ovations Model (based on Miller & Page, 2004)

Performance	Scientific Study
Action: Stand up	Action: Trust without replication
Quality	f (power , effect size)
Actions of others	Actions of colleagues

Aggregators: Stand up if the literature supports the finding.
Skeptics: Stand up only If their own replication attempts succeed.
Trend-followers: Stand up if their close colleagues stand up.

Dynamics of different distributions of these types (speed/accuracy)

Summing up

Is science an enterprise that corrects its mistakes?

Self-correction is fragile: The social structure in which frequentist statistics is deployed affects its long run performance.

This is an example that puts pressure on the idea that we can ground self-correction primarily in properties of inference methods.

Philosophical attention to methods can only take us so far. We have to study the interaction between methods and social structures to understand how error infects scientific communities.

Thank you!

Felipe Romero
Washington University in St. Louis
cfromero@wustl edu