Small Telescopes: *Detectability and the Evaluation of Replication Results*

Uri Simonsohn

Wharton
UNIVERSITY *of* PENNSYLVANIA

Replications → New Data

Question 1. Combined: d_{All} ?

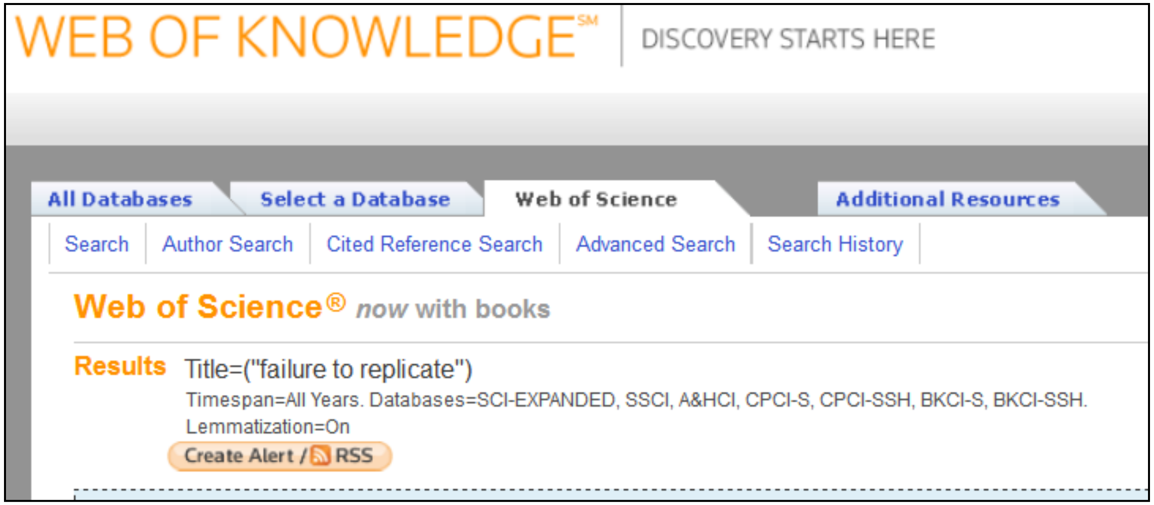
Question 2. $\mathrm{d}_{\mathrm{original}}$ vs. $\mathrm{d}_{\mathrm{replication}}$?

Question 3. Effect is zero or negligible?

Currently: is replication p<.05?

Top-10 cited

1. n.s.
2. n.s.
3. n.s.
4. n.s.
5. n.s.
6. n.s.
7. n.s.
8. n.s.
9. n.s.
10. n.s.



Results Title=("failure to replicate")

Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH.

Lemmatization=On

Next: 3 examples of bad inferences with it.

Example 1. Embodiment of morality

Original (study 1)

Zhong Liljenquist, 2006

- Recall (un)ethical
- Word completion

S _ _ P

- **Results:**

  N=60

  p<.05

  d = .54

Replication

Gamez et al (2011)

Results:

N=45

p=.77 ("failure")

Power = 40%

Example 2: Endowment effect

Original: Kahneman Knetsch and Thaler

WTA/WTP ~ 2.5

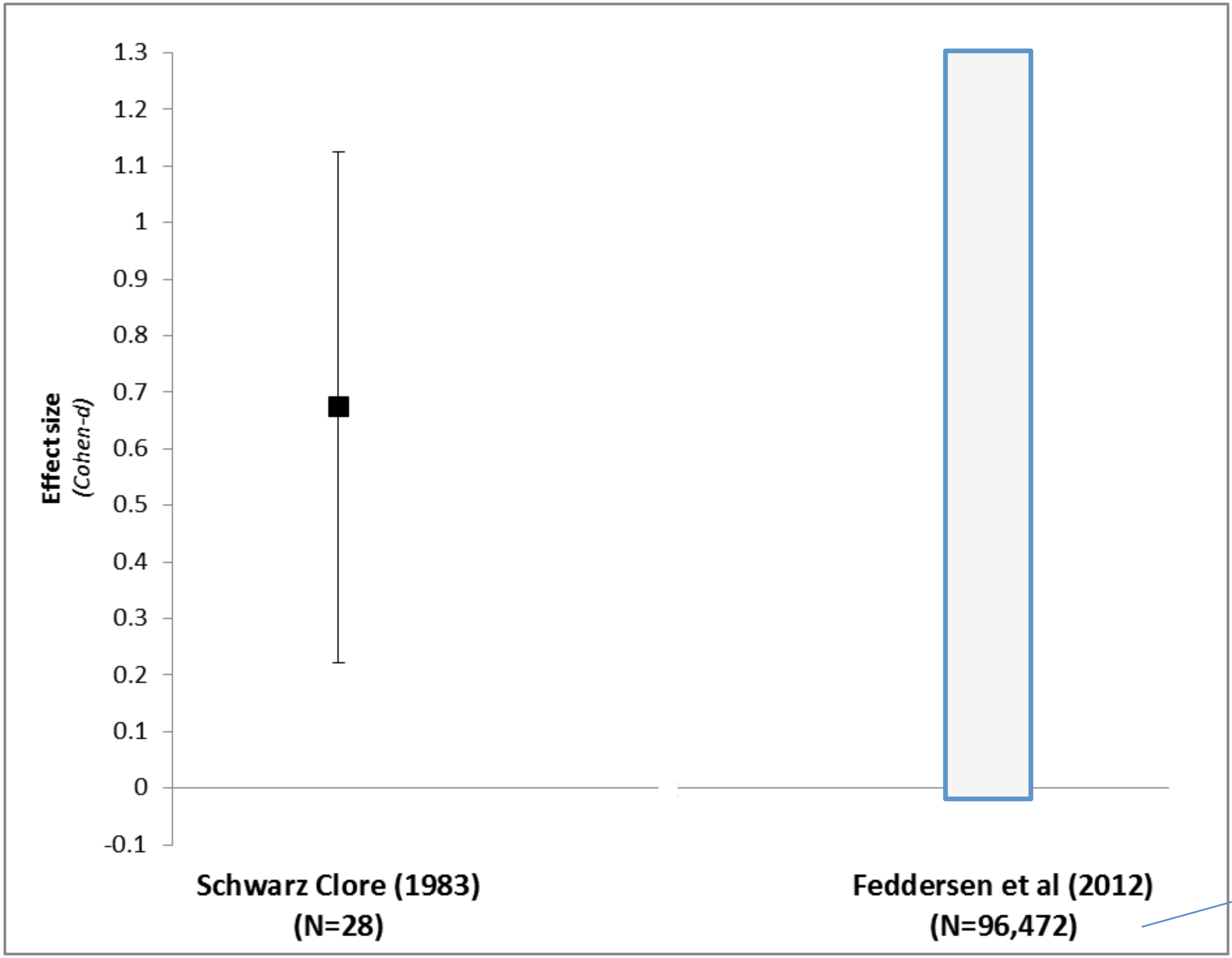
Coursey, Hoviz and Schulze

WTA/WTP = 2.6

n.s.

“Market experience eliminates endowment effect”

Example 3: Sunshine and happiness



"Despite this difference in magnitude, we do confirm Schwarz and Clore's (1983) finding that cloudiness matters." (pp.6)

10 years worth

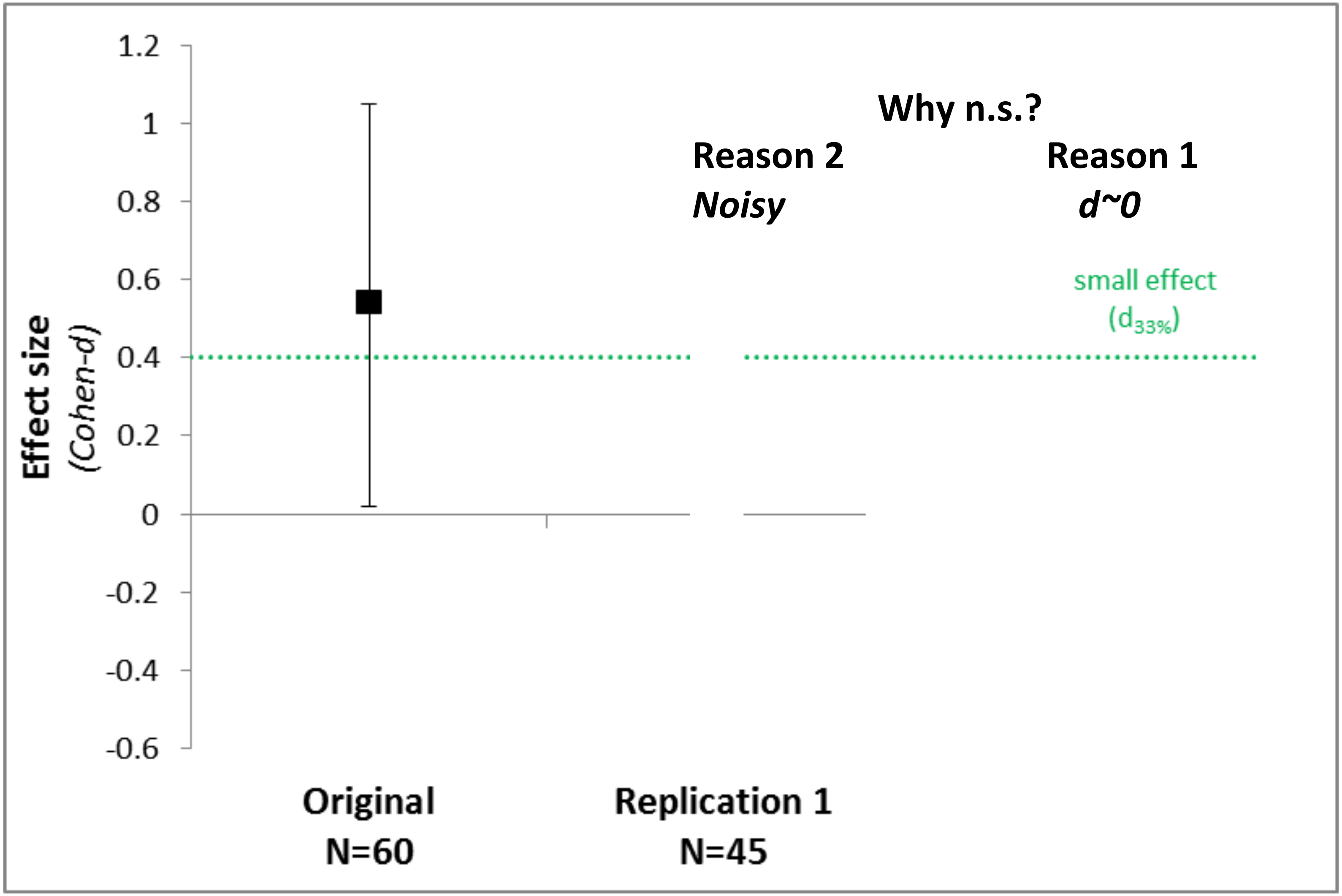
Why is a result n.s.?

- **Reason 1.** Effect is very small or 0
  - Answers Question 3
- **Reason 2.** Effect is noisily estimated
  - Does not answer Question 3
- How to distinguish?
- Test null of small effect
- Combines hypothesis testing and effect size estimation into single test.

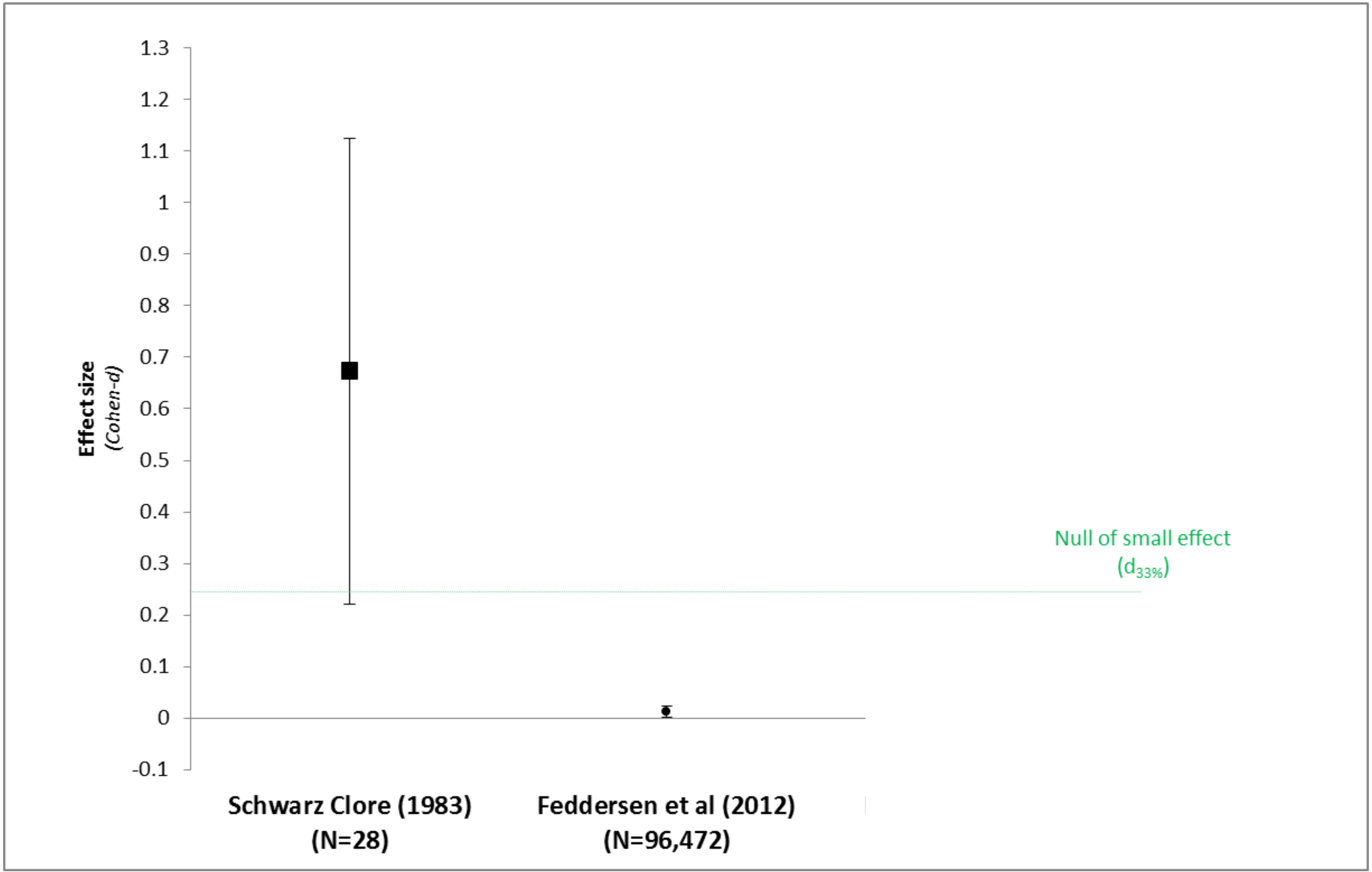
What's "small"?

- Typical answer
  - So small, we subjectively do not *care* about
    - $d<.1$
    - $R^2<5\%$
    - WTA-WTP<$1
    - <10% of people show effect
  - There's a reason we've ignored it so fa
- New answer (for replications)
  - Objectively difficult *to detect*
  - $d_{33\%}$

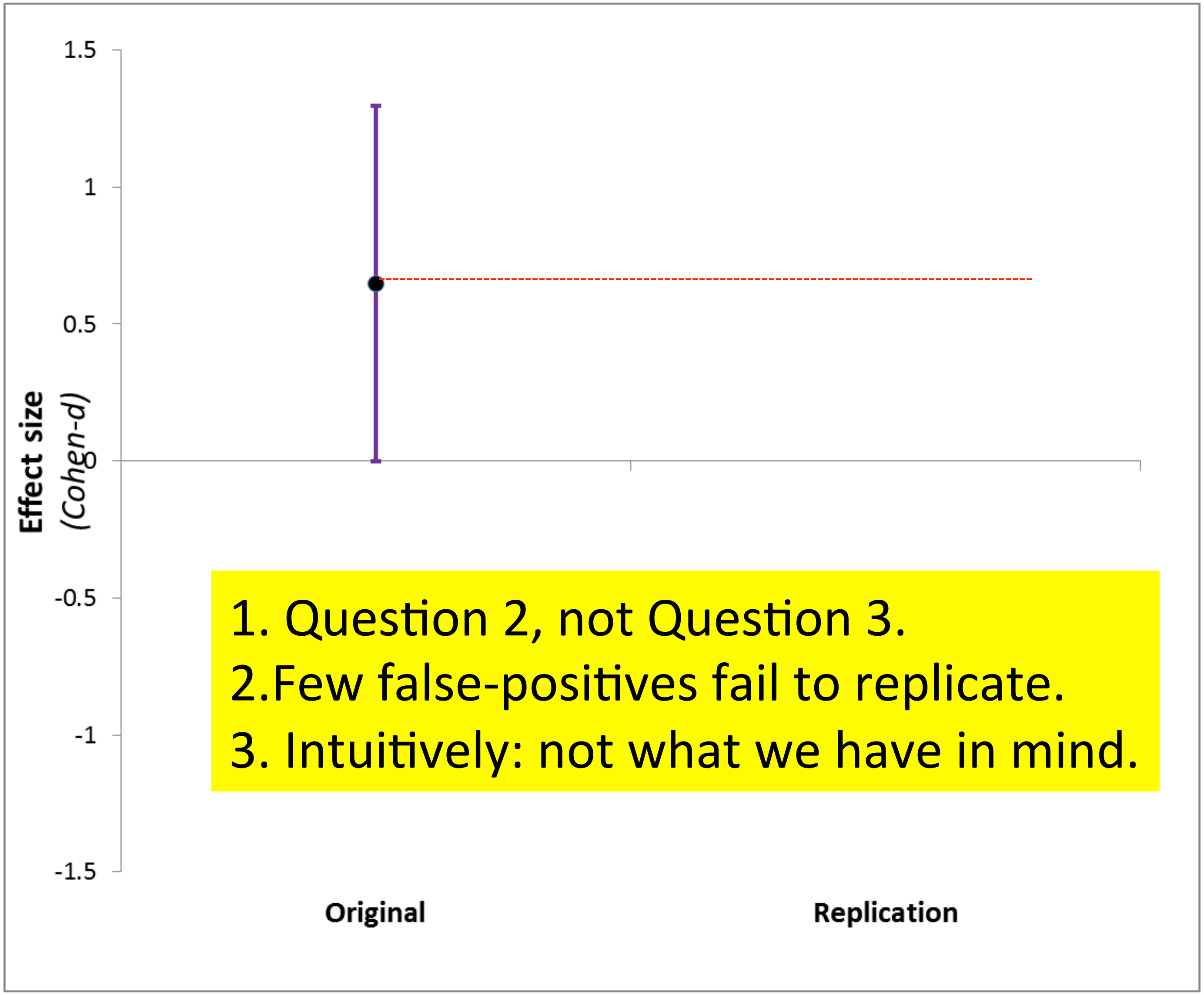
Example 1. Morality and cleanliness



Example 3. Rain and happiness



What about comparing effect size?



1. Question 2, not Question 3.

2.Few false-positives fail to replicate.

3. Intuitively: not what we have in mind.

Approach in context

- Early on, predictions are qualitative.
- “People can levitate”
  - Original: 9”
  - Replication: 0 “
  - Average is 4.5”.
  - So?
- “People can levitate”
  - Original: 9”, n=100
  - Replication: 7” n=5000
  - Replication < Original, p=.0001
  - So?

Approach in context

- Result sections aren't bumperstickers
- Report
  - Effect size
    - in d
    - $
    - %
  - Confidence intervals
  - d_{33}
  - *p*-values
- An useful contrast
- Not the only useful contrast

Paper statistics

Abstract Views:	3,933
Downloads:	688
Download Rank:	17,868
References:	42

Small Telescopes: Detectability and the Evaluation of Replication Results

Uri Simonsohn
University of Pennsylvania - The Wharton School

December 10, 2013