Issues with Scientific Evidence

Understanding Political Numbers

April 24, 2019

Science is human

...all too human

Science is human

...all too human

The *way science is done* is affected by the professional context in which researchers operate

Professional context

Academic Research

- "Publish or perish"
- What gets rewarded? Inquiry vs. accomplishment
- Statistical vs. domain expertise
- Prestige and overwork

Professional context

Academic Research

- "Publish or perish"
- What gets rewarded? Inquiry vs. accomplishment
- Statistical vs. domain expertise
- Prestige and overwork

Campaigns, advocacy, industry

- Organizations need **information** to make **decisions**
- Quest for "good enough"
- Honesty vs. advocacy
- Machine learning: performance vs interpretation

Science (ideally)

Science (ideally)

Conducting Research

- Researchers identify interesting puzzles
- Use reliable body of *scientific literature* to develop theoretical explanations
- Devise studies to test theories
- Collect and analyze data, evaluate evidence w/r/t theories

Science (ideally)

Conducting Research

- Researchers identify interesting puzzles
- Use reliable body of *scientific literature* to develop theoretical explanations
- Devise studies to test theories
- Collect and analyze data, evaluate evidence w/r/t theories

Disseminating Research

- Researchers write a study
- Peer review: study is evaluated by other experts
- Reliable studies are accepted into scientific literature
- Knowledge accumulates over time

Science ("reality bites" version)

Science ("reality bites" version)

Research is high-stakes career output

- Other researchers judging your work "interesting" is major factor in career survival
- Citations to existing science is very political (peer review)
- Studies are "low power" tests of theory
- Data analysis is biased toward favorable findings

Science ("reality bites" version)

Research is high-stakes career output

- Other researchers judging your work "interesting" is major factor in career survival
- Citations to existing science is very political (peer review)
- Studies are "low power" tests of theory
- Data analysis is biased toward favorable findings

Disseminating Research

- Non-scientific (peri-scientific?) considerations
- Peer reviewers have idiosyncratic and inconsistent opinions (low n)
- Flashy results vs. Careful methodology
- Published record is a biased

Statistical significance

p-values are useful but abused

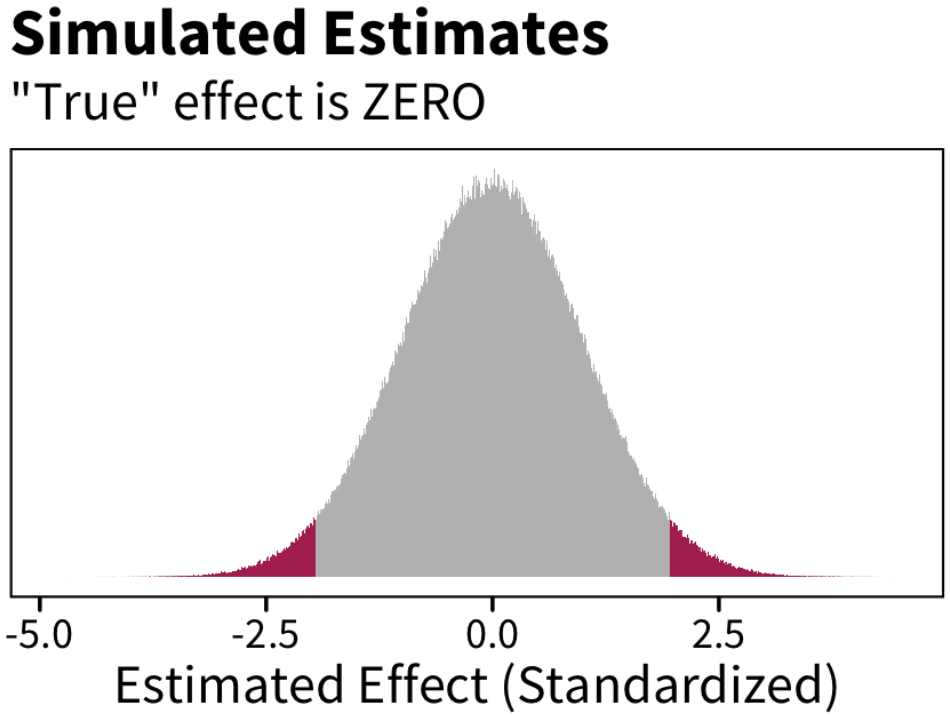
p-value: probability of a "more extreme" effect (if no true relationship)

Dichotomania: splits the world into "zero" and "non-zero" effects

p-hacking

Data analysis is a "garden of forking paths"

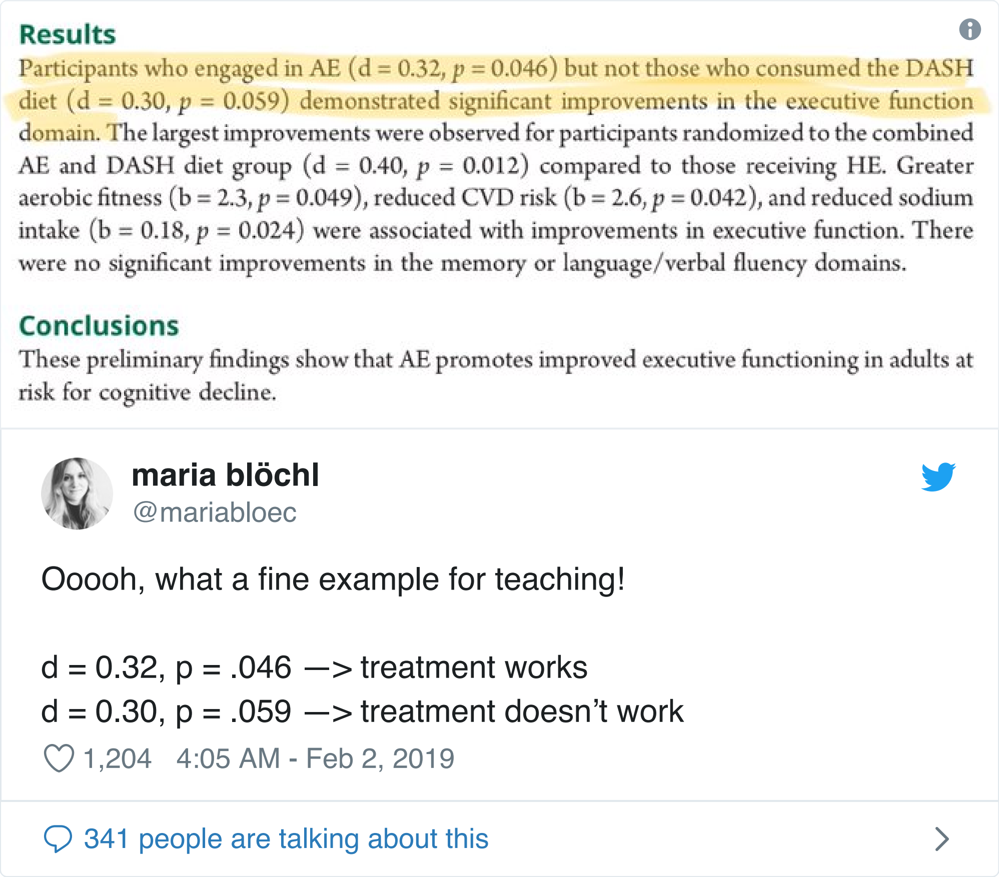
The difference between "significant" and "not significant" is not itself statistically significant



The "statistical significance filter"

> Got reviewer comments back.
>
> We report a P-values of 0.051 and 0.062. Reviewer: "If it’s not significant, it’s not significant. Delete."
>
> Here's a response I've used in the past, sharing for anyone who might find it useful pic.twitter.com/dQIxBCdsV8
>
> — Kevin Kohl (@KevinDKohl) January 29, 2019

Dichotomania



Results

Participants who engaged in AE (d = 0.32, *p* = 0.046) but not those who consumed the DASH diet (d = 0.30, *p* = 0.059) demonstrated significant improvements in the executive function domain. The largest improvements were observed for participants randomized to the combined AE and DASH diet group (d = 0.40, *p* = 0.012) compared to those receiving HE. Greater aerobic fitness (b = 2.3, *p* = 0.049), reduced CVD risk (b = 2.6, *p* = 0.042), and reduced sodium intake (b = 0.18, *p* = 0.024) were associated with improvements in executive function. There were no significant improvements in the memory or language/verbal fluency domains.

Conclusions

These preliminary findings show that AE promotes improved executive functioning in adults at risk for cognitive decline.

maria blöchl
@mariabloec

Ooooh, what a fine example for teaching!

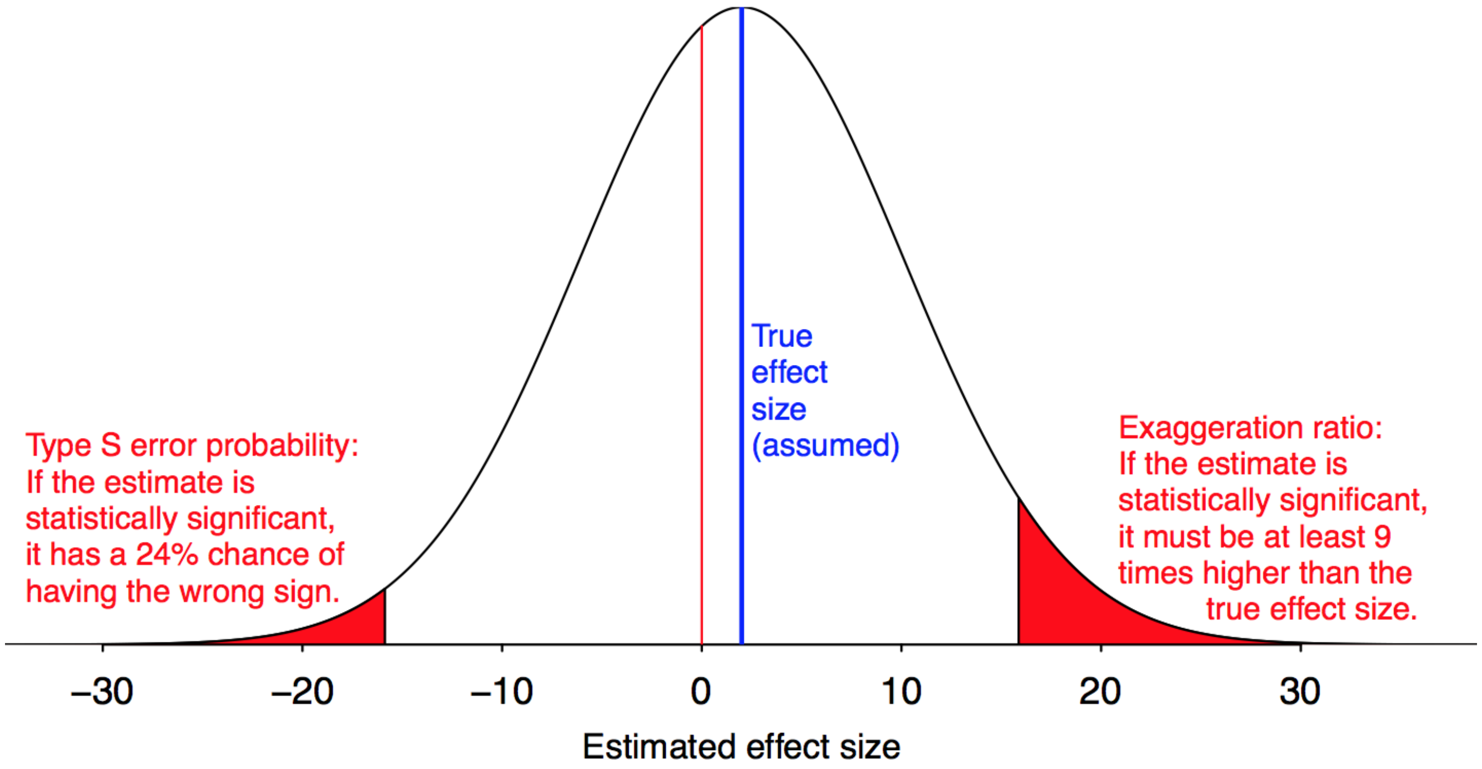
d = 0.32, p = .046 —> treatment works
d = 0.30, p = .059 —> treatment doesn’t work

♡ 1,204 4:05 AM - Feb 2, 2019

341 people are talking about this

Statistical "power"

True effects rarely zero, but need lots of data to estimate small effects



(source)

Publication Bias

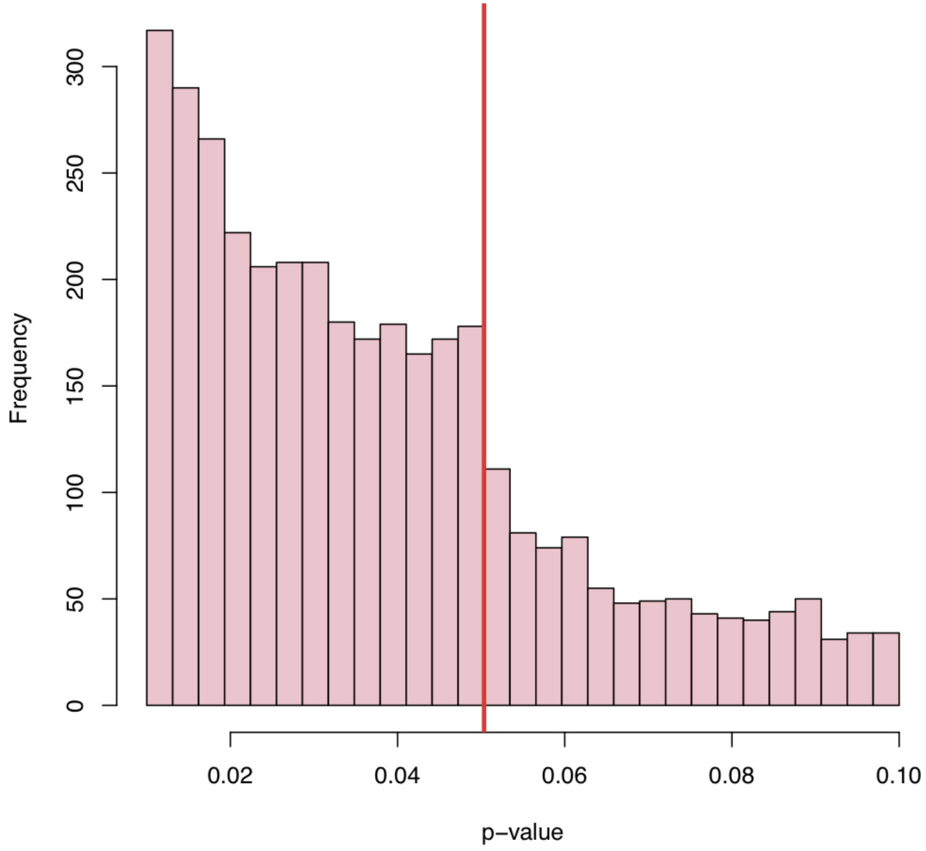
Publication bias

Most published findings are over-estimates or false (video)

File drawer problem

Replication (and failure to replicate)

Do journals care? (or hiring committees, tenure committees...?)



(source)

Very little reward for *improving the conduct* of science

Falsifying hypotheses

Are we learning from science?

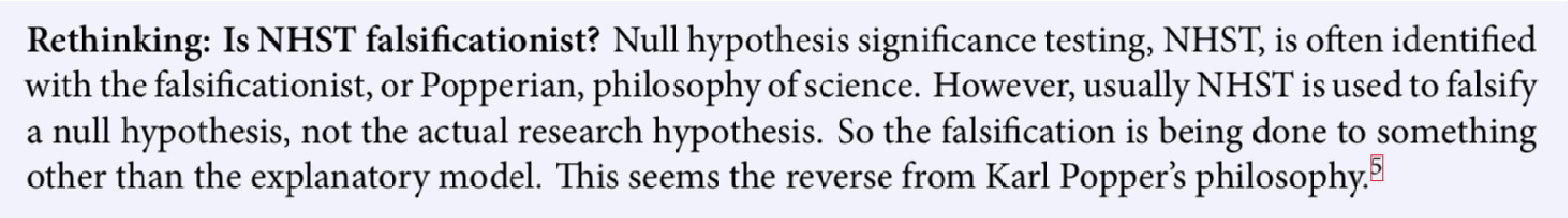
Verification vs. Falsification

Falsification and the *null hypothesis*

Reject serious, competitive hypotheses!

McElreath "Evolution of Statistical Methods" talk

© BNPS.CO.UK



Rethinking: Is NHST falsificationist? Null hypothesis significance testing, NHST, is often identified with the falsificationist, or Popperian, philosophy of science. However, usually NHST is used to falsify a null hypothesis, not the actual research hypothesis. So the falsification is being done to something other than the explanatory model. This seems the reverse from Karl Popper's philosophy.[5]

Teaching Evaluations

`aefis.wisc.edu`

Specifics are better!

Constructive is better!