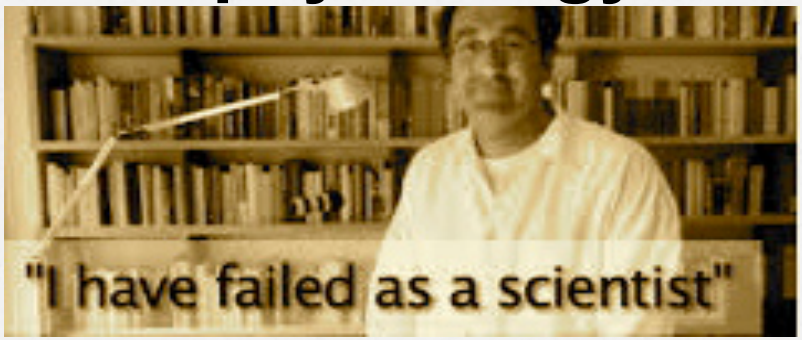
Notes on the replication revolution in psychology



- Diederik Stapel, the social psychologist who fabricated his data (2011)
- Investigating Stapel revealed a **culture of verification bias: plunder your data**
- A string of high profile cases followed, as did replication research

Stapel describing his "first time":

I was alone in my tastefully furnished office at the University of Groningen. …I opened the file with the data that I had entered and changed an unexpected 2 into a 4; then, a little further along, I changed a 3 into a 5. …When the results are just not quite what you'd so badly hoped for; when you know that that hope is based on a thorough analysis of the literature; ... then, surely, you're entitled to adjust the results just a little?…I looked at the array of data and made a few mouse clicks to tell the computer to run the statistical analyses. When I saw the results, the world had become logical again. (Stapel 2014, p. 103)

Failed Replication

- Failed replication: Results found statistically significant are not found significant when an independent group tries to replicate the finding with new subjects, and more stringent protocols
- Note on terminology p. 97

“I see a train-wreck looming,” Daniel Kahneman, calls for a “daisy chain” of replication in Sept. 2012

OSC
OPEN SCIENCE
COLLABORATION

OSC: Reproducibility Project: Psychology: 2011-15 (Led by Brian Nozek, U. VA)

Main problems are *fallacies of rejection p. 94*

- The reported (nominal) statistical significance result is *spurious* (it's not even an actual P-value). This can happen in two ways: biasing selection effects, or violated assumptions of the model.

- The reported statistically significant result is genuine, but it's an isolated effect not yet indicative of a genuine experimental phenomenon. (Isolated low P-value ≠> *H*: statistical effect)

- There's evidence of a genuine statistical phenomenon but either (i) the magnitude of the effect is less than purported, call this a *magnitude error*, or (ii) the substantive interpretation is unwarranted. ($H \not\Rightarrow H^*$)

An *audit* of a P-value: a check of any of these concerns, generally in order, depending on the inference.

So I place the background information for auditing throughout our 'series of models' representation (figure 2.3, p. 87.

Biasing selection effects:

One function of severity is to identify problematic selection effects (not all are) p. 92

- ***Biasing selection effects***: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that **the minimal severity requirement is violated, severity is seriously altered or incapable of being assessed**

- Replication researchers (re)discovered that data-dependent hypotheses and stopping are a major source of spurious significance levels.

- Statistical critics, Simmons, Nelson, and Simonsohn (2011) place at the top of their list the need to block flexible stopping

"Authors must decide the rule for terminating data collection before data collection begins and report this rule in the articles"
(Simmons, Nelson, and Simonsohn 2011, 1362).

- "[W]e need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable which will rarely fail to give us a statistically significant result." (Fisher 1935, 14) **(low P-value ≠> H: statistical effect)**

- "[A]ccording to Fisher, rejecting the null hypothesis is not equivalent to accepting the efficacy of the cause in question. The latter...requires obtaining more significant results when the experiment, or an improvement of it, is repeated at other laboratories or under other conditions." (Gigerentzer 1989, 95-6) (*H ≠> H**)

SEV is applied informally

- If flaws in the substantive alternative *H** have not been probed by, the inference from a statistically significant result to *H** fails to pass with severity
- Generally goes beyond statistics
- Largely ignored in today's replication research

People may want to believe claims (for political or ethical reasons)

- Diederik Stapel says he always read the research literature extensively to generate his hypotheses.

- *"So that it was believable and could be argued that this was the only logical thing you would find*." (E.g., eating meat causes aggression.)
- (In "The Mind of a Con Man," NY Times, April 26, 2013[4])

To return to the OSC: Reproducibility Project:

- Crowd sourced: Replicators chose 100 articles from three journals (2008) to try and replicate using the same method as the initial research: **direct replication**

Does a negative replication mean the original was a false positive?

- Preregistered, avoid P-hacking, designed to have high power
- Free of “perverse incentives” of usual research: guaranteed to be published

But there may also be biases

- But might they be influenced by replicator's beliefs? (Kahneman emphasized getting approval from original researcher)
- Subjects (often students) often told the purpose of the experiment
- Other COIs, p.99

- One of the non-replications: cleanliness and morality: *Do cleanliness primes make you less judgmental?*

"*Ms. Schnall had 40 undergraduates unscramble some words.* ***One group unscrambled words that suggested cleanliness*** (pure, immaculate, pristine), while the **other group unscrambled neutral words.** ***They were then presented with a number of moral dilemmas, like whether it's cool to eat your dog after it gets run over by a car****."*

Turns out it did. Subjects who had unscrambled clean words weren't as harsh on the guy who chows down on his chow."
(Bartlett, *Chronicle of Higher Education*)

- By focusing on the P-values, they ignore the larger question of the methodological adequacy of the leap from the statistical to the substantive.
- Are they even measuring the phenomenon they intend? Is the result due to the "treatment"?

Nor is there discussion of the multiple testing in the original study

- Only 1 of the 6 dilemmas even in the original study showed statistically significant differences in degree of wrongness–not the dog one
- No differences on 9 different emotions (relaxed, angry, happy, sad, afraid, depressed, disgusted, upset, and confused)
- Many studies are coming into experimental philosophy: philosophers of science need to critique them
- Replicators are blocked by pushback, repligate

Researcher degrees of freedom in the Schnall study

After the priming task, participants rated six moral dilemmas : ‘‘Dog’’ (eating one’s dead dog), ‘‘Trolley’’ (switching the tracks of a trolley to kill one workman instead of five), ‘‘Wallet’’ (keeping money inside a found wallet), ‘‘Plane Crash’’ (killing a terminally ill plane crash survivor to avoid starvation), ‘‘Resume’’ (putting false information on a resume), and ‘‘Kitten’’ (using a kitten for sexual arousal). Participants rated how wrong each action was from 0 (perfectly OK) to 9 (extremely wrong).

Replication revolution won't be revolutionary enough until…

- Methods and measurements are tested (bott p. 100)
- many may be falsified

We should worry much more about

- Links from experiments to inferences of interest
- I've seen plausible hypotheses poorly tested
- Macho men p. 101-4