Replicability crisis in science? A View from Philosophy of Science

---

Branden Fitelson
*v*2 (09/19/23)

1up version of slides

4up version of slides

- Most studies in various sciences have *low power* [probability of a "positive" result R, *if* there is a true effect P: $\Pr(R \mid P)$], despite the vast prevalence of "positive" results in those sciences.
  - Sedlmeier & Gigerenzer [87] estimate the median power of psych. experiments (in *JAP*) were 0.47 in 1960 and 0.37 in 1984.
  - Fraley & Vazire [28] estimate that, in social-personality psychology (across a large swath of journals from 2006–2010), the median power to detect the typical effect size was ≈ 0.5.
  - Fanelli [24] provides evidence that "positive" (published) results are 2-5 times more likely in "soft" sciences *vs.* "hard" sciences — and the proportion of "positives" is often *much* greater than 50%.
- The surprising number of "positive" results (and lack of "negative" results [23]) is one source of concern (as we'll see, this was anticipated by some researchers in the 1970's).
- These concerns seemed to be confirmed by later attempts to replicate various experiments in various fields.

- Bargh, Chen, and Burrows's [4] social priming study in which, *e.g.*, subjects primed with elderly stereotypes were perceived to walk more slowly when exiting the lab [88].
- Bem's [6] ESP studies, which claimed to generate "positive" results supporting the existence of subjects with ESP [79].
- These sorts of cases raised concerns about the reliability of the research methodologies being employed (esp in psych research).
- Larger scale studies began to be performed.
- Nosek *et al* [67] attempted to replicate 100 contemporary psychology experiments, 97 of which had reported "positive" results. Only 36 of these replicated (with *half* the effect sizes).
- Scientists in biotech companies [5, 69] were only able to replicate less than 20% of studies in preclinical research (oncology).
- Similar studies have been performed in other fields [13].

- There does seem to be a problem (call it "the crisis"). Supposing that there is, what might be some good *explanations* of it?
- **Fraud** is *sometimes* the best explanation of *individual* cases of unreliability. There have been some high-profile cases recently.
  - The president of Stanford recently resigned and retracted a couple of scientific papers, after allegations of fraud [2].
  - Two famous "honesty" researchers in psychology have also recently retracted papers, after allegations of fraud [72, 57].
- Be that as it may, I suspect that fraud is *not* a primary cause (or explanation) of "the crisis." I'd guess it's $< 10\%$ of cases [75, 34].
- I would like to focus on another set of causes (or explanations): **Questionable Research & Publication Practices** (QRPPs).
- And, to make things more tractable, I will take a more specific *explanandum*: the fact that *there seem to be far more "positive" results published* than one would reasonably expect.

- Dorothy Bishop's [9] "Four Horsemen" of the Replication Crisis.
  - **Publication (& Citation) Bias (& Spin)** (and their interactions)
    - Publication bias: journals tend to publish far fewer studies with "negative" results than ones with "positive" results.
    - Citation bias: people tend to cite their own work, and the work of others that confirms theirs (it's a kind of confirmation bias [100]).
    - Spin: misleading/exaggerated claims in abstracts.
  - p-**Hacking** (and **outcome reporting bias**)
    - Various things that can make your finding of a "significant" or "positive" result (R wrt P) less ***epistemically*** *significant*.
  - **Low Statistical Power**
    - Problems arising from the fact that $\Pr(R \mid P) = 1 - \beta$ is often low.
  - **HARKing** (**H**ypothesizing **A**fter **R**esults are **K**nown [52])
    - Look at *a lot* of data, pluck out a "significant" finding R for an "exciting" P and write a paper to tell a story around it (as if P were of initial interest to you, and "E" were designed to test it).

- The first potential explanation I will discuss has been known for decades: **Publication Bias** (*a.k.a.*, The "File Drawer Problem").
- In 1975, Greenwald [35] hypothesized that there was a "bias against the null hypothesis." He conjectured that

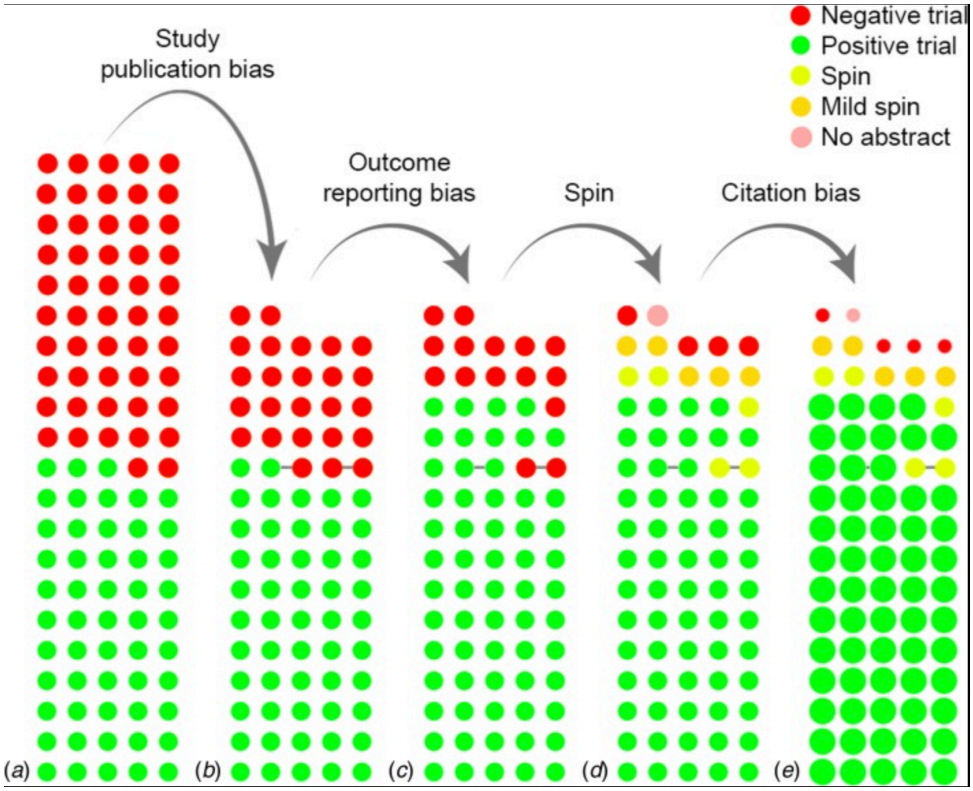
  > there may be relatively few publications on problems for which the null hypothesis is . . . true, and of these, a high proportion will erroneously reject the null hypothesis.

- In 1979, Rosenthal [82] described a "worst-case scenario" of publication bias in the following (eerily prescient) way.

  > For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results.

- Publication bias and citation bias can be mutually reinforcing.

- de Vries *et al* [19] argue that publication and citation bias have interacted (with outcome reporting bias and spin) to make "positive" trials far more prevalent (anti-depressant trials).



- ***p*-Hacking** (and **outcome reporting bias**) are umbrella terms for *things that make the fact that you reported a "significant" R (which is supposed to support P) less* ***epistemically*** *significant.*

- Some specific examples of ***p*-Hacking** include [92, 36].

  - conducting analyses midway through experiments to decide whether to continue collecting data
  - recording many response variables and deciding which to report after the analyses are done
  - deciding whether to include or drop outliers post-analyses
  - excluding, combining, or splitting treatment groups post-analysis
  - including or excluding covariates post-analysis
  - stopping data exploration if an analysis reaches "significance"

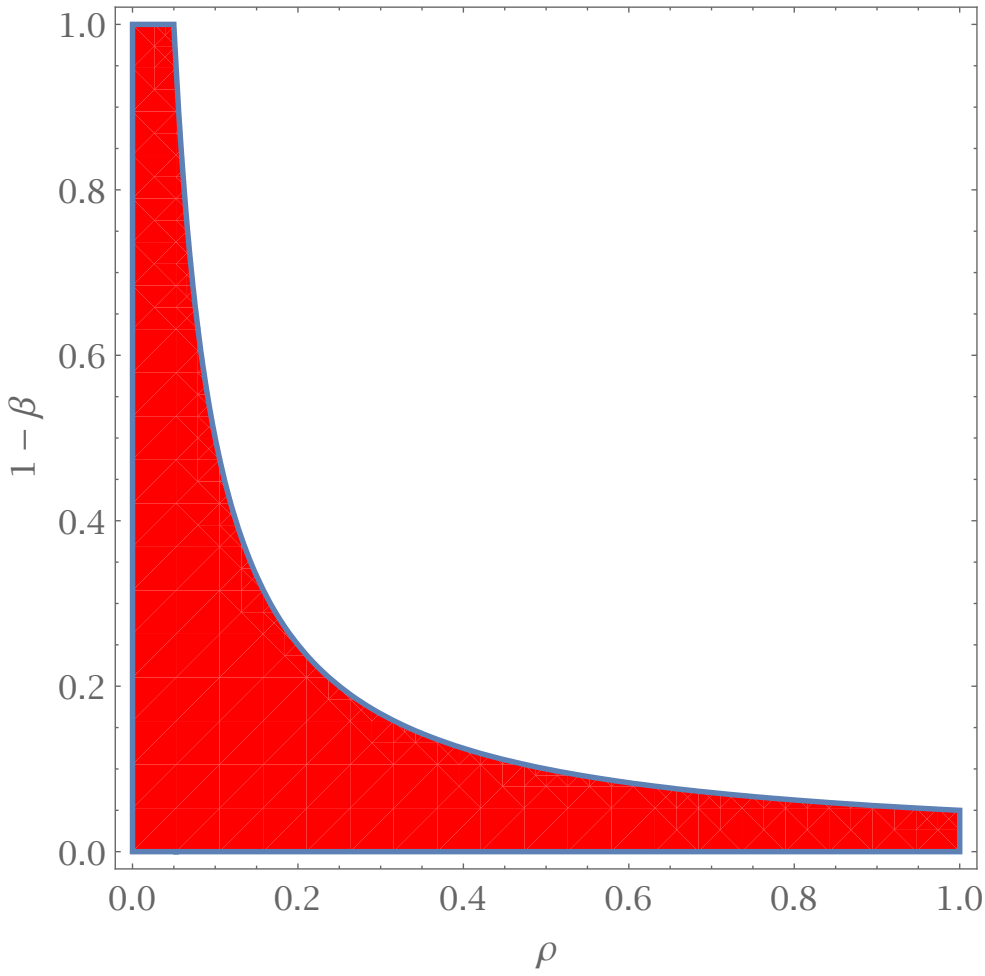
- Many studies in various sciences have low power [12, 20]. In some circumstances, low power can make it probable that published "positive" results are misleading [96, 44].
- Let $R \stackrel{\text{def}}{=}$ a "positive/significant" result is generated by E (wrt P), $\neg P \stackrel{\text{def}}{=}$ the "null hypothesis" (P asserts that there *is* a true effect).

	$\neg P$	P
$\neg R$	$\Pr(\neg R \mid \neg P) = 1 - \alpha$	$\Pr(\neg R \mid P) = \beta$
R	$\Pr(R \mid \neg P) = \alpha$	$\Pr(R \mid P) = 1 - \beta$

- If ρ is the ratio of "true relationships" to "no relationships" among those tested in the field [$\rho \stackrel{\text{def}}{=} \frac{\Pr(P)}{\Pr(\neg P)}$], then $\Pr(P) = \frac{\rho}{\rho+1}$.
- And, $\Pr(P \mid R)$ (*a.k.a.*, PPV) can be calculated *via* Bayes's Theorem.

$$\Pr(P \mid R) = \frac{\rho(1-\beta)}{\rho - \beta\rho - \alpha} < \frac{1}{2} \text{ iff } (1-\beta)\rho < \alpha$$

- We can plot this *misleading region* of $\langle \rho, 1-\beta \rangle$-space ($\alpha = 0.05$).


1.0
0.8
0.6
0.4
0.2
0.0
$1-\beta$
0.0
0.2
0.4
0.6
0.8
1.0
ρ

- HARKing [52] involves deciding what hypotheses ($\neg P$) you are "testing" *after* a bunch of data have been generated, and a "significant" effect (R, for some P) has been discovered.
- *Mere* HARKing — especially if done *secretly* (SHARKing [41]) — is a questionable research practice, *if* one's aim is to provide *confirmation* or *justification* for P (*using* R).
- Here, a classic philosophical distinction between the *context of justification* and the *context of discovery* [18] can be useful.
- Philosophers (and methodologists) have done more systematic work on the nature of justification/confirmation [16, 97].
- Scientific discovery [89] is receiving more attention lately (in philosophy and science). Specifically, *exploratory data analysis* can certainly be a useful tool in the process of discovery [43].
- But, HARKing can be epistemically problematic — especially if it is used non-transparently in the context of justification [41, 39].

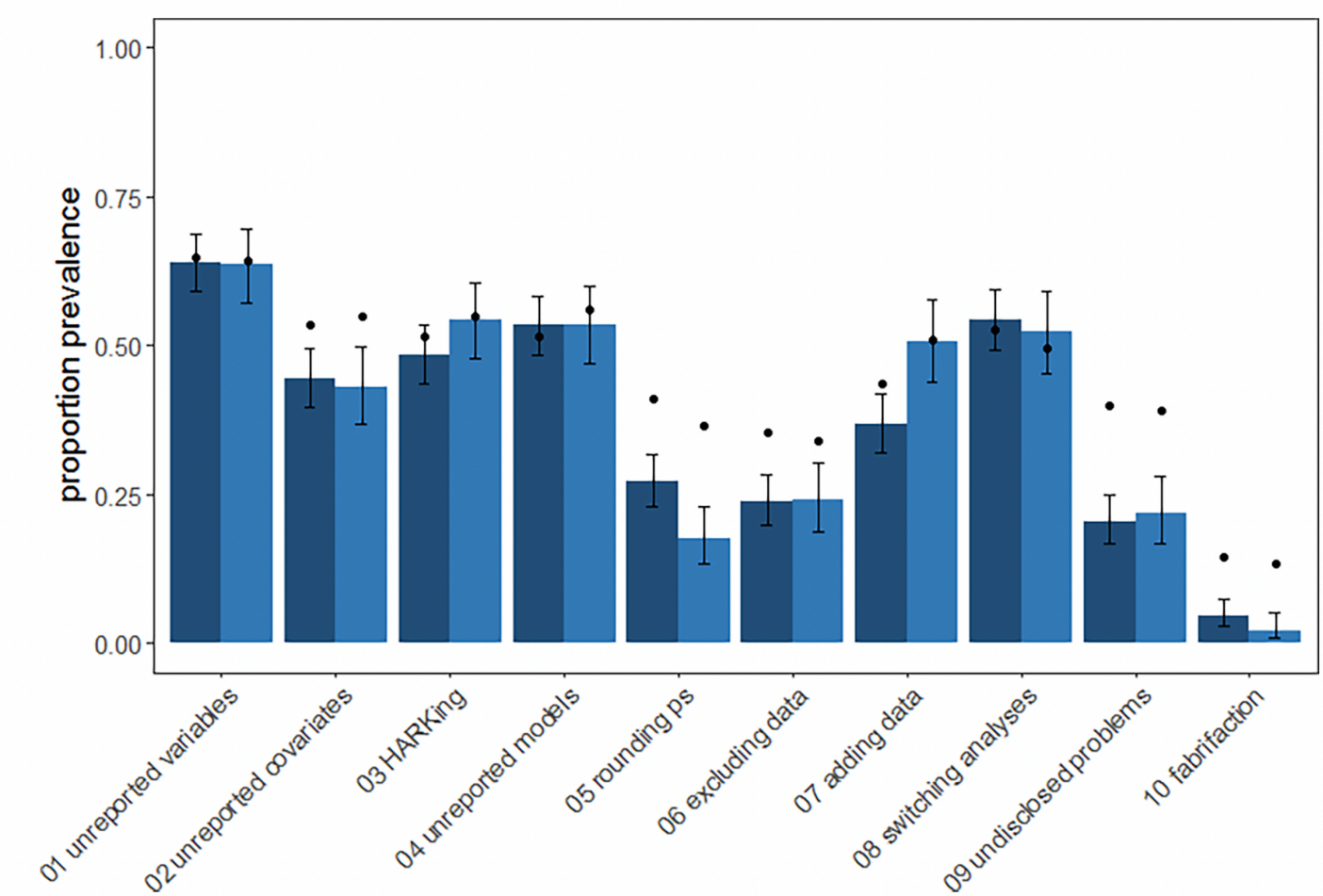
- Here are survey results from 500 US psychologists in 2012 [48].

Item	Self-admission rate (%)		Odds ratio (BTS/control)	Two-tailed p (likelihood ratio test)	Defensibility rating (across groups)
	Control group	BTS group			
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5	1.14	.23	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0	1.08	.46	1.79 (0.44)
3. In a paper, failing to report all of a study's conditions	27.7	27.4	0.98	.90	1.77 (0.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5	1.57	.00	1.76 (0.48)
5. In a paper, "rounding off" a p value (e.g., reporting that a p value of .054 is less than .05)	22.0	23.3	1.07	.58	1.68 (0.57)
6. In a paper, selectively reporting studies that "worked"	45.8	50.0	1.18	.13	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4	1.23	.06	1.61 (0.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0	1.45	.00	1.50 (0.60)
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	4.5	1.52	.16	1.32 (0.60)
10. Falsifying data	0.6	1.7	2.75	.07	0.16 (0.38)

- . . . *vs* 220 Italian psychologists in 2014 [1].

	US		Italian Association of Psychology	
QRP	**Self-admission rate (*M*)**	**95% CI**	**Self-admission rate (*M*)**	**95% CI**
1. In a paper, failing to report all of a study's dependent measures	63.4 (486)	59.1–67.7	47.9 (219)	41.3–54.6
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9 (490)	51.5–60.3	53.2 (222)	46.6–59.7
3. In a paper, failing to report all of a study's conditions	27.7 (484)	23.7–31.7	16.4 (219)	11.5–21.4
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6 (499)	12.4–18.8	10.4 (221)	6.4–14.4
5. In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .054 is less than .05)	22.0 (499)	18.4–25.7	22.2 (221)	16.7–27.7
6. In a paper, selectively reporting studies that "worked"	45.8 (485)	41.3–50.2	40.1 (217)	33.6–46.6
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2 (484)	33.9–42.6	39.7 (219)	33.3–46.2
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0 (489)	23.1–30.9	37.4 (219)	31.0–43.9
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0 (499)	1.5–4.5	3.1 (223)	0.9–5.4
10. Falsifying data	0.6 (495)	0.0–1.3	2.3 (220)	0.3–4.2

- 800 Australian ecology/evolution researchers in 2016 [29].



- 7000 Dutch researchers in 2020 [34].

		Disciplinary field			
QRP	**Description (In the last three years.)**	**Life and medical sciences**	**Social and behavioural sciences**	**Natural and engineering sciences**	**Arts and humanities**
Any frequent QRP	Score 5, 6 or 7 on at least 1 of the 11 QRPs	55.3 (53.4, 57.1)	50.2 (48.0, 52.5)	49.4 (46.8, 52.0)	42.1 (38.3, 46.1)
Fabrication	Making up of data or results	5.5 (3.2, 7.7)	4.8 (2.2, 7.5)	2.5 (0, 5.5)	0.7 (0, 5.1)
Falsification	Manipulating research materials, data or results	4.9 (2.7, 7.2)	2.0 (0, 4.6)	5.3 (2.2, 8.4)	6.1 (1.4, 10.9)
Any FF	Fabrication and/or Falsification	10.4 (7.1, 13.7)	5.7 (1.8, 9.5)	7.6 (3.1, 12.1)	8.4 (1.6, 15.3)

- So, a major *proximal* cause of the "crisis" would seem to be questionable research & publication practices (QRPPs).
- But, what are the *distal* causes/explanation? That is, why are researchers and journals engaging in QRPPs in the first place?
- Various distal explanations have been proposed.
  - **Perverse incentives** for actual scientists. Specifically, incentives for productivity, eminence, and influence [25, 47, 11, 59, 38].
    - *p*-hacking, *etc.*, requires *fewer resources*
    - encourages simpler, media-friendly narratives
    - incentivizes bold, surprising, attention-grabbing hypotheses [95]
    - leads to fewer negative results
  - **The structure of peer-review**, *e.g.*, peer-review happens *after* results have been produced, analyzed, finessed, reported [66].
  - **Journal editorial practices** above & beyond peer review [11, 30].
    - "encouragement by journals to publish negative results,"
    - "notion in editors' minds that findings should be unexpected."

- Three kinds of remedies have been proposed: **social reforms**, **methodological reforms**, and **statistical reforms**.
- Under the category of **social reforms**, we have (*inter alia*):
  - **education** (in, *e.g.*, statistical inference and methodology [85, 30])
  - **incentives for replication & confirmatory research** [54, 78]
- Proposed **methodological reforms** have included:
  - **pre-registration** of studies and their data analysis plan [66]
    - + **reforming peer review** so that it takes place at the stage of *proposals of questions and experimental methodologies for investigating them*, rather than "post-analysis" [66]
  - **transparency** — sharing *all* data — for both "successful" and "failed" studies [64]. As Jacot said way back in 1937 [46]

    All data should be published. One of the fundamental characteristics of scientific work is that it is so conducted that it may be repeated or checked up — gone over by others as often as necessary and with whatever variations may be deemed advisable.