Replicability crisis in science? A View from Philosophy of Science

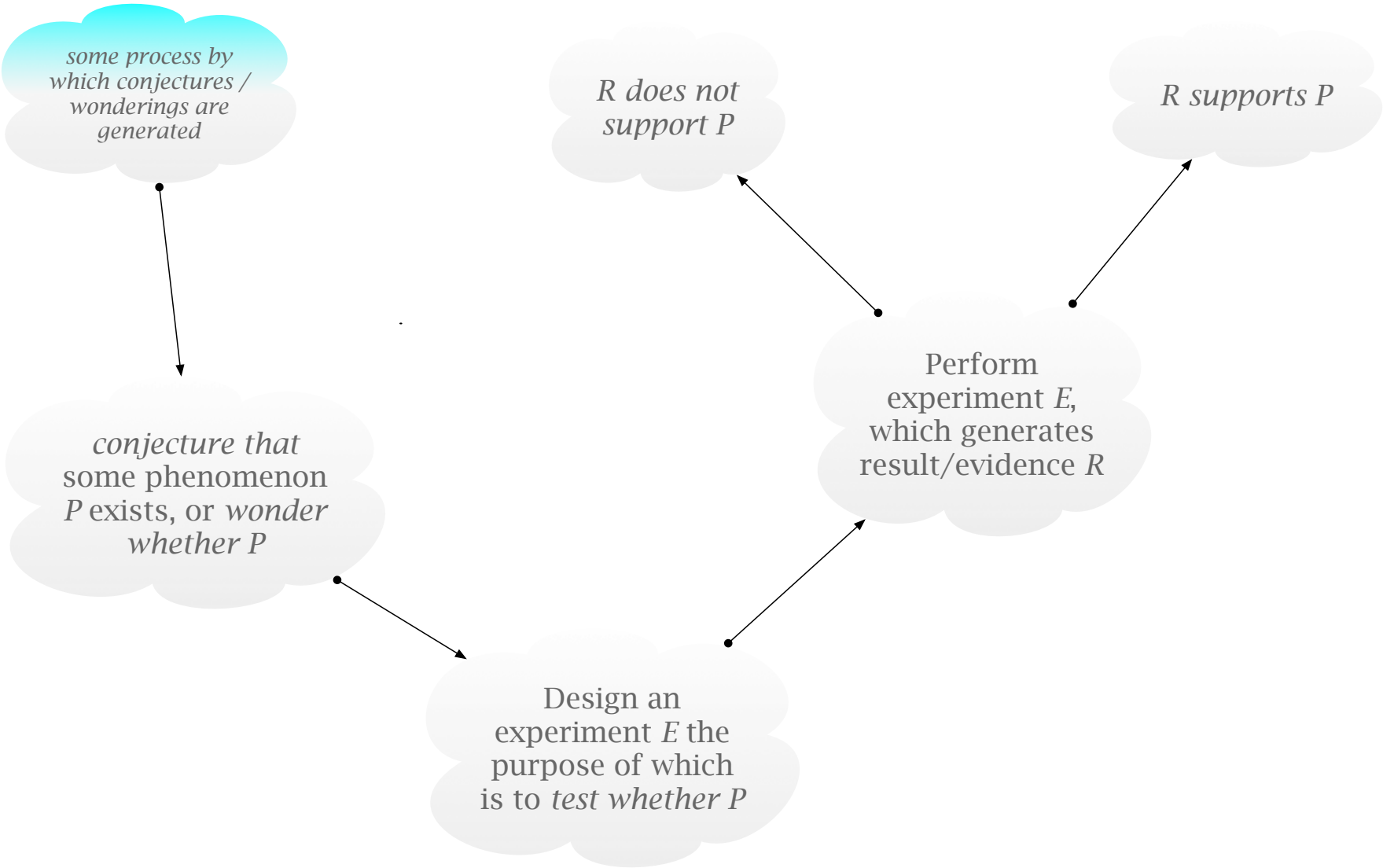
---

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

1up version of slides

4up version of slides

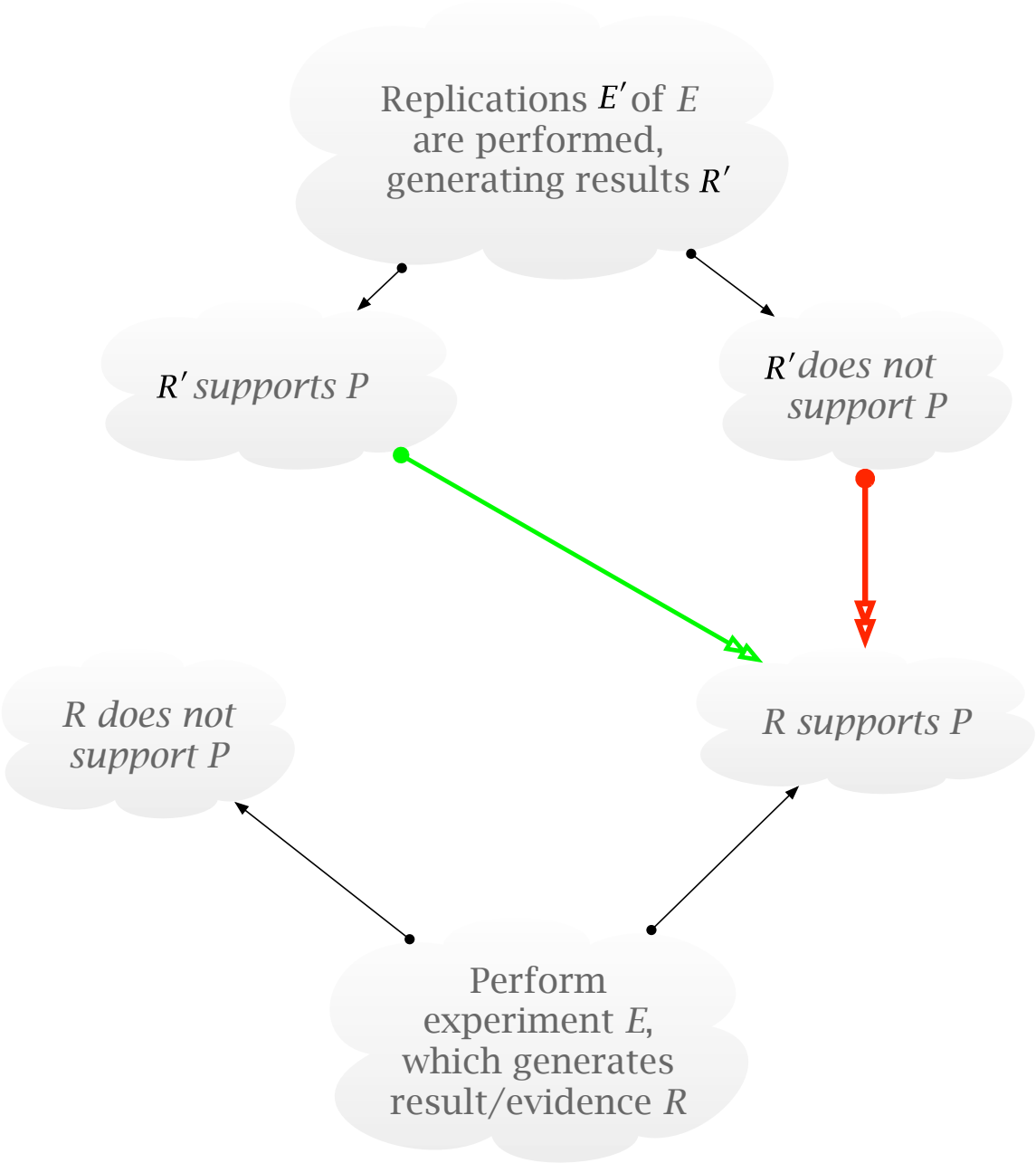
- In these lectures, I will explore how (some) philosophers of science have been thinking about the so-called "replicability crisis" in science [77]. But, first, I will take a step back.
- Science is a human activity. So, how it is *actually* practiced is *itself* a matter of (social/psychological) science — sometimes called "science of science" [56, 27, 65, 60].
- *Philosophers of* science are interested in both how science *is* practiced (**descriptive** philosophy of science) and how it *ought to be* practiced (**normative** Φ of science) [32].
- I have been primarily interested in the normative questions. Mainly, I've worked on understanding the *justification/confirmation* [16] of scientific hypotheses.
- In these lectures, that will also be my main focus. But, later on, I will also talk about the *discovery* of scientific hypotheses [89]. I will start with a "big picture."


some process by which conjectures / wonderings are generated
conjecture that some phenomenon P exists, or wonder whether P
Design an experiment E the purpose of which is to test whether P
Perform experiment E, which generates result/evidence R
R does not support P
R supports P

- Let's begin with a first approximation of replication's (or the lack thereof's) *epistemic* significance in science.
  - Suppose an experiment E produces a result R, which is claimed to *support* P (think of $\neg P$ as the "null hypothesis" and think of R as a "significant" result wrt $\neg P$).
  - R will count as supporting P *only if* replication(s) E' of E reliably produce results R' which also support P.
  - Suppose E produces a "significant" result R, which is claimed to *support* P (*i.e.*, a "positive" result). *But*

    (D) Replications E' of E tend to produce results R', which do *not* support P.
  - Then, D is a *defeater* [71] of R's support for P.

  ☞ *Failures* of replication serve as *defeaters of the evidential support of (the results of) experiments*.
- In the next section, I will outline a simple Bayesian theory of evidential support and defeat, which I will use throughout.


Replications E' of E are performed, generating results R'
R' supports P
R' does not support P
R does not support P
R supports P
Perform experiment E, which generates result/evidence R

- We will also improve upon this approximation, by saying more about what experimentation and replication *are*. But, first, some ramifications of failures of replication.
- Generally, if one discovers a defeater of R's support for P, this should cause one to *discount* R as a reason to believe P.
- *Some* failures of replication are inevitable (fallibilism [37]).

☞ But, if most (or many) experiments in a discipline fail to replicate, then this (justifiably) *undermines our belief in the reliability of that experimental discipline*.

- And, if many disciplines experience widespread failures of replication, then this would (rightly) undermine our belief in the reliability of experimental science, generally.
- The ramifications of *that* would be myriad (and not good). In that case, we would plausibly be facing a *crisis of confidence* in empirical science as a producer of knowledge.

- I will adopt a broadly Bayesian way of thinking about evidential support (or confirmation [16, 94]) & defeat [55].
- I will suppose that (at least, ideal) agents come equipped with *probabilistic degrees of belief* (or *credences*) [94].
- Specifically, we will suppose that our scientist has a probabilistic credence function $\Pr(\cdot)$, which reflects their (unconditional) degrees of confidence (in some context C).
- We'll also assume our scientist has *conditional* (or *suppositional*) probabilistic credences $\Pr(\cdot \mid X)$, which reflect their degrees of confidence *on the supposition that X is true*.
- With these conventions in place, we can explicate "support."
  - R **supports** P iff $\Pr(R \mid P) > \Pr(R \mid \neg P)$.
- We can also define measures of *degree of support* (or confirmation). *E.g.*, the likelihood ratio (or Bayes factor) $l(P,R) = \frac{\Pr(R|P)}{\Pr(R|\neg P)}$ is often used for this purpose [33, 81].

- With these basic conventions in hand, we can explicate two notions of evidential defeat: rebutting and undercutting.
- Suppose we are in a case in which R supports P (for S in C).
  - D is an **undercutting defeater** of R's support for P iff
    $$\Pr(R \mid P \,\&\, D) \leqslant \Pr(R \mid \neg P \,\&\, D).$$
  - D is a **rebutting defeater** of R's support for P iff
    $$\Pr(P \mid R \,\&\, D) < t, \text{ for some threshold } t.$$
- Here is an example to illustrate the difference.
  - A forecaster has (R) predicted that (P) it will rain on Tuesday. The forecaster is highly reliable [$\Pr(R \mid P) = 0.9$, $\Pr(R \mid \neg P) = 0.1$], and Tuesday rain is a 50/50 bet. Therefore, $\Pr(P \mid R) = 0.9$. Consider these two defeaters:

    (D_1) The forecaster is actually a fair coin (heads R, tails $\neg R$).

    (D_2) It will not rain on Tuesday (*viz.*, $\neg P$).

    - $\Pr(R \mid P \,\&\, D_1) = 0.5 = \Pr(R \mid \neg P \,\&\, D_1)$.
    - $\Pr(P \mid R \,\&\, D_2) = 0$.

- For our purposes, it is undercutting defeat that is most important. Failures of replication do not (generally) rebut experimental findings, they merely undercut them.
- Here is a whimsical (but true) example, which also gives a sense of how difficult it can be to diagnose the ultimate explanations of failures of replication [68, 99, 98].

> In radio astronomy, a fast radio burst (FRB) is a transient radio pulse of length ranging from a fraction of a millisecond to 3 seconds, caused by some high-energy astrophysical process not yet understood. The first FRBs were discovered in 2007. In 2015, it was believed that a new kind of FRB had been discovered at Parkes Observatory. But, upon closer inspection, there was something different about these radio bursts. Failures to detect them at other observatories led, ultimately, to an investigation which discovered that it was the microwave ovens in the observatory that were causing the observed radio bursts. Such man-made bursts have become known as "perytons."

Edouard Machery
Department of History & Philosophy of Science
University of Pittsburgh

- Before getting to what a replication of an experiment is, we need to have a general, working notion of "experiment." Here, I will follow Machery's general approach [58].
- A **token experiment** E (of **type** $\mathcal{E}$) is a sequence of events $e_1, \ldots, e_j$ brought about in order to produce data (R) relevant to inferring the reality of a phenomenon (P), where the sequence $E = e_1, \ldots, e_j$ belongs to $\mathcal{E} = \mathcal{E}_1, \ldots, \mathcal{E}_j$.
- Two token experiments E and E' are **equivalent** if and only if their constitutive sequences $e_1, \ldots, e_j$ and $e'_1, \ldots, e'_j$ fall under the same sequence type $\mathcal{E} = \mathcal{E}_1, \ldots, \mathcal{E}_j$.

☞ Event types can be individuated more or less coarsely, and depending on how these types are individuated two token experiments count or fail to count as equivalent.

  - *e.g.*, an experiment collects data from 100 participants from a given population. The corresponding type could specify the exact number of participants, or (more coarsely) just that some participants were sampled from the population.

- An **experimental component** is an aspect of an experiment that can be independently modified. Psychologists, *e.g.*, distinguish four different experimental components.
  - **experimental units**,
  - **treatments** (*i.e.*, "independent variables"),
  - **measurements** (*i.e.*, "dependent or response variables"), and
  - **settings**
- A treatment is an exogenous cause that changes the state of some aspect of the experimental units.

☞ Scientists aim to determine whether and how this change influences some other aspect of the experimental units.

- When an experiment has several conditions (*e.g.*, drug *vs.* placebo), the treatment can be in one of several states; psychologists often say that it has several "levels."
- Typically, participants assigned to different levels of the treatment are presented with different stimuli.
- An experiment can involve several treatments.

- Two treatments are **crossed** when measurement happens for each combination of the levels of these two treatments.
- They are **nested** when measurement happens for some combinations of the levels of these two treatments.
- Measurement is a causal interaction with the experimental units aimed at determining what state a particular aspect of the experimental units is in.
- The setting is a vague and umbrella construct, which includes the identity of the experimenter/lab conducting the experiment, whether it is done online or in a lab, *etc.*

☞ Experimental components are constitutive of an experiment type's defining event types. They determine "what the experiment is" (*i.e.*, which experiments are equivalent to it).

- Psychology journals require authors to describe the experimental components that constitute their experiments, with a focus on units, treatments, and measurements.

PSYCHOLOGICAL REVIEW

VOL. 80, NO. 4 JULY 1973

ON THE PSYCHOLOGY OF PREDICTION [1]

DANIEL KAHNEMAN [2] AND AMOS TVERSKY

Hebrew University of Jerusalem, Israel, and Oregon Research Institute

Psychological Review

VOLUME 90 NUMBER 4 OCTOBER 1983

Extensional Versus Intuitive Reasoning:
The Conjunction Fallacy in Probability Judgment

Amos Tversky
Stanford University

Daniel Kahneman
University of British Columbia, Vancouver,
British Columbia, Canada

- Kahneman & Tversky wrote a series of papers in the 70's and 80's, which inspired many experiments that became exemplars of *replicable* behavioral science [21, 40, 3, 93, 17].

Psychological Review

VOLUME 102 NUMBER 4 OCTOBER 1995

How to Improve Bayesian Reasoning Without Instruction: Frequency Formats

Gerd Gigerenzer
University of Chicago

Ulrich Hoffrage
Max Planck Institute for Psychological Research

Is the mind, by design, predisposed against performing Bayesian inference? Previous research on base rate neglect suggests that the mind lacks the appropriate cognitive algorithms. However, any claim against the existence of an algorithm, Bayesian or otherwise, is impossible to evaluate unless one specifies the information format in which it is designed to operate. The authors show that Bayesian algorithms are computationally simpler in frequency formats than in the probability formats used in previous research. Frequency formats correspond to the sequential way information is acquired in natural sampling, from animal foraging to neural networks. By analyzing several thousand solutions to Bayesian problems, the authors found that when information was presented in frequency formats, statistically naive participants derived up to 50% of all inferences by Bayesian algorithms. Non-Bayesian algorithms included simple versions of Fisherian and Neyman–Pearsonian inference.

- In the mid-90's, another paper on the so-called "base rate fallacy" appeared, which also became a classic exemplar of replicable research in behavioral science [14].

- The infamous "mammogram prompt" [21] is as follows

> The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ____%

- Even experts (*e.g.* doctors) tend(ed) to give the wrong answer to this question [21]. This is an oft replicated result.
- Hoffrage & Gigerenzer wondered whether performance on this task could depend on the *format/way* in which probabilistic information is represented.
- We will use their experiment to illustrate Machery's notion of replication (and as an exemplar replicability/reliability).

- Here's how Gigerenzer & Hoffrage's experiment [31] fits into Machery's framework for characterizing experiments.
  - **experimental units**.
    - 60 students, 21 men, 39 women from 10 disciplines (mostly psychology) from the University of Salzburg, were paid for their participation. The median age was 21 years. None of the participants was familiar with Bayes's theorem.

    ☞ It is not clear which aspect of the sample is meant to characterize this population (*e.g.*, students, people of a certain age, people ignoring Bayes's theorem).
  - **treatments** (*i.e.*, "independent variables").
    - The study crosses 2 treatments ("format" and "menu"), each with 2 levels (frequency *vs.* probability format and short *vs.* standard menu), resulting in 4 conditions.
    - A third treatment ("vignette") is nested under these 2.
    - Depending on which of the four conditions a participant is in, she will be exposed to different versions of 15 vignettes for each condition, together with the dependent variable.

Format and menu	Description of problem
Standard probability format	The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ____ %
Standard frequency format	10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ____ out of ____
Short probability format	The probability that a woman at age forty will get a positive mammography in routine screening is 10.3%. The probability of breast cancer *and* a positive mammography is 0.8% for a woman at age forty who participates in routine screening. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ____ %
Short frequency format	103 out of every 1,000 women at age forty get a positive mammography in routine screening. 8 out of every 1,000 women at age forty who participate in routine screening have breast cancer *and* a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ____ out of ____

- **measurement** (*i.e.*, "dependent variables"). Surveys, as above.
- **setting.** Participants were studied individually or in groups of 2 or 3 (in two cases, 5). We informed participants that they would need approximately 1 hr for each session but that they could have more time if necessary. On average, students worked 73 min in the first session (25–180 min) and 53 min in the second (30–120 min).

- Experimental components are either **fixed** or **random**. If they are random, their levels are randomly sampled from a population.
- For instance, other participants could have been sampled (if experimental unit is a random factor) or other stimuli could have been used (if treatment is a random factor).
- In a randomized controlled trial testing a new drug, treatment is often conceived as a fixed factor. The new drug is not conceived as one of the many drugs participants could have received, and the experimenter does not intend to generalize to other drugs.

☞ Whether an experimental component should be treated as a random or a fixed factor depends on context. [We'll see a subtle case study below in which this question is crucial.]

- Units are typically treated as the levels of a random factor since they are typically meant to stand for a population, although the identity of the relevant population is rarely made explicit.

- Treatment is sometimes properly conceived as fixed (*e.g.*, drug testing). In other cases it should be viewed as random.
  - When participants are exposed to some particular stimuli (*e.g.*, words in psycholinguistics experiments) that are meant to stand for a broader class of stimuli (*e.g.*, all the words participants could have been presented with), treatment should be conceived as a random factor. Psychologists rarely explicitly do so [49].
- The same goes for measurement.
  - In psychology, treatments and measurements are rarely treated as the value of a random sample. But, they could be. *E.g.*, Gigerenzer & Hoffrage could have specified a *population of vignettes* by describing a recipe for producing more vignettes.
- The distinction between fixed and random factors determines which statistical generalizations are allowed by an experiment.

☞ *One can only generalize statistically from observed to unobserved levels of a component if that component is random.*

- Machery's "Resampling Account" of replication.

> **The Resampling Account of Replication**. E' **replicates** E if and only if (a) E' and E are *equivalent* (*i.e.*, they are constituted by sequences of events of the same type), and (b) E' involves *resampling* some of E's *random* experimental components $\{e^r\}$, while E's fixed components $\{e^f\}$ are also held fixed in E'.

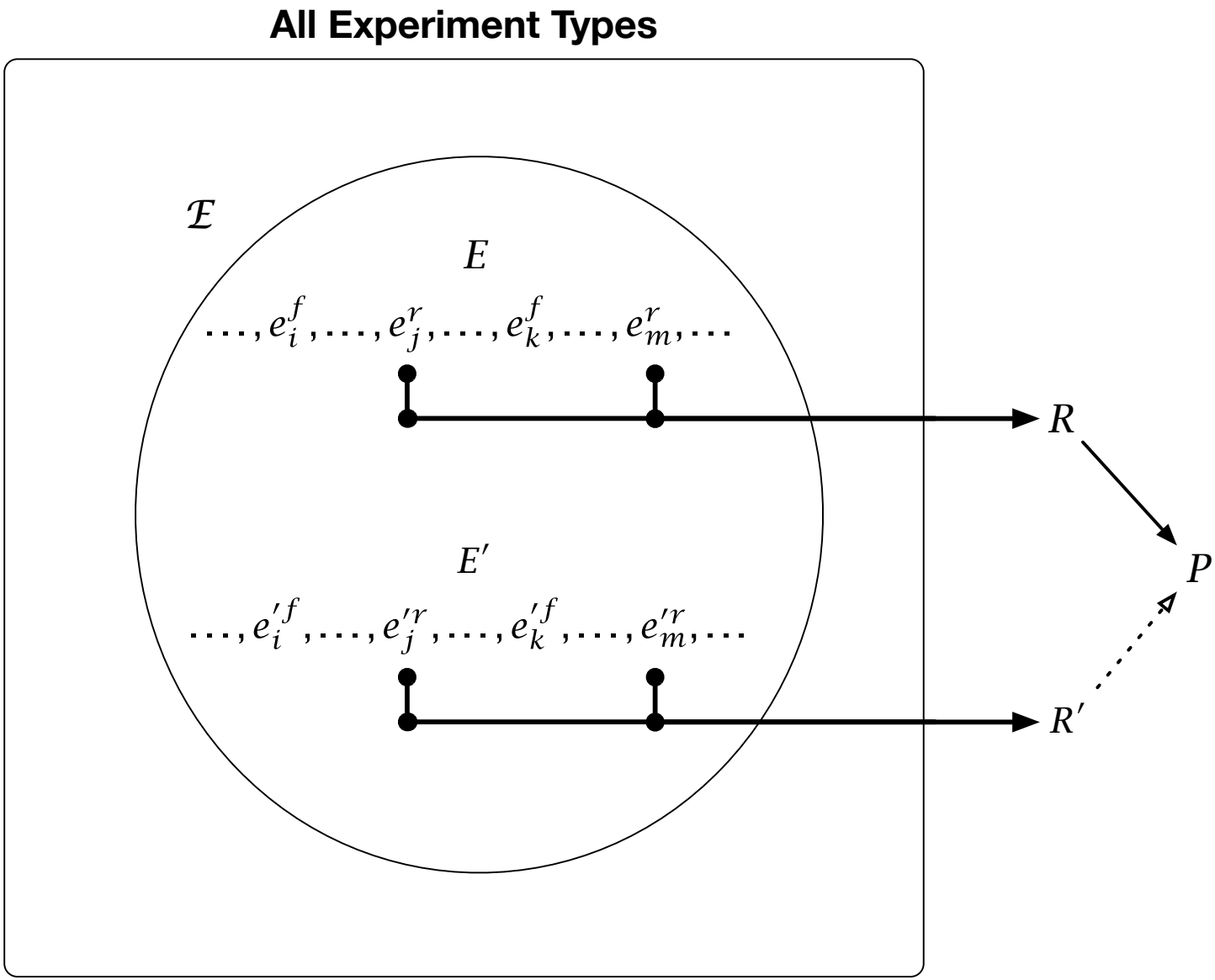
- Machery also uses these ideas to explicate a notion of *reliability*.

> E is **reliable** iff repeated sampling of new values for E's random experimental components, with all fixed components held fixed, yields the same experimental outcome (with high probability).

- E is reliable iff replications E' of E are likely to yield the same results. E is *un*reliable if its experimental outcome is an *outlier*.

☞ Clarification of *un*relaibility: E' is likely to produce $R' \neq R$, where R and R' ***have different epistemic significance regarding*** P.

- Here is a pictorial representation of Machery's framework.


All Experiment Types
$\mathcal{E}$
E
$\ldots, e_i^f, \ldots, e_j^r, \ldots, e_k^f, \ldots, e_m^r, \ldots$
E'
$\ldots, e_i'^f, \ldots, e_j'^r, \ldots, e_k'^f, \ldots, e_m'^r, \ldots$
R
R'
P

- The case of Gigerenzer & Hoffrage's experiment E is a *success story* in terms of replication (we'll get plenty of *bad* news later!). Their experiment is clearly *reliable* in Machery's sense.
- Their finding was that (P) presenting probabilistic information in the form of natural frequencies (rather than probabilities) improves performance on the classic "base-rate fallacy" task.
- That experiment E has been replicated E' [14], yielding similar results R', which also support the existence of the phenomenon.
- While there has been some disagreement about *why* P is true [3], there is now widespread agreement regarding the truth of P. And, this agreement seems to be well justified.
- Indeed, there have even been *extensions* of their experimental protocol to *other* probabilistic reasoning tasks [40].

☞ The representation of probabilistic information as natural frequencies seems to be *generally* helpful for reasoning.

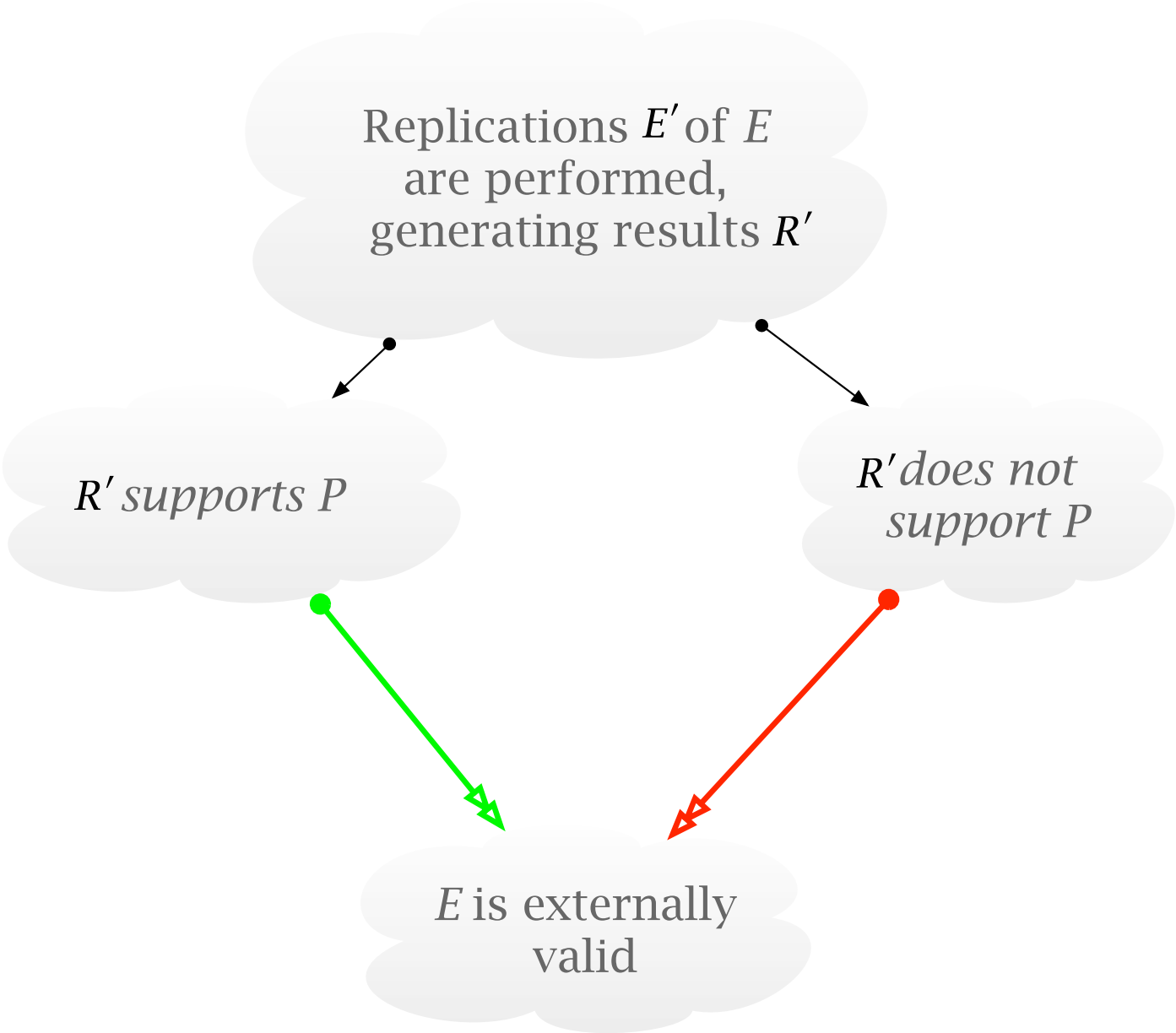
- Unreliability of experiments can have various causes.
  - **nondirectional error**. Deviation between the estimate of a quantity and its true value that is as likely to result in overestimation as in underestimation of its true value. This has various causes.
    - **measurement error**
    - **sampling error**
    - **imprecise manipulation/intervention**
  - **honest mistakes**
    - For instance, the case of the microwave-caused apparent FRB's (perytons) at Parkes Observatory [68, 99, 98].
  - **frauds**
    - Some think this is a major cause of unreliability [76]. I doubt it [75]. If you're interested in lying and deception, see, *e.g.*, [22].

☞ **questionable (*viz.*, biased) research practices** (QRPs), *e.g.*, **outcome reporting bias** [53, 45].

- selective reporting of *some* of the analyzed outcomes in a study,
- selective reporting of a *specific* outcome,
- *incomplete* reporting of an outcome

- Reliability (in Machery's sense) is to be distinguished from *validity*. *E* is *valid* just in case its results *R* provide *warrant* [70, 63] for the conclusion it claims to establish (*viz.*, *P*).
  - *E* is ***internally*** **valid** iff *R* provides *warrant* for the claim that the treatment caused the measured difference between the conditions. [Plausible but uncontrolled causal confounds can serve to undermine the internal validity of an experiment.]
  - *E* is ***externally*** **valid** iff *R* provides *warrant* for a conclusion (*P*) about a situation/phenomenon outside the lab that is of interest to scientists and which motivated the research in the first place.
- Of course, ultimately, it is *external validity* that we really want from an experiment. Above, we said that *E*'s failure to replicate is an *undercutting defeater of the support R provides for P*.

☞ Another way to put this is to say that *E*'s unreliability (in Machery's sense) is a *rebutting defeater of E's external validity*. *I.e.*, *E*'s unreliability undermines belief in *E*'s external validity.


Replications E' of E are performed, generating results R'
R' supports P
R' does not support P
E is externally valid

- **Machery Applied**: if an apparent replication failure occurs, it may be because some factor was "mis-classified" as random [15].
  - Simard *et al* [91] performed experiments which involved inducing spinal cord injuries in mice, to determine whether PHN (a type of necrosis) can be treated with glibenclamide (a drug).
  - When some other researchers failed to replicate their (positive) results, they worked with Simard *et al* to try to diagnose the cause. It turned out that *a very slight difference* in the procedure for inducing the injury was making a difference to the outcome.
  - Simard *et al* had originally assumed this was a random factor (*i.e.*, these components can be *resampled*). But, after discovering that these minute differences *make a difference for the outcome*, these were re-classified as "fixed" features of the experiment, thus disqualifying the follow-up experiments as replications [90].

☞ The problem is that Simard *et al* were *not* studying whether PHN caused by *the specific injury they induced* can be treated with glibenclamide. They [91] report "significant improvements in *all* of the characteristic manifestations of PHN" with glibenclamide.

- "direct replication" is a term that is sometimes used in the literature. It has been said that A "direct replication" of E is
  - a "study that aspires to be comparable to E in all aspects." [42]
  - a "repetition of E's experimental procedure." [84]
  - an "experiment whose design is identical to E's design in all factors that are supposedly causally relevant to the effect." [77]
- I think Machery's explication of replication is clearer and more probative than these (the last one is getting closer to Machery's).
- "Conceptual replication" is used in an even more perplexing way. *E.g.*, "the repetition of a test of a hypothesis or of a result of an earlier research work with different methods." [84]
- "Reproducibility" of a computational (*e.g.*, data mining or machine learning) "experiment" implies *yielding the same numerical results when repeating the analysis using the original data and the same computer code* [73]. Even this fails sometimes.

- 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

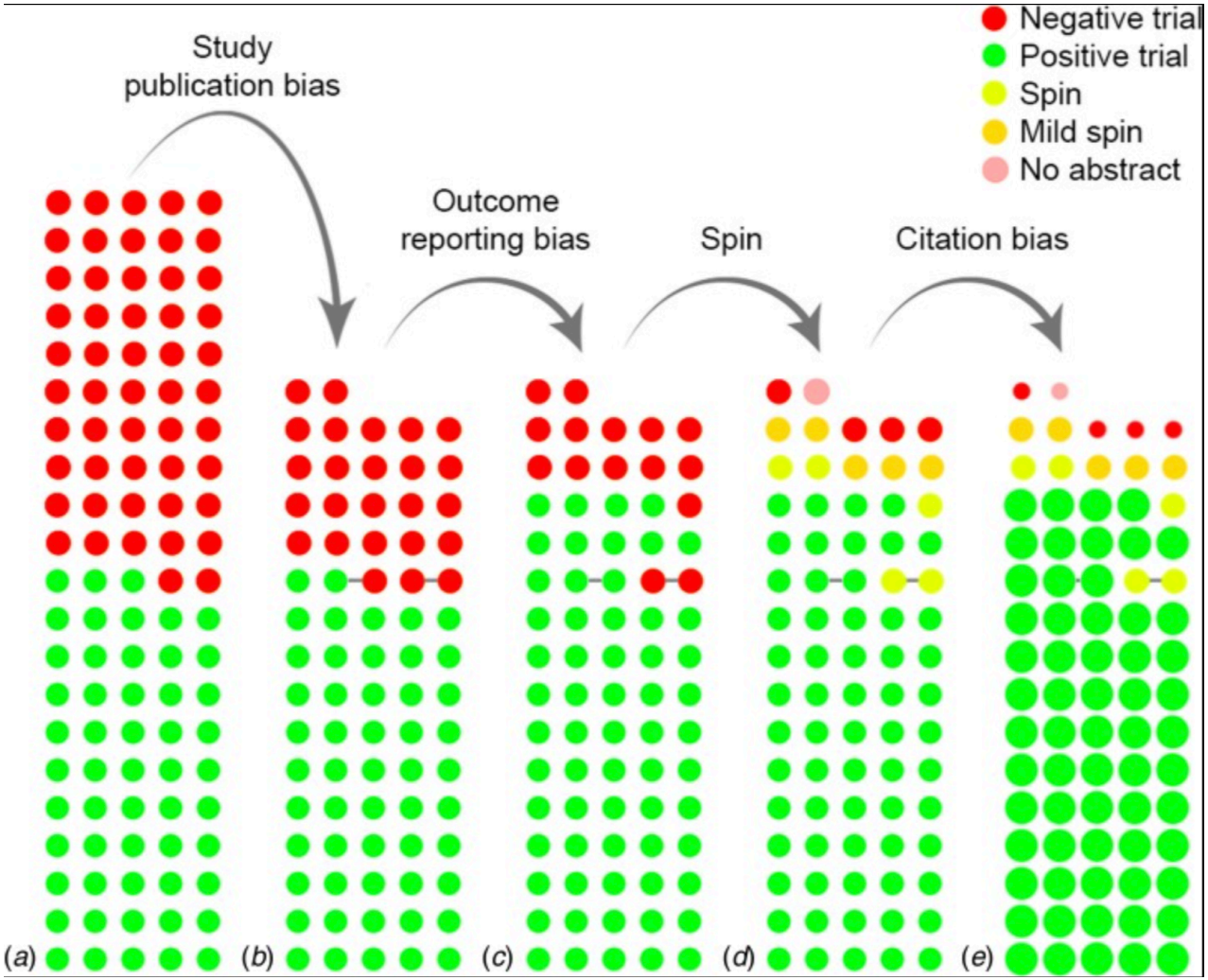
  > there may be relatively few publications on problems for which the null hypothesis is . . . true, and of these, a high proportion will erroneously reject the null hypothesis.

- 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"

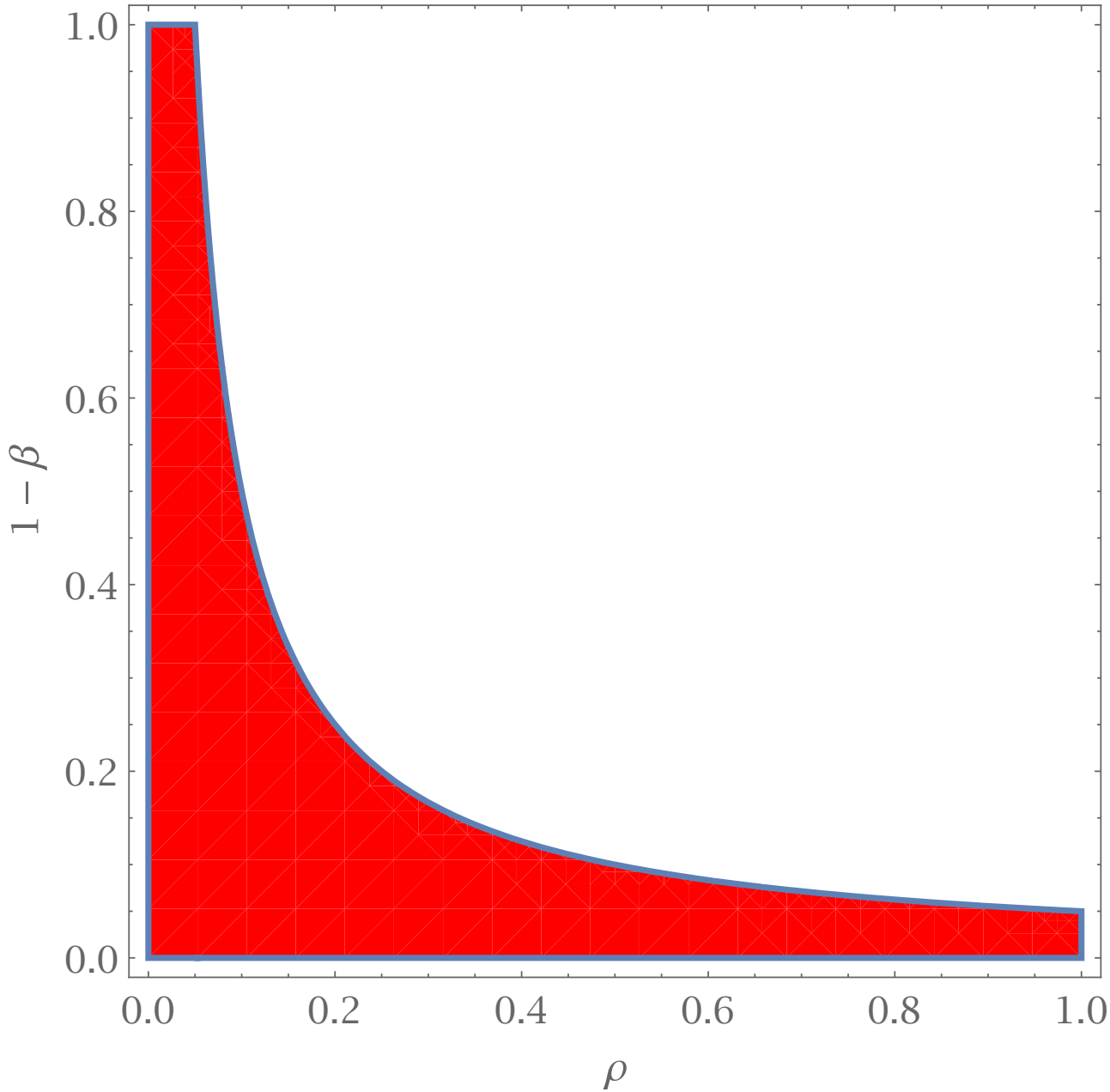
- Many studies in various sciences have low power [12, 20]. In some circumstances, low power can make it probable that published "positive" results are misleading [96, 44].
- Let $R \stackrel{\text{def}}{=}$ a "positive/significant" result is generated by E (wrt P), $\neg P \stackrel{\text{def}}{=}$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

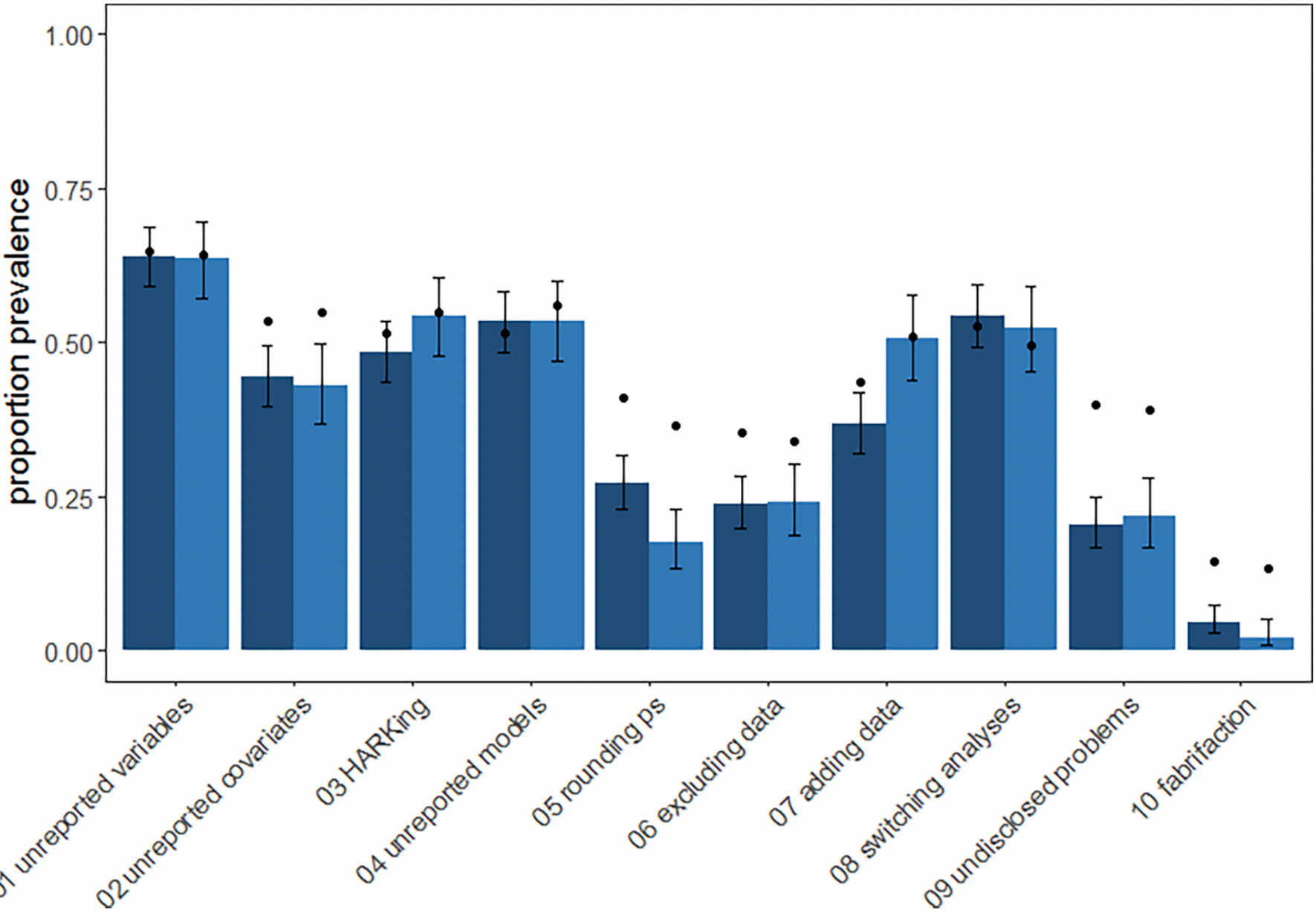
- Here are survey results from 500 US psychologists in 2012 [48].

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.

Felipe Romero
Faculty of Philosophy
University of Groningen

Jan Sprenger
Center for Logic, Language & Cognition
Department of Philosophy & Education
University of Turin

Synthese (2021) 198 (Suppl 23):S5803–S5823
https://doi.org/10.1007/s11229-020-02697-x

S.I.: RELIABILITY

Scientific self-correction: the Bayesian way

Felipe Romero[1] · Jan Sprenger[2]

Received: 11 March 2019 / Accepted: 12 May 2020 / Published online: 29 June 2020

- Romero & Sprenger sent me their R code yesterday!

`http://fitelson.org/rs_code.zip`

- When it comes to **statistical reforms**, I would favor a move in the direction of **Bayesian** [8] approaches, as opposed to **null hypothesis [101] significance testing** (NHST).
- I will discuss a really neat recent paper by Romero & Sprenger [81], which provides evidence (in the form of computer simulations) that this move would be beneficial for the "crisis."
- The basic idea behind the work of (Romero [80]) and Romero & Sprenger [81] is to see whether a Bayesian *versus* NHST approach to statistical inference leads to better **self-correction**.

SCT⋆ Given a series of replications of E, the meta-analytical aggregation of their effect sizes will converge on the true effect size as the length of the series of replications increases.

- SCT⋆ can be checked (*via* computer simulation), for various statistical approaches, on simple, well-understood hypothesis testing problems. Bayesianism *vs* NHST is the competition [81].

- First, a brief review of NHST *vs* Bayesian statistical inference.
- We will be looking at a classic inference problem: trying to determine whether two normally distributed random variables X_1 and X_2 with means μ_1 and μ_2 have the same mean value.
- We assume, for simplicity that the variance of X_1 and X_2 are the same. Hence, $X_1 = N(\mu_1, \sigma^2)$ and $X_2 = N(\mu_2, \sigma^2)$.
- The **null hypothesis** H_0 asserts that $\mu_1 = \mu_2$ and the **alternative hypothesis** $H_1 = \neg H_0$ asserts that $\mu_1 \neq \mu_2$.
- To test H_0 and H_1 against each other, NHST-ers conduct a **two-sided hypothesis test**: a design where large (estimated) deviations in either direction from the "null value" ($\mu_1 - \mu_2 = 0$) count as evidence against H_0, and in favor of H_1.
- It is customary for NHST-ers to use a t-**statistic** here [102]. If t is sufficiently large (p-value $= 0.05$), then the null is "rejected" and this is interpreted as "strong evidence" against the null.

- The **true effect size** δ is given by the standardized difference of the unknown means: $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. This is estimated using Cohen's (d) measure of **observed effect size.**
- Bayesians, on the other hand, will use the **Bayes Factor** as their measure of *degree of evidential favoring* [26].

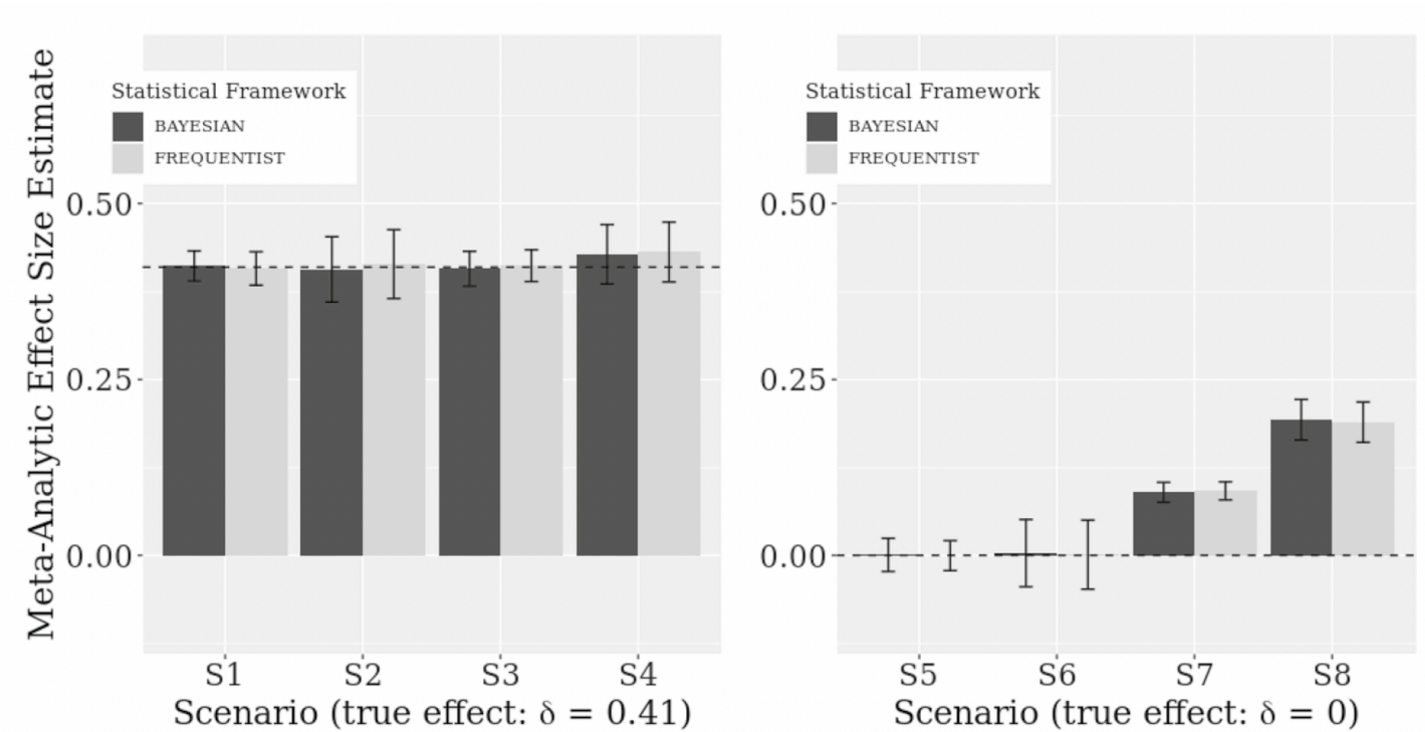
$$BF_{10} = \frac{\Pr(D \mid H_1)}{\Pr(D \mid H_0)}$$

- If $BF_{10} > 1$ then D is evidence against the null hypothesis, and if $BF_{10} \leqslant 1$, then D is not evidence against the null hypothesis.
- $\frac{1}{3} < BF_{10} < 3$ is considered weak or inconclusive evidence (Bayesian re-analysis of data with an observed significance level of $p \approx .05$ often corresponds to a Bayes factor $BF_{10} \approx 3$ [7]).
- BF_{10} seems to be a more adequate measure of the degree to which D favors H_1 over H_0 than the p-value is [83, 86, 51].

- Romero & Sprenger explore three factors ("variables") that are thought to be contributors to the "crisis."
  - **low power** (which they call "insufficient resources")
  - **direction bias** (suppression of negative effect size results), and
  - **suppression of inconclusive (or "weak evidence") results**.
- To test for effects of low power, they run simulations in two conditions: **high power**, $n = 156$, which corresponds to a power of 95% for a true effect size of 0.41 [28]; and, **low power**, $n = 36$, which corresponds to a power of only 40%.
- To test the effects of direction bias, they look at: **all results published** (direction bias absent) *vs* **all results with negative effect size magnitude suppressed** (direction bias present).
- To test the effects of suppression of inconclusive results:
  - NHST: **all results** *vs* **only results at 5% significance**
  - Bayes: **all Bayes factors** vs **non-weak Bayes factors**

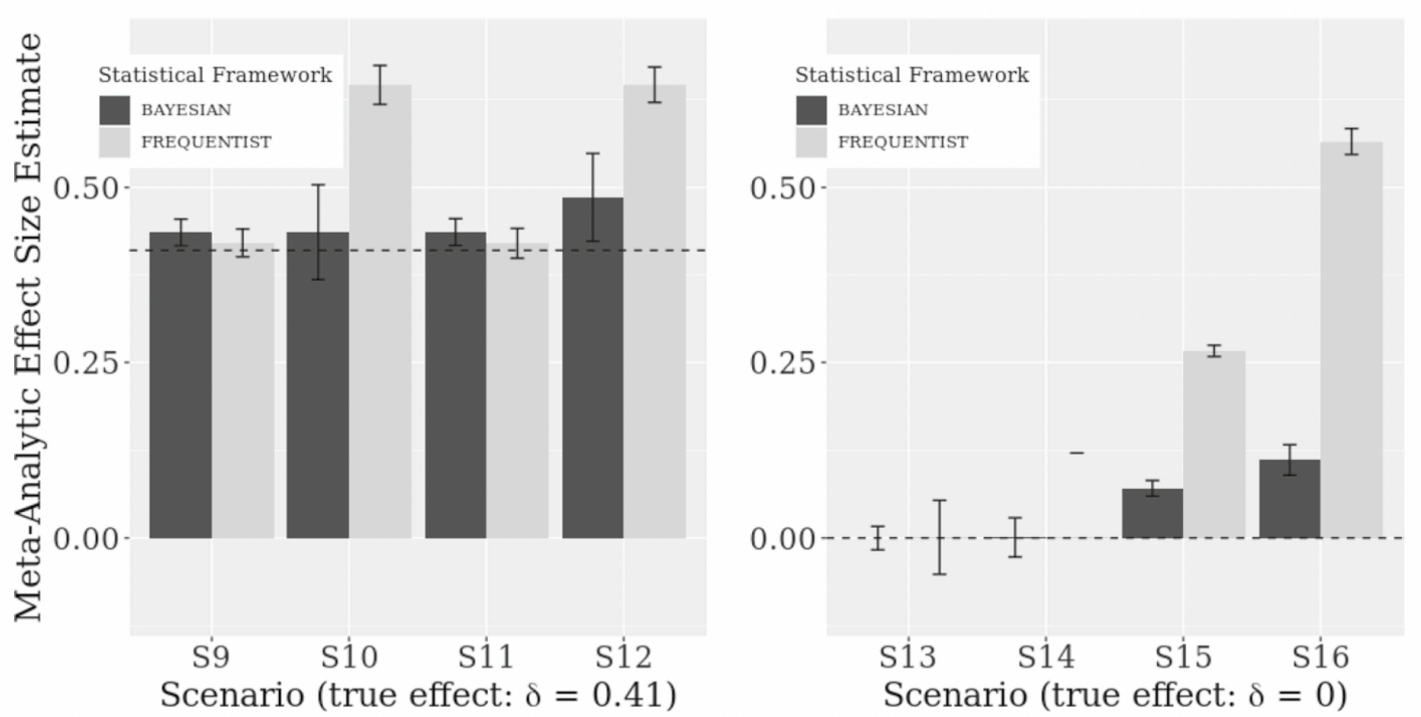
	$\delta = 0.41$				$\delta = 0$				$\delta = 0.41$				$\delta = 0$			
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
SUFFICIENT RESOURCES	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
NO DIRECTION BIAS	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗
INCONCLUSIVE EVIDENCE IS PUBLISHED	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

- When all inconclusive evidence is included, R & S's simulations yield *similar* aggregate effect sizes for NHST & Bayes.

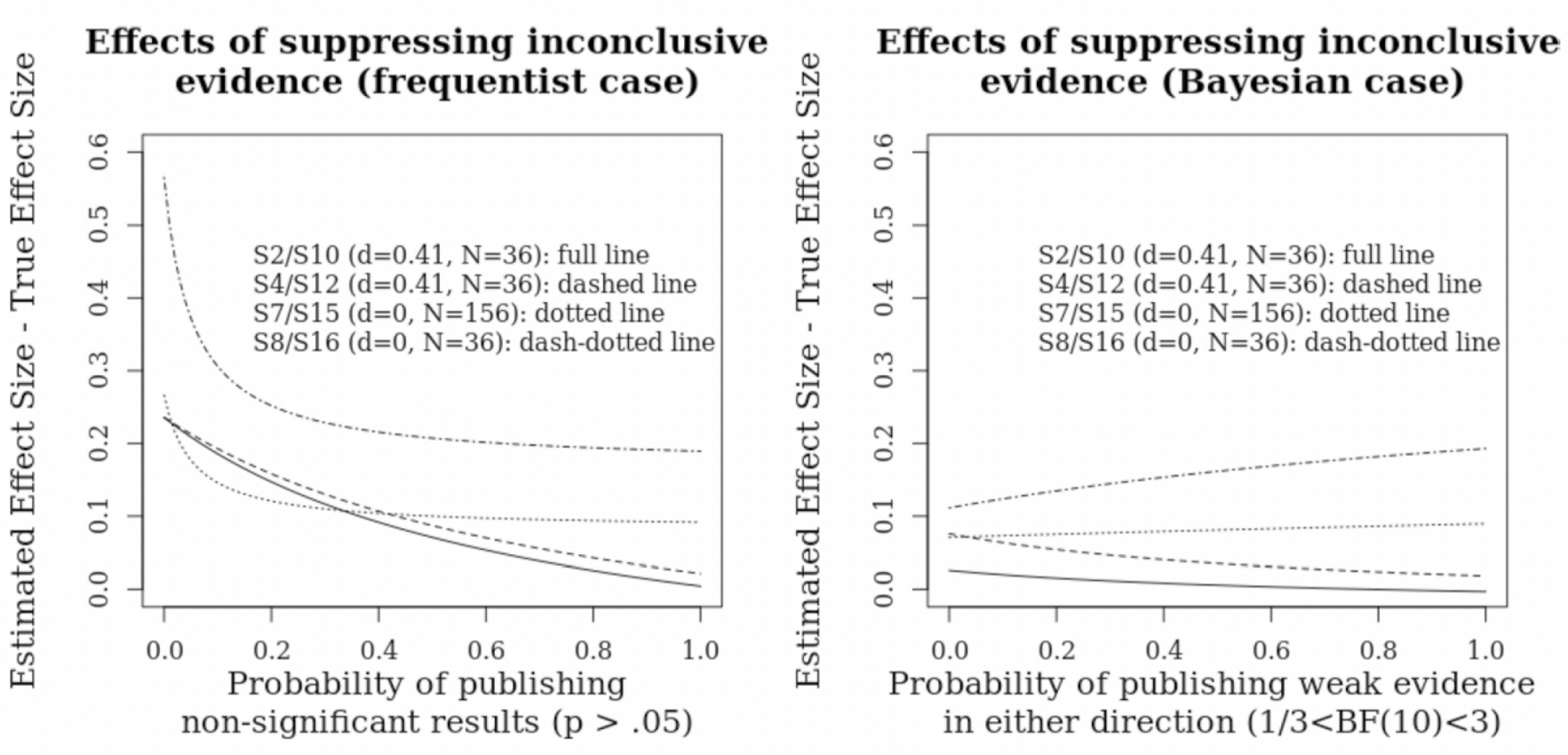


	$\delta = 0.41$				$\delta = 0$				$\delta = 0.41$				$\delta = 0$			
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
Sufficient Resources	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
No direction bias	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗
Inconclusive evidence is published	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

- When all inconclusive evidence is suppressed, the simulations yield *dissimilar* aggregate effect sizes for NHST & Bayes.

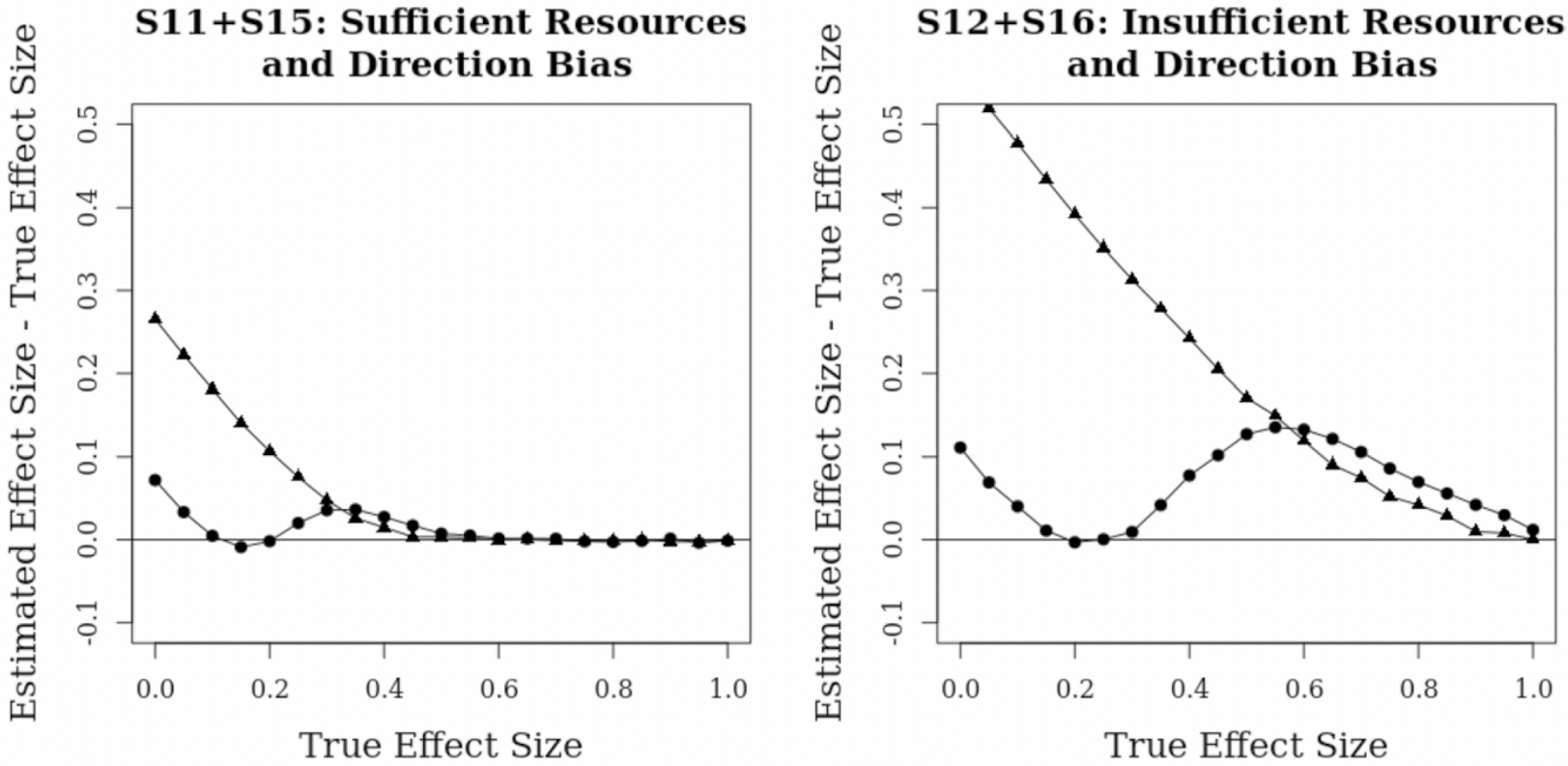


- **Extension #1**: *probabilistic* file drawer effect — how does the estimated effect (in relation to the true effect size) size depend on the *probability* that weak evidence is included?



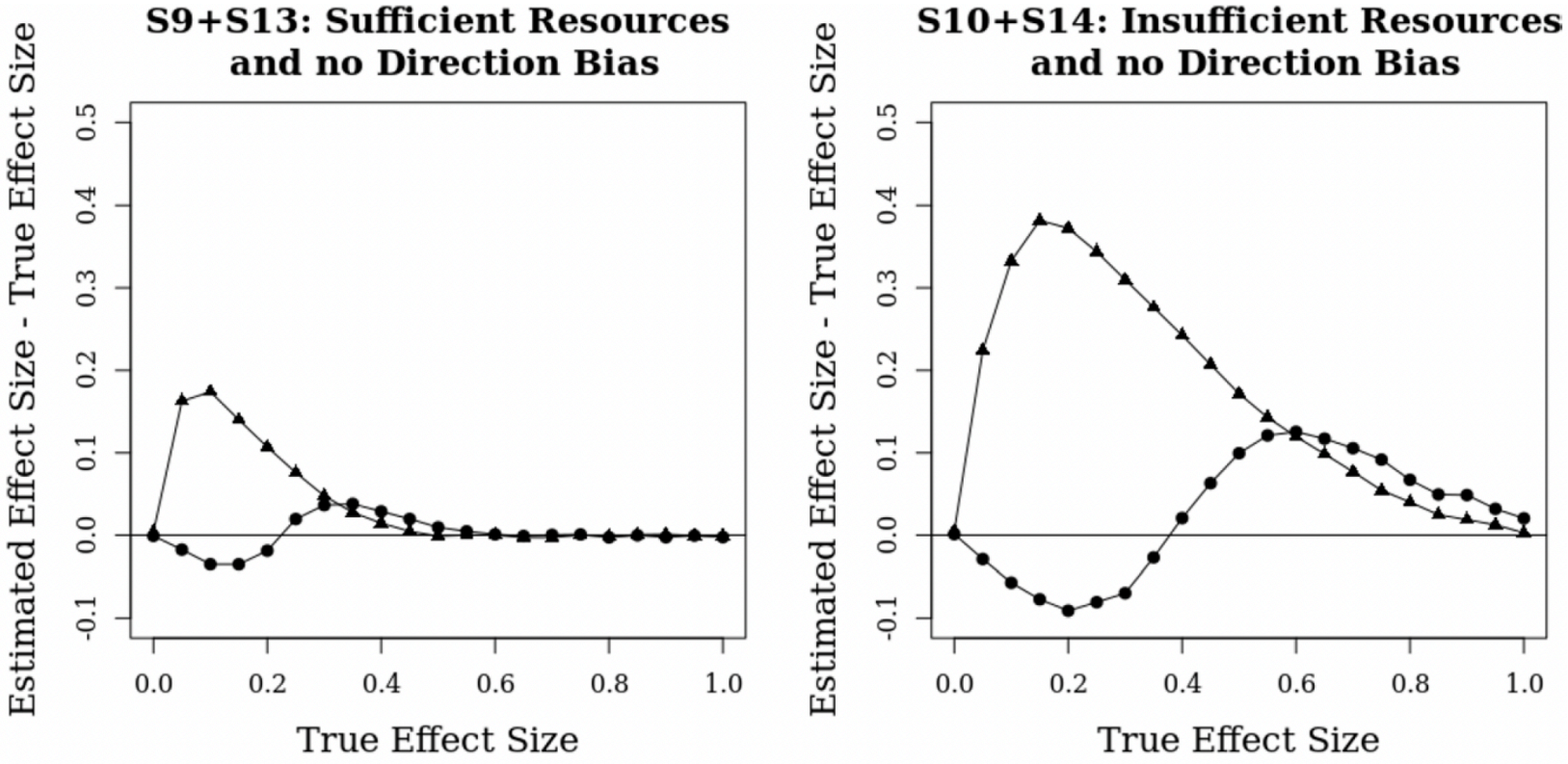
☞ If even 20-30% of of statistically non-significant results are published, frequentist estimates become similarly accurate.

- **Extension #2.1**: wider effect sizes — how does the estimated effect (in relation to the true effect size) size depend on the true effect size — in the cases where there *is* direction bias?



☞ NHST (▲) largely overestimates small effects due to the combination of direction bias and suppressing inconclusive evidence, but it estimates large effects accurately.

- **Extension #2.2**: wider effect sizes — how does the estimated effect (in relation to the true effect size) size depend on the true effect size — in the cases where there is *no* direction bias?



☞ NHST (▲) is *no longer monotonically decreasing*: small effects are substantially overestimated while null effects are estimated accurately. And, Bayes (●) *underestimates some small effects*.

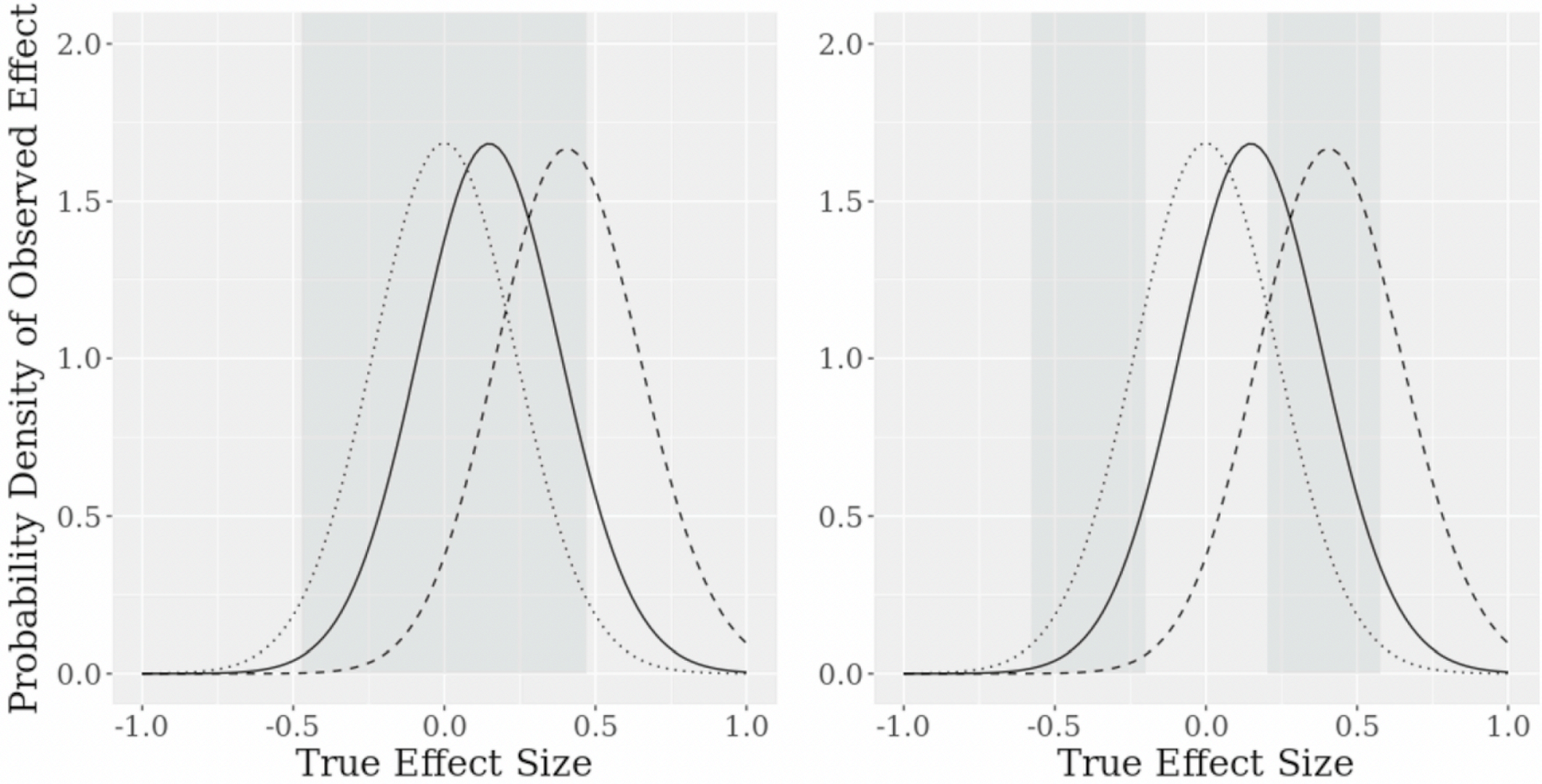


Fig. 7 Probability density functions for the standardized sample mean in a single experiment for $N = 36$ and different values of the real effect size. Full line: $\delta = 0.15$, dashed line: $\delta = 0.41$, dotted line: $\delta = 0$. The suppressed regions (i.e., observations that do not enter the meta-analysis because $p > .05$ or $\frac{1}{3} < BF_{10} < 3$) are shaded in dark. Left graph: frequentist case, right graph: Bayesian case

☞ Omitting weak evidence in favor of either hypothesis leads to more accurate meta-analytic estimates than omitting statistically non-significant results. This is especially true for small effects.

- SCT* — the thesis about the self-corrective nature of science in sequential replications of an experiment — therefore holds for a wider range of effect sizes when replacing NHST with Bayes.
- This agrees with the distribution of effect sizes in OSC replication project for behavioral research [67].
- That is, replications of experiments with large observed effects usually confirm the original diagnosis, while moderate effects often turn out to be small or nonexistent in the replication.

☞ Statistical frameworks where evidence for the null can be expressed on the same scale as evidence for the alternative would likely lead to more "null" results being reported.

- Future studies should include non-NHST frequentism(s) [61].

Kelter *BMC Medical Research Methodology* (2020) 20:88
https://doi.org/10.1186/s12874-020-00968-2

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research

Riko Kelter

- Kelter [51] performs similar analyses to Romero & Sprenger, and his results are similar to theirs. He also examines several other Bayesian alternatives to NHST (besides the Bayes Factor approach). To my mind, more studies like these would help…

[1] F. Agnoli *et al*, *Questionable research practices among italian research psychologists*, 2017.
[2] T. Baker, *Two Tessier-Lavigne papers retracted on his last day as president*, 2023.
[3] A. Barbey and S. Sloman, *Base-rate respect: From ecological rationality to dual processes*, 2007.
[4] J. Bargh *et al*, *Automaticity of social behavior…*, 1996.
[5] C. Begley and L. Ellis, *Raise standards for preclinical cancer research*, 2012.
[6] D. Bem, *Feeling the future…*, 2011.
[7] D. Benjamin *et al*, *Redefine statistical significance*, 2018.
[8] J. Bernardo and A. Smith, *Bayesian Theory*, 1994.
[9] D. Bishop, *Rein in the four horsemen of irreproducibility*, 2019.
[10] M. Borenstein *et al*, *Introduction to meta-analysis*, 2009.
[11] S. Bruton *et al*, *Personal Motivations and Systemic Incentives…*, 2020.
[12] K. Button et al, *Power failure: why small sample size undermines the reliability…*, 2013.
[13] C. Camerer et al, *Evaluating the replicability of social science experiments…*, 2018.
[14] G. Chapman and J. Liu, *Numeracy, frequency, and Bayesian reasoning*, 2023.
[15] D. Colaço et al, *When should researchers cite study differences…*, 2022.
[16] V. Crupi, *Confirmation*, 2020.
[17] V. Crupi *et al*, *Probability, Confirmation, and the Conjunction Fallacy*, 2008.
[18] M. Curd, *The logic of discovery: An analysis of three approaches*, 1980.
[19] Y. de Vries *et al*, *The cumulative effect of reporting and citation biases…*, 2018.
[20] E. Dumas-Mallet *et al*, *Low statistical power in biomedical science …*, 2017.
[21] D. Eddy, *Probabilistic reasoning in clinical medicine*, 1982.
[22] D. Fallis, *Lying and Deception*, 2010.
[23] D. Fanelli, *Negative results are disappearing from most disciplines and countries*, 2012.
[24] ———, *"Positive" Results Increase Down the Hierarchy of the Sciences*, 2010.

[25] ———, *Do Pressures to Publish Increase Scientists' Bias?*, 2010.
[26] B. Fitelson, *Contrastive Bayesianism*, 2010.
[27] S. Fortunato *et al*, *Science of Science*, 2018.
[28] R. Fraley and S. Vazire, *The N-Pact Factor: Evaluating the Quality of...*, 2014.
[29] H. Fraser *et al*, *Questionable research practices in ecology and evolution*, 2018.
[30] G. Gigerenzer, *Statistical Rituals: The Replication Delusion...*, 2018.
[31] G. Gigerenzer and U. Hoffrage, *How to improve Bayesian reasoning without...*, 1995.
[32] P. Godfrey–Smith, *Theory and Reality: An Introduction to the Philosophy of Science*, 2021.
[33] I.J. Good, *Good Thinking*, 1983.
[34] G. Gopalakrishna *et al*, *Prevalence of QRPs ... among ... researchers in The Netherlands*, 2022.
[35] A. Greenwald, *Consequences of Prejudice Against the Null Hypothesis*, 1975.
[36] M. Head *et al*, *The Extent and Consequences of p-Hacking in Science*, 2015.
[37] S. Hetherington, *Fallibilism*, 2005.
[38] A. Higginson and M. Munafò, *Current Incentives for Scientists...*, 2016.
[39] C. Hitchcock and E. Sober, *Prediction versus Accommodation and the Risk of Overfitting*, 2004.
[40] U. Hoffrage *et al*, *Natural frequencies improve Bayesian reasoning...*, 2015.
[41] J. Hollenbeck and P. Wright, *Harking, Sharking, and Tharking...*, 2016.
[42] J. Hüffmeier et al, *Replication as a Sequence of Different Studies...*, 2016.
[43] J. Hullman and A. Gelman, *Designing for Interactive Exploratory Data Analysis...*, 2021.
[44] J. Ioannidis, *Why Most Published Research Findings Are False...*, 2005.
[45] J. Ioannidis *et al*, *Publication and other reporting biases in cognitive sciences...*, 2015.
[46] A. Jacot, *Principles of Scientific Publication*, 1937.
[47] S. Janke *et al*, *Dark Pathways to Achievement in Science...*, 2018.
[48] L. John *et al*, *Measuring the Prevalence of Questionable Research Practices...*, 2012.

[49] C. Judd *et al*, *Treating Stimuli as a Random Factor in Social Psychology . . .*, 2012.
[50] D. Kahneman and A. Tversy, *On the Psychology of Prediction*, 1973.
[51] R. Kelter, *Analysis of Bayesian posterior significance and effect size indices. . .*, 2020.
[52] N. Kerr *et al*, *HARKing: Hypothesizing After the Results are Known*, 1998.
[53] J. Kirkham *et al*, *The impact of outcome reporting bias in randomised. . .*, 2010.
[54] S. Koole and D. Lakens, *Rewarding replications*, 2012.
[55] M. Kotzen, *A Formal Account of Epistemic Defeat*, 2019.
[56] T. Kuhn, *The Structure of Scientific Revolutions*, 1962.
[57] S. Lee, *A Famous Honesty Researcher Is Retracting A Study Over Fake Data*, 2023.
[58] E. Machery, *What is a Replication?*, 2020.
[59] L. Maggio *et al*, *Factors Associated with Scientific Misconduct. . .*, 2019.
[60] L. Malich and C. Rehmann-Sutter, *Metascience Is Not Enough . . .*, 2022.
[61] D. Mayo, *Statistical inference as severe testing. . .*, 2018.
[62] M. McDowell and P. Jacobs, *Meta-Analysis of the Effect of Natural . . .*, 2017.
[63] T. Merricks, *Warrant Entails Truth*, 1995.
[64] M. Munafò *et al*, *A manifesto for reproducible science*, 2017.
[65] D. Peterson and A. Panofsky, *Metascience as a Scientific Social Movement*, 2023.
[66] B. Nosek *et al*, *The preregistration revolution*, 2018.
[67] B. Nosek *et al*, *Estimating the reproducibility of psychological science*, 2015.
[68] R. Petroff *et al*, *Identifying the source of perytons at the Parkes radio telescope*, 2015.
[69] F. Prinz *et al*, *Believe it or not: how much can we rely on published data. . .*, 2011.
[70] J. Pryor, *Highlights of Recent Epistemology*, 2001.
[71] L. Moretti and T. Piazza, *Defeaters in current epistemology . . .*, 2018.
[72] C. O'grady, *After honesty researcher's retractions, colleagues expand scrutiny. . .*, 2023.

[73] R. Peng, *Reproducible research in computational science*, 2011.
[74] A. Potochnik, *The Diverse Aims of Science*, 2015.
[75] M. Reisig, *Assessing the perceived prevalence of research fraud...*, 2020.
[76] S. Ritchie, *Science Fictions...*, 2020.
[77] F. Romero, *Philosophy of science and the replicability crisis*, 2019.
[78] ———, *Who should do replication labor?*, 2018.
[79] ———, *Novelty vs. replicability: Virtues and vices...*, 2017.
[80] ———, *Can the Behavioral Sciences Self-Correct?...*, 2016.
[81] F. Romero and J. Sprenger, *Scientific self-correction: the Bayesian way*, 2020.
[82] R. Rosenthal, *The file drawer problem and tolerance for null results*, 1979.
[83] R. Royall, *Statistical Evidence: A Likelihood Paradigm*, 1997.
[84] S. Schmidt, *Replication*, 2017.
[85] F. Schmidt, *Statistical significance testing and cumulative knowledge in psychology*, 1996.
[86] F. Schönbrodt and E. Wagenmakers, *Bayes factor design analysis...*, 2018.
[87] P. Sedlmeier and G. Gigerenzer, *Do Studies of Statistical Power Have an Effect...*, 1989.
[88] D. Shanks *et al*, *Priming Intelligent Behavior: An Elusive Phenomenon*, 2013.
[89] J. Shickore, *Scientific Discovery*, 2022.
[90] J. Simard and V. Gerzanich, *When replication teaches more...*, 2012.
[91] J. Simard *et al*, *Endothelial sulfonylurea receptor 1–regulated NC Ca–ATP...*, 2007.
[92] J. Simmons *et al*, *False-Positive Psychology: Undisclosed Flexibility in Data Collection...*, 2011.
[93] K. Tentori *et al*, *The conjunction fallacy: a misunderstanding about conjunction?*, 2004.
[94] M. Titelbaum, *Fundamentals of Bayesian Epistemology*, 2022.
[95] S. Vosoughi *et al*, *The spread of true and false news online*, 2018.
[96] S. Wacholder et al, *Assessing the Probability That a Positive Report is False...*, 2004.
[97] E. Wagenmakers *et al*, *An Agenda for Purely Confirmatory Research*, 2012.
[98] Wikipedia, *Fast Radio Burst*, 2023.
[99] ———, *Peryton (astronomy)*, 2023.
[100] ———, *Confirmation Bias*, 2023.
[101] ———, *Null Hypothesis*, 2023.
[102] ———, *Student's t-test*, 2023.