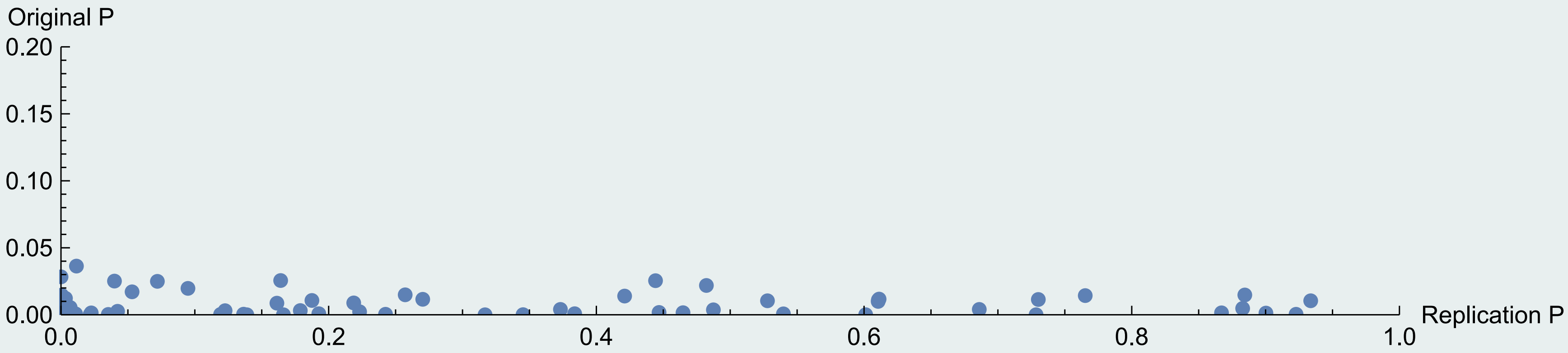
Methodological hypocrisy and effectism in psychology

J.P. de Ruiter

Depts. of Computer Science & Psychology
Tufts University

The replication crisis

- It is not controversial that large areas of Psychology are having a *replication crisis*.
- Some people still in denial, esp. Ivy League professors (e.g. Gilbert et al. 2016, Fiske 2016).
- This is what the replication crisis looks like in terms of p-values:



- Note: NHST P-values are (by definition) distributed uniformly under H_0

This crisis has many causes

- Some prime suspects:

A. Perverse incentive structure

B. Publication bias

C. Dysfunctional statistical paradigm: *Null Hypothesis Significance Testing* (NHST)

D. Illegitimate use of NHST (extremely common)

E. The way we develop and test theories: our *scientific logic*

F. The interaction between C, D, and E.

Psychology's scientific logic

- Officially, we are still Popperian Falsificationists.
- Classical ("naïve") Popper in a nutshell:
  - We come up with a *theory/hypothesis*
  - We derive a *prediction* from the theory
  - We try to *falsify* that prediction in an experiment
  - If the prediction is falsified, we *ditch* the theory
  - If the prediction is not falsified, the theory *can stay* (for now)

This is *normative* reality

- We *try* to be falsificationist in the jargon used in articles and in the review process, where we are urged to:
  - Specify hypotheses
  - Test using a null-hypothesis and and “alternative” hypothesis
  - Try to *reject* a hypothesis (H_0), not *confirm* it
- Not strictly enforced, but we see a strong *normative orientation*.
- But what do we actually *do* in psychology (and in most other social and behavioral sciences)?

What we actually do

- What we actually do, at best:
  - Formulate a theory
  - Derive a prediction from theory: an *effect* of IV on DV
  - Perform a random controlled experiment
    - H0: the IV has no effect on DV
    - H1: the IV does have some effect on DV
  - Perform a significance test
  - If the probability of the recorded difference between the levels of IV (or an even larger difference) under H0 is lower than α (usually .05), then we REJECT H0. (We do NOT confirm H1, because we are falsificationists!)

This is the wrong way around

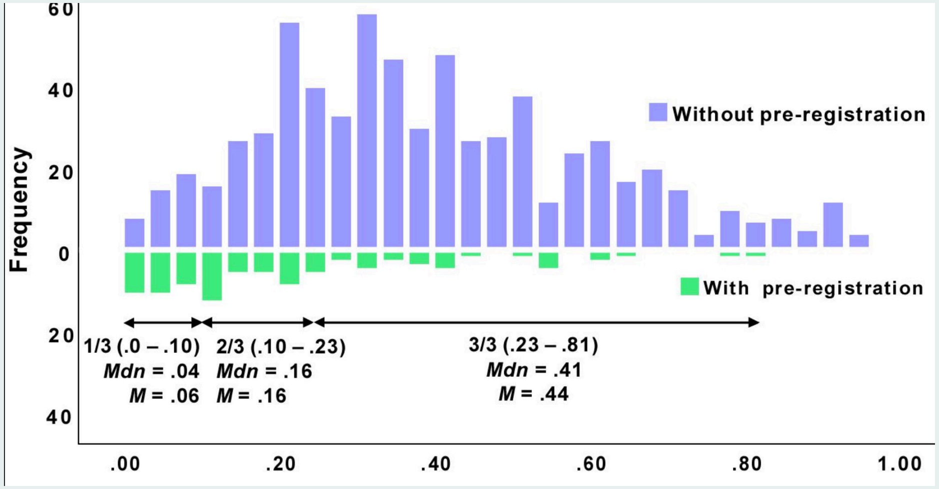
- This is the *reverse* of what Falsificationism requires.
  - Falsificationist: try to falsify your prediction (which is **H1**)
  - NHST: try to falsify H0 (which is **negation** of H1)
- This has been noted before (McElreath 2015, De Ruiter & Albert 2017)
  - Note: there are still people (e.g. Deborah Mayo, Daniel Lakens) who insist that NHST is the statistical implementation of Falsificationism.
- A correctly formulated Popper/NHST result for a “successful” experiment would therefore be:
  - *It is **unlikely** that these data (or more extreme data) would occur under the assumption that the **negation** of the prediction that we have tried to **falsify** is true. We therefore conclude that our **falsification** attempt has **failed**, so we do **not reject** our theory.*
- That’s a lot of chained negatives, and what we really mean by it is:

We confirmed our theory!

In practice, it is even worse…

While this merely sounds a bit Kafkaesque, reality is even more worrying, due to:

- HARKing: Hypothesizing After Results are Known.
  - Still very common (often even required)
  - Could be improved by requiring preregistration



Schäfer & Schwarz 2019

In practice, it is even worse than that…

- For technical reasons, we cannot *accept* H0 in NHST, so we cannot *reject* H1
  - Our statistical paradigm *does not allow us* to falsify our theory.
  - So much for falsificationism using NHST!
- We can't publish our falsifications, because "null findings" (where p > .05) are not accepted by journals.
  - Nobody is interested in the fact that someone had a theory which predicted something that they failed to reject the negation of.
  - When someone has a null finding, people start suggesting that maybe the researcher is not good enough to "evoke" the effect. (Baumeister's *flair* factor, Zwaan's "shy animal" model.)

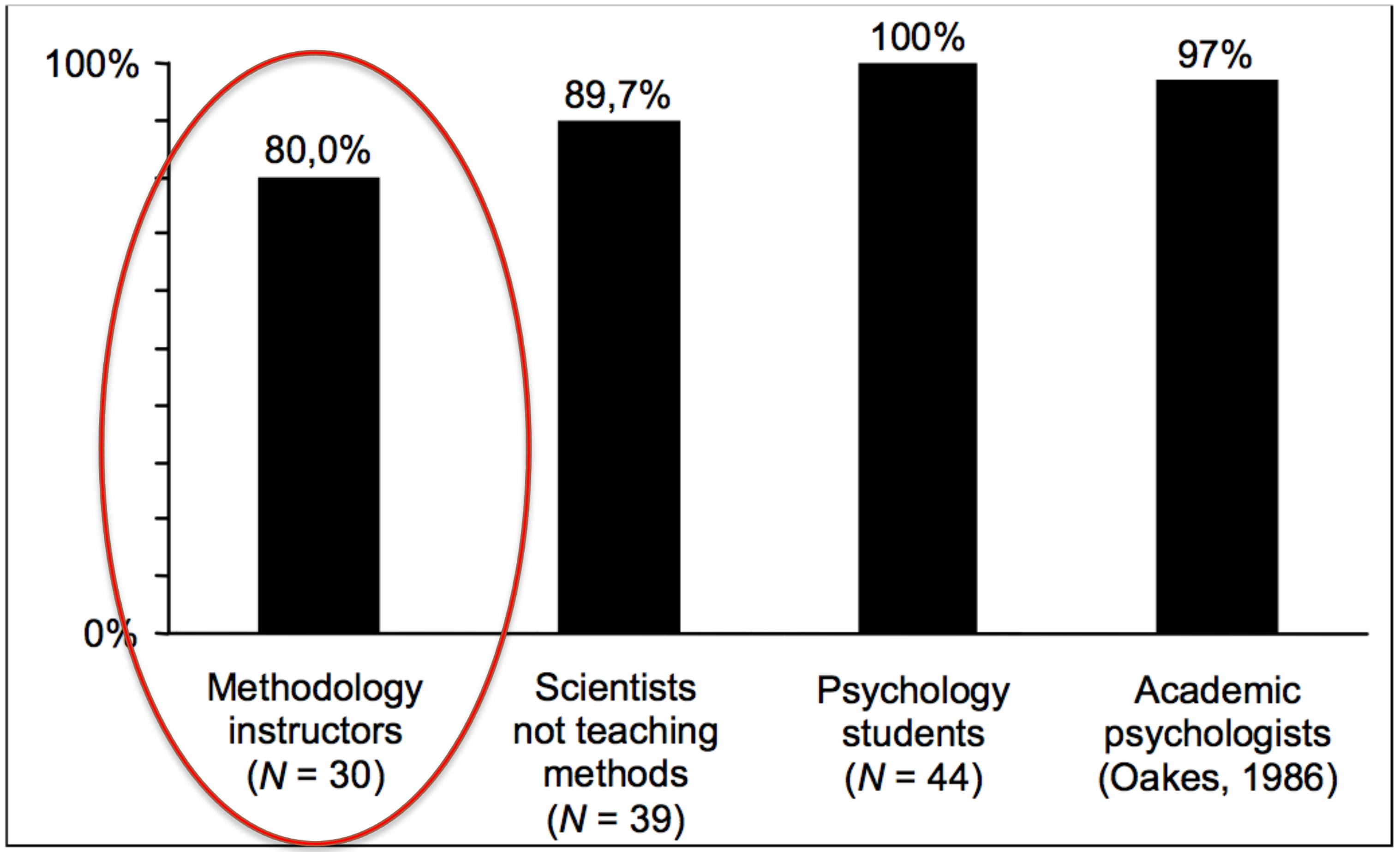
So to recap

- (Naïve) Popperian Falsificationism + NHST, officially:
  - Theory -> Prediction -> Experiment -> Result:
    - IF failed to reject H0 -> Falsification (statistically incorrect, but hey…)
    - IF H0 rejected -> Failure to falsify -> Keep theory
- Reality:
  - Experiment -> Results -> Theory:
    - IF H0 rejected -> Prediction -> Theory that predicted finding confirmed
    - IF failure to reject H0 -> study ends up in *file drawer*
  - So now we also get severe *publication bias*
- *What could possibly go wrong?*

To make matters even worse…

- NHST does not give us what any reasonable scientist is interested in, which is: **P(hypothesis | data).**
- Instead, it gives us
  P(data or more extreme | not our hypothesis)
  but we still act as if that gives us **1 – P(hypothesis | data)** because that's what we *want* it to mean so much (Gigerenzer, 2004).
- Evidence for this:
  - the *abundance* of articles still claiming that P > .05 so there is no effect
  - Haller & Krauss (2002) who checked with 6 very simple questions if Psychology Students, Psychologists, and Methodology Instructors understood NHST.

Percentage of people making at least one error



Freudian model (inspired by Gigerenzer 2004)

SUPEREGO:

We should try to falsify our own theory!

EGO:

Publish effect supported by NHST but then use falsificationist language to report them.

ID:

We want to find cool significant effects and publish them!

Underlying cause: *Effectism*

Effectism:

The assumption that a statistically significant effect is evidence for the theory that most intuitively explains it.

Irony

- It all started with Popper pointing out that *induction* is strictly speaking not valid in empirical arguments.
- So we were persuaded to use falsificationism, which relies solely on *deduction*.
- But in practice, we end up with *abduction*, which is arguably even less valid than induction.

Examples

- Interactive Alignment Theory
  - **Finding**: structural priming (Pickering & Branigan 1999)
  - **Theory**: Dialogue processing = mutual priming of linguistic representations (Pickering & Garrod 2004)
- The Mirror Neuron System
  - **Finding**: same neuron fires both when “participant” *perceives* and *performs* an action (Pellegrino et al. 1992)
  - **Theory**: There is a “mirror neuron system” (Iacoboni et al 2005) that is responsible for intention recognition, empathy, Theory of Mind, communication, partner selection, etc…

Examples (cont'd)

- Embodied Language Understanding
  - **Finding**: Language processing activates semantically related sensory/motoric areas in the brain (Pulvermüller 1999, 2002).
  - **Theory**: We understand language using motor simulation (Pecher & Zwaan, 2005)

Probably not limited to cognitive psychology

- Gender effect in grant funding (Albers 2015)
  - Finding: men get more funding than women fromDutch Research Council
  - Theory: gender discrimination
  - In fact: women tend to apply to fields with less funding (Albers 2015)

What is the problem with Effectism?

- **An effect is not its own explanation.**
- Take last example of embodied cognition:
- Activation of (conceptually related) sensory/motoric brain areas is at best <u>necessary</u> but never <u>sufficient</u> evidence for Embodied Language Understanding.
  - "Disembodied" (abstract, symbolic) processing could *also* activate these regions through cross-modal *priming* (e.g. Collins & Loftus 1975: semantic networks).
  - In order to activate the relevant motor cortex region, the system needs to first *recognize* the verb. So it's a circular explanation.
- The fact that processing the concept of "walking" activates leg-regions does not prove that conceptual processing is **based on** (constituted by) motoric representations/simulations.

Illustrative example

- The logic
  - Perceiving “walking” activates the leg-region in the motor-cortex, therefore understanding of verbs is based on motor-programs.
- The underlying rule
  - Perceiving P activates representation R, therefore understanding of P-things is based on R-information.
- Example
  - Perceiving “America” activates “hamburger”, and perceiving ”Italy” activates “pizza”, therefore understanding countries is based on food information.
  - “Embellied” cognition?

Effects of Effectism

- It leads to theories that only predict the effect that inspired them.
- It rewards fishing expeditions, at the expense of coherent theory building.
- It underestimates the fact that effects can have alternative causes.
- It creates a false sense of progress.
- It contributes to the replication crisis.

Why does this not happen in the Natural Sciences?

- Far more detail in the predictions
  - If I drop a ball from height *h*, it will have speed $g \sqrt{(h/0.5\, g)}$ m/s when it hits the ground. This can be tested for range of *h*'s and *g*'s
  - If all Newton could have worked with is that balls dropped from high hit the ground significantly faster than from low ($p < .05$) we would still live in the Stone Age.
- This is not to blame social science
  - Our units of analysis are much more complex, and our measurements are much more noisy, both conceptually as well as quantitatively.
  - People are far more complex, noisy, and unpredictable than atoms or billiard balls.

Summary of issues

- We (as a field) like to think of ourselves as Falsificationists, but in practice we are trying to find interesting effects and then take it from there.
- **Effectism**:
  - formulating theories that are suggested by the effects we found
  - explaining the effects with that theory
- This leads to very weak and circular theories
- It also encourages behavior that leads to publication bias and false positives. [Replication crisis]

What can we do to improve?

- Formulate theory at a higher level of abstraction than the data that have inspired it.
- Derive and test new and risky (= implausible) predictions as well. E.g.,
  - Alignment theory: will L2 speakers cause L1 speakers to copy their (L2) mistakes? [No]
  - Embodied language comprehension: If we process “the duck is swimming”, do we activate our feet-area? [?]
- Specify actual computational mechanisms (AI approach)
  - This has so far failed spectacularly in the example cases. That tells us something.

What can we do to improve?

- Generate *differential* predictions based on competing accounts (if these exist).
  - Machery (2019): *"Typically psychologists compare two theories, one, but not the other, predicting a (causal or not) relation between two or more variables."*
  - Most "competing theories" that are tested are null models.
  - It is much better if both theories predict a different effect!
  - Whatever the outcome, we learn something (and can publish it).
- Use Bayesian methods (modeling, inference).
  - We can quantify relative evidence for different theories (including the null "theory")
  - At least we get a reliable estimate of our uncertainties.

Thank you for your attention