Why theory matters:
New arguments for an old conclusion

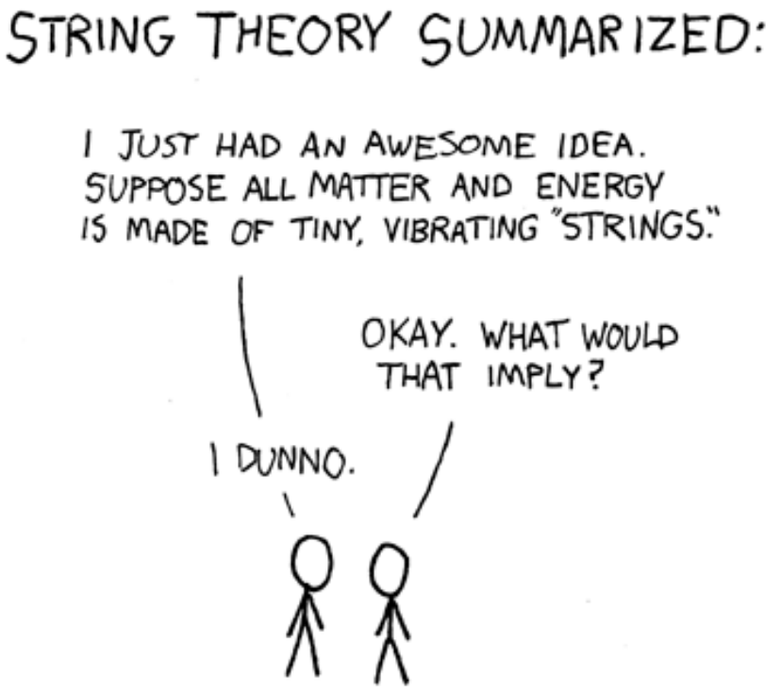
ERIM Research Transparency Campaign
Erasmus University Rotterdam
March 15, 2022

NTNU

Giosuè Baggio
Language Acquisition and Language Processing Lab
Department of Language and Literature
Norwegian University of Science and Technology

The value of theory

- A formal inferential system: representation, explanation, prediction, comparison etc.
- Weak, absent, or latent theories reduce our capacity to make valid inferences to and from data
- Rationally convincing, practically ineffective: one can generate results with little/no theory, so why bother?



van Rooij & Baggio (2020, 2021)

The fear factor

- False *negative* reports are bad, e.g., in medicine, public health, geology, ecology etc.: undetected harm is bad
- In basic science, false *positive* reports are much worse:
  - They imply more false hypotheses ('stacking')
  - They can misdirect research programs and fields
  - They fail replication: confidence/credibility crisis
- What are false positive reports? How could we keep them in check? What does theory have to do with this?

False positive report probability

Prior probability: $\pi = P(T)$
Type I error: $\alpha = P(pos \mid F)$
Type II error: $\beta = P(neg \mid T)$
Power: $1 - \beta = P(pos \mid T)$

$$P(F \mid pos) = \frac{P(pos \mid F) \cdot P(F)}{P(pos)} \quad = \frac{\alpha \cdot (1 - \pi)}{?}$$

$$\begin{aligned} P(pos) &= P(pos \wedge T) + P(pos \wedge F) \\ &= P(pos \mid T) \cdot P(T) + P(pos \mid F) \cdot P(F) \\ &= (1 - \beta) \cdot \pi + \alpha \cdot (1 - \pi) \end{aligned}$$

$$\text{FPRP} = P(F \mid pos) \quad = \frac{\alpha \cdot (1 - \pi)}{(1 - \beta) \cdot \pi + \alpha \cdot (1 - \pi)}$$

Wacholder et al. (2004); Ioannidis (2005); Colquhoun (2019); Bird (2021)

False positive report probability

$$\text{FPRP} = \frac{\alpha \cdot (1-\pi)}{(1-\beta)\cdot\pi + \alpha\cdot(1-\pi)} = \frac{1}{\frac{1-\beta}{\alpha}\cdot\frac{\pi}{1-\pi}+1}$$

«The prior odds for my hypothesis are 50/50. If I minimize Type I and Type II error rates (i.e., maximize power), then FPRP would be minimized, too.» (Assuming no bias etc.)

$$\text{FPRP} = \frac{1}{\frac{1-\beta}{\alpha}\cdot\frac{\pi}{1-\pi}+1} = \frac{1}{\frac{1-0.2}{0.005}\cdot\frac{0.5}{1-0.5}+1} = 0.006$$

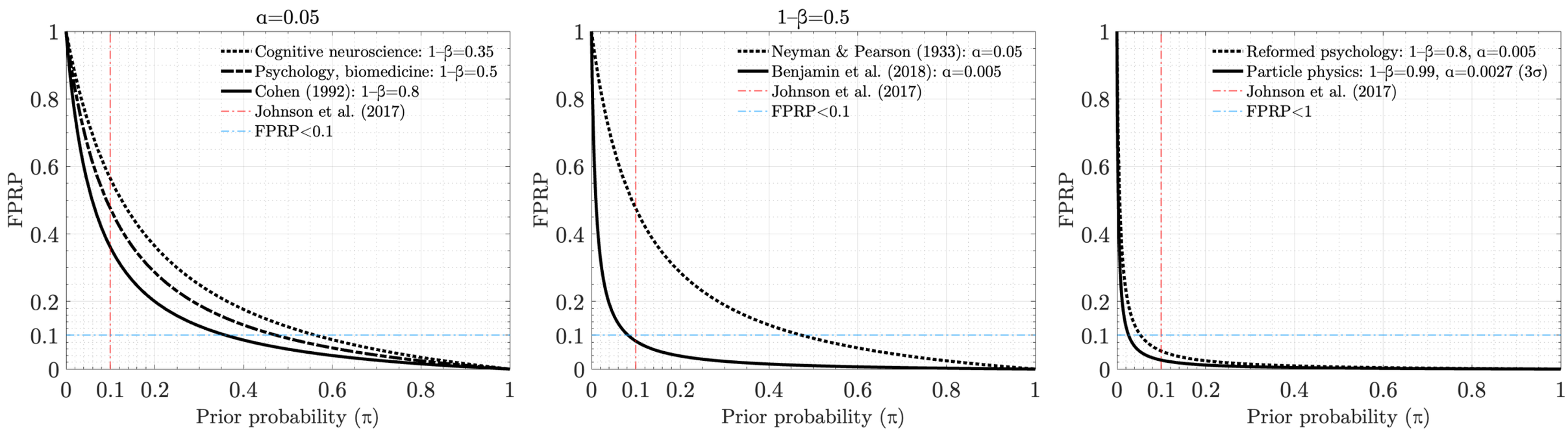
π in the sky?

- A (null) hypothesis may be true or false, but that does not make the prior odds 50/50 (principle of indifference)
- Prior probability
  - Degree of belief in a hypothesis [0 1], based on local data (e.g. prior experiments) and wider web of belief
  - Ratio of true hypotheses to tested hypotheses:
    - GWAS: 30,000 genes; 50,000-250,000 functional variants; 1-5 SNPs contribute to the condition; prior probability is between 0.0001 and 0.00001
    - 'Big data': similar problems in neuroimaging etc.

The replication crisis revisited

- High FPRP and low replication rates from high-quality science, if there is a high rate of false hypotheses before testing and insufficient Type I error control (Bird 2021)
- Evidence of low π, e.g., 0.1 in psychology (Johnson et al. 2017), low replication rates, current statistical practices
- Possible (not mutually exclusive) ways out?
  1. Readjust belief in published results (skepticism)
  2. Methodological reform, e.g., Type I/II error control
  3. Generate hypotheses more likely to be true (theory)

Methodological reform



- Increasing power and redefining significance reduce FPRP, but not for extreme values of prior probability ('the π trap')
- Known problems may be worsened by methodological reform: Meehl's paradox, nonlinearity of FPRP function for low π etc.

Button et al. (2013); Fraley & Vazire (2014); Szucs & Ioannidis (2017); Dumas-Mallet et al. (2017)

randomized controlled trials.

Finally, instead of chasing statistical significance, we should improve our understanding of the range of R values—the pre-study odds—where research efforts operate [10]. Before running an experiment, investigators should consider what they believe the chances are that they are testing a true rather than a non-true relationship. Speculated high R values may sometimes then be ascertained. As described above, whenever ethically acceptable, large studies with minimal bias should be performed on research findings that are considered relatively established, to see how often they are indeed confirmed. I suspect several established "classics" will fail the test [36].

Nevertheless, most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds. We should then acknowledge that statistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the report and in the relevant field at large. Despite a large statistical literature for multiple testing

One potential limitation to the FPRP approach is the challenge of assigning a range for prior probability. However, investigators informally use prior probabilities already to decide which experiments to perform, which studies to field, and which specific hypotheses to test in those studies, and for interpreting results. With more experience, investigators should be better able to determine a prior probability. In the meantime, however, a crude classification of prior probabilities into low, medium, and high or a sensitivity analysis of FPRP across a range of prior probabilities should be an improvement over the current practice of relying entirely on statistical significance. An additional important benefit of considering prior probability is that investigators are forced to evaluate the existing evidence before seeing the results of their own study.

5.2. Higher quality hypotheses

The second approach advocates trying to generate hypotheses with a greater probability of truth. This is, of course, rather easier said than done. In medicine this amounts to reaffirming the importance of basic research. It may also encourage the continuation of basic research beyond the point when a new therapeutic idea becomes plausible. Rather than moving straight to a trial, as in the case of Bapineuzumab, further basic research may be a better use of resources, so that hypothesized mechanisms of action can be better confirmed before being used to generate or support a proposal regarding an intervention. In some cases greater use of observational data may be used to supplement mechanistic hypotheses. In social psychology matters are different. Here the problem is not so much an incomplete basic theory, but lack of any basic theory at all; and the prospect of one capable of generating high-quality hypotheses is some way off. In any case the route to that theory will most likely be via the collection first of a set of reliable results of the general kind that social psychology is currently trying to produce. So it will be difficult directly

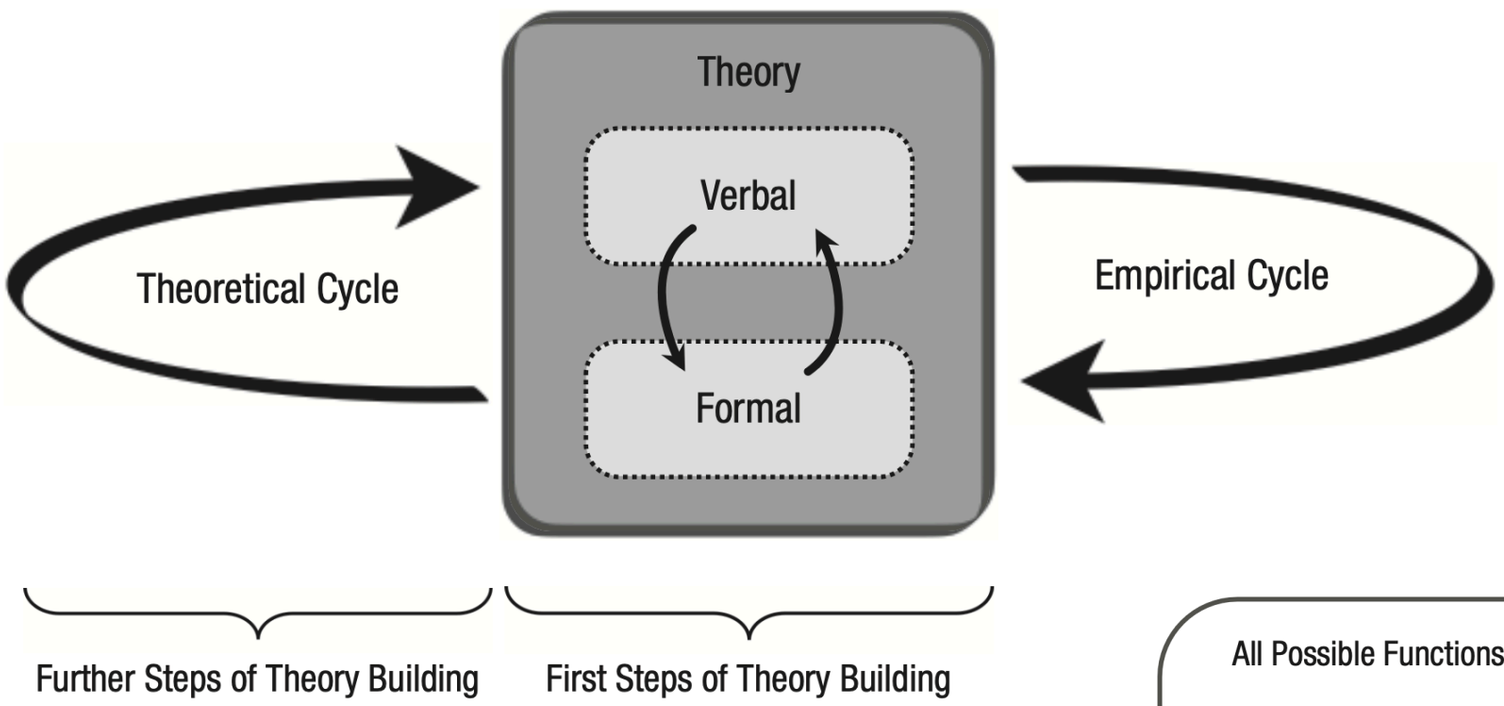
Ioannidis (2005); Wacholder et al. (2004); Bird (2021)

Escaping the π trap

- Testing our way out of the π trap is not an option: few positive results, a high false/true positive ratio etc.
- 'Improve understanding of the prior odds' (Ioannidis), '(In)formally develop prior probabilities' (Wacholder), 'Try to generate higher quality hypotheses' (Bird)
- 'More theory' is not the answer, rather 'better theory'
- Good theory is formal, representational, explanatory, predictive etc.: the traditional 'inferential system' view
- Good theory is also aimed at avoiding low π: a new view

van Rooij & Baggio (2020, 2021)

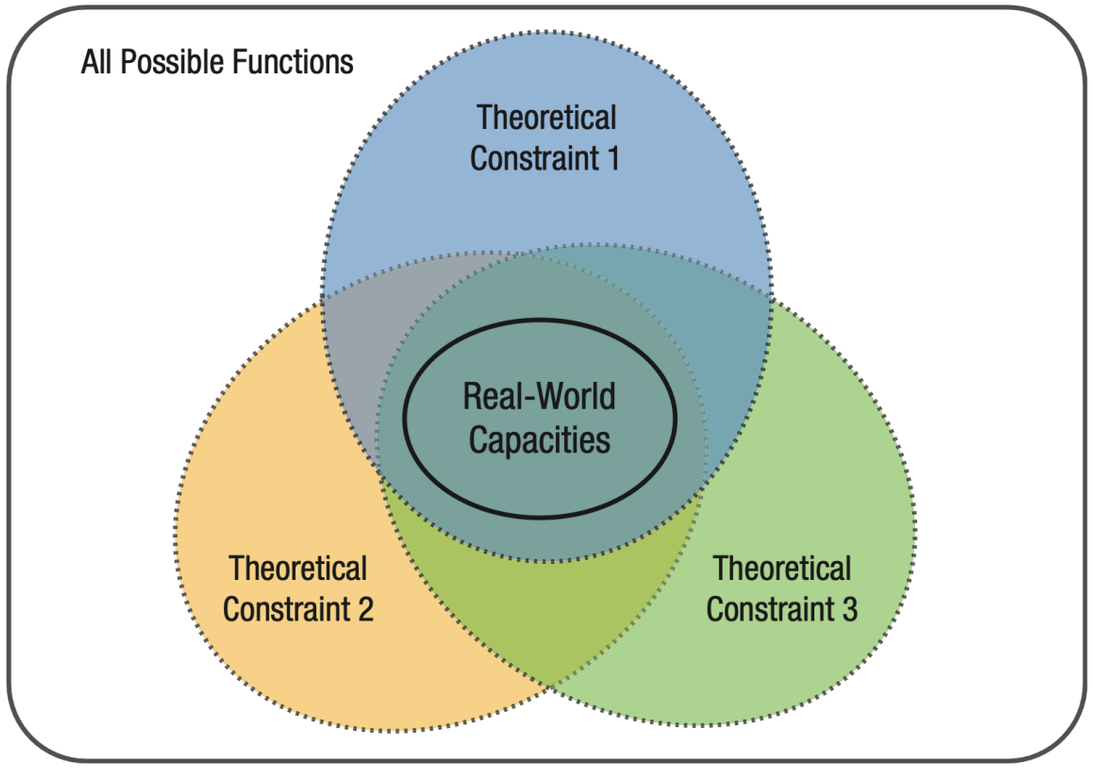
Escaping the π trap



Constraints in cognitive science:

- Computability, tractability of function
- Learnability, evolvability of function
- Substrate compatibility (e.g., biology)

...



van Rooij & Baggio (2020, 2021)

Take home messages

- Low prior probability of hypotheses (π) is a problem
- Methodological reform is important and useful but is not sufficient: it is ineffective for low π, unnecessary for high π
- The empirical cycle is not a way out of the π trap: too many false positive reports (unreformed science) or too few positive reports (reformed science) for low π values
- Theoretical cycle: formal plausibility analyses to discard low prior probability hypotheses before empirical testing
- Traditional arguments for theory qua inferential system must be used parsimoniously; new arguments are needed

Thank you

Based on joint work with Iris van Rooij:

- van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, 31(4), 321-325.
- van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 16(4), 682-697.

Twitter: @giosuebaggio, @IrisVanRooij