Crisis of experimental science: roots & remedies

Michael Franke

A noxious melange of psychological and sociological factors undermines optimal scientific practices. Direct replication, preregistration, data sharing, Bayesian data analysis and adversarial collaborations are some of the possible remedies to alleviate the problems.

The publication generating process

- Just as we can think about the observation generating process in the 3-cards problem in a (false) naïve or a (correct) sophisticated way, we can have a more or less accurate picture of the *publication generating process*.
  - It would be too naïve to take as true every published result, even if supported by empirical data and seemingly sound statistical analyses.
  - The nature of the publication generating process leads to many false research claims, despite empirical data and sound statistics.
- Problems can arise at all stages:
  1. research topic/question/hypothesis
  2. study design & materials
  3. data collection
  4. data processing
  5. statistical analyses
  6. reporting results
  7. narrative integration
  8. publication

Figure 1: Sin of Bias.

Confirmation bias

- Tendency to select, favor, recall or interpret information in ways supportive of currently held beliefs or opinions.
- Wason's selection task (see Figure 2)
- Confirmation bias could be a rational adaptation in a world where what matters is the winning of arguments (social status) not true knowledge gain (Mercier & Sperber).
- Consequences for the publication generating process:
  - tendency to favor conceptual over direct replication[1]

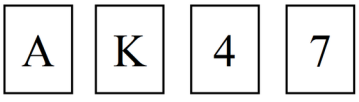

Figure 2: Example item from Wason's card selection task. Participants must select all cards relevant to testing the rule "If there is a vowel on one side, the other side is an even number" where all cards have a number on one side and a letter on the other.

[1] *Conceptual replication* examines predictions of a general idea which was previously tested in one scenario in a different setting. *Direct replication* tries to recreate the exact conditions *C* from a previous experiment believed necessary for effect *X* and tests whether *X* is observed in a new experiment which implements *C*.

- publication bias towards "confirmatory results," away from "ambiguous results" or "negative results"[2]
- narrative over substance: a convincing and coherent grand story is more important than methodological soundness

[2] An *ambiguous result* is one which does not clearly speak for a concrete conclusion. A *negative result* is one which does not give a significant test result that would refute the null hypothesis.

Hidden flexibility: researcher degrees of freedom

aggressive design create a design and stimulus material so as to promote the likelihood of the desired outcome

- Jones wants to test the hypothesis that surface scope readings are most salient. He measures the reading difficulty on an anaphoric pronoun. He picks the first sentence, not the second:

(1) Every ten minutes a man gets mugged in NY city. He is one miserable bastard.

(2) Every ten minutes a light blinks on the machine. It indicates full functionality.

garden of forking paths getting lost despite honest intentions, ending up with unintentional *p*-hacking

p-hacking intentionally trying to turn a non-significant test result into a significant test result, e.g., by:

- trying different tests
  - two-sided instead of one-sided test
  - regression instead of ANOVA
  - Bayes vs. frequentist
- excluding data points
  - all data from subjects who made too many mistakes
  - all data from subjects who took too long
  - all data from subjects who said "bla" in the post-questionnaire[3]
- reinterpreting the dependent measure
  - ordinal rating scale data as metric
  - proportional data as metric
- choosing a dependent measure
  - eye-tracked reading study: first-pass, regressions, ...
  - EEG: time window, region of interest, preprocessing
  - mouse-tracking: AUC, XNeg, TTT, Entropy, ...
- including additional factors
  - gender, handedness, ...
  - interaction terms in regression analyses

[3] Subjects' post-survey comments may make it seem very legitimate to exclude their data, e.g., those guys *obviously* did not understand the experiment.

  - no, smaller or bigger mixed-effects structure

biased stopping freedom to stop data collection based on test results guarantees a significant outcome in the limit (see Figure 3)

biased debugging double-check only in case of non-significant result

HARKing changing the hypothesis after the results are known[4]

- *post hoc* analyses
- hindsight bias

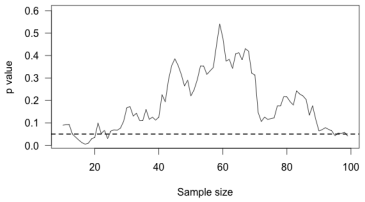

Figure 3: Development of the p-value as more and more data trickles in.

[4] It's here that psychology is particular vulnerable.

Potential remedies

- wide-spread direct replication
  - career incentives
  - grants
  - reproducibility index[5]
  - pottery barn rule[6]
- simple preregistration
  - commitment before data collection on details of data processing, analysis and interpretation
  - upload declaration of intention with dummy analysis scripts to, e.g., `https://osf.io`
- peer-reviewed registered reports (see Figure 4)
- disclosure statements
  - the 21-word solution:

    ```
    We report how we determined our sample size, all
    data exclusions (if any), all manipulations, and
    all measures in the study.
    ```

- Bayes factors instead of p-values[7]
- adversarial collaborations[8]
- open data
  - supply all data, experimental scripts and materials at all stages during review and after publication
  - maximally possible transparency of choices[9]
- raising awareness from the earliest point during education[10]

[5] Keeping track how many of a journal's published results replicate; similar to the impact factor, this could become a sign of good quality research.

[6] Journal that publishes a paper is committed to publish any direct replication.

[7] Bayes factors quantify evidence (also in favor of the null hypothesis). Adopting Bayesian methods might transform the way we think about "publishable results".

[8] Teams of researchers with opposing preconceptions, beliefs or opinions. Contra confirmation bias.

[9] Full transparency is impossible to achieve in practice. Cheaters will cheat. Liars will lie.

[10] This class.

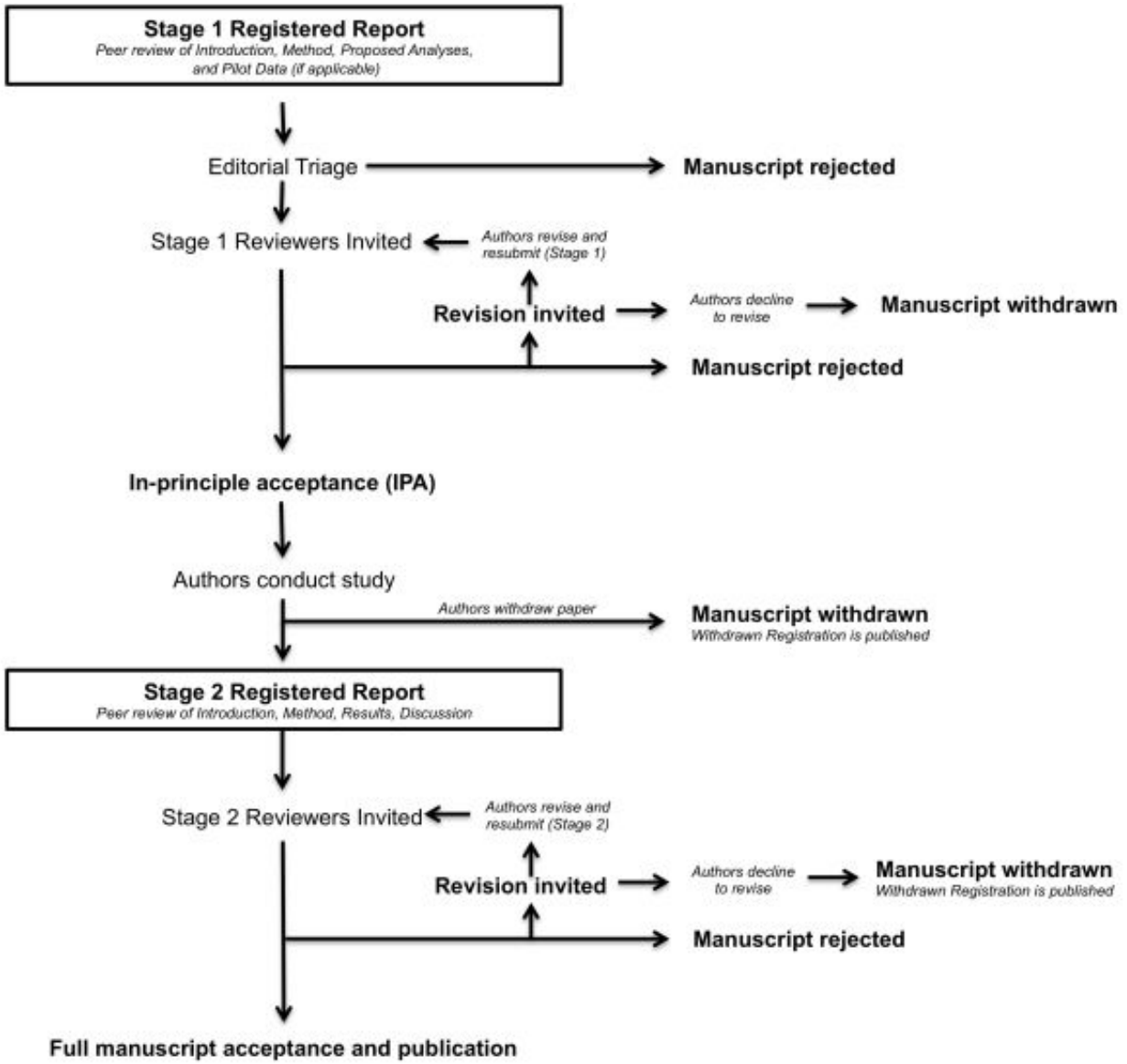

Figure 4: Process of peer-reviewed registered reports. See Chapter 8 of Chalmbers "The Seven Deadly Sins of Psychology"