XPLab 2019

Michael Franke

LabPrac, April 3 2019

> Experimental work is hard. Opportunities for suboptimality and failure abound. This course is all about avoiding pitfalls and cultivating a mindset aimed at continually improving practices. We will execute the whole process of implementation, execution and data analysis during this course, based on a replication of an existing experiment, which we will preregister.

The experimental method

Let's look at some fictitious case studies (*science fiction* if you wish). We will use them to remind ourselves of the benefits of experimental methods and of the perils of naivety about their limitations.

Objective evidence

Smith has talked to a lot of people during the last 10 years and made extensive notes. He claims that using the right toothpaste makes you smarter. Smith knows this because he talked to a lot of people and made extensive notes. Why do we not believe him?

Observation vs. manipulation

Smith subjected 700 people to an IQ test. He also recorded for each participant which toothpaste they use regularly. (There are only two brands: *bling* and *shiny*.) Here's a visualization of his data:

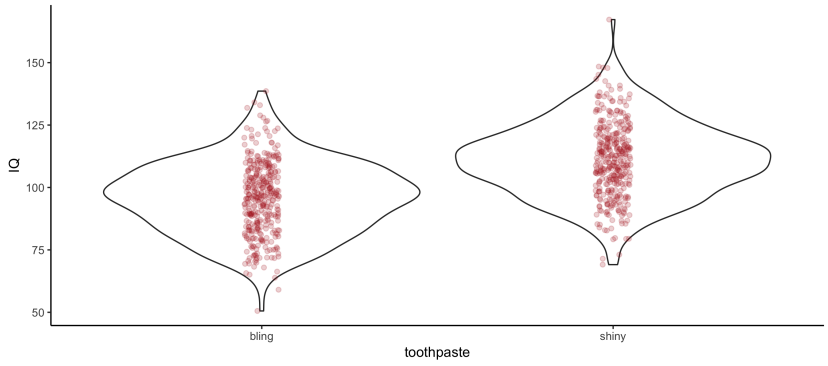


Figure 1: Data from an observation study

A statistical test reveals that there is a significant difference between the two groups of toothpaste users. Smith publishes a paper with the title: "*shiny* makes you smart." Why do you strongly dislike this paper?

The publication-generating process

Smith recruited 50 participants. Each used one brand of toothpaste for 4 weeks before taking an IQ test. A statistical test reveals that there is significant difference between the two groups. Smith submits a research paper with the title "You are what you brush: *shiny* makes you smart." to a top-tier journal.

Meanwhile, Smith has independently carried out the same experiment. A statistical test on her data reveals no significant difference between groups. Jones still submits a research paper to a top-tier journal with the title "Expect the expected: toothpaste does not influence IQ scores."

Three months later, Smith's paper gets published, Jones' doesn't. Why is this disturbing?

Researcher degrees of freedom

Jones is frustrated by the rejection. She looks at the data again. She realizes that toothpaste does have a significant effect on IQ scores after all, but only for right-handed participants and the subset of IQ-questions related to language. She also realizes that this ties in with professor Brainstawn's work on lateralization. She submits a paper to a different journal. The paper is accepted as: "Brush up your language the right way: toothpaste influences on IQ and lateralization in the brain." Why is this bad for science?

Crisis of experimental science: roots & remedies

A noxious melange of psychological and sociological factors undermines optimal scientific practices. We will only look at publication bias and researcher degrees of freedom here. Direct replication, preregistration, data sharing and other practices of open science are some of the possible remedies to alleviate the problems.

Publication bias

Significant findings have a higher possibility of being published. Negative results go into the file drawer. This may result in many published research findings being actually false.

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[1]
- reinterpreting the dependent measure
  - ordinal rating scale data as metric
  - proportional data as metric

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

- 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
  - 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 2)

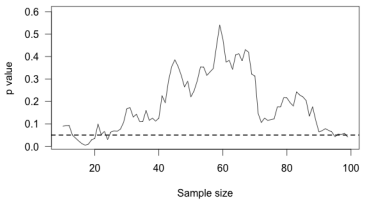


Figure 2: Development of the p-value as more and more data trickles in.

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

HARKing changing the hypothesis after the results are known[2]

- *post hoc* analyses
- hindsight bias

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

Potential remedies

- wide-spread (direct) replication[3]
  - career incentives
  - grants
  - reproducibility index[4]
  - pottery barn rule[5]
- 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 3)
- 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[6]

[3] *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*.

[4] 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.

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

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

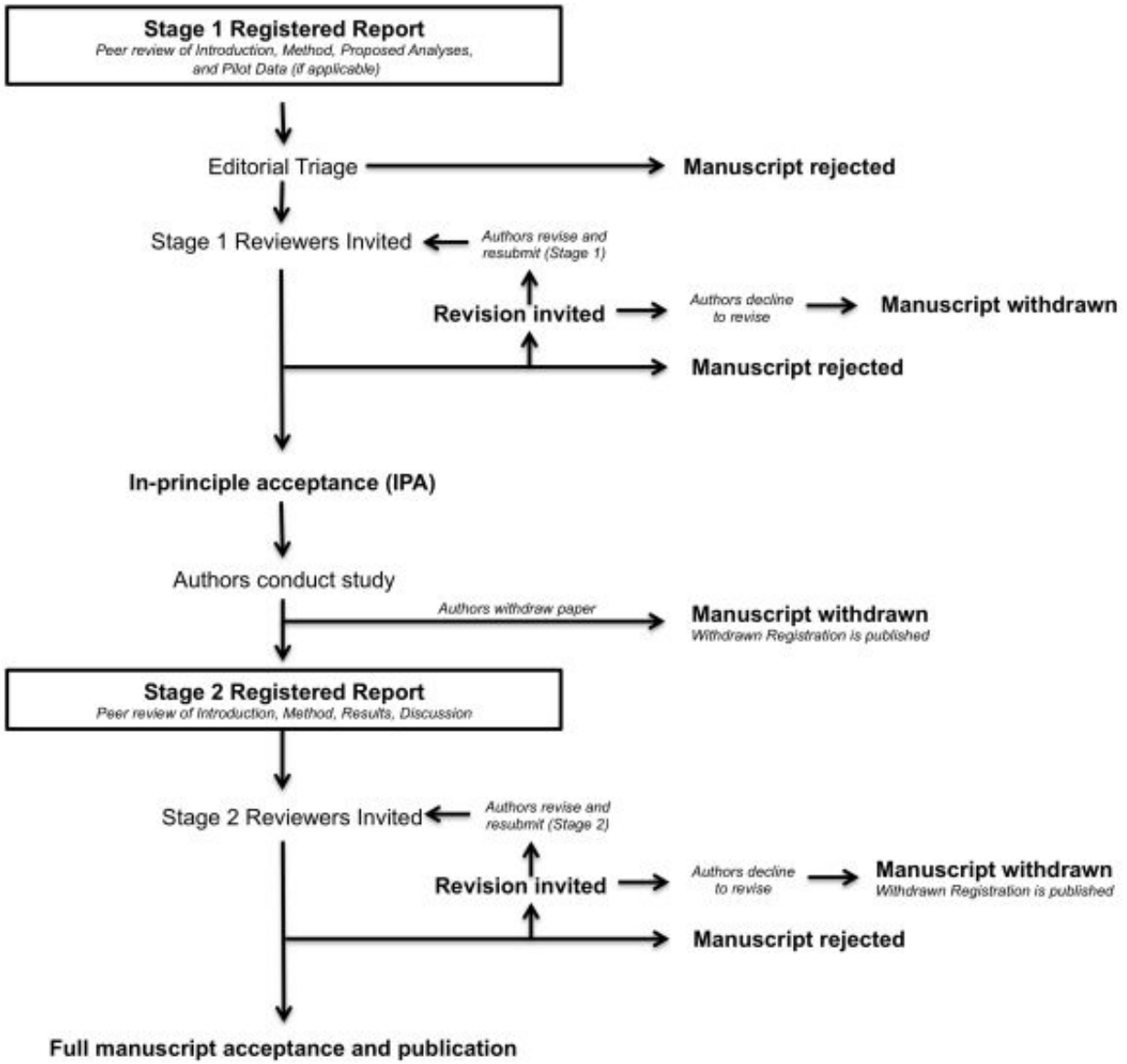


Figure 3: Process of peer-reviewed registered reports. See Chapter 8 of Chalmbers "The Seven Deadly Sins of Psychology"

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

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

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

[9] This class.

Learning goals

We will conquer new concepts and tools.

Concepts

- replication ::: preregistration ::: open science
- experiment design
- cooperation ::: version control ::: issue tracking
- data wrangling ::: visuals ::: analysis
- tidiness
- crowdsourcing

Tools

- git & markdown
- HTML, CSS & Javascript
- R, tidyverse, Rmarkdown
- ggplot

Procedure

The course has two parts. In the first part we will:

1. discuss key ideas to motivate what we are doing;
2. go through the whole cycle of implementation, preregistration, execution and analysis once together.

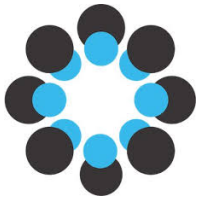
In the second part, teams of 2-5 members pick an existing study and try to replicate it.

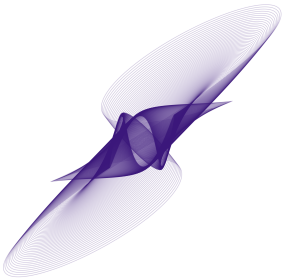
`_babe`

The focus of this course is on **browser-based experiments**. `_babe` provides templates and functionality for implementation and deployment. More information is here: `https://babe-project.github.io/babe_site/`

Using `_babe` in this course has advantages and disadvantages.

You need not care about open science and reproducibility. But this course teaches you that it is no more complicated that staying in hiding.





`_babe`
basic architecture for browser experiments

Disadvantages

- slightly steeper learning curve
- less accuracy of measurement

Advantages

- versatile ::: accessible ::: non-proprietary
- offline and online deployment
- domain-general skills (web-app!)

`_babe` is work in progress. We need your input: feedback, criticism, user stories, active development ...

Homework for next class

- install git for the command line from `https://git-scm.com`
- familiarize yourself with git by exploring the documentation resources available from `https://git-scm.com/doc`

You might just enjoy a video!

- open an account on GitHub at `https://github.com`
- read and execute the instructions in the guide for GitHub at `https://guides.github.com/activities/hello-world/`
- clone the repository that contains the website and material for this course from:
  `https://github.com/michael-franke/XPLab2019`
- read the guide on markdown available at `https://guides.github.com/features/mastering-markdown/`