The reproducibility crisis in the life sciences

Workshop outline

I) Introduction to the “reproducibility crisis”

II) Exchange of experiences

- Is there a reproducibility crisis?
- Implications for different research areas
- Reasons for poor reproducibility

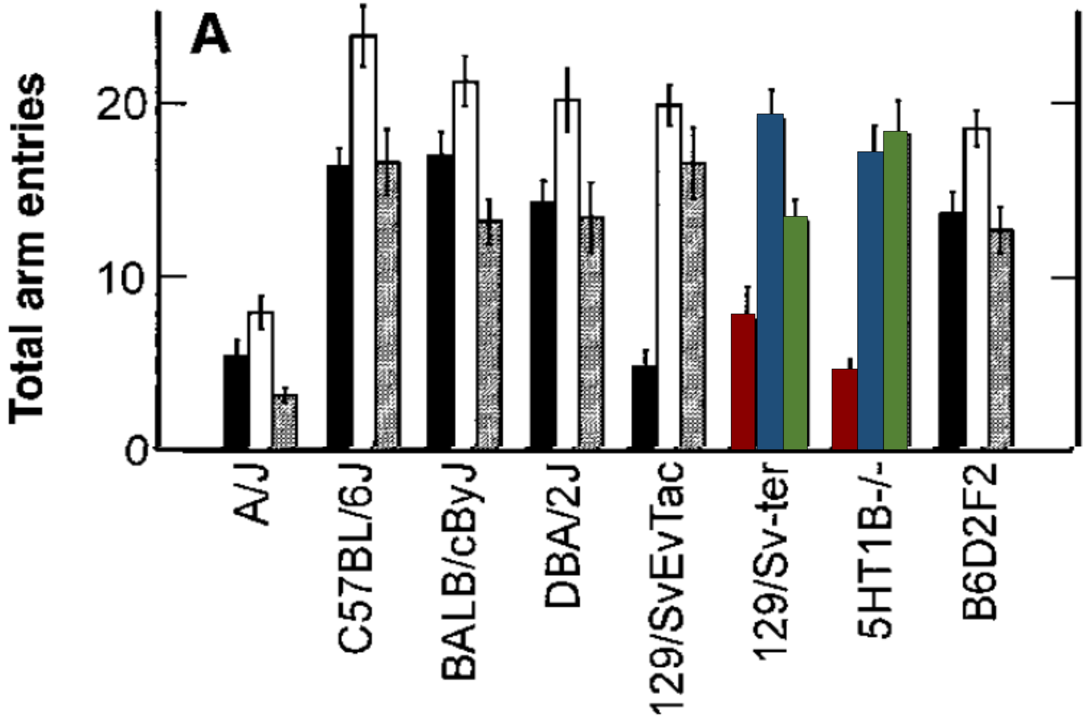
III) Journal Club: Strategies to improve reproducibility?

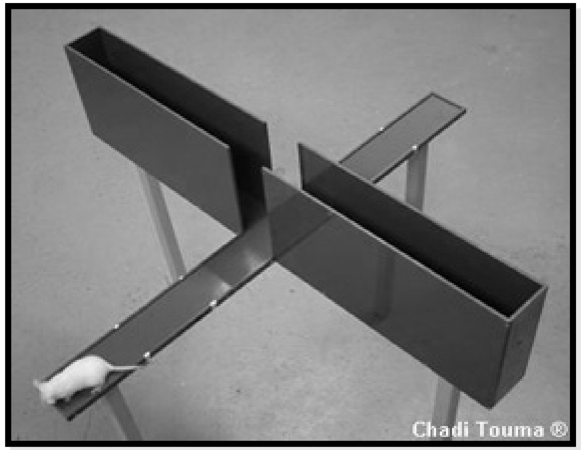
- 4 publications, 4 different ideas
- Think-pair-share
- Discussion

Introduction to the “reproducibility crisis”

Never replicate a successful experiment?

Reproducibility in animal experimentation





🟩 Portland
🟦 Edmonton
🟥 Albany

- ➢ Study testing behavioural differences between inbred and mutant strains of mice
  - Exactly the same procedures in three different laboratories (mice ordered from same breeders, housing under the same conditions, using the same protocols for testing)
  - Crabbe et al. concluded: “Experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory”

Crabbe et al., Science, 284: 1670-1672, 1999.

- After these very initial findings about poor reproducibility in animal experimentation, several other studies have been conducted afterwards, confirming the overall problems with reproducing the same results

- Current estimated rates of irreproducible results range between 50-90 % in preclinical animal research

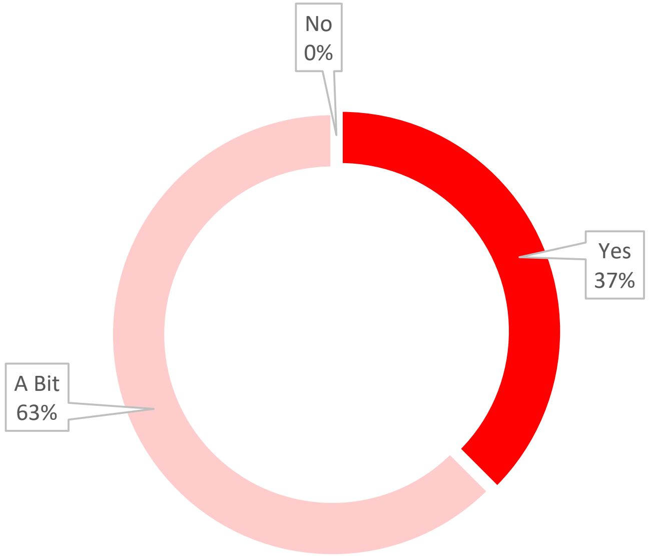
- US $ 28B/year for irreproducible preclinical research in the United States

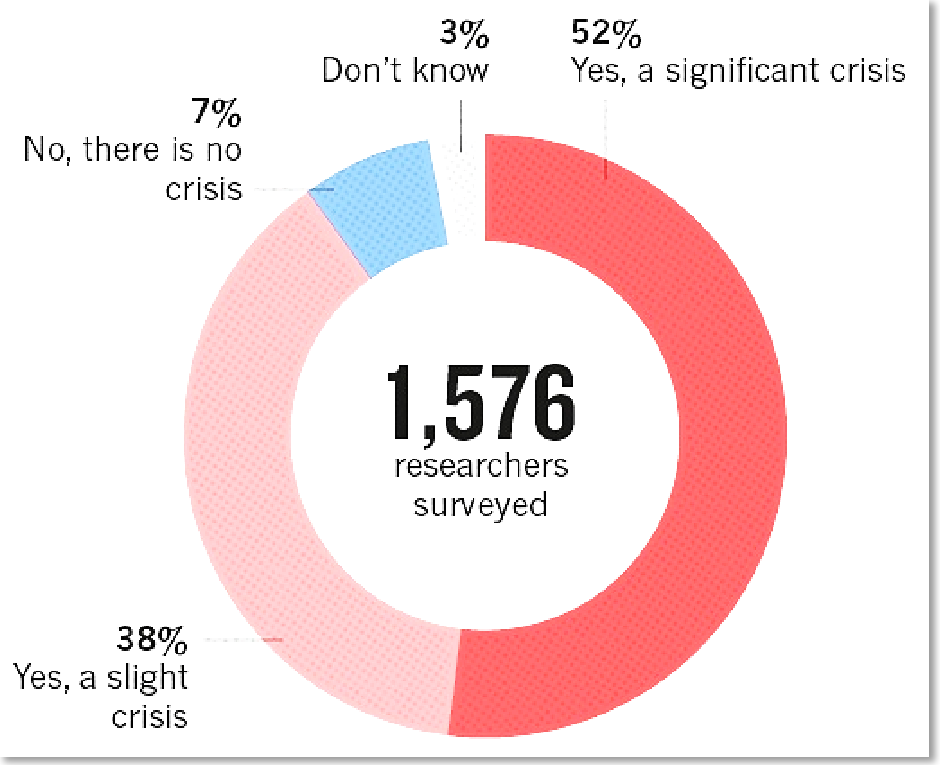
Freedman et al., PLoS Biology. 13(6): e1002165, 2015.

Never replicate a successful experiment?

Is there a reproducibility crisis?

Your Answers!



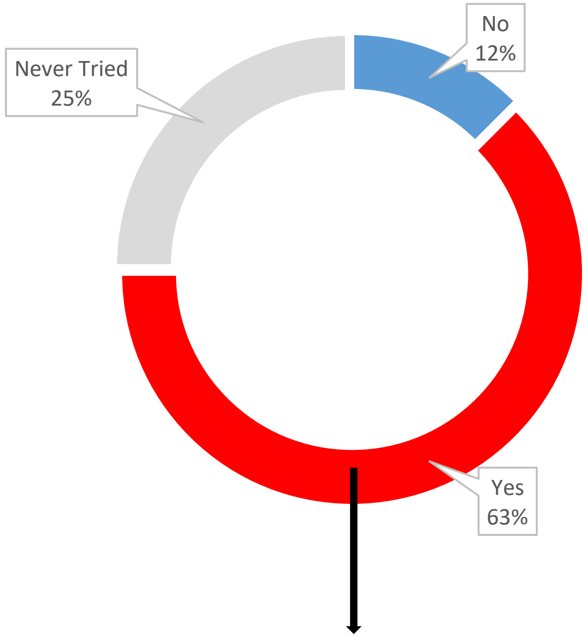


Baker, Nature, 533: 452-454, 2016.

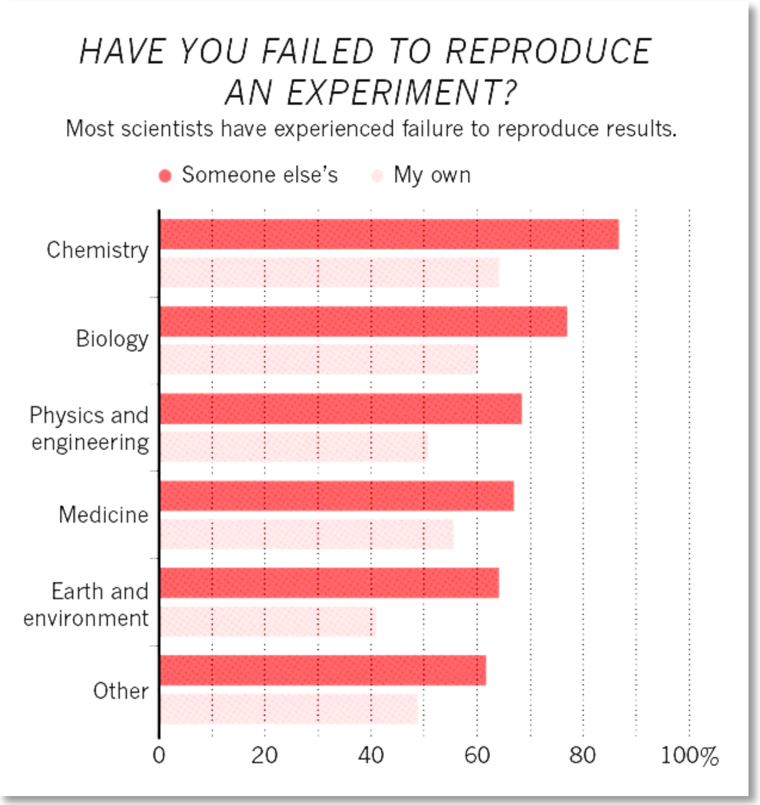
Never replicate a successful experiment?

Have you failed to reproduce an experiment?

Your Answers!



How many times: 1 to “many”


HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.

Baker, Nature, 533: 452-454, 2016.

Never replicate a successful experiment?

Some definitions…

- Reproducibility, Repeatability, Replicability:

  different and sometimes conflicting meanings
- Industrial systems:
  - Reproducibility -> Difference between testers under different conditions
  - Repeatability -> Repeated evaluations under identical conditions
- Genome studies:
  - Replicability -> Repetition by same lab or researchers but with a different technology or dataset
- Preclinical studies:
  - Reproducibility -> Recreating the same numbers by different labs

- How to make sense of this?

Kenett & Shmueli 2015, Nature Methods.

Never replicate a successful experiment?

Some definitions…

- It depends on the intended generalization of the study
  - Statistical generalizability -> Inferring from a sample to a target population
  - Scientific generalizability -> Applying a model based on a particular target population to other populations

- Example – Industrial system:
  - Repeatability assesses measurement error of a device for future use -> test conditions constant, different testers -> statistical generalization
  - Reproducibility aims at generalizing to future use under different testing conditions -> scientific generalization

Kenett & Shmueli 2015, Nature Methods.

Never replicate a successful experiment?

Some definitions…

- External validity:

Applicability of a result to other conditions, populations or species

The extent to which a result can be generalized

Richter et al 2009, Nature Methods.

Never replicate a successful experiment?

What kind of problems are being discussed?

- Poor experimental design / analysis of experiments (internal validity)

✓ Experimenter bias (> blinding)

✓ Selection bias (> randomization)

✓ Detection bias (> sample size calc.)

✓ Statistical bias (> e.g., multiple testing correction)

✓ Choice of wrong experimental unit

✓ Inclusion of wrong control groups

- Poor welfare of laboratory animals (e.g., stereotypies)
- Choice & suitability of animal model
- Publication bias, selective reporting & p-hacking
- Standardization (external validity)

Van der Worp et al., PLoS medicine, 7: e1000245, 2010; Richter, Lab Animal 46: 343-349, 2017.

Never replicate a successful experiment?

What solutions? Example 1 – Pre-registration

- Preregistration separates hypothesis-generating (exploratory) from hypothesis-testing (confirmatory) research

- Confirmatory Research
  - Hypothesis testing
  - Results are held to the highest standards
  - Data-independent
  - Minimizes false positives
  - P-values retain diagnostic value
  - Inferences may be drawn to wider population

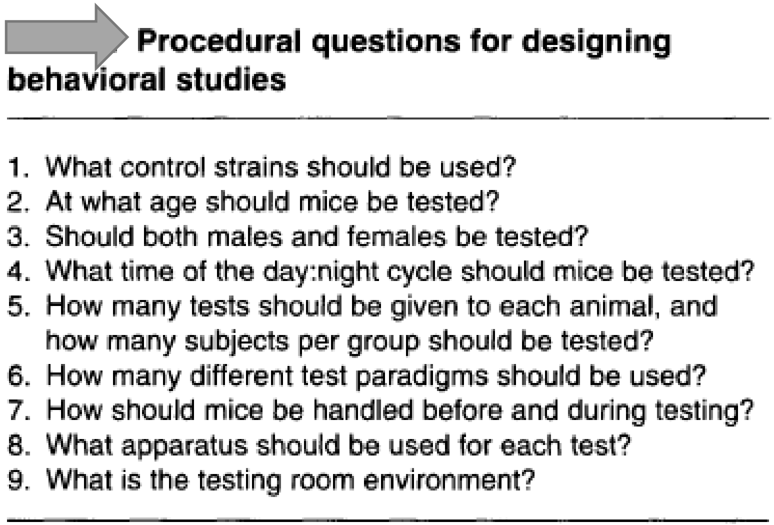
- Exploratory Research
  - Hypothesis generating
  - Results deserve to be replicated and confirmed
  - Data-dependent
  - Minimizes false negatives in order to find unexpected discoveries
  - P-values lose diagnostic value
  - Not useful for making inferences to any wider population

Nosek et al. 2017.

Never replicate a successful experiment?

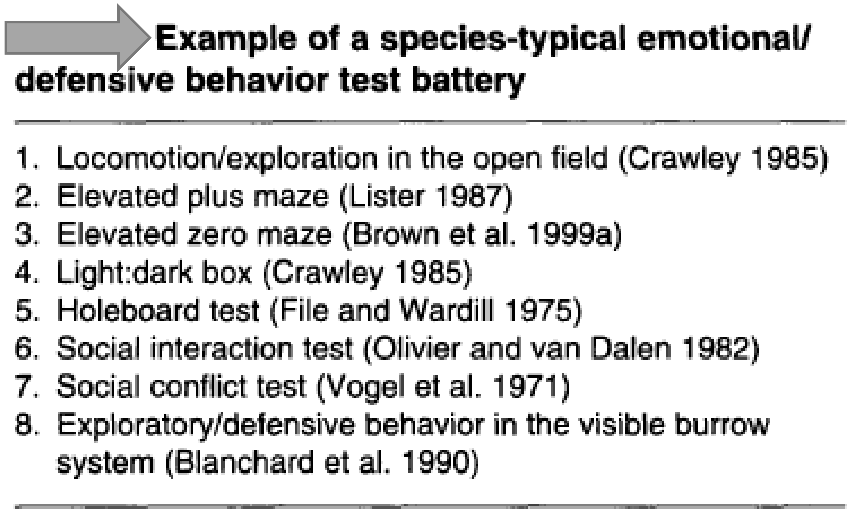
What solutions? Example 2 – Test Batteries

- Current technology allows for simultaneous testing of multiple outcome measures



Procedural questions for designing behavioral studies

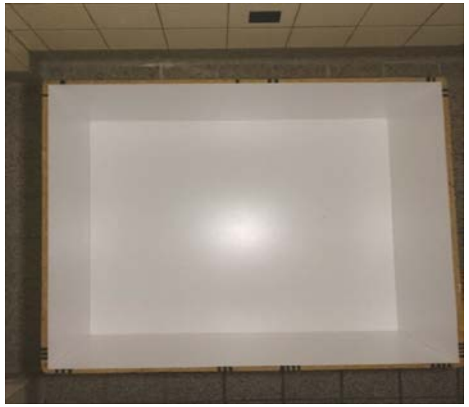
1. What control strains should be used?
2. At what age should mice be tested?
3. Should both males and females be tested?
4. What time of the day:night cycle should mice be tested?
5. How many tests should be given to each animal, and how many subjects per group should be tested?
6. How many different test paradigms should be used?
7. How should mice be handled before and during testing?
8. What apparatus should be used for each test?
9. What is the testing room environment?



Example of a species-typical emotional/ defensive behavior test battery

1. Locomotion/exploration in the open field (Crawley 1985)
2. Elevated plus maze (Lister 1987)
3. Elevated zero maze (Brown et al. 1999a)
4. Light:dark box (Crawley 1985)
5. Holeboard test (File and Wardill 1975)
6. Social interaction test (Olivier and van Dalen 1982)
7. Social conflict test (Vogel et al. 1971)
8. Exploratory/defensive behavior in the visible burrow system (Blanchard et al. 1990)

Brown et al 2000.

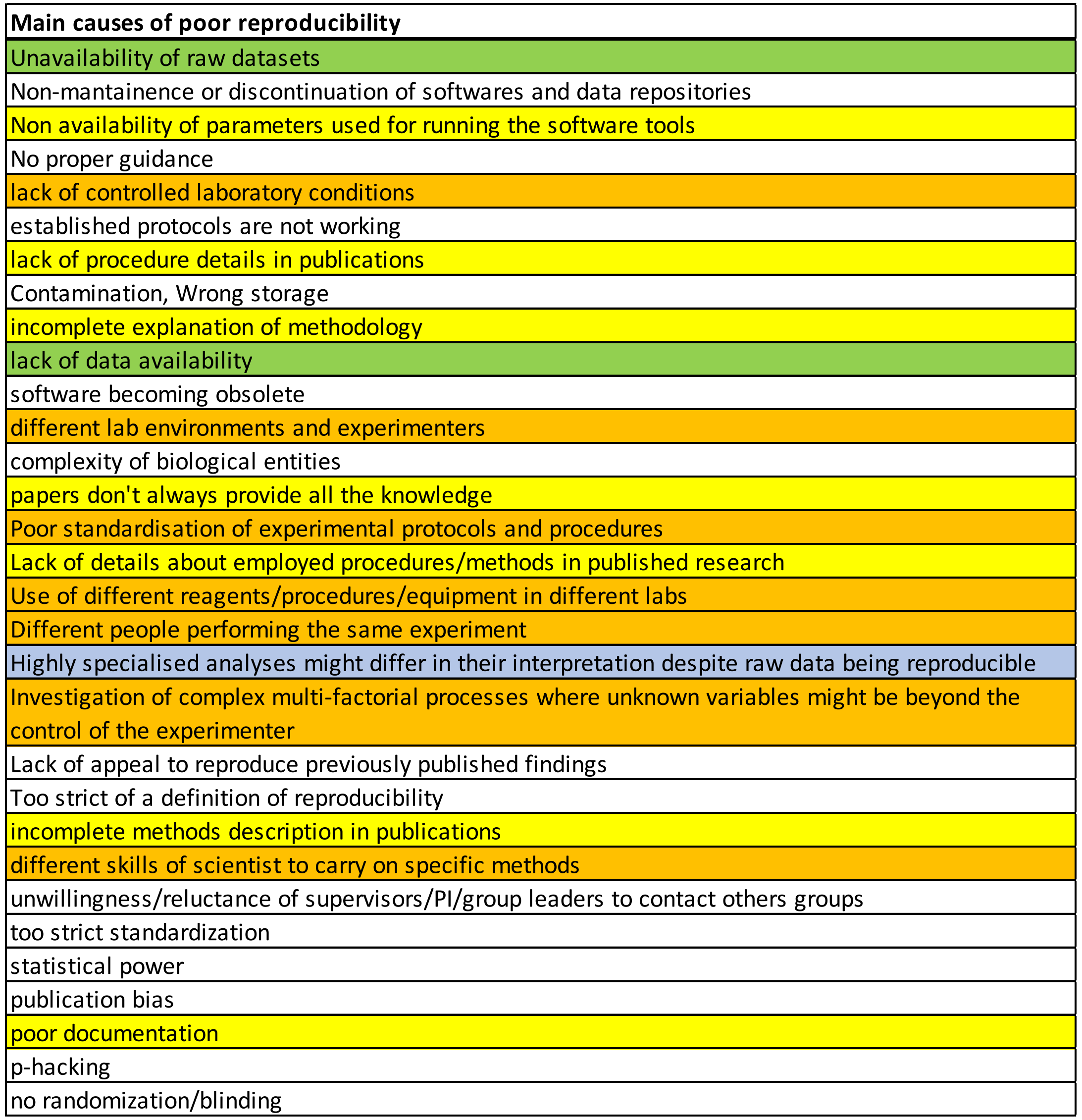


- ➢ More robust trait measurement with appropriate correction (e.g. Benjamini-Hochberg)

Exchange
# of experiences

Never replicate a successful experiment?

What are the main causes of poor reproducibility in your research field?



Main causes of poor reproducibility
Unavailability of raw datasets
Non-mantainence or discontinuation of softwares and data repositories
Non availability of parameters used for running the software tools
No proper guidance
lack of controlled laboratory conditions
established protocols are not working
lack of procedure details in publications
Contamination, Wrong storage
incomplete explanation of methodology
lack of data availability
software becoming obsolete
different lab environments and experimenters
complexity of biological entities
papers don't always provide all the knowledge
Poor standardisation of experimental protocols and procedures
Lack of details about employed procedures/methods in published research
Use of different reagents/procedures/equipment in different labs
Different people performing the same experiment
Highly specialised analyses might differ in their interpretation despite raw data being reproducible
Investigation of complex multi-factorial processes where unknown variables might be beyond the control of the experimenter
Lack of appeal to reproduce previously published findings
Too strict of a definition of reproducibility
incomplete methods description in publications
different skills of scientist to carry on specific methods
unwillingness/reluctance of supervisors/PI/group leaders to contact others groups
too strict standardization
statistical power
publication bias
poor documentation
p-hacking
no randomization/blinding

1. Incomplete information from Methods
2. Lack of standardization
3. Insufficient data sharing
4. Different analysis – different results
5. Lack of appeal to replicate same study

Never replicate a successful experiment?

Exchange of experiences

- Do you think that the reproducibility crisis is specific to some research areas (e.g. behavioural research) or is it a more general phenomenon?
- What specific problems with reproducibility did you have?

- Do you know all of the described pitfalls / reasons for poor reproducibility?
- Are there other or additional reasons for poor reproducibility?

- Have you heard of the “Standardization Fallacy” term?
- Would “heterogenization” also be an approach for your field?
- What kind of alternative strategies (beyond “heterogenization”) would you see to improve reproducibility in your field?

Never replicate a successful experiment?

The “Standardization Fallacy” problem

- Standardisation bad for reproducibility? Counterintuitive for many researchers at first

- However, differences between laboratories are *unavoidable* (e.g., the animals are different, the experimenters interacting with the animals are different, the gut flora of the animals varies, etc.)

- These differences can affect the animals’ phenotype and thus the outcome of the study

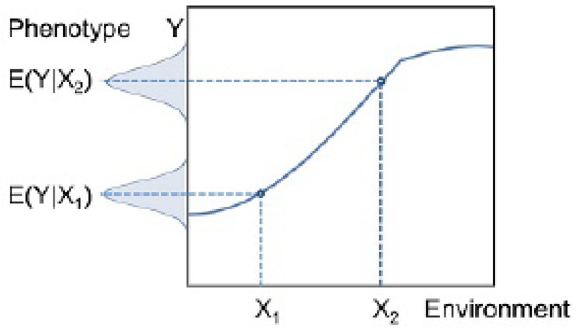
- Different labs inherently standardise to different local study conditions

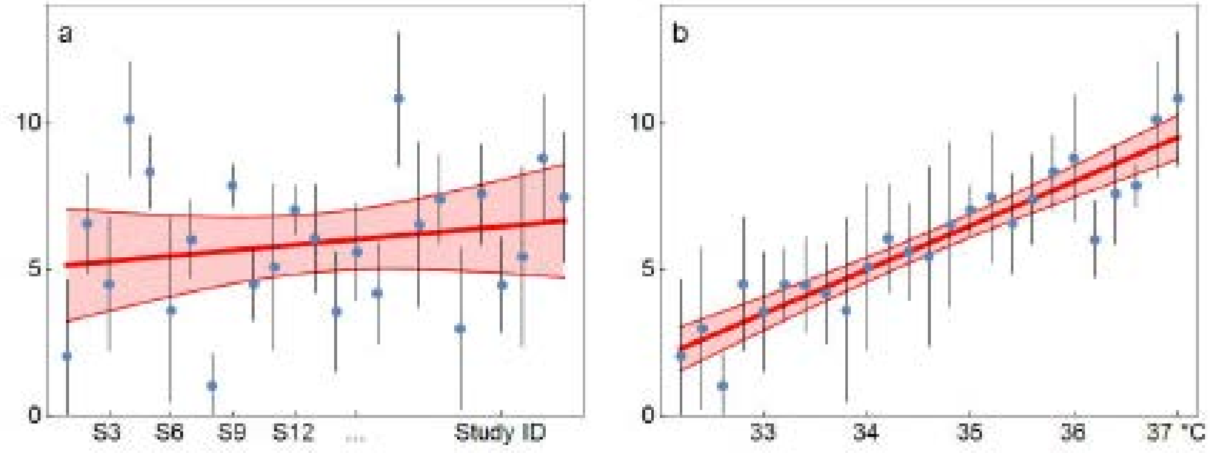
    - ➢ For results to be reproducible across independent studies, research should be conducted in a way to include and generalise across such unavoidable differences between study conditions. This requires heterogenisation of study conditions, not standardisation

Würbel, Nature Genetics 26: 263, 2000.

Never replicate a successful experiment?

A reaction norm perspective on reproducibility…





Example on the effect of dominating factors on effect size estimates and reproducibility.

- In this simulation, between-study variability is relatively large in comparison to within study variability (a), thus several studies would not cover the CI for the combined effect size estimate, suggesting ”replication failure”

- When accounting for a dominating environmental factor (e.g. temperature; b) however, all those studies capturing the predicted value for the respective ambient temperature should be considered successful replications

Voekl & Würbel 2019.

Journal Club

Never replicate a successful experiment?

Journal Club: Alternative ways to improve reproducibility

- Test batteries („measure as much as possible“)
- Meta-analyses
- Pre-registration of studies
- Statistical approaches (e.g. r-value)

- Automated test systems
- Multi-centre studies
- Reporting guidelines (e.g. ARRIVE)
- A world beyond the p-value

Never replicate a successful experiment?

Journal Club: Alternative ways to improve reproducibility

Beyond the p-value

- Amrhein et al. (2019): Retire statistical significance. Nature, 567, 305-307.

Reporting guidelines

- Baker et al. (2014): Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies. PLoS Biology, 12(1): e1001756.

Multi-centre studies

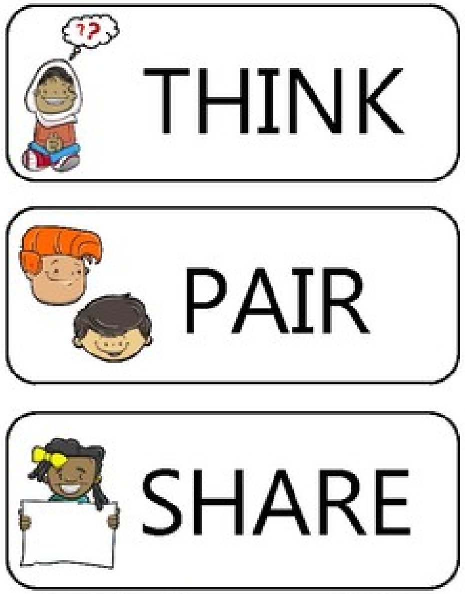
- Voelkl et al. (2018): Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biology 16(2): e2003693.

Variability / heterogeneity

- Milcu et al. (2018): Genotypic variability enhances the reproducibility of an ecological study. Nature Ecology & Evolution 2, 279–287.

Never replicate a successful experiment?

Journal Club: Alternative ways to improve reproducibility



Please…

… read the paper carefully

… share the main thoughts with your group member(s)

… discuss it under the light of the reproducibility crisis (pros & cons)

… prepare a poster with the main thoughts

… present it to the plenum.