Empirical Software Engineering Research

Replications

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

Learning Goals

- Understand the purpose of replicating and reproducing experiment (results)
- Be aware of the different types of replications and the corresponding terms

Replication

What do you think about replications?

Are they important? If so, why?

Replication

- Replication is repeating an experiment under the same or similar conditions
- Replications are fundamentally important as they show that results of an experiment can be reproduced/replicated to increase confidence. A „true“ finding should be found again and again by studies with the same method and a high degree of reliability.
- But, if the result of a study is not reproducible/replicable this does not mean that it is a bad study. It is part of the scientific process and helps to identify additional factors and biases.

Replication: Example

- Experiment on whether first-semester computer science students prefer learning programming with Python or Java
- Replication 1
  - We repeat the same experiment one year later, again with first-semester computer science students (but different people)
  - If the sampling was representative, the results should be the same
- Replication 2
  - We repeat the same experiment with first-semester students of psychology
  - Do the findings generalize to other populations?

If the results cannot be replicated, our experiment design appears to be missing a variable capturing an important aspect!

Replication: Terms

- Reproducibility versus replicability
  - Reproducibility refers to drawing the same conclusion with the same or different methods based on the same data set
    - If successful, it shows that the used analysis method(s) are appropriate and applied correctly
  - Replicability refers to collecting a new data set but with the same design and methods
    - If successful, it shows that the results can be found reliably

For example, we survey all 18 participants of this lecture whether this was a good course.

- 15 say “yes”.
- Someone analyzing the data again will be able to *reproduce* the result.

Next year, we survey all new 18 participants of this lecture whether this was a good course.

- 10 say “yes”.
- The results could not be *replicated*.

Replication: Terms (in SE)

There are many synonyms classifying replications

- Exact vs. non-exact replication
  - An exact (*strict, close*) replication means the study was repeated identically (or as close as possible)
  - A non-exact (*differentiated*) replication means the study was replicated with the same research questions but (purposefully) varied in some small way, while still being very similar
    - For example, the replication changed the population from novice programmers to intermediate programmers
    - Unlike an exact replication, the non-exact replication has the issue that a different result is hard to interpret. Did we find a different result because the findings of the original study were spurious or due to the change(s)?
- Factors to consider for replications
  - Experiment site, experimenters, experiment design, experiment material, experiment variables and their operationalization, participant sample, …

Replication: Terms (in SE)

- Internal vs. external replication
  - Internal replication refers to replicating an experiment by the **same** team
  - External replication refers to replicating an experiment by a **different**, often independent team

- Internal replications are more typical in SE for non-human studies
  - For example, applying some analysis to several open-source software projects (rather than just one)
  - Internal replications are much more common than external replications

Replication Crisis

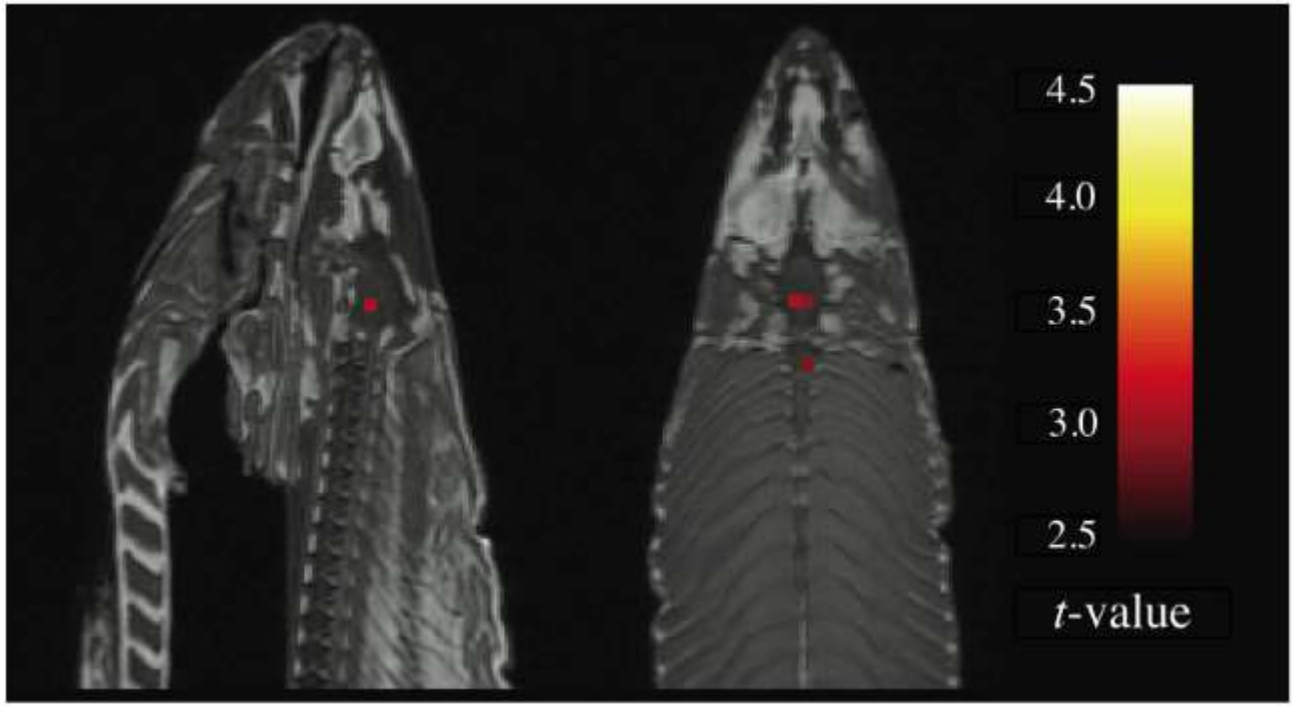
- When attempted in a systematic manner, the results of many studies cannot be replicated
  - Originally from psychology, but affects many fields
- Can have many causes from unclear terms, experiment design, methods, lack of transparency, scientific misconduct (fraud), …
  - If you report on a study, be sure to be detailed, specific, and transparent
  - Ideally, provide raw data (reproducibility) and a replication package (replicability)
  - One major attempt to relief these issues is open science (which we discuss in the next lecture)

Replication Crisis & Media

- (Social) media often reports on single studies showing some phenomena and present them as absolute truths
  - "Study shows XYZ"
  - But a single study generally is not sufficient to be accepted as "truth" in science
  - Authors of studies often report/discuss certain limitations, but those are often dropped in a media report
  - Only if a study can be replicated under different conditions (which takes years), we can have confidence in the findings
  - → Be critical of media reporting on scientific studies and, if in doubt, take a close look at the actual paper

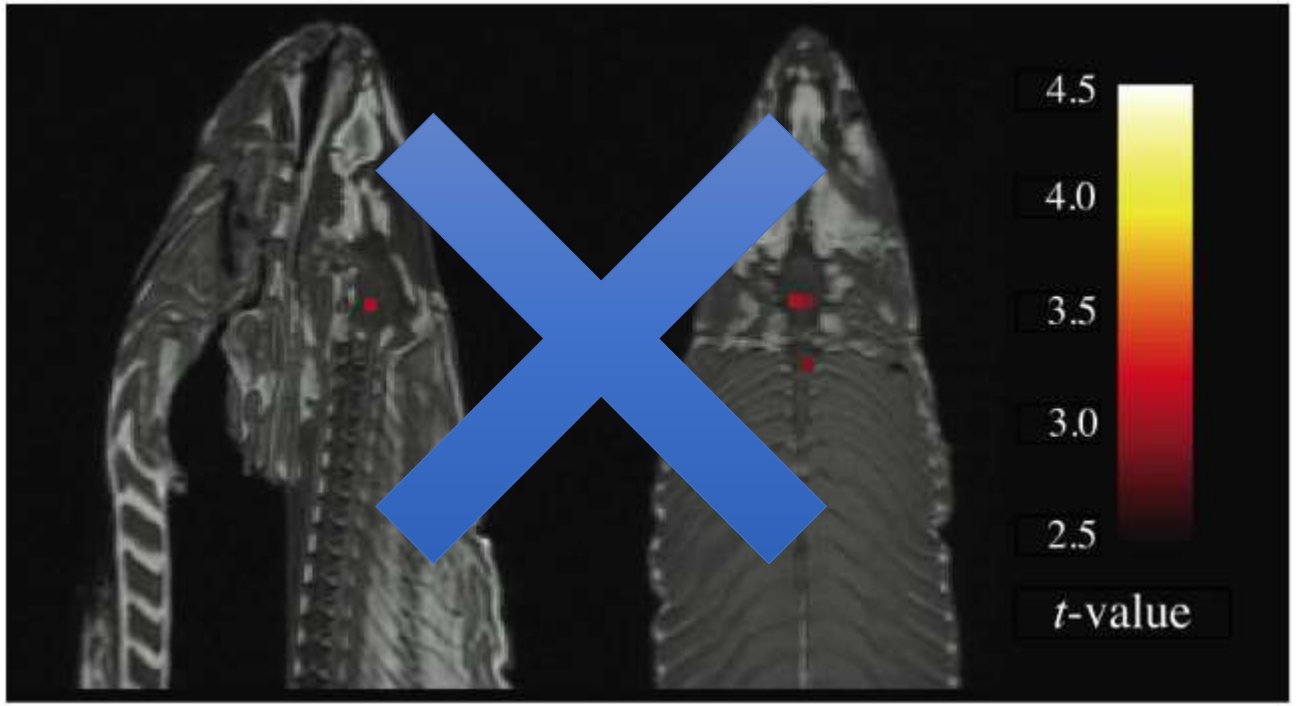
Replication Crisis & Media: Example

Bennet et al. : “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon”



Replication Crisis & Media: Example

Bennet et al. : “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: **An argument for multiple comparisons correction**”



When the analysis was controlled for false discovery rate (FDR), there was no active voxels left, even at $p < 0.25$.

Replication Crisis & Media: Example

Have you experienced overhyped media presentations of a (single) scientific study?

What can we do?

Replication: Issues

- Little incentives for researchers to conduct replications
  - Less interesting regarding their career
  - Journals and conferences tend to bias towards new results (rather than replications)
  - Funding is focused on novel research (rather than replications)
- Some publication vendors offer tracks for replications
  - But they may be seen as less "reputable" than research tracks