Experimentation in Software Engineering

Hands-on FAST-Workshop by:
Iflaah Salman
Rahul Mohanani

13.8.2025

What is an Experiment?

An experiment models key characteristics of reality in a controlled environment and manipulating them iteratively to investigate the impact of such variations and get a better understanding of a phenomenon.

Credits: Sira Vegas (Universidad Politécnica de Madrid)

Why to Experiment?

- To investigate a cause-effect relationship
  - To gain evidence of a presumed cause-effect relationship

    OR

  - To validate a hypothesis

Credits: Sira Vegas (Universidad Politécnica de Madrid)

SE Experiments

- Identify and understand
  - The variables that play a role in software development
  - The connection between variables
- Learn cause-effect relationships between the development process and the obtained products.
- Establish laws and theories about software construction that explain development behaviour.

Credits: Sira Vegas (Universidad Politécnica de Madrid)

Conditions for Causality

- Association
- Direction of influence (X => Y, and not Y => X)
- Elimination of rival explanations

- Two empirical strategies
  - Randomised Controlled Experiment
  - Statistical modelling (controlling) of alternative explanations

Controlled Experiment

- Control Group AND
- Treatment Group

Orientation:

- Technology (different tools are applied to the object)
- Humans (humans apply different treatments to objects)

Experimental Process

1. Definition
2. Planning
3. Operation
4. Analysis and Interpretation
5. Presentation and Packaging

Step by Step Execution

“Keep It Simple” Approach

1. Definition

Group Activity

1. Form groups
2. Choose a topic for an Experiment

Definition

Goal Definition Template:

Analyse <Object of the study: what is studied in the experiment>

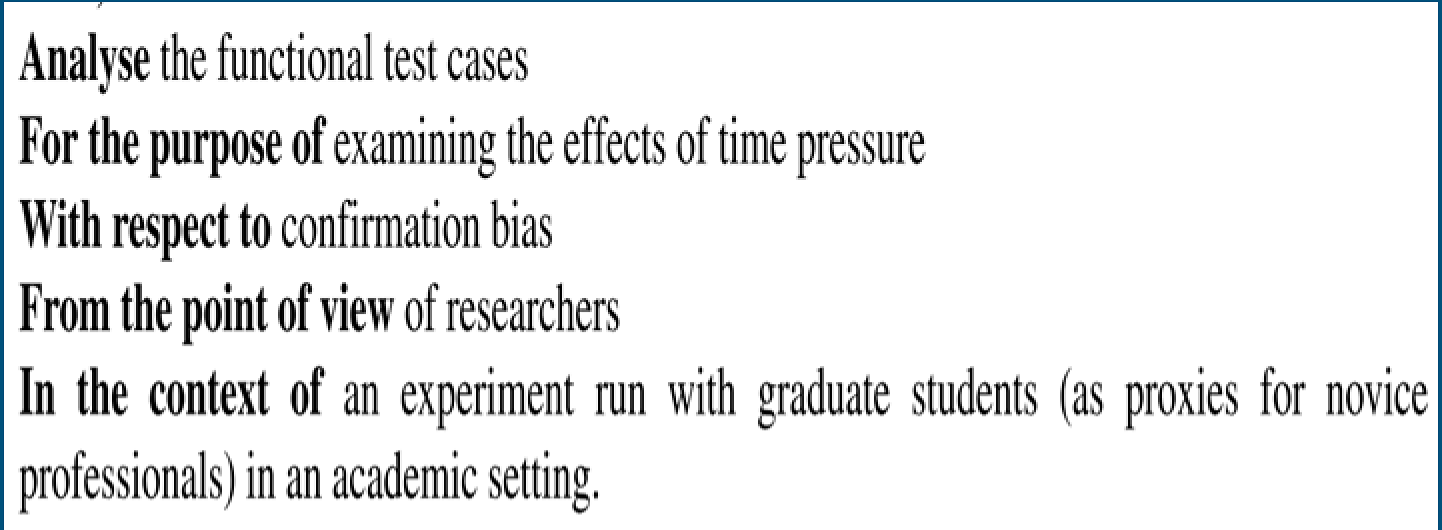
For the purpose of <Purpose: what is the intention of the experiment>

With respect to their <Quality focus: the primary effect under study in the experiment>

From the point of view of the <Perspective: the viewpoint from which the experiment results are interpreted>

In the context of <Context: the environment in which the experiment is run. Who are the subjects?>

Example – Define Yours



Analyse the functional test cases

For the purpose of examining the effects of time pressure

With respect to confirmation bias

From the point of view of researchers

In the context of an experiment run with graduate students (as proxies for novice professionals) in an academic setting.

2. Planning

Planning: Context – Choose Yours

- In order to achieve the most general results, experiment should be executed in large real software projects.
- Realism is not always possible, and the costs are usually quite high.
- We can characterise the context according to four dimensions:
  1. Online vs offline
  2. Student vs. professional
  3. Toy vs. real problems/tasks
  4. Specific vs. general

Planning: Variables Selection

- Independent Variables:
  - the variables the we can control and change in the experiment. OR
  - the variables we manipulate.
    - Define the manipulation, i.e., the conditions/levels/treatments
- Dependent Variables:
  - The effect of the treatment is measured on these variables, the values of which are expected to vary by varying the independent variable. OR
  - The outcome we measure against the manipulation of IV.

Planning: Variables Selection – Choose Yours!

- Example:
  - Independent Variable = time-based condition
    - levels: time pressure, no time pressure
  - Dependent Variable = external code quality

Planning: Hypothesis Formulation

- A core component of the experiment.
- It is developed based on the definition of our experiment.

- Two types of hypotheses are formed:
    - Alternative hypothesis – H1: There are underlying trends or patterns.
        - This is the hypothesis in favour of which we reject the Null hypothesis.
    - Null hypothesis – H0: There are no real underlying trends or patterns in the experiment setting.

Planning: Formulate Hypothesis – Formulate Yours!

- Two types of hypothesis Testing
  - One tailed or Unidirectional: one side/group is better than the other
  - Two tailed or bidirectional: there's a difference between the two sides

Example of two tailed hypothesis: (KIS)

Alternative hypothesis: The external code quality differs between time pressure and no time pressure groups.

Null hypothesis: The external code quality does not differ between time pressure and no time pressure groups.

Planning: Metrics – Choose Yours!

- Metric(s) for Dependent Variable
  - How are we going to measure our dependent variable?
    - We need to quantify this!
  - On what scale do we measure our dependent variable?
- Values or States for Independent Variable
  - How do we define/create/implement our conditions/treatments?
- Example:
  - IV = time-based condition (time pressure [30 min], no time pressure [60 min])
  - DV = external code quality (no. of bugs [found on execution]

Planning: Selection of Subjects (Sampling)

- This is closely connected to the generalisation of results to the desired population.
  - The sample must be representative!
- There are two types of techniques
  - Probability sampling: probability of selecting each subject is known.
  - Non-probability sampling: contrary to the above.
- The most popular one is:
  - Convenience Sampling – A non-probability sampling.
  - We can choose this for now!

Convenience Sampling – Choose Yours!

1. Choose your sample!
2. What are the characteristics of your sample?
   a. Are their characteristics relevant to the objectives of the experiment?
   b. Can their characteristics act as confounding factors for the experiment?

Planning: Experiment Design

- The choice of Design is closely related to the statistical analysis and interpretation.
- The objective is to conduct a series of tests of the treatments.
- The Design tells how the tests are organised and run.

Planning: Experiment Design

- Standard Design types are:
    - One factor with two treatments
        - Factor is *time-base condition*, 2 treatments are *time pressure* and *no time pressure*
    - One factor with more than two treatments
    - Two factors with two treatments
        - Factor A= Programming language; 2 levels are: C++, Java
        - Factor B = IDE, 2 levels are: Eclipse, IntelliJ
    - More than two factors each with two treatments

Planning: Experiment Design – Design Yours

- One Factor with two treatments
  - We want to compare two treatments against each other.

	Treatment 1 (TP)	Treatment 2 (NTP)
Experimental Session	Group 1	Group 2

Completely Randomised Design:

Between Subjects Design!

Subjects	Treatment 1	Treatment 2
1	x	
2		x
3		x
4	x	

Planning: Instrumentation – Decide Yours!

Goal is to provide means for performing the experiment and monitor the execution.

There are multiple types of Instruments:

- Object(s): what sort of object/artefact subjects would work on?
  - For example, requirements document of a minesweeper game
- Guidelines:
  a. A set of instructions for the participants OR
  b. A set of instructions for the experimenters (usually this one)

Planning: Instrumentation – Decide Yours!

Goal is to provide means for performing the experiment and monitor the execution.

There are three types of Instruments:

- Measurement: conducted via data collection
  - For example, may need to prepare questionnaires/surveys
- Tools/setups
  - For example, Virtual machine setup.

Planning: Validity Evaluation

There are four types of validity threats:

1. Conclusion Validity ensures there’s a statistical relationship, i.e., with a given significance.
2. External Validity is about the generalizability of results.
3. Internal Validity ensures that the relationship observed between the treatment and outcome is a causal relationship.
4. Construct Validity is concerned with the relation between theory and observation.

Planning: Validity Evaluation – Assess Yours!

1. Internal Validity ensures that the relationship observed between the treatment and outcome is a causal relationship.
    a. Social Threats to internal validity
        i. Diffusion or imitation of treatments: when a control groups learns about the treatment from the group in the experiment study.
        ii. Resentful demoralisation: A subject receiving less desirable treatment may give up and not perform as good as it generally does.

Planning: Validity Evaluation – Assess Yours!

1. Construct Validity concerns generalising the results of the experiment to the concept of theory behind the experiment.
    a. Design threats
        i. Mono-operation bias: if the experiment includes a single independent variable, case, subject or treatment, the construct is possibly underrepresented.
    b. Social threats:
        i. Hypothesis guessing: Participants may try to figure out the purpose and intended results of the experiment. They might base their behaviour on their guesses.

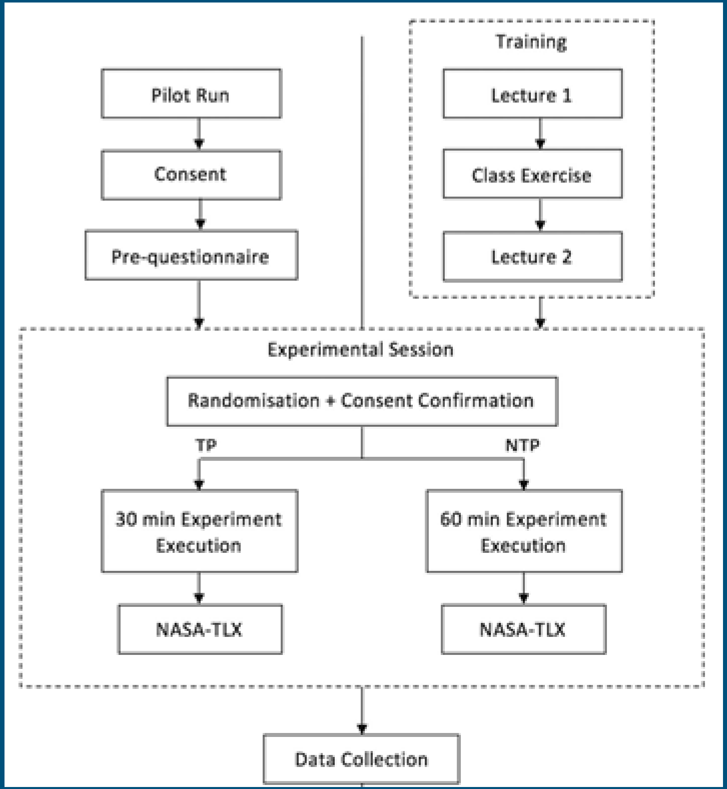
Operation

Operation: Preparation – Define Yours!

- Commit Participants
  a. Obtain Consent
  b. Sensitive Results: ensuring the confidentiality
  c. Inducement: Incentivising the participants
  d. Instrumentation concerns

Operation: Execution – Develop Yours!

- Execution Protocol
  a. Planning the actual execution
    i. How to apply randomisation?
    ii. How will the decided number of sessions be run?



Operation: Execution – Think Yours!

Data Collection

- Collecting the data for or according to our dependent variables.
- It can involve transforming data into the desired values for the variables.

Thank you!

References

1. Salman, I., Turhan, B., & Vegas, S. (n.d.). A Controlled Experiment on Confirmation Bias and Time-Pressure in Software Testing. *Empirical Software Engineering*.

2. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer. https://doi.org/10.1007/978-3-642-29044-2