Basics on Design and Analysis of SE Experiments: Widespread Shortcomings

Natalia Juristo

Universidad Politecnica de Madrid (Spain)
&
University of Oulu (Finland)

Content

1. Experiment distinguish hallmark
2. Good practices for running a SE experiment
   1. Definition & Operationalization
   2. Design
   3. Implementation & Execution
   4. Analysis
   5. Interpretation
   6. Packaging & Publication
3. Summary of advices

Experiments Distinguishing Hallmark

- Causality
- Control

Scientific Knowledege

- Scientific laws are patterns of behaviour
- Describe cause-effect relationships
- Explain
  - why some events are related
  - how the mechanism linking the events behaves

Why Experiments Are Needed

- We cannot perceive laws directly through our senses
- Two activities are necessary
  - Systematic objective observation
  - Inference of links between cause & effect

A Scientific Method

- Collection of Empirical **Data**
  - Systematic observation to appreciate the nexus

- Theoretical **Interpretation** of Data
  - Form a hypothesis (right or wrong) about the mechanisms relating the events
- Collection of Empirical **Data**
  - Hypothesis are permanently tested against reality to know if they are true or not

SE Experiments

- Identify and understand
  - the variables that play a role in software development
  - the connections 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

Experiment Definition

- Experiment
  - Models key characteristics of a reality in a controlled environment and manipulating them iteratively to investigate the impact of such variations and get a better understanding of a phenomenon
- Laboratory
  - Simplified and controllable reality where the phenomenon under study can be manipulated

Other type of empirical studies do not manipulate reality, just observe it

Control Is The Key For Causality

- The key aspect of a controlled experiment is…

  Control!!!

- Causality is discovered through the following reasoning
  - Control voids the effect of all irrelevant variables
  - The impact we observe in the response variable is only due to the manipulated variables

Factors & Response Variables

- To gain evidence of a presumed cause-effect relationship, the experimenter
    - Manipulates
        - the independent variables
        - or factors
    - Observes changes in
        - the dependent variable
        - or response variable

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Interpretation
6. Packaging and Publication

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Interpretation
6. Packaging and Publication

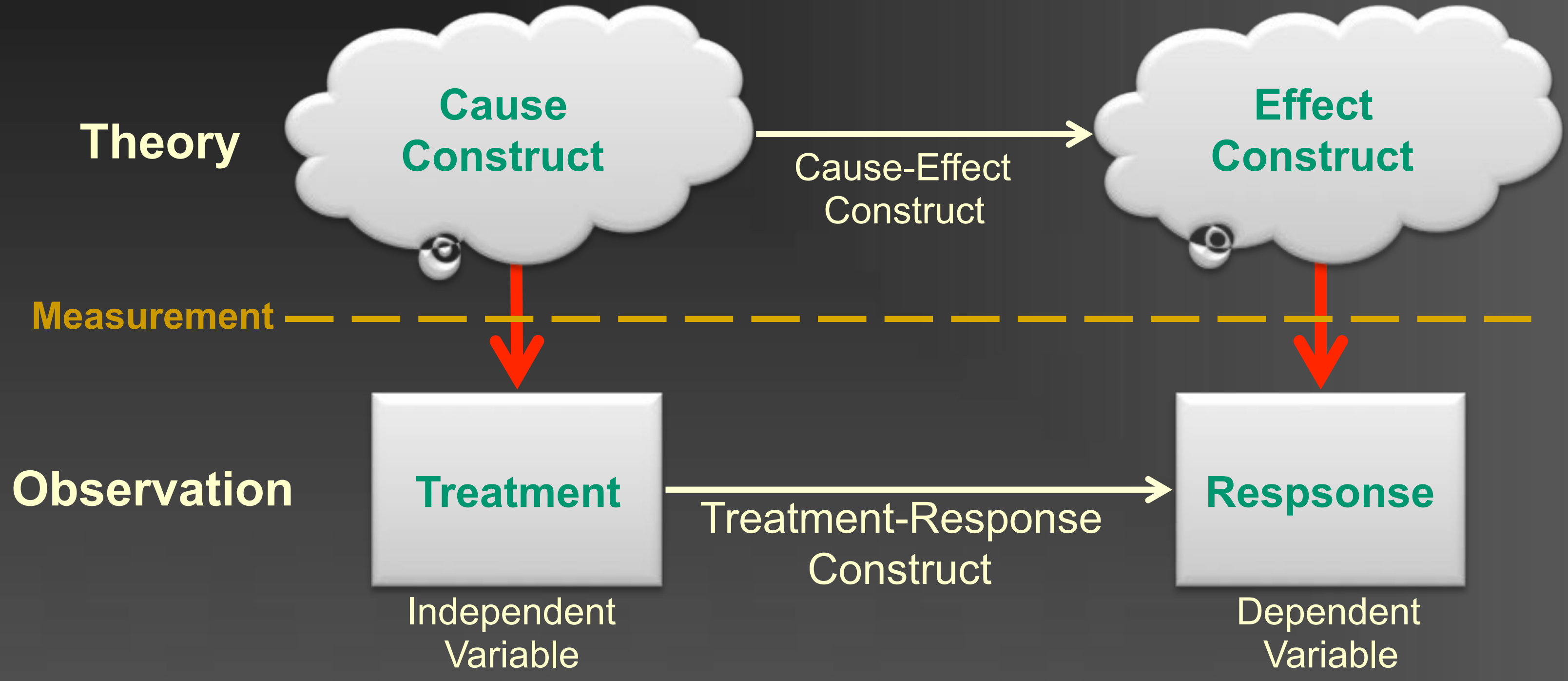
Definition Goal

- Problem Definition
  - As in any other research choose an open problem
- Research goals and questions
  - Causal research question
    - Does X cause Y?
    - Does X1 cause more of Y than X2 causes of Y?
  - Example
    - Does MDD cause higher quality software than other development paradigm?

MDD Experiment Example

- Run a subjects-based experiment on MDD
  - in the context of a course about MDD
- Factor
  - Development approach
- Treatments
  - MDD
  - Control?
- Response variable
  - Quality

Constructs Operationalization



Effect Operationalization

1. Effect Variables into Response Variables
   - Higher quality software => less defects in it => Testing techniques help to identify defects
   - Effectiveness
2. Metrics Definition
   - Number of defects found
     - More defects found = more effective testing technique
     - Proportion of defects found out of those seeded
3. Instrumentation
   - Seed defects into programs
     - Which type of defects?
     - How do we generate such defects?
   - Need one or more programs
   - Subjects applying the testing techniques
     - Which type of subjects?
   - Form where subjects write down the test cases generated OR the defects found
     - Do we want the subjects running their test cases OR the experimenter?
4. Data Collection procedure
   - Number of defects identified by subjects
     - Subjects writing down the defects founded
   - Number of defects exercise by the test cases generated by the subjects
     - Subjects writing down the test cases generated
5. Measurement procedure = Metrics collection procedure

Cause Operationalization

1. Cause variables into treatments
   - Factor
     - Testing techniques
   - Treatments
     - White box / Black box applied by subjects
2. Treatments definition
   - Version of the technique
   - How treatment is administer
     - Teaching?
   - Description in a “reminder sheet”
   - Otros?

Effect Operationalization: Size Example

1. Response Variable
2. Metrics Definition
3. Instruments
4. Data Collection procedure
5. Measurement (metrics collection) procedure

Effect Operationalization: Size Example

1. Variables
   - Table length
2. Metrics Definition
   - Centimeters
3. Instruments
   - Measuring tape
4. Data Collection procedure
   1. Place the beginning of the tape just at one end of the table
   2. Pull the tape until the other end
5. Measurement procedure (metrics collection)
   - Look at the number printed on the tape that matches the extreme of the table

Effect Operationalization: Quality Example

1. Variables
   - Code quality -> Functionality -> Accuracy [ISO9126]
2. Metrics Definition
   - Percentage of acceptance test cases that are successfully fulfilled
     - 1 test case per atomic requirement
     - Each test case subdivided in items
     - All items need to be passed to consider a test case satisfied
3. Instruments
   - IDE where the code developed by subjects is stored
4. Data Collection procedure
   1. For each test case
      1. Run the code
5. Measurement (metrics collection) procedure
   1. For each test case decide if it is passed
   2. Sum up the number of test cases passed
   3. Convert such a number into a proportion

Cause Operationalization Treatment Definition

- Version of the treatment
  - What exactly is MDD?
    - NDT, WebRatio, OOHDM, OO-Method, etc
  - What exactly is traditional?
    - Model-centric?; Code-centric?; other?
- How treatment is administer
  - Teaching?
- Are treatments applied through tools?
  - Which?

Formulate Hypothesis

MDD (OO-Method w/ Integranova tool)

satisfies different amount of test cases

for small problems implemented in java

than

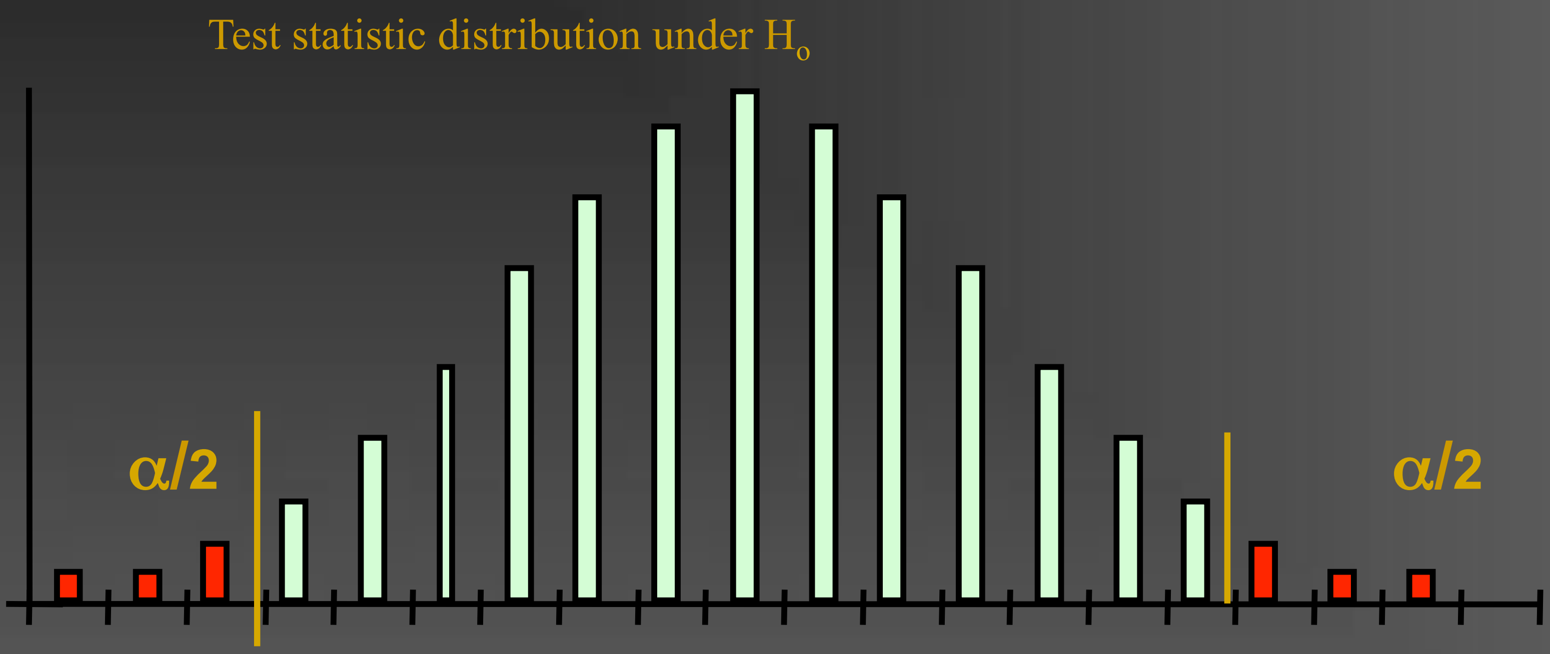
A model-centric (UML w/ Eclipse)

when applied by novice developers

One-tailed vs Two-tailed

- Two-tailed hypothesis = Non directional
  - Predicts a difference between two variables
  - Not the direction or the nature of their relationship
    - Quality(MDD) <> Quality(Model-centric)
- One-tailed hypothesis = Directional
  - Predicts the direction of the difference between two variables
    - A positive or negative correlation
    - Quality(MDD) > Quality(Model-centric)
  - Requires previously obtained knowledge about the effect
    - Theory or evidence

Two-tailed Tests



Good Practices

- Think carefully about which metrics to use
  - Metrics are not yet a solved issue in SE
- Remember to decide on the measurement process beforehand!
  - This influences the instruments
- Use two-tailed hypothesis, better than one-tailed

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Interpretation
6. Packaging and Publication

Experimental Design

- Describe how the study is organized
- Identify undesired sources of variability
- Iterate improving design evaluating threats and confounding variables

Types of Design

- Depending on the number of factors and treatments a type of design is chosen
  - One factor w/ 2 treatments
  - Blocked design
  - Factorial design
    - Completely randomized design
    - Blocked factorial design
    - Fractional factorial design
    - …
  - Repeated-measures randomized controlled trial

Design and Control

- The key aspect of a controlled experiment is…

  Control!!!

- The design of a controlled experiment is a set of strategies aiming to control
  - The relevant variables (under study)
  - The irrelevant variables but with known values
  - The irrelevant variables with unknown values

Main Design Strategies

- Treatments
  - Equality inside treatments
  - Similar conditions among treatment
- Irrelevant variables with known values
  - Blocking
    - The non-desired variable has effect on the dependent variables, but similar effect on every treatment group
  - Block as many variables as you can
- Irrelevant variables with unknown values
  - Randomization
    - Assign treatments at random to experimental units to avoid the undue influence of any possible variables
  - Randomize for the rest

Example:

Blocking

- The MDD experiment with two groups
  - Factor
    - Development paradigm
  - 2 Levels
    - MDD & Traditional

		MDD	Traditional
Session 1	P1	G1	G2

- Imagine we have experts and novices
  - We blocked by experience

		MDD		Traditional	
Session	P1	Novices	Experts	Novices	Experts
		G1		G2	

Main Design Strategies

- Treatments
  - Equality inside treatments
  - Similar conditions among treatment
- Irrelevant variables with known values
  - Blocking
    - The non-desired variable has effect on the dependent variables, but similar effect on every treatment group
  - Block as many variables as you can
- Irrelevant variables with unknown values
  - Randomization
    - Assign treatments at random to experimental units to avoid the undue influence of any possible variables
  - Randomize for the rest

Blocking

- Blocking is the arrangement of experimental units into groups (blocks) consisting of units that are similar to one another
- Blocking reduces known but irrelevant sources of variation between units and thus allows greater precision in the study output

Blocking

- Blocking is the arrangement of experimental units into groups (blocks) consisting of units that are similar to one another
- Blocking reduces known but irrelevant sources of variation between units and thus allows greater precision in the study output
- Purposely assign every value of the non-desired variable to every experimental group
- The non-desired variable has effect on the dependent variables, but similar effect on every group (treatment)

Blocking

- Purposely assign every value of the non-desired variable to every experimental group
- The non-desired variable has effect on the dependent variables, but similar effect on every group (treatment)

Randomization

- To assign treatments at random to the experimental units
- Aims to avoid the undue influence of any possible confounders (known or unknown)
- The presence of uncontrolled confounders will tend to increase the experimental error

Randomization

- To assign treatments at random to the experimental units
- Aims to avoid the undue influence of any possible confounders (known or unknown)
- The presence of uncontrolled confounders will tend to increase the experimental error
- The importance of randomization cannot be over stressed
- Randomization is necessary for conclusions drawn from a experiment to be correct, unambiguous and defensible

Randomization

- The importance of randomization cannot be over stressed

- Randomization is necessary for conclusions drawn from a experiment to be correct, unambiguous and defensible

Iterating for Design

- Designing an experiment is an iterative task to reaching a trade-off among validity threats
  1. Design
  2. Evaluate issues that threaten validity

- Several design choices need to be made to limit threats to validity
  - There is not such a thing as The Perfect Experiment that avoids all validity threats

Threat to Validity

- Experimenters must weigh the threats to validity and design the experiment trying to avoid them
- Those threat to validity which the experimenter suspect has failed to prevent has to be made explicit
- Good design try to avoid confounding variables

MDD Experiment Example

- Run a subjects-based experiment on MDD
  - in the context of a course about MDD
- Factor
  - Development approach
- Treatments
  - MDD
  - Traditional
- Response variable
  - Quality

1 Factor Design
2 Treatments

		MDD	Traditional
Session 1	P1	G1	G2

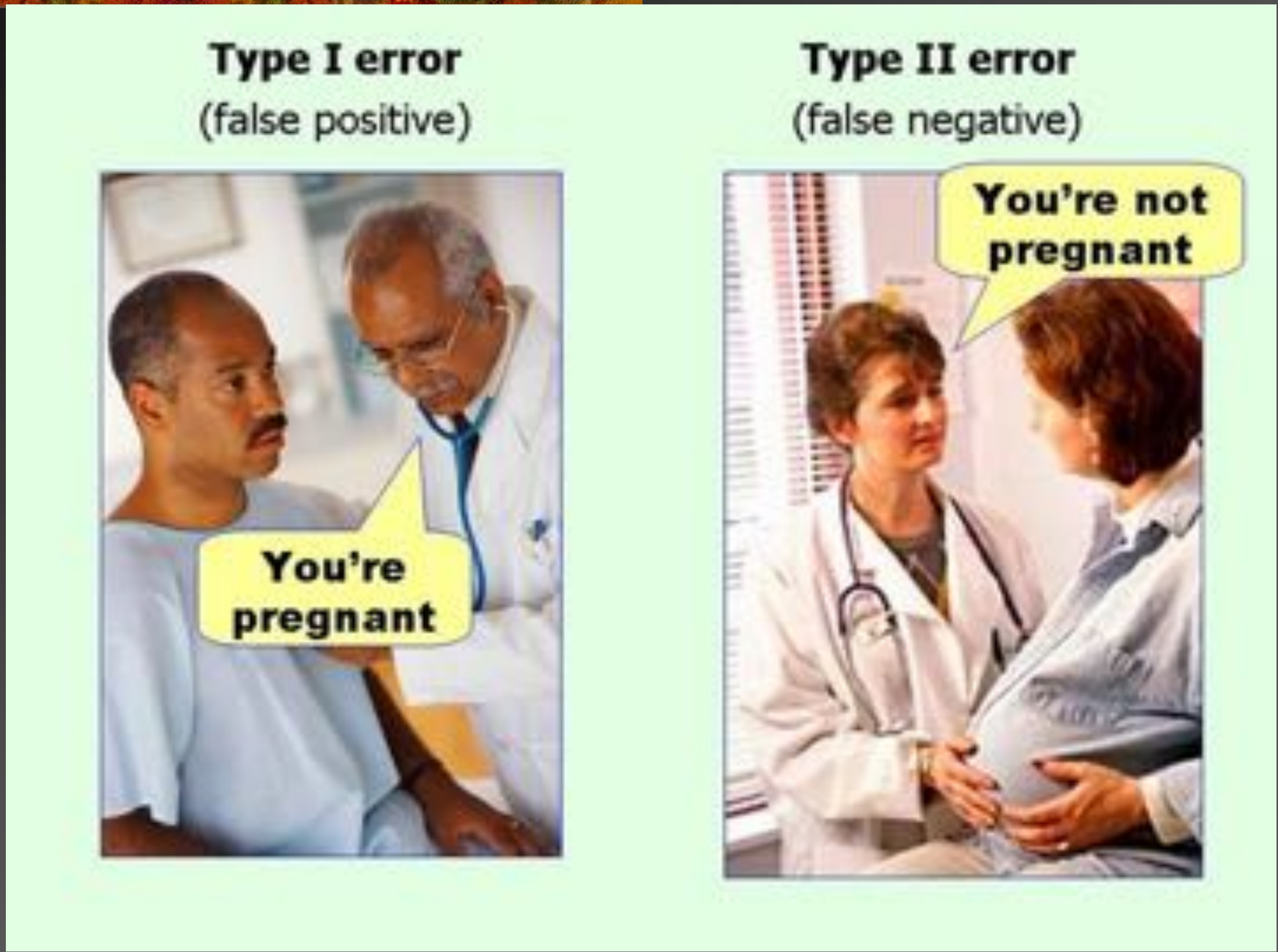
- 1 session, 2 groups, 1 experimental
- Cons
  - Divide by two the number of subjects
    - Decreasing the sample size and therefore lowering power
  - Training perspective, pairs will only practice MDD or traditional method
    - Not viable alternative in a MDD course
  - Very low generalization
    - Only to one problem
- Pros
  - Treatments comparison done through identical conditions

Live with this depends on the sample size we have

We cannot live with this in this context

We can hardly live with this

Power relates with Type II error



Paired Design 1 Object

		P1
Session 1	Traditional	G1
Session 2	MDD	G1

- 2 sessions, 1 group, 1 object
- Cons
  - Threat: Learning effect on object
    - Subjects might learn the problem in the first
  - Treatments comparison in not identical conditions
    - Similar conditions: Different sessions
    - Very dissimilar conditions: Different order
- Pros
  - Biggest sample size
    - Highest power
  - Same subject under each treatments
    - Better control of subjects differences

We cannot live with this

We can hardly live with this

Great!!! ☺

Paired Design 2 objects

		P1	P2
Session 1	Traditional	G1	
Session 2	MDD		G1

- 2 sessions, 1 group, 2 objects
- Cons
  - Treatments compared in different conditions
    - Similar conditions
      - Different sessions
    - Dissimilar conditions
      - Different order
      - Different problem
- Pros
  - Biggest sample size
  - Better control of subjects differences
  - Avoid learning effect on object

We can live with this

We cannot live with this

Great!!! ☺

Cross-over
2 objects

		MDD	Traditional
Session 1	P1	G1	G2
Session 2	P2	G2	G1

- 2 session, 2 groups, 2 objects
- Cons
  - Session and object is confounded
    - But does not affect treatments
  - Hard to sell alternative in a MDD course
    - Specially the MDD-T order
- Pros
  - Avoid the influence of session on trea
  - Biggest sample size
  - Better control of subjects differences
  - No learning effect on object

We can live with this

We can hardly live with this
in our context

Paired
Blocked by object

		P1	P2
Session 1	Traditional	G1	G2
Session 2	MDD	G2	G1

- 2 session, 2 groups, 2 objects
- Cons
  - Session and development paradigm confounded
    - But adheres to the regular way it happens
  - Weak cheating effect on object
    - Since different treatments are being applied
- Pros
  - No learning effect on object
  - Biggest sample size
  - Better control of subjects differences
  - Make sense from an educational point of view

We can live with this

We can live with this

Cross-over
1 object

	P1	
	MDD	Traditional
Session 1	G1	G2
Session 2	G2	G1

- 2 sessions, 2 groups, 1 object
  - First, half subjects MDD, the other half T; Then, the other way around
  - Same problem in both sessions
- Cons
  - Threat : Learning effect on object
    - Subjects might learn the problem in the first session and the results obtained in the second one may depend on the knowledge obtained in the first one
  - Threat : Cheating effect
  - Low generalization for other objects
    - Results are valid for only one problem
- Pros
  - We use the biggest sample size we can
    - Highest power
  - Avoid the influence of session on treatments

Just an Example!

- Noticed these are all not the only designs
  - Cross-over with 2 objects
  - Cross-over blocked by object
  - Matched pairs designs
  - …..
- We could have followed other reasoning

Design is Experiment-dependent

- The best design for certain situation can be the worst in others
  - Sample size was a problem in our experiment
    - If it is not, then first design could work
  - Sequential application of treatments is ok in our context (due to technology being tested)
    - For others, for example testing, application of treatment in only one order would be a big threat

Good Practice

- Do not copy your design from others!!
  - The sources of variability is particular to every experiment
  - You need to iteratively think about your design, evaluate threats and modify it selecting the best you can
  - Include the iterative process and decision in the paper!

Good Practice

- Replicate your own experiment
  - If you do it identically
    - Sample size is increased
  - If change something
    - Some threats to validity can be mitigated
      - In the example
        - Order threat
        - Low generalizability
        - …

Good Practice

- Make always a previous demographic questionnaire
  - Helps on blocking
  - For post-hoc analysis

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Interpretation
6. Packaging and Publication

Implementation & Execution Goals

- Implementation
  - Instantiate the experimental design, so can be executed
  - Tasks
    - Design all required instruments
      - Questionnaires, protocols and tools
    - Prepare all necessary material
      - Guidelines, document templates, specifications, codes and tools
- Execution
  - Run the experiment

Good Practice

- Run a Pilot
  - To be sure instruments work well
  - To assure explanations are clear
  - …
  - Things usually do not go out as expected ☹

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Results Interpretation
6. Packaging and Publication

Analysis Goal & Tasks

- Analyze collected data for
  - Describing sample
  - Testing hypothesis

- Tasks
  1. Descriptive statistics
  2. Select statistical test
  3. Hypothesis testing
  4. Power analysis
  5. Effect size calculation

Statistical Test Selection

- Statistical tests
  - Exist for different purposes
  - Have different preconditions
  - Have different power
- Your data set must fulfill the test assumptions on
  - Experimental design
  - Distribution of data
- Choosing appropriate statistical test is key to get a reliable rejection or not rejection of the null hypothesis

Statistical Test Selection

Number of variables	Subjects in condition	Parametric Test	Non parametric Test
One variable: two treatments	Independent	Independent t-test	Mann-Whitney U test
	Dependent	Paired t-test	Wilcoxon matched pairs test
One variable: > 2 treatments	Independent	One factor independent ANOVA	Kruskal-Wallis-One way ANOVA
	Dependent	One factor repeated measures ANOVA	Friedman ANOVA
Two or more treatments	Independent/ Dependent	Variation of ANOVA-Analysis	

Parametric vs. Non-parametric

- Select statistical test considering data distribution
  - Normal distribution
    - Parametric tests
  - Non-normal or ordinal/nominal distribution
    - Non-parametric tests
- Do not assume normality (using the Central Limit Theorem)
    - Irrespective of the distribution of the parent population - given that its mean m and a variance s2, and so long as the sample size n is large, the distribution of sample means is approximately normal with mean m and variance s2 /n
  - Consider non-parametric tests
    - SE experiments have small sample sizes
- But neither use always non-parametric test

Hypothesis Testing

1. Formulate the alternative and null hypothesis
2. Select statistical test considering data distribution
   - Normal distribution
     - Parametric tests
   - Non-normal or ordinal/nominal distribution
     - Non-parametric tests
3. Select significance level (α-value) and perform power analysis
   - α conventionally 0.05 or 0.01
   - Power = 1- β (β conventionally 0.2)
     - Determine optimal sample size based on α, effect size and power
     - Determine α based on sample size, effect size and power

Perform Power Analysis

		In the population …	
		H_0 is true	H_0 is false
Decision	H_0 is not rejected	**Correct** outcome True negative	Type II **error** False negative
	H_0 is rejected	Type I **error** False positive	**Correct** outcome True positive

Sample Size & Statistical Power

- The foolish astronomer
  - *An astronomer decides to build a telescope to study a distant galaxy*
  - *He foolishly builds it on the basis of available funds, rather than on the calculations of the needed power to actually see the galaxy*
  - *He orders the biggest telescope he can afford and hopes for the best…*

Understanding the Outcome

- If null-hypothesis is **rejected**
  - There is an effect
- If null-hypothesis is **not rejected**
  - It is not possible to conclude there is no effect!
  - There is not sufficient evidence to accept there is an effect

Three Critical Parameters

- Statistical significance
  - A result is significant because it is predicted as unlikely to have occurred by chance alone
  - The observed effect seems to have a cause
- Power
  - The probability that a test finds there is no difference between treatments when there is
- Effect size
  - Magnitude of the results
  - Which is the size of the improvement?

Good Practice

- Learn about analysis
  - Get the advice of an expert
- Check the proper analysis for your design
- Do not always apply the same type of tests
  - Check tests assumptions on data distribution
- Provide the three parameters
  - Significance, power, effect size

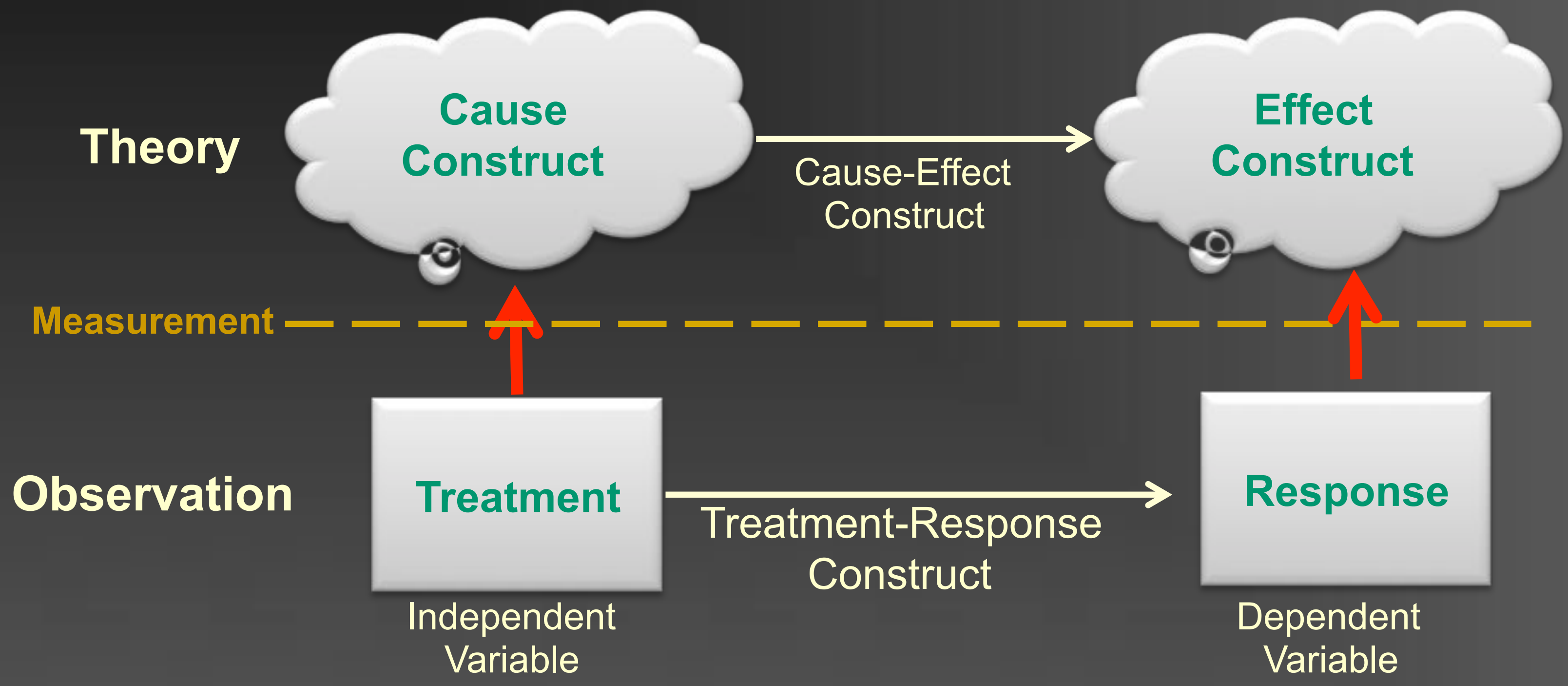
Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Results Interpretation
6. Packaging and Publication

Interpretation Goal

- Answering research questions
- Statistical testing is just the means to an end
  - Not an end in itself!!
- More difficult than running statistical tests
  - Interpretation of the results
  - What does the results mean?

Results Interpretation



Good Practice

- Do not forget to interpret the results and close the circle!
  - An experiment does not only give an output of a statistical test, you need to give an answer to the research question taking into account
    - The statistical issues
      - hypothesis test output, power, effect size
    - But also
      - Populations (subjects, objects), experiment protocol, observation of subjects, acontecimientos,…

Good Practices for Running a SE Experiment

1. Definition & Operationalization
2. Design
3. Implementation & Execution
4. Analysis
5. Interpretation
6. Packaging and Publication

Laboratory Package

- Motivating and enabling replication
  - Enabling independent confirmation of results
  - Making study design available for further investigation in different contexts
- Detailed account that allows replication
  - Measures, questionnaires, surveys, interview protocols, observational protocols, transcriptions, tape records, video record, pictures, …

Make your Results Public

- Presenting, sharing and spreading results
  - For community building a body of knowledge
  - Enabling review, discussion and challenge of results
- Follow guidelines to compose your manuscript
  - Jedlitschka

Good Practice

- Make an experimental package for others to replicate your experiment
  - The proper content for a lab package in SE is not solved yet
  - Not only materials should be there but more info on the experiment to be repeated
- Follow guidelines when reporting an experiment

Summarizing

Good Practices

- Operationalization
  - Think carefully about metrics to use
  - Decide before hand on the measurement process
  - Use two-tailed hypothesis
- Design
  - Do not copy your design from others!
  - Replicate your experiment
  - Make always a demographic questionnaire
- Implementation
  - Run a pilot
- Analysis
  - Learn about tests or get the advice of an expert
  - Be sure to correctly interpret the tests outcome
  - Provide significance, power and effect size
- Interpretation
  - Give answer to the research question
- Packaging
  - Made public at the web a replication package
- Publication
  - Follow guidelines

Basics on Design and Analysis of SE Experiments: Widespread Shortcomings

Thanks

Natalia Juristo

Universidad Politecnica de Madrid (Spain)
&
University of Oulu (Finland)

Free at:
https://sites.google.com/site/basicsofsoftwareengineeringexp/