The Experimental Process

Empirical Methods in Software Engineering (01QTFIU)

http://softeng.polito.it/EMSE/

SoftEng
http://softeng.polito.it

Version 1.6.0

BY NC ND

creative commons

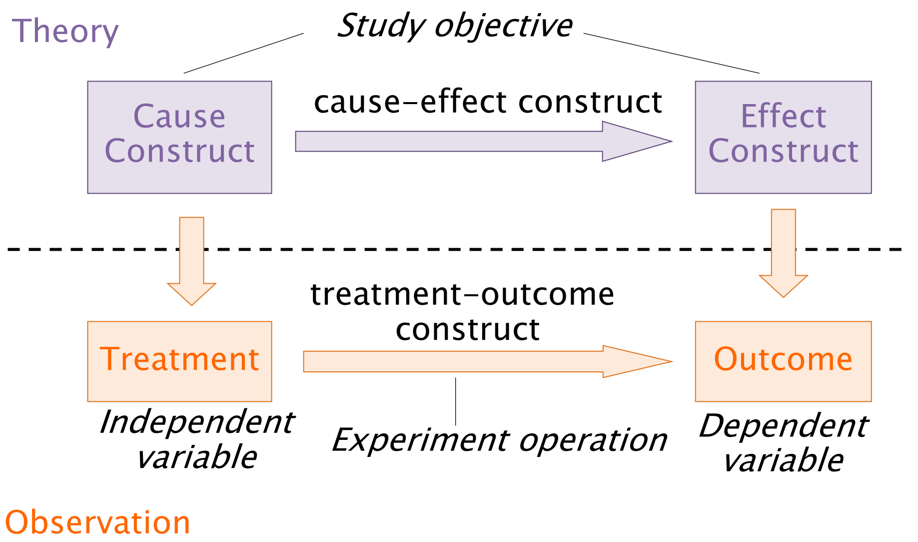
THE EXPERIMENTAL PROCESS

Wohlin et. al. 2000

Empirical method

- From observation find an explanation
- Formalize it into a theory
- Formulate an hypothesis
- Test it with a study

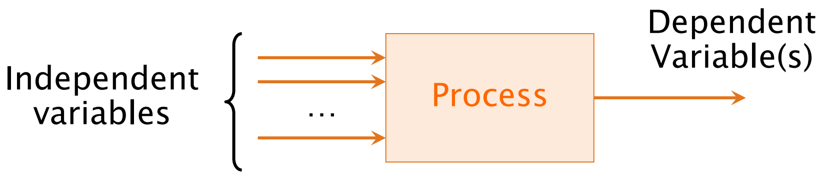
Experimental principles



Glossary

- Construct (conceptual)
  - Broad concepts or topics of study
    - Abstract
    - Not directly observable
    - May be complex (have multiple parts)
  - E.g. quality, productivity, skill
- Variable (Operational Construct), Measure, Metric
  - Precise definition
  - Procedure to measure

Glossary



- Dependent (output, response) variable:
  - Quantities observed in the study
  - E.g. LOC/day
- Independent (input) variable:
  - Quantities controlled and monitored
  - E.g. years of experience, development method

Glossary

- Factor
  - An input variable whose effect on the output we want to study
  - E.g. development method
- Treatment (Level)
  - A particular value of a factor
  - E.g. upfront design vs. incremental design

Glossary

Subject performs a task with an object

- Experimental unit of observation

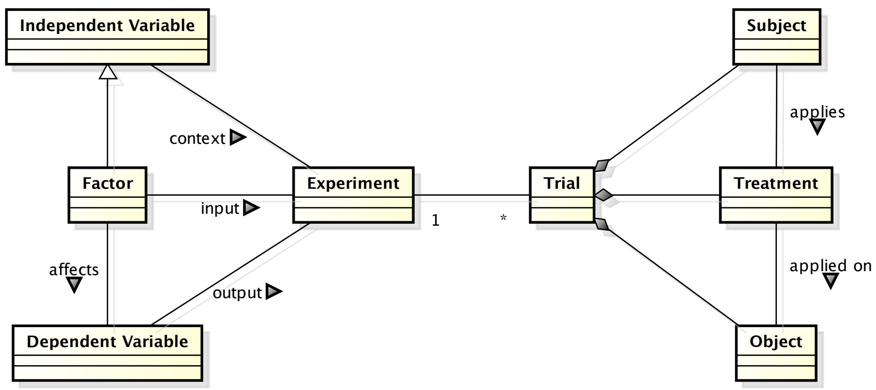
- The treatment may be applied to:
  - Task
    - E.g. Develop using a given ***methodology***
  - Object
    - E.g. Requirement with a given ***notation***
  - Subject
    - E.g. Developer with a particular ***skill/training***

Glossary

- Trial (experimental unit)
  - A combination of
    (Subject, Task, Object, Treatment)
  - Subject + Treatment
    - Task and Object counted as part of the treatment
    - Object: Software artifact

- An experiment typically involves several trials

Experimental Process Concepts

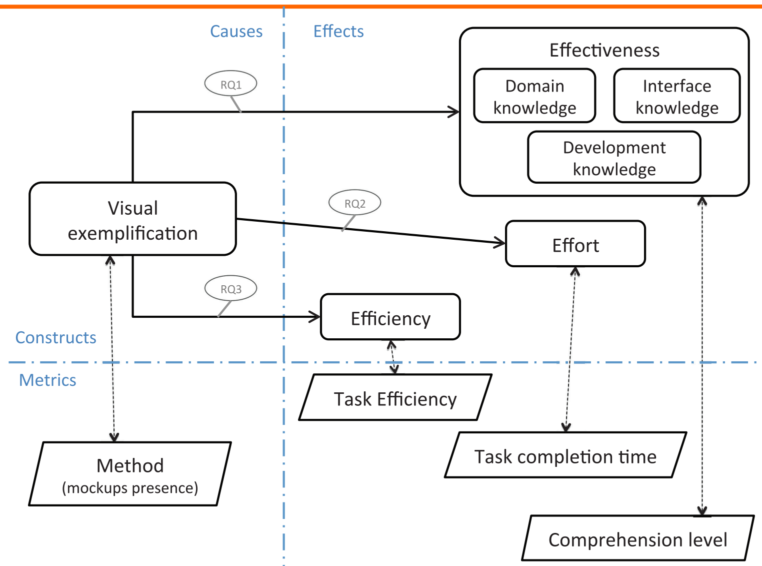


A running example

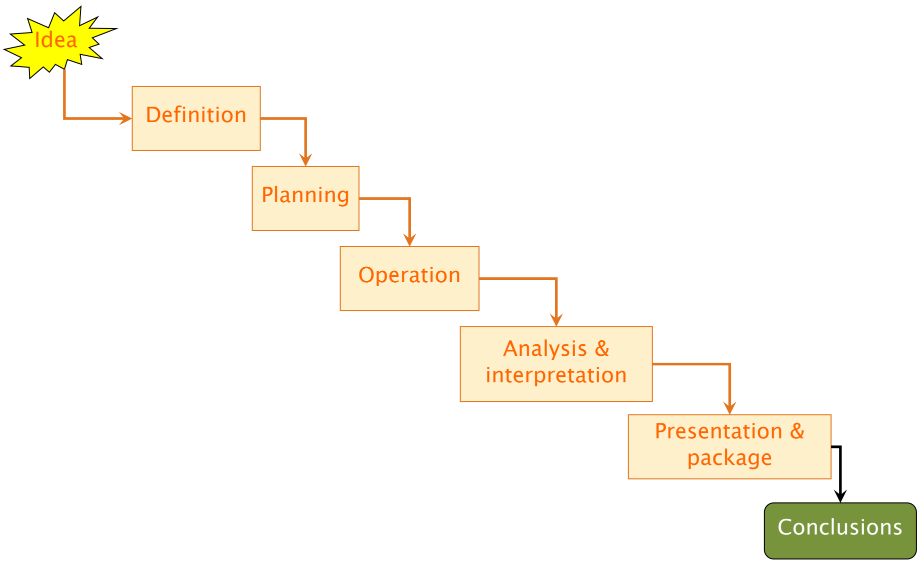
Ricca et al., Assessing the Effect of Screen Mockups on the Comprehension of Functional Requirements. ACM Transactions on Software Engineering and Methodology, 24(1), 38pp, 2014

http://doi.acm.org/10.1145/2629457

Conceptual model



Experimental process



Experimental process steps

- Definition
  - Goals and objectives of the study
- Planning
  - Define context
  - Formulate hypotheses
  - Identify input and output variables
  - Design the study
  - Analyze threats to validity

Experimental process steps

- Operation
  - Preaparation
  - Execution
  - Data validation
- Analysis and interpretation
  - Data understanding
    - Descriptive statistics, EDA
  - Possible data reduction
  - Hypothesis testing
  - Results interpretation

Experimental process steps

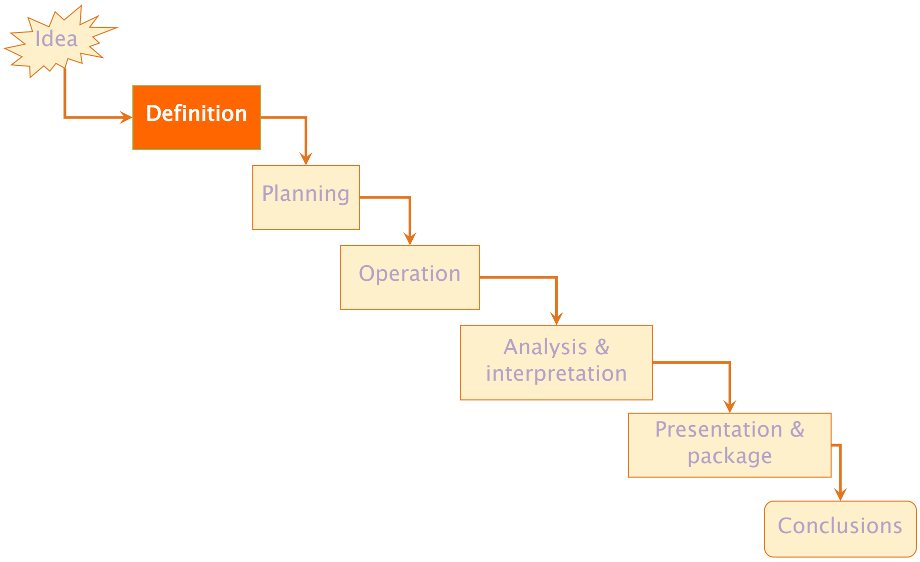
- Presentation and package
  - Document results
  - Prepare lab-package to enable replications
  - Sum up lessons learned

SoftEng http://softeng.polito.it

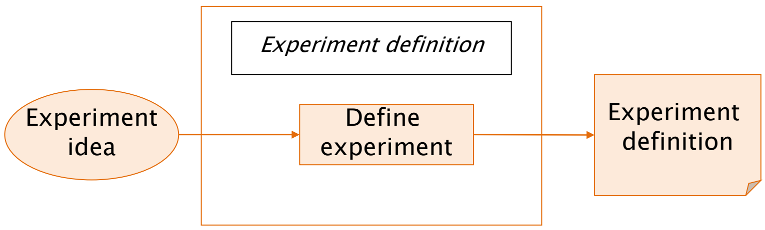
EXPERIMENT DEFINITION

SoftEng http://softeng.polito.it

Definition



Experiment definition: overview

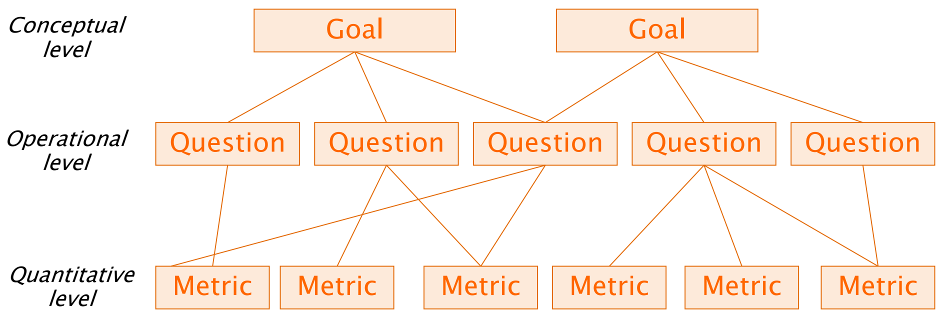


The definition determines the foundation of the experiment (*what* and *why*).

At this level, hypotheses sohuld be clear but not formally described

Goal-Question-Metric (GQM)

- Research approach



[Basili94b][Solingen99]

Goal definition template

Analyze	Objects(s) of study
for the purpose of	Purpose
with respect to their	Quality focus
from the point of view of	Perspective
in the context of	Context

Goal definition examples

Object of study	Purpose	Quality focus	Perspective	Context
Product	Characterize	Effectiveness	Developer	Subjects
Process	Monitor	Cost	Modifier	Objects
Model	Evaluate	Reliability	Maintainer	
Metric	Predict	Maintainability	Project manager	
Theory	Control	Portability	Corporate manager	
	Change	Comprehension	Customer	
			User	
			Researcher	

Goal definition example

"*Analyze the PBR and checklist techniques for the purpose of evaluation with respect to effectiveness and efficiency from the point of view of the researcher in the context of M.Sc. And Ph.D. students reading requirements documents*"

Regnell et al. Are the Perspectives Really Different?

Goal definition example

Analyze	Objects(s) of study **PBR and checklist techniques**
for the purpose of	Purpose Evaluation (and comparison)
with respect to their	Quality focus effectiveness and efficiency
from the point of view of	Perspective the researcher
in the context of	Context M.Sc. and Ph.D. students reading requirements docs

SoftEng http://softeng.polito.it

Goal definition example

Analyze	Objects(s) of study stereotyped UML diagrams
for the purpose of	Purpose evaluating their usefulness
with respect to their	Quality focus comprehension
from the point of view of	Perspective researcher and project manager
in the context of	Context research associates, undergraduate students graduate students

SoftEng http://softeng.polito.it

Goal definition example

- *Analyze* the use of screen mockups for the purpose of understanding their utility with respect to the effectiveness in comprehending requirements and the effort and the efficiency in performing comprehension tasks from the point of view of the requirements analyst, developer, and customer in the context of students reading two requirements specification documents for desktop applications.

Goal definition example

Analyze	Objects(s) of study the use of screen mockups
for the purpose of	Purpose understanding their utility
with respect to	Quality focus effectiveness in comprehending requirements and the effort and the efficiency in performing comprehension tasks
from the point of view	Perspective the requirements analyst, developer, and customer
in the context of	Context students reading requirements specification for desktop applications

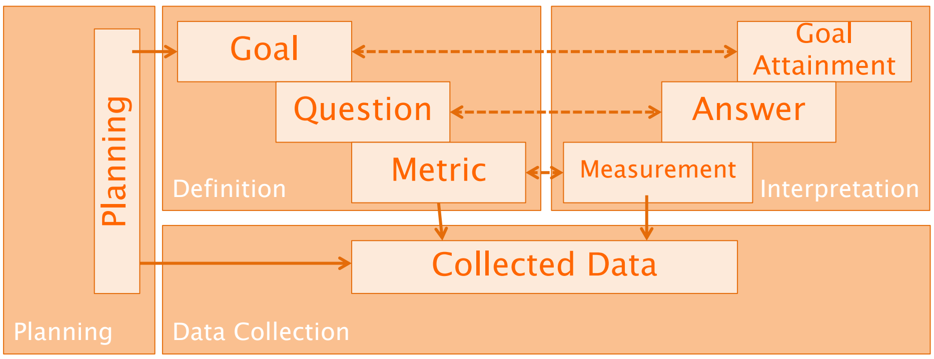
Example questions

RQ1. Does the requirements comprehension *effectiveness* vary when use cases are provided in conjunction with screen mockups?

RQ2. Does the *effort* required to complete a comprehension task vary when use cases are provided in conjunction with screen mockups?

RQ3. Does the *efficiency* in performing a comprehension task vary when use cases are provided in conjunction with screen mockups?

GQM in perspective



Questions

- Refinement of goals to a more operational level
  - By answering the question one should be able to conclude whether the goals has been achieved
- Expected answers can be formulated as (high level) hypotheses
- Questions may focus on different aspects of the goal

Example – Goal

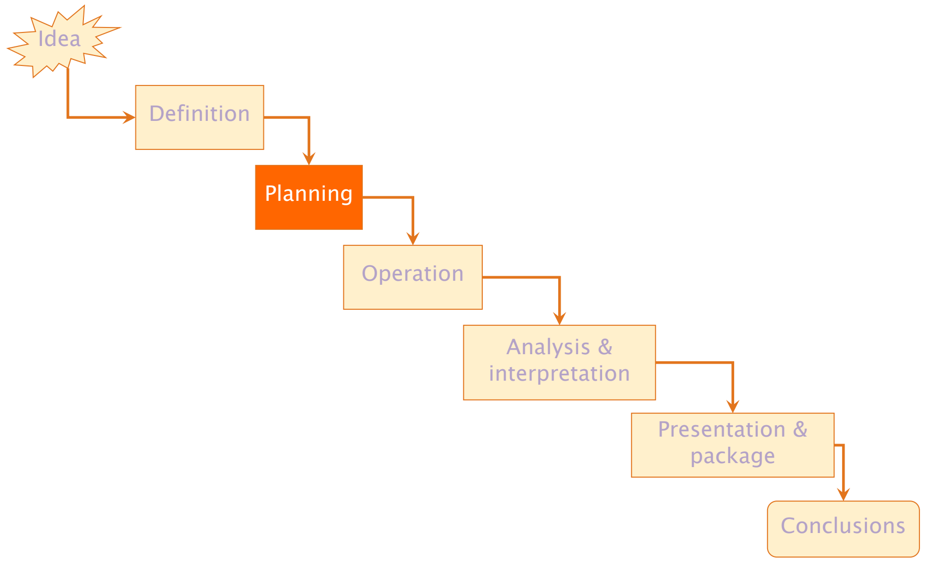
- Goal: reliability
  - Analyze the product and process
  - For the purpose of characterizing
  - With respect to reliability and its causes
  - Form the point of view of the software development team
  - In the context of project A

Example – Questions

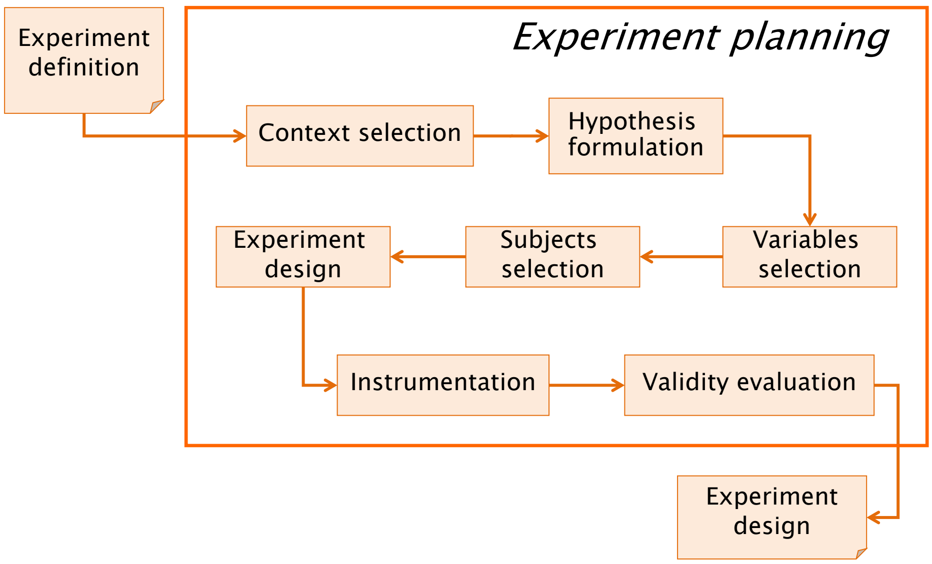
- Product definition
  - Does the sw adhere to coding standards
  - What is the complexity of sw?
- Quality
  - What is the distribution of failures?
  - What is the distribution of faults?
  - What was the distribution of failure handling effort?
  - What is the relationship between code reviews and reliability?

EXPERIMENT PLANNING

Experiment planning



Experiment planning: 7 steps



Context selection

- The content of the experiment can be characterized according to 4 dimensions:
  - Offline vs. Online
  - Student vs. Professional
  - Toy vs. Real
  - Specific vs. General

Hypothesis formulation

- Two hypotheses must be formalized
  - Null hypothesis H_0
    - no real underlying trends or patterns in the experiment setting
  - Alternative hypothesis H_a
    - There exists real underlying trends or patterns in the experiment setting
  - If we investigate the existence of a pattern, the null hypothesis must state that no patterns exist

Scientific method (reminder)

- Conjecture (P)
  - Administration of treatment has influence on some feature
- Consequence (Q)
  - We observe a difference in terms of some feature

If P, then Q

Falsification (*modus tollens*)

- We aim at detecting ~Q
  - The opposite of the consequence
  - We test the null hypothesis
  - If verified we can conclude the conjecture is false
- Aiming at verifying Q: is wrong
  - Provides no insight on the conjecture
  - *Affirming the consequent* fallacy

Example

- Question
  - Do code reviews affect quality?
- Conjecture (P)
  - Code reviews reduce defects
- Hypothesis – alternative (Q)
  - (When code reviews are applied) we observe fewer defects than when they are not applied
- Hypothesis – null (~Q)
  - We observe no difference in terms of defects when code reviews are applied or not

Example

- Outcome of the experiment
  - We confirm the null hypothesis (~Q)
    - We conclude that code reviews do not reduce defects (~P)
    - ~Q => ~P
    - The conjecture has been falsified
  - We reject the null hypothesis
    - We are more confident that it is likely that code reviews reduce defects (P)
    - Q ~~=>~~ P
    - The conjecture has been corroborated

Hypotheses example

H_{c0}. The presence of screen mockups *does not significantly improve* the comprehension level of functional requirements.

H_{t0}. The presence of screen mockups *does not significantly improve* the time to accomplish a comprehension task.

H_{e0}. The presence of screen mockups *does not significantly affect* the efficiency of the comprehension task.

Variable selection

- Independent variables
  - Variables that we can control
    - Treatment
  - Variables that we can monitor
    - Context and domain
    - Possible confounding factors
- Dependent variables
  - Allow measure the effect of treatments
    - Sometimes they cannot be measured directly
    - Use of proxies

Example variables

Variable	Type	Description	Scale
Method	Indep.	whether requirements are augmented with screen mockups	Nominal $\in \{S, T\}$
Application	Indep.	experimental object used in task	Nominal $\in \{AMICO, EasyCoin\}$
Lab	Indep.	order of the experiment unit within the experiment for the participant	Ordinal $\in \{1, 2\}$
Experiment	Indep.	participants' profiles and experiment	Nominal $\in \{UniBas1, UniGe, PoliTo, UniBas2\}$
Comprehension level	Dep.	comprehension achieved by a participant on the functional requirements	Ratio $\in [0, 1]$
Task completion time	Dep.	time spent by a participant to complete a comprehension task	Interval $\in (0, \infty)$
Task efficiency	Dep.	task time efficiency	Ratio $\in [0, 600]$
Source	Dep.	(main) source of information used to perform comprehension task	Nominal $\in \{G, K, UC, UCD, S\}$

Subjects selection

- Subjects are selected from a population:
  - Probability sampling
    - Simple random, systematic, stratified
  - Non probability sampling
    - Convenience, quota
- Related to the level of generalization of the experiment

Subjects selection

- General principles
  - The larger the variation of the population is, the larger is the sample size needed
  - Analysis of data may influence choice of sample size: consider how to analyze data since design stage

SoftEng http://softeng.polito.it

Subjects Example

	UniBas1	UniGe	PoliTo	UniBas2
Date	Nov. 2009	Dec. 2009	Apr. 2010	Dec. 2010
Location	Univ. Basilicata	Univ. Genova	Poly. Torino	Univ. Basilicata
Degree	Computer Science	Computer Science	Computer Engineering	Math/Telecom
Level	Undergrad	Undergrad	Graduate	Graduate
Year	2nd	3rd	2nd	1st
Semester	1st	1st	2nd	1st
Number	33	51	24	31

SoftEng http://softeng.polito.it

Design

- Experiment = series of trials
  - Number of factors and treatments determines design type and data analysis

- Memento:
  - Design and interpretation of results are closely related: the choice of design affects the analysis and vice versa

Design

- General design principles
  - Randomization
  - Blocking
  - Balancing

Randomization

- Used to
  - average the effects of a factor that may otherwise be present
  - select representative subjects for the population they come from
- Applies on
  - allocation of the objects
  - allocation of the subjects
  - order the tests are performed

SoftEng
http://softeng.polito.it

Blocking

- Used to eliminate the undesired effect of a (confounding) factor we are not interested in
- Blocks are built separating by factor
  E.g.:
  - Block 1 : subjects with experience
  - Block 2 : subjects with no experience
- Blocks studied separately
- Effects between blocks not studied

SoftEng
http://softeng.polito.it

Balancing

- Experiment design is balanced when treatments are assigned so that each treatment has equal number of subjects
  - Viceversa also subjects should have a burden as far as possible similar
  - Desirable because it both simplifies and strengthens the statistical analysis of data, but not necessary

Fully randomized design

- The levels of the primary factor are randomly assigned to the experimental units.
- Balance
  - Same number of replications for each level

Fully randomized design

- Model
  - $Y_{ij} = \mu + T_i + \text{error}$
  - Where
    - i level of the main factor
    - j replication (subject) for that level
- Estimates
  - $\mu = \overline{Y}$
    - the average of all the data
  - $T_i = \overline{Y}_i - \overline{Y}$
- Hypothesis
  - H_0: $T_i = T_j \Leftrightarrow \overline{Y}_i = \overline{Y}_j \Leftrightarrow \mu_i = \mu_j$

1 Factor, 2 Treatment Levels

- Most typical and simple case
  - One of the treatments can be “absence of”
    - Does the introduction of a technique affects some output variable?
- Example
  - Factor: design notation
    - Treatment 1: UML
    - Treatment 2: UML w/stereotypes

Randomized – 1 F, 2 T

- Each subject is randomly assigned to one of the two treatments
  - Balancing
  - Comparison of subjects in two groups
- Example of hypothesis
  - H0: $\mu_1 = \mu_2$
  - H1: $\mu_1 \neq \mu_2$, $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$
  - Where:
    - μ_t = mean of dependent variable for subject that received treatment t

SOftEng
http://softeng.polito.it

Randomized – 1 F, >2 T

- Each subject is randomly assigned to one of the treatments
  - Balancing
  - Comparison of subjects in groups
- Example of hypothesis
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$
  - H_1: $\mu_i \neq \mu_j$ for at least one pair (i,j)

SOftEng
http://softeng.polito.it

Randomized – 1 F, >2 T

Subjects	Treatment 1	Treatment 2	Treatment 3
1	X		
2		X	
3			X
4			X
5	X		
6		X	
	μ_1	μ_2	μ_3

Example design

Table IV. Experiment Design. S is for requirements specification with screen mockups and T without them

	Group 1	Group 2	Group 3	Group 4
First laboratory run	S, EasyCoin	T, EasyCoin	T, AMICO	S, AMICO
Second laboratory run	T, AMICO	S, AMICO	S, EasyCoin	T, EasyCoin

Instrumentation

- Goal of instrumentation :
  - provide means for performing the experiment and monitor it, without affecting the control of the experiment
- Instruments are of three types
  - Objects
  - Guidelines
  - Measurement tools

Example instrumentation

3.4. Procedure

The participants accomplished comprehension tasks using computers equipped with MS Word. An Internet connection was available, while performing the comprehension tasks. We provided the participants with the following material.

—The requirements specification documents in electronic format (MS Word) of EasyCoin and AMICO. In particular, each document contained
  (i) the system mission, namely, a textual description of both the functionality of the future system and the environment in which it will be deployed;
  (ii) a UML use case diagram summarizing the use cases of the systems;
  (iii) functional requirements expressed as use cases specified according to the chosen template. Depending on the experiment design, use cases were or were not complemented with screen mockups;
  (iv) a glossary of the terms.

—A paper copy of the comprehension questionnaires of EasyCoin and AMICO.

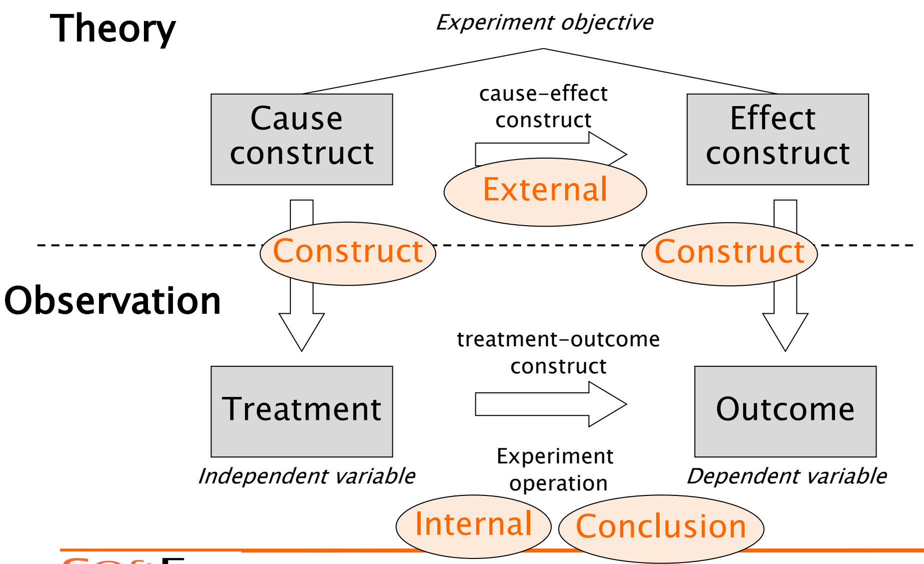
—A paper copy of the postexperiment questionnaire shown in Table VI.

—The training material, which included a set of instructional slides describing the template employed for the specification of the use cases, some examples not related with the experiment objects, and a set of slides describing the procedure to follow in the task execution.

Validity evaluation

- Adequate validity is obtained when results are valid for the population to which we would like to generalize
- Threats to validity are limitations to the adequate validity
- There are 4 types of threats:
  - Conclusion
  - Internal
  - Construct
  - External

Validity evaluation



Conclusion validity

- Conclusion validity
  - Threats concerning the statistical issues that can affect the ability to draw the correct conclusion about the relationship between treatments and outcome

SOftEng http://softeng.polito.it

Internal validity

- Internal validity
  - Threats concerning issues that lead to indicate a causal relationship, when there is none
  - The extent to which the behavior observed in the experiment could be due to disturbing factors instead of the treatments

SOftEng http://softeng.polito.it

Construct validity

- Construct validity
  - Threats concerning issues related to the relationship between
    - cause construct and treatment
    - effect construct and outcome
  - They refer to the extent to which the experiment settings actually reflect the construct under study

SOftEng http://softeng.polito.it

External validity

- External validity
  - Can the result of the study be generalized outside the scope of the study ?

SOftEng http://softeng.polito.it

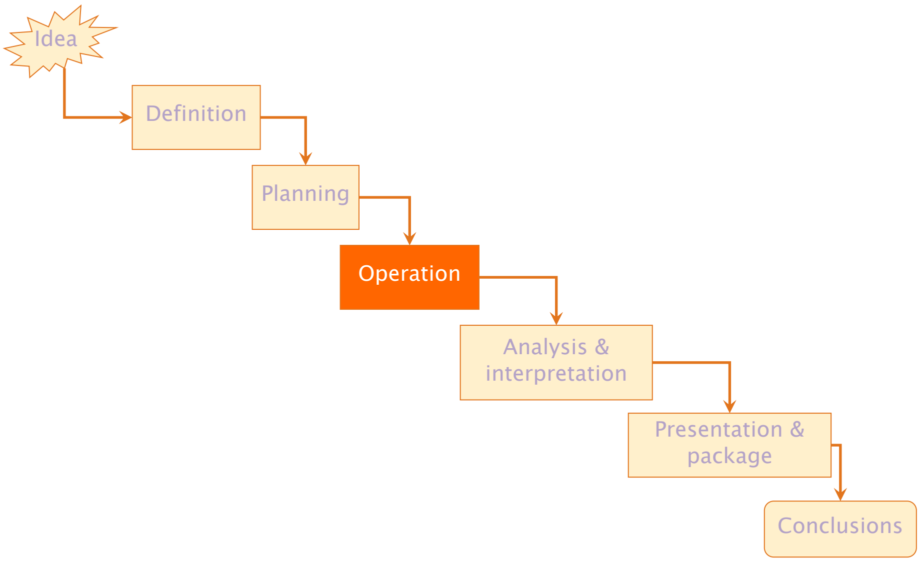
Validity evaluation

- For each threat type, a list of threats is available in [Cook79] and [Campbell63]
- Priority among the threats is a matter of optimization

- Possible rank in theory testing:
  - Internal – construct – conclusion – external
- Possible rank in applied research:
  - Internal – external – construct – conclusion

OPERATION

Operation



Operation

- Preparation
  - Get participants
    - Ethical issues
    - Privacy issues
- Execution
  - Data collection
- Data validation

Ethical issues

- Professionals
  - Paid to perform a task
- Students
  - Recruitment
  - Experiment as part of a course
    - Is it integrated?
    - Do all participants go through equivalent experience (counterbalanced design)
    - Credits
- Ethical board

Privacy

- In Italy there is quite a strict law:
  - http://www.garanteprivacy.it/garante/document?ID=1219452
  - Section 7 provides a list of the rights of the subject,
  - Section 13 details the information to be provided to the subjects

Privacy

- Information to be provided
  - Purposes and modalities of the processing for which the data are intended
  - Nature of providing the requested data
  - Consequences of denial to reply
  - Entities or categories of entity of data communication and dissemination
  - Rights
  - Responsible for the data

Privacy information

- Purposes and modalities of the processing for which the data are intended
  - The data you provide will be handled for statistical and scientific purposes, aimed at investigating the details of software development. The handling will be carried on by electronic means.
- Nature of providing the requested data
  - The participation in the investigation is voluntary.
- Consequences of denial to reply
  - Denying to answer will have no consequence.
- Entities or categories of entity of data communication and dissemination
  - Personal data collected during the investigation will be shared only among the researchers involved in the project.

Privacy information

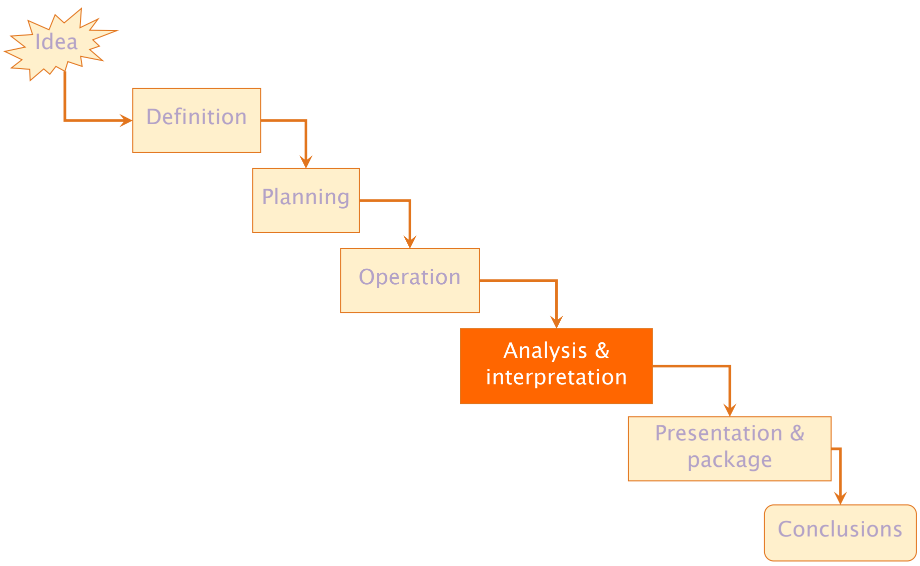
- Rights
  - At any time you will be able to exert your rights with the responsible for the data handling, according to section 7 of D.lgs.196/2003, which we copy integrally:
  - …
- Responsible for the data
  - The responsible for data treatment is …

Execution

- Data collection
  - Manually entered by participants
  - Tool supported
  - Interviews
  - Automatic
- Experimental environment

ANALYSIS AND INTERPRETATION

Analysis



Analysis and interpretation

- Descriptive statistics
  - Distribution
  - Central tendency
  - Dispersion
  - Visualization
- Data reduction
- Hypothesis testing

Error types

- Type I error
  - When we conclude there is a trend/pattern but actually there isn't
  - α = P(reject H_0 | H_0 is true)
- Type II error
  - When we don't see any relation between factors and outcome, but actually there is a trend/pattern
  - β = P(accept H_0 | H_0 is false)

Error types – Hypothesis testing

		H_0 True	H_0 False
Test decision	Reject	Type I Error	OK
	Fail to reject	OK	Type II Error

SoftEng http://softeng.polito.it

Error types – Justice system

Reject presumption of innocence

		Defendant Innocent	Defendant Guilty
Trial decision	Guilty	Type I Error	OK
	Not Guilty	OK	Type II Error

Fail to reject presumption of innocence

SoftEng http://softeng.polito.it

Power

- The power of a test is the probability that the test reveal a true pattern if H_0 is false
- Power $= P(\text{ reject } H_0 \mid H_0 \text{ is false})$
  $= 1 - P(\text{accept } H_0 \mid H_0 \text{ is false})$
  $= 1 - P(\text{type II error}) = 1 - \beta$

SOftEng http://softeng.polito.it

Hypothesis testing

- Steps
  - Fix the significance level (α)
    - Typically in planning phase
  - Select the statistical tests
    - Typically in planning phase
  - Perform the tests
  - Decide about null hypotheses
    - Reject
    - Fail to reject

SOftEng http://softeng.polito.it

Significance level α

- What is the acceptable α level in our study?
  - Level of confidence: 1−α
- Standard levels:

Significance (α)	Confidence (1−α)
5%	95%
1%	99%

Test

- P-value
  - probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true
- Decision
  - Reject when p-value < α
  - Fail to reject when p-value > α

Example – Hypothesis

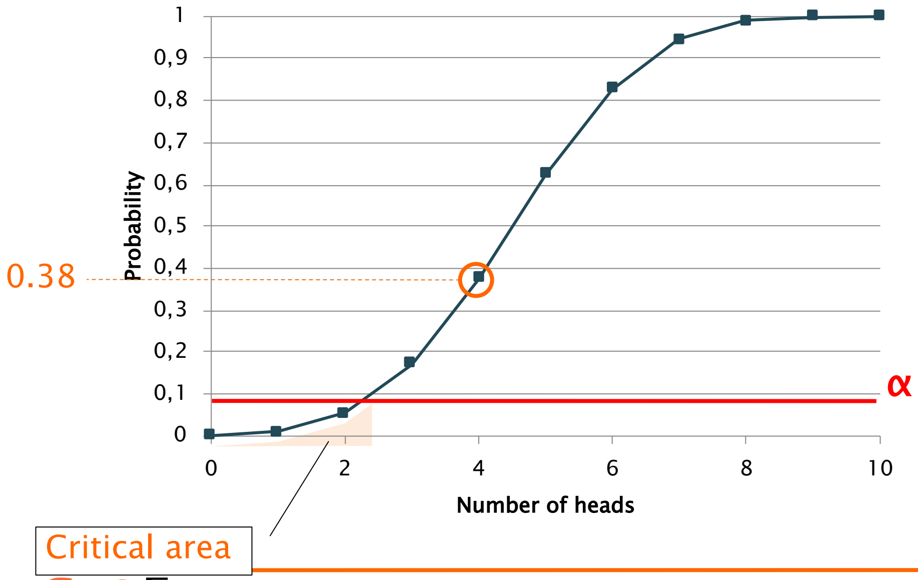
- Conjecture:
  - coin is “tricky” and disfavors heads
- Consequence:
  - as a result of a series of tosses the number of heads is smaller than the number of tails.
- Hypotheses
  - H_0: Heads = Tails = # Tosses / 2
  - H_a: Heads < Tails
- We assume $\alpha = 10\%$

Example – Experiment

- Experiment result: 4 heads in 10 trials

- Assuming H_0 is true, what is the probability of having 4 or less heads in 10 trials?
  - Binomial distribution
    - Cumulative function

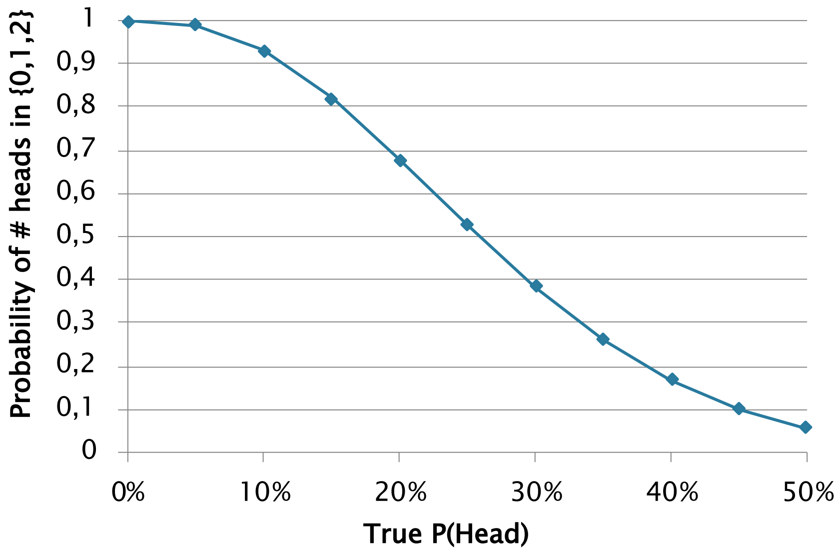
Example – Testing



Example – Power

- What is the real capability of the previous experiment to discover a tricky coin?
  - Power
- Assuming H_0 is false (H_a is true) what is the probability of rejecting H_0 ?
  - We reject H_0 if we are in the critical area
    - In the example above: # Heads in {0, 1, 2}
  - H_a is true if P(Head) < 0.5

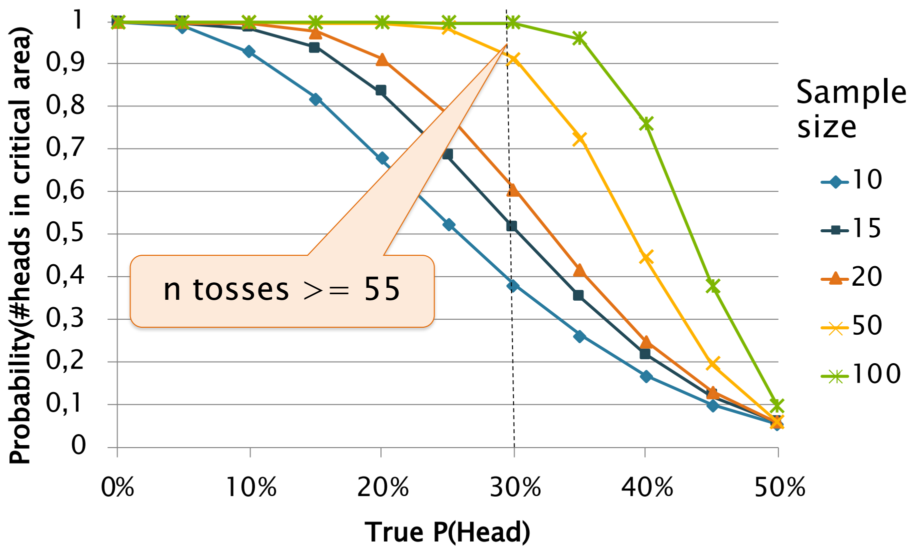
Example – Power



Example – Power

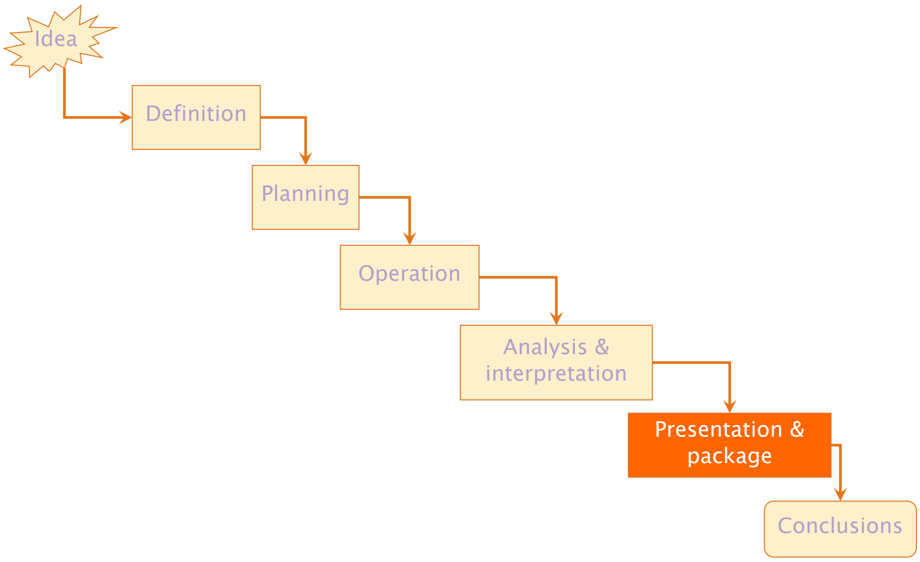
- Let's suppose we suspect that heads show up just 30% of the times
  - P(Head) = 30%
- How many trials should we run to have at least 95% of chances to discover such a bias?
  - Power > 0.95

Example – Power vs. Sample



PRESENTATION AND PACKAGING

Reporting



Reporting

- Introduction
- Problem statement
- Experiment planning
- Operation
- Data analysis
- Interpretation
- Discussion and conclusions

APA Guidelines

- Abstract
- Introduction
- Method
  - Design
  - Subjects/Participants
  - Apparatus/Materials
  - Procedure
- Results
- Discussion

Replication package

- "Laboratory packages" (Basili et al., 1999; Shull et al.; 2002; Ciolkowski et al., 2002)
- Fundamental to allow replication
  - Analysis and goals of the experiment
  - Motivation for the design decisions
  - Experimental design, including validity threats and strengths
  - Context in which the experiment was carried on
  - Procedure
  - Analysis methods

Online databases

- Promise
  - http://promisedata.org/?cat=11
- Floss metrics
  - http://melquiades.flossmetrics.org/
- Sw Lifecycle Empirical DB
  - http://www.thedacs.com/databases/sled/

Online repositories

- Zenodo
  - Link to GitHub
- FigShare

Bibliography

- Claes Wohlin, Per Runeson, Martin Host, Magnus Ohlsson, Bjorn Regnell, Anders Wesslen. 2012. Experimentation in Software Engineering – An Introduction. Kluwer.
- Basili, Victor; Gianluigi Caldiera, H. Dieter Rombach (1994). "The Goal Question Metric Approach"
  - ftp://ftp.cs.umd.edu/pub/sel/papers/gqm.pdf
- Van Solingen, Rini; Egon Berghout (1999). The Goal/Question/Metric Method. McGraw-Hill Education.

Bibliography

- Regnell, Björn, Per Runeson, and Thomas Thelin. "Are the perspectives really different?-further experimentation on scenario-based reading of requirements." *Empirical Software Engineering* 5(4) (2000): 331-356.
- Ricca, et al. "How Developers' Experience and Ability Influence Web Application Comprehension Tasks Supported by UML Stereotypes: A Series of Four Experiments." *IEEE Transactions on Software Engineering,* 36(1) (2010): 96-118.
- **Ricca et al., Assessing the Effect of Screen Mockups on the Comprehension of Functional Requirements.** ACM Transactions on Software Engineering and Methodology, 24(1) (2014)