



## Experiment Process: Step 2. Planning

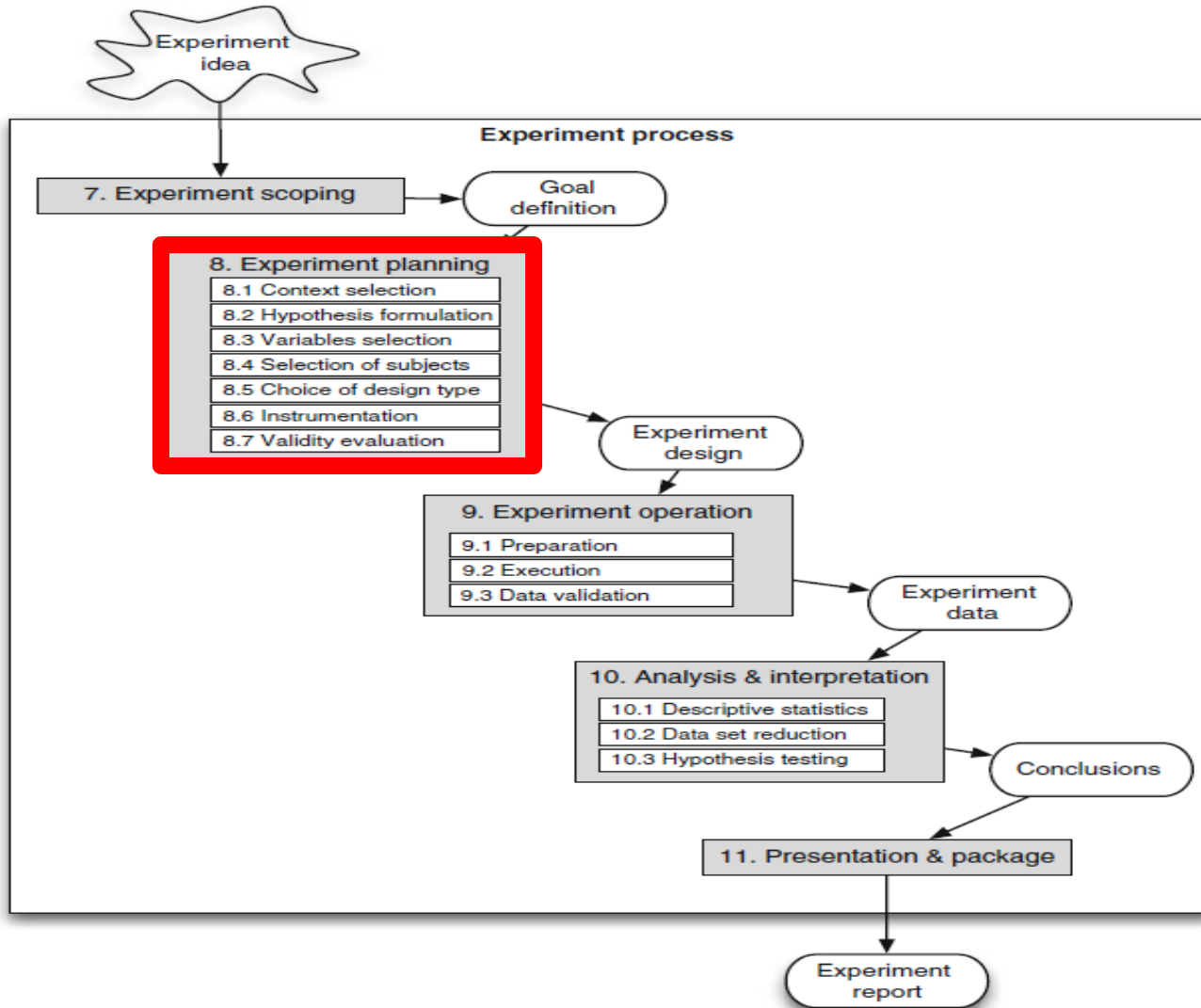
*The slides are prepared by: Dr. Mahmood Niazi, Dr. Malak Baslyman, Dr. Hamoud Aljamaan & Dr. Mohammad Alshayeb*

# Lecture Objectives

- ✓ Context Selection
- ✓ Hypothesis formulation
- ✓ Variables selection
- ✓ Selection of subjects
- ✓ Experiment Design
- ✓ Instrumentation
- ✓ Validity evaluation




# Experiment process



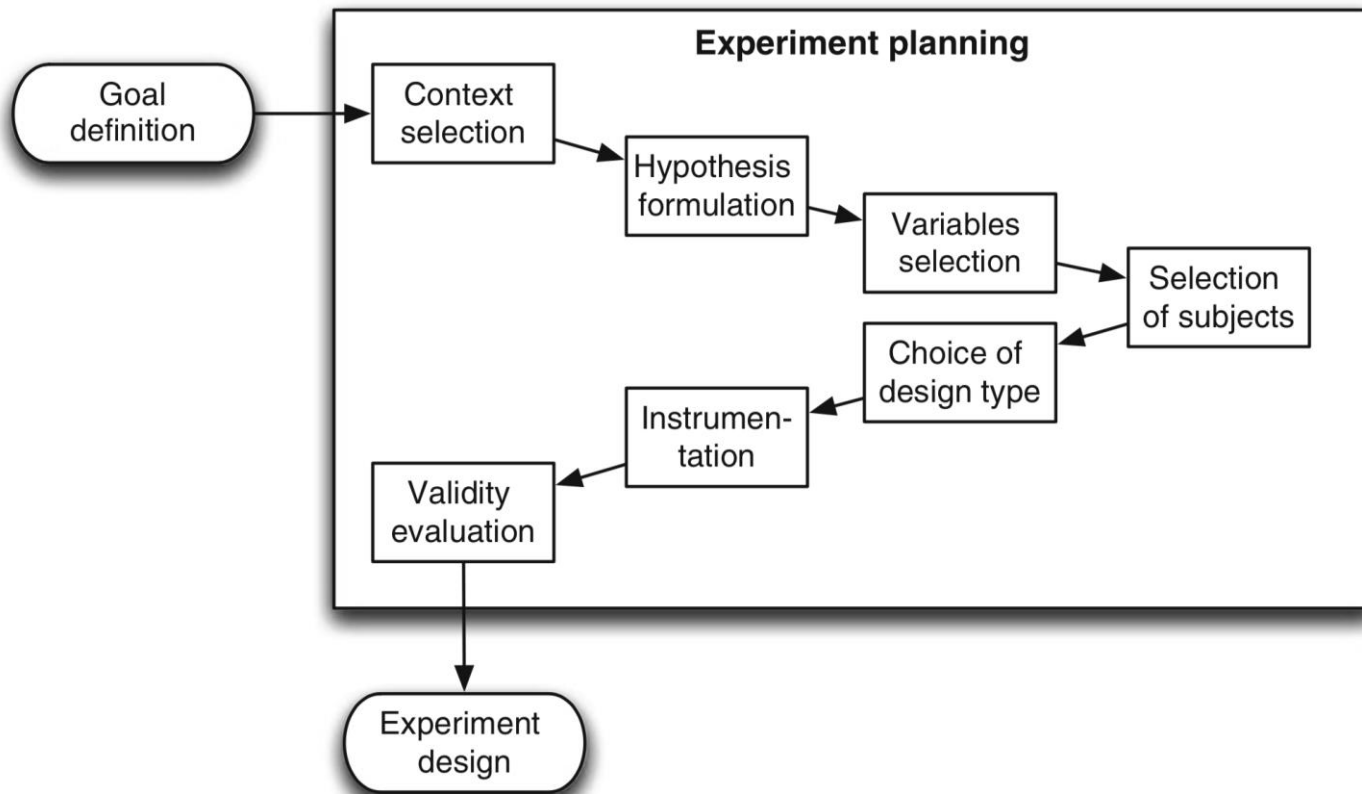
# Introduction

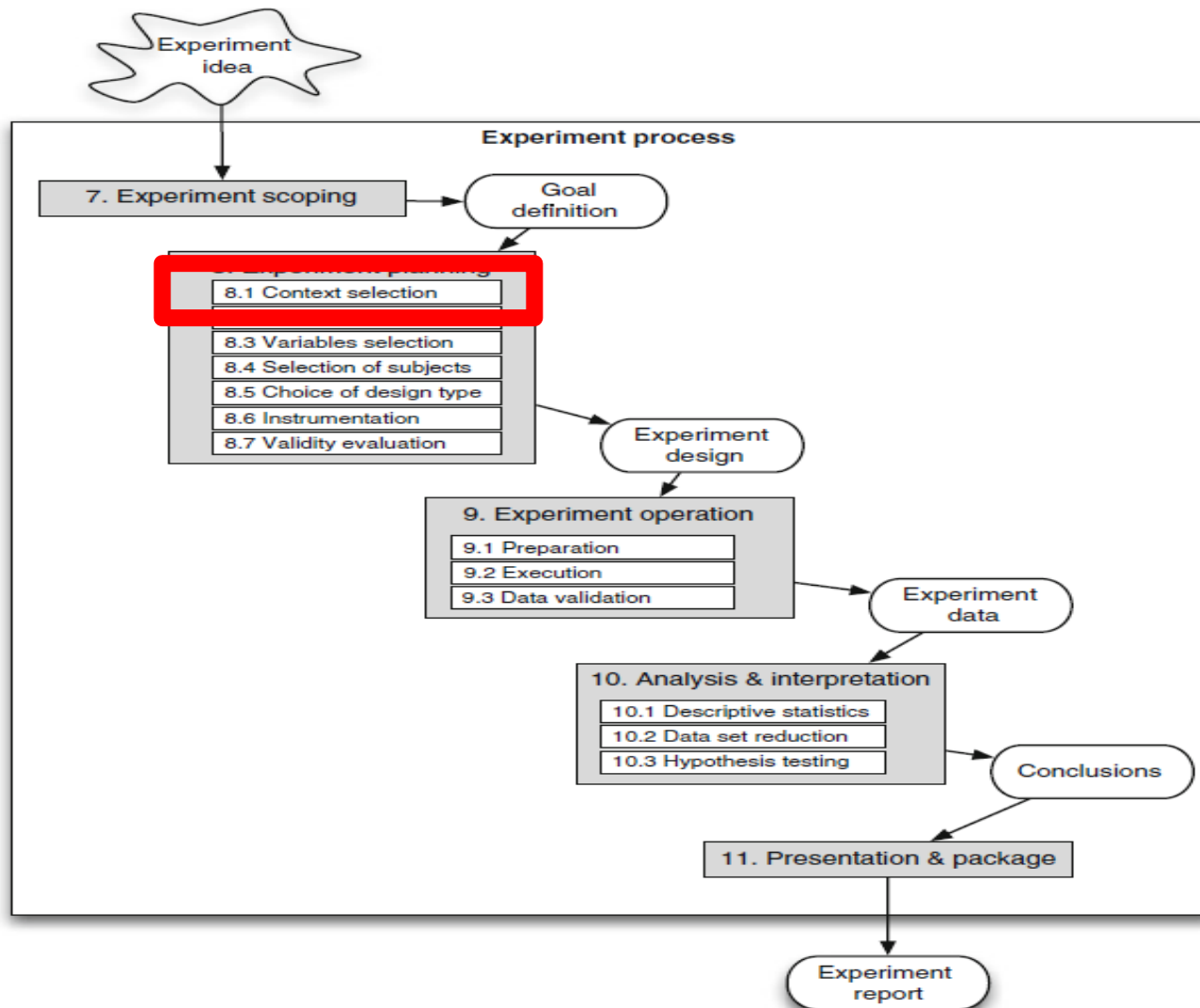


- After the scoping of the experiment, the planning takes place.
  - The **scoping** determines the foundation for the experiment – *why the experiment is conducted* – while the **planning** prepares for *how the experiment is conducted*.
  - As in all types of engineering activities, the experiment must be planned, and the plans must be followed-up in order to control the experiment.
  - The result of the experiment can be disturbed, or even destroyed if not planned properly.
- 

# Planning phase overview

- The planning phase of an experiment can be divided into seven steps

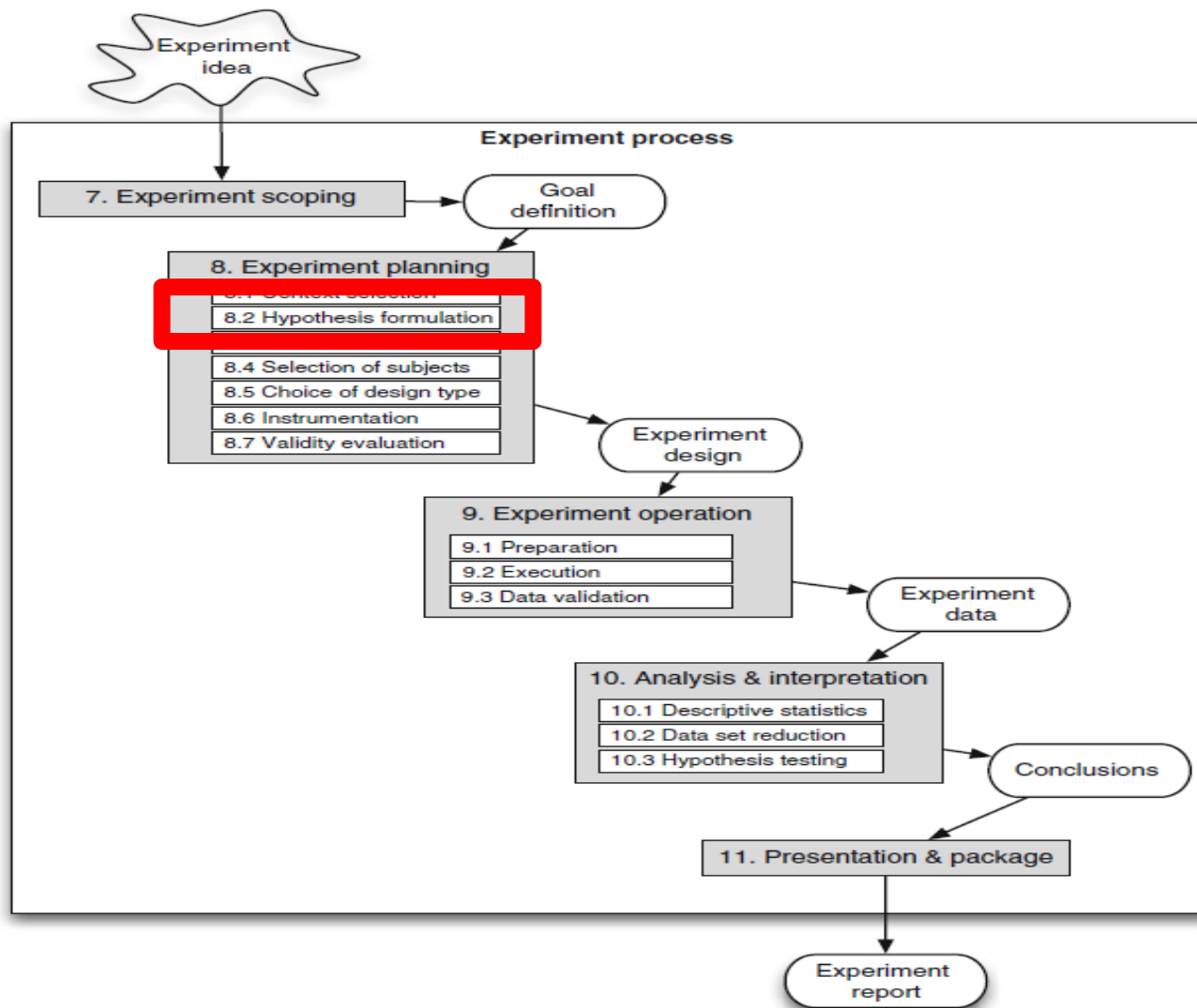




# Context selection




- In order to achieve the most general results in an experiment, it should be executed in **large, real** software projects, with **professional** staff.
- The context of the experiment can be characterized according to four dimensions:
  - Off-line vs. on-line
  - Student vs. professional
  - Toy vs. real problems
  - Specific vs. general
- A common situation in an experiment is that something existing is compared to something new, for example an existing inspection method is compared to a new one.





# Hypothesis formulation (1/3)



- The basis for the **statistical analysis** of an experiment is **hypothesis testing**.
  - A hypothesis is stated formally, and the data collected during the course of the experiment is used to, if possible, reject the hypothesis.
  - If the hypothesis can be **rejected** then conclusions can be drawn, based on the hypothesis testing under given risks.
  - In the planning phase, the experiment definition is formalized into hypotheses
- 

# Hypothesis formulation (2/3)

- A **null hypothesis**,  $H_0$ 
  - It states that there are no real underlying trends or patterns in the experiment setting; the only reasons for differences in our observations are coincidental (by chance).
  - This is the hypothesis that the experimenter wants to reject with as high significance as possible.
  - An example hypothesis is that a new inspection method finds on average the same number of faults as the old one.

$$H_0 : \mu_{N_{old}} = \mu_{N_{new}}$$

- Where  $\mu$  denotes the **average** and N is the number of faults found

# Hypothesis formulation (3/3)

- An **alternative hypothesis**,  $H_a$ ;  $H_1$ , etc.
  - It is the hypothesis in favor of which the null hypothesis is rejected.
  - An example hypothesis is that a new inspection method on average finds more faults than the old one

---

$$H_1 : \mu_{N_{old}} < \mu_{N_{new}}$$

- Where  $\mu$  denotes the average and N is the number of faults found

# Types of errors

Hypothesis testing involves two main types of risks:

- Type-1-error (false positive)
  - the risk of rejecting a true null hypothesis

$$P(\text{type-I-error}) = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- Type-2-error (false negative)
  - the risk of not rejecting a false null hypothesis

$$P(\text{type-II-error}) = P(\text{not reject } H_0 \mid H_0 \text{ false})$$

# Types of errors

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome ! (True positive)
Fail to reject null hypothesis	Correct outcome ! (True negative)	Type II Error (False negative)

# Tips to write good hypothesis (1/3)

- Develop either research questions or hypotheses to reduce redundancy.
- Include your variables in your hypotheses.
- Use the same pattern of word order in the hypotheses to help a reader to easily identify your major variables.
- Develop your research hypotheses based on a theory.
- Measure your independent and dependent variables separately.  
This process supports the cause-and-effect logic of quantitative research.

# Tips to write good hypothesis (2/3)









- Checklist to check the hypothesis:
  - Is the language clear and focused?
  - Does the hypothesis introduce the research topic?
  - Does the hypothesis include both an independent and dependent variable?  
Are they easy to identify?
  - Can the hypothesis be tested through experimentation?
  - Does the hypothesis explain what you expect to happen during your experiment?

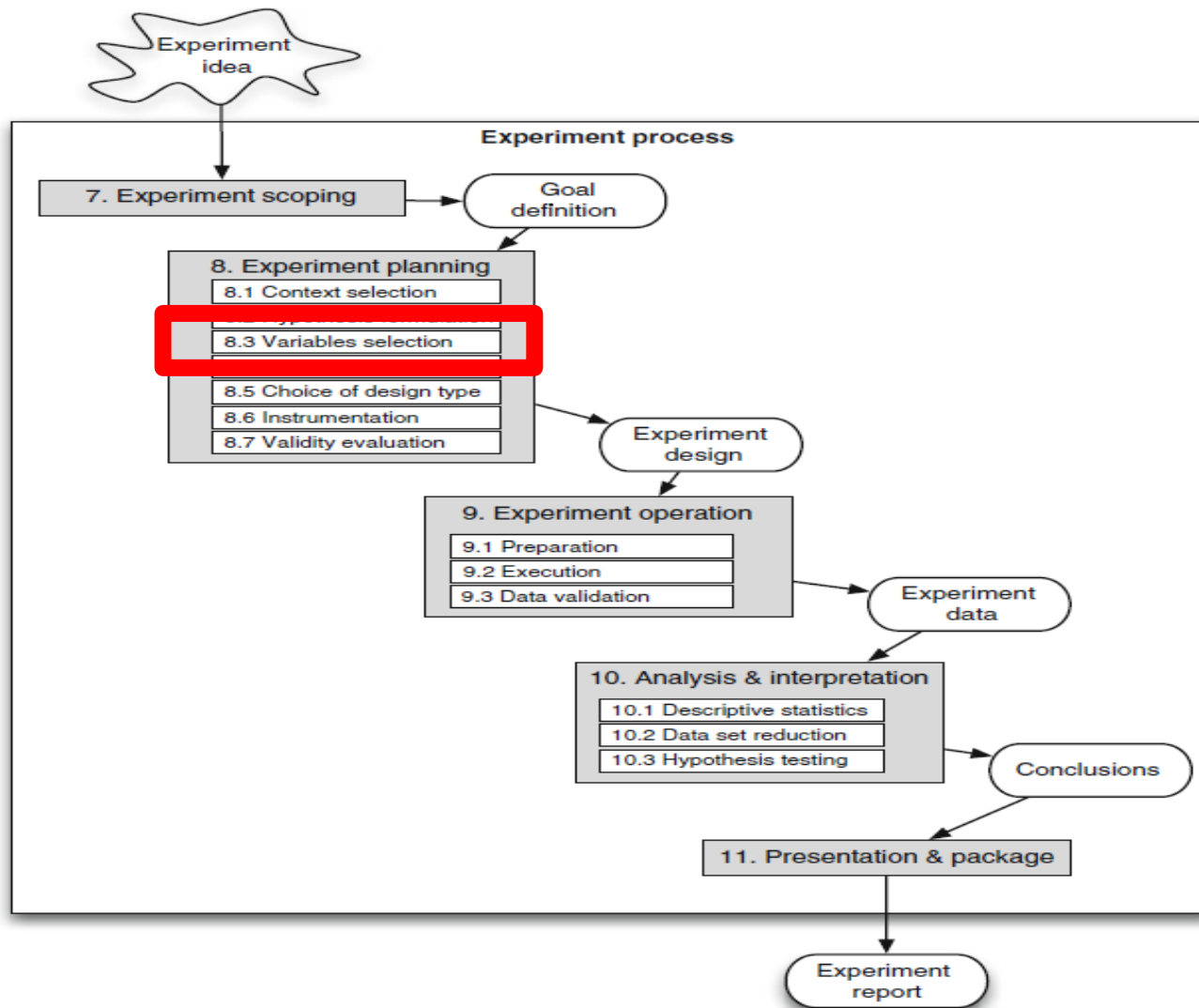
# Tips to write good hypothesis (3/3)

- To identify the variables, you can write a simple prediction in *if...then form*. The first part of the sentence states the independent variable, and the second part states the dependent variable.
  - If a first-year student starts attending more lectures, then their exam scores will improve.
- In academic research, hypotheses are more commonly phrased in terms of correlations or effects, where you directly state the predicted relationship between variables.
  - The number of lectures attended by first-year students has a positive effect on their exam scores.
- If you are comparing two groups, the hypothesis can state what difference you expect to find between them.
  - First-year students who attended most lectures will have better exam scores than those who attended few lectures.



# Hypotheses examples

- “Using a deep learning model will improve the accuracy of real-time speech recognition systems by at least 15% compared to traditional statistical models.”  

- “Machine learning can improve computer performance.”  

- “Implementing algorithm X in data sorting will reduce the average processing time by 20% compared to the current leading algorithm Y under similar conditions.”  

- “Using machine learning techniques for network intrusion detection will increase the detection rate of zero-day exploits by 30% over traditional signature-based methods.”  

- “Using more complex algorithms will solve problem Y.”  

- “Algorithm X is better.”  




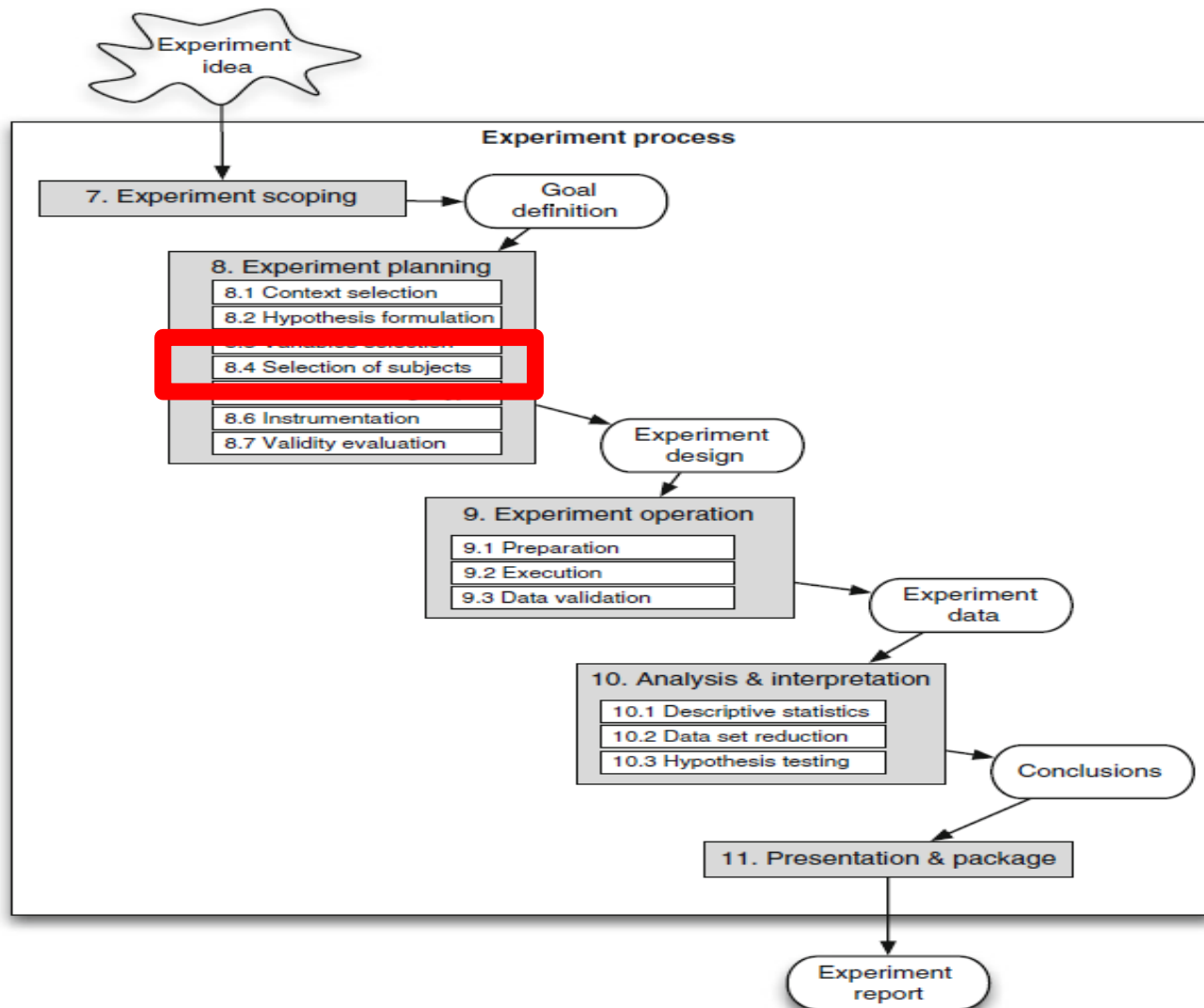
# Variables selection

- Independent variable(s)
  - The independent variables are those variables that we can control and change in the experiment.
  - Choosing the right variables is not easy and it usually requires domain knowledge.
  - The variables should have some effect on the dependent variable and must be controllable.
    - Example: experience in Java (5 years, 10 years etc)
- Dependent variable(s)
  - The effect of the treatments is measured in the dependent variable(s).
  - Often there is only one dependent variable, and it should therefore be derived directly from the hypothesis (Otherwise there will be threat to conclusion validity).
  - The variable is mostly not directly measurable, and we have to measure it via an indirect measure instead
    - Example: Fault density in Java coding

# Variables selection example



- If you are conducting an experiment to see how different sorting algorithms affect the time it takes to sort an array of numbers, your variables might be:
- Independent Variable: Dataset, resources, etc...
- Treatment: The type of sorting algorithm used (e.g., quicksort, mergesort, bubblesort).
- Dependent Variable: The amount of time it takes to sort the array.



# Selection of subjects (1/6)



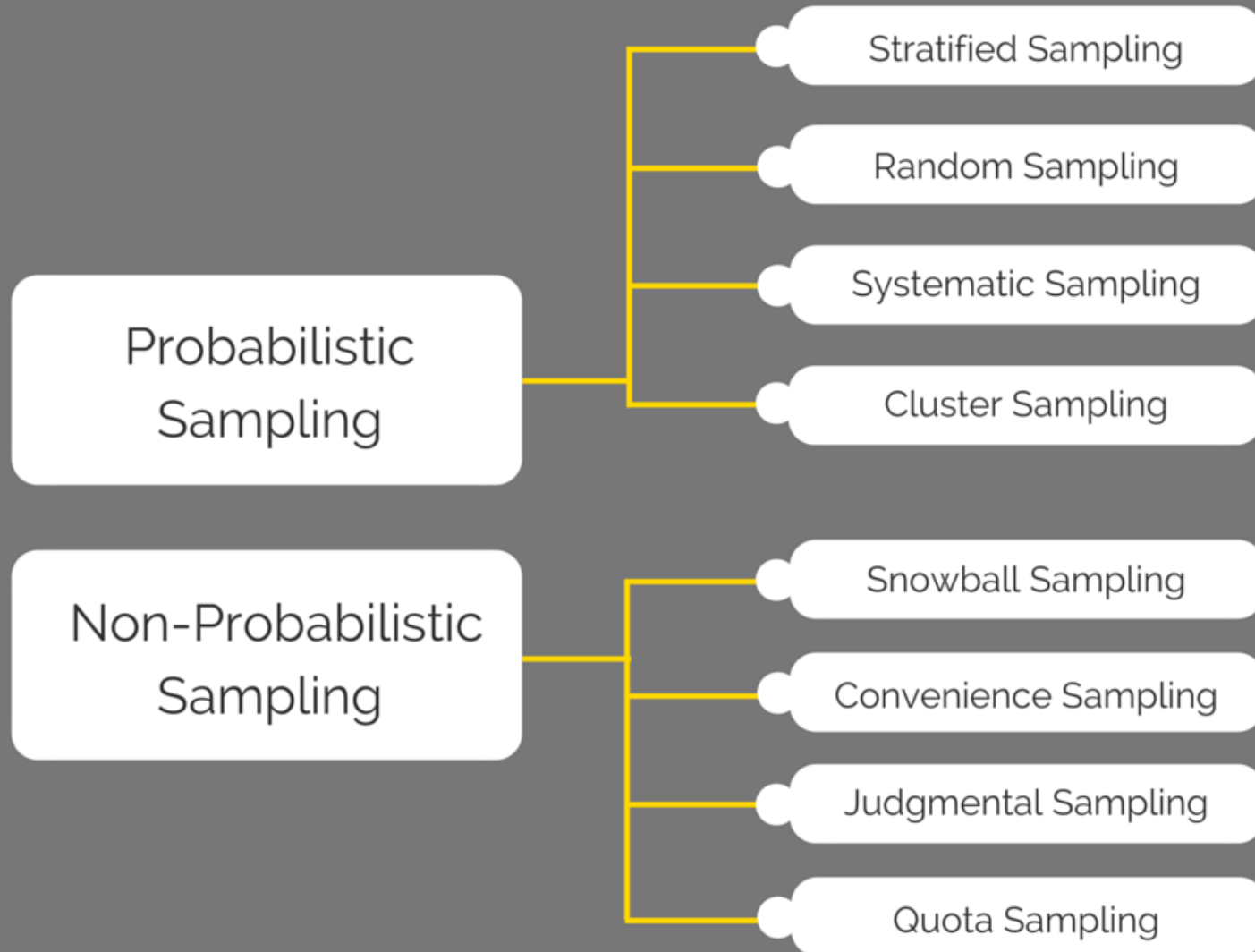
- The selection of the subjects is closely connected to the **generalization** of the results from the experiment.
- In order to generalize the results to the desired population, the selection must be representative for that population.
- The selection of subjects is also called a **sample from a population**.
- The size of the sample also impacts the results when generalizing.
  - **The larger the sample is, the lower the error becomes when generalizing the results.**

# Selection of subjects (2/6)



- The sampling of the population can be either a **probability** or a **non-probability** sample.
  
- In the probability sampling, the probability of selecting each subject is known
  - Simple random sampling
  - Stratified random sampling
  - Systematic sampling
  - Cluster Sampling
  
- In the non-probability sampling the probability of selection is unknown.
  - Convenience sampling
  - Quota sampling
  - Volunteer sampling
  - Snowball Sampling

# SAMPLING TECHNIQUES






# Selection of subjects (3/6)

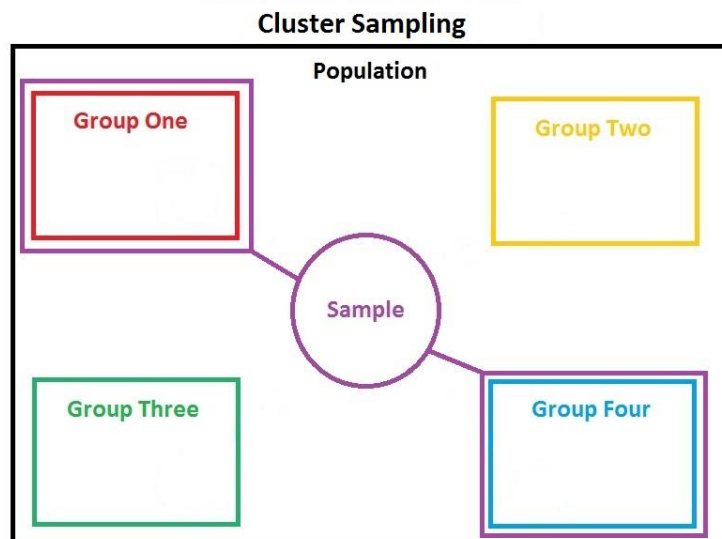


## Probability sampling techniques

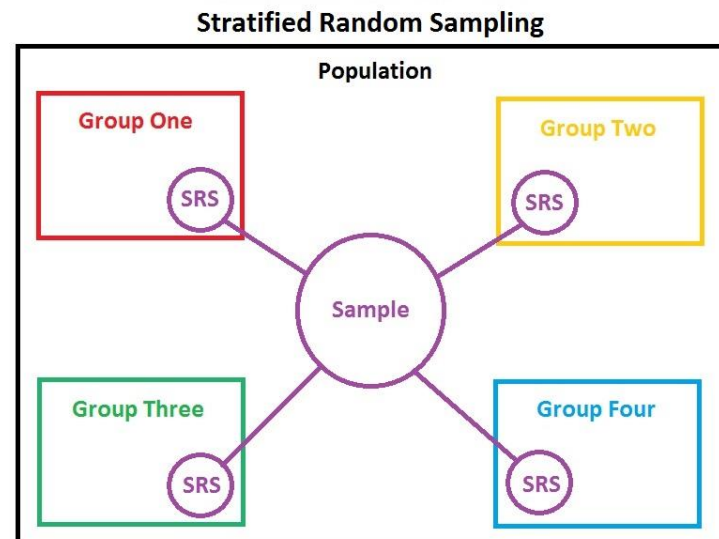
- Simple random sampling
    - Subjects are selected from a list of the population at random.
  
  - Stratified random sampling
    - The population is divided into a number of groups or strata with a known distribution between the groups.
    - Random sampling is then applied within the strata.
  
  - Systematic sampling
    - The first subject is selected from the list of the population at random and then every  $n^{\text{th}}$  person is selected from the list.
  
  - Cluster sampling
    - Researchers divide a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample.
- 

# Selection of subjects (4/6)

- The main difference between cluster sampling and stratified sampling is that in cluster sampling the cluster is treated as the sampling unit so sampling is done on a population of clusters (at least in the first stage).
- In stratified sampling, the sampling is done on elements within each stratum. In stratified sampling, a random sample is drawn from each of the strata, whereas in cluster sampling only the selected clusters are sampled.



[https://en.wikipedia.org/wiki/Cluster\\_sampling](https://en.wikipedia.org/wiki/Cluster_sampling)

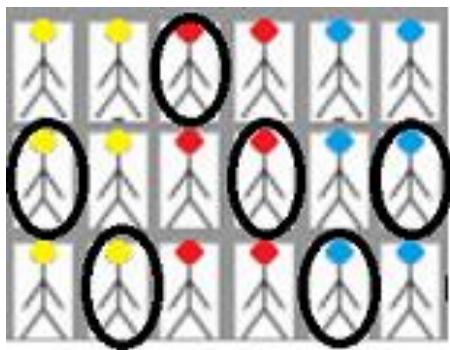


[https://en.wikipedia.org/wiki/Stratified\\_sampling](https://en.wikipedia.org/wiki/Stratified_sampling)

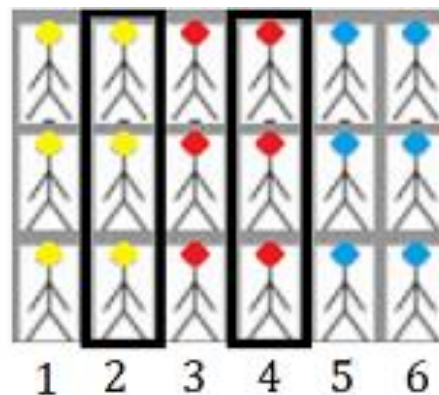
# Selection of subjects (5/6)

## Non-probability sampling techniques

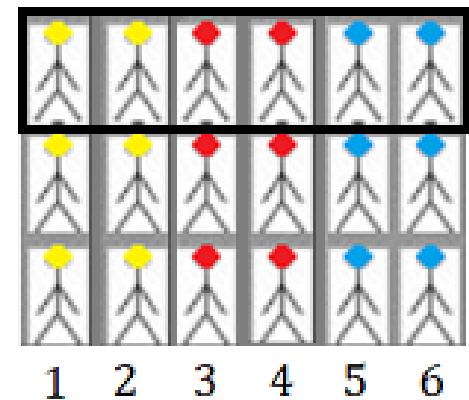
- Quota sampling
  - Divide the population into mutually exclusive subgroups
  - This type of sampling is used to get subjects from various elements of a population.
  - Convenience sampling is normally used for each element.
  - Quota sampling is somewhat similar to stratified sampling in that similar units are grouped together. However, it differs in how the units are selected. In probability sampling, the units are selected randomly while in quota sampling it is usually left up to the interviewer to decide who is sampled.



**Stratified Sampling**



**Cluster Sampling**

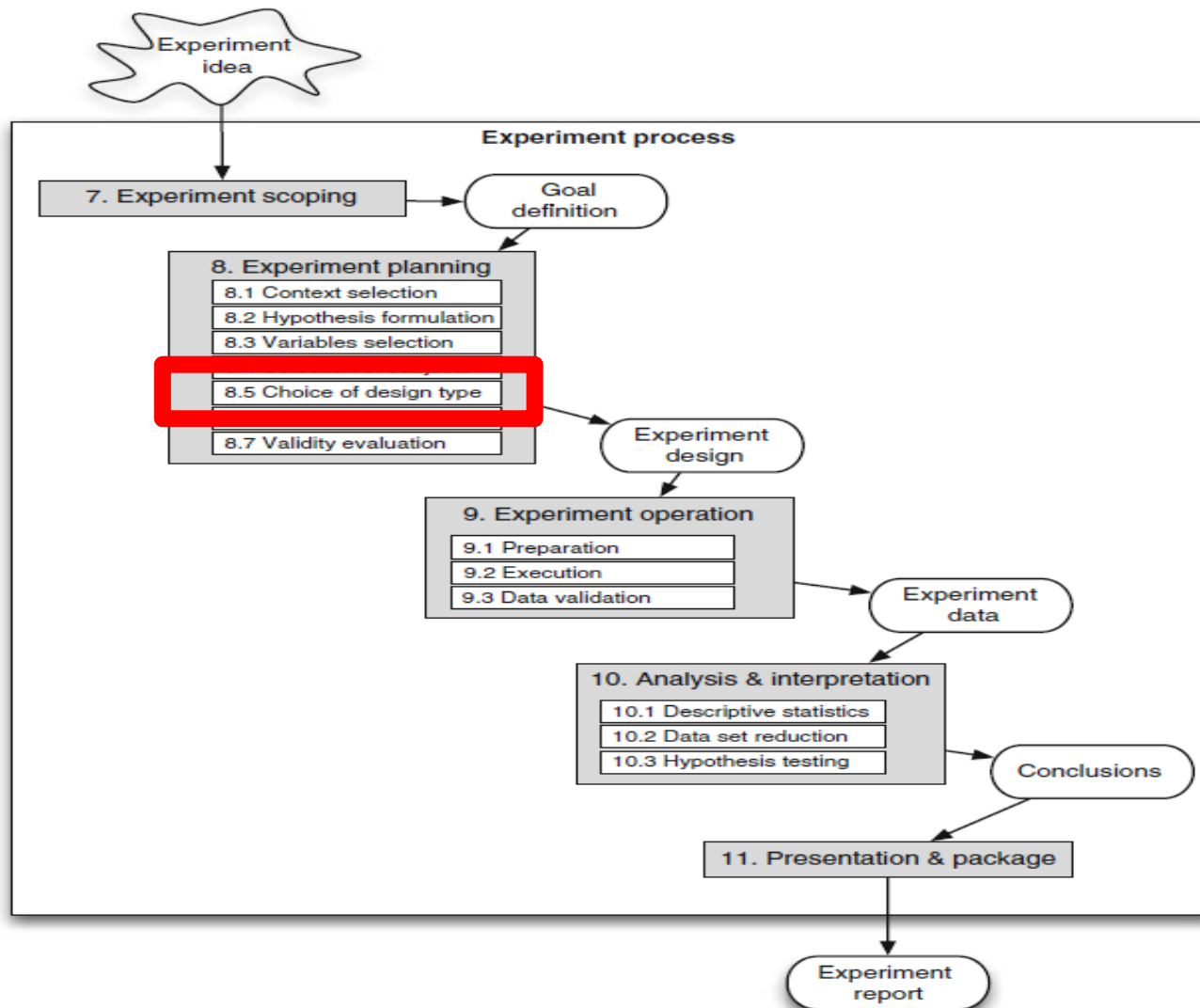


**Quota Sampling**

# Selection of subjects (6/6)


## Non-probability sampling techniques

- Convenience sampling
  - The nearest and most convenient persons are selected as subjects.
  - E.g., the first 100 customers to enter a department store.
- Judgement sampling
  - A sample is taken based on certain judgements about the overall population
- Snowball Sampling
  - Existing subjects provide referrals to recruit samples required for a research study.
- Volunteer sampling
  - People volunteer their services for the study



# Experiment design



- An experiment consists of a series of tests of the treatments
  - To get the most out of the experiment, the series of tests must be carefully planned and designed
  - A design of an experiment describes how the tests are organized and run.
  - To design the experiment, we have to look at the hypothesis to see which statistical analysis we have to perform to reject the null hypothesis.
  - During the design we determine how many tests the experiment shall have to make sure that the effect of the treatment is visible.
- 

# General design principles



- The general design principles are:
  - Randomization
  - Blocking
  - Balancing
  
- Consider the following example.

A company will conduct an experiment to investigate the effect on the reliability of a computer program when using object-oriented design instead of the standard company design principle. The experiment will use computer program A as the experiment object. The experiment design is of type “multi-test within object study” (i.e., examines a single object across a set of subjects.)

# Randomization

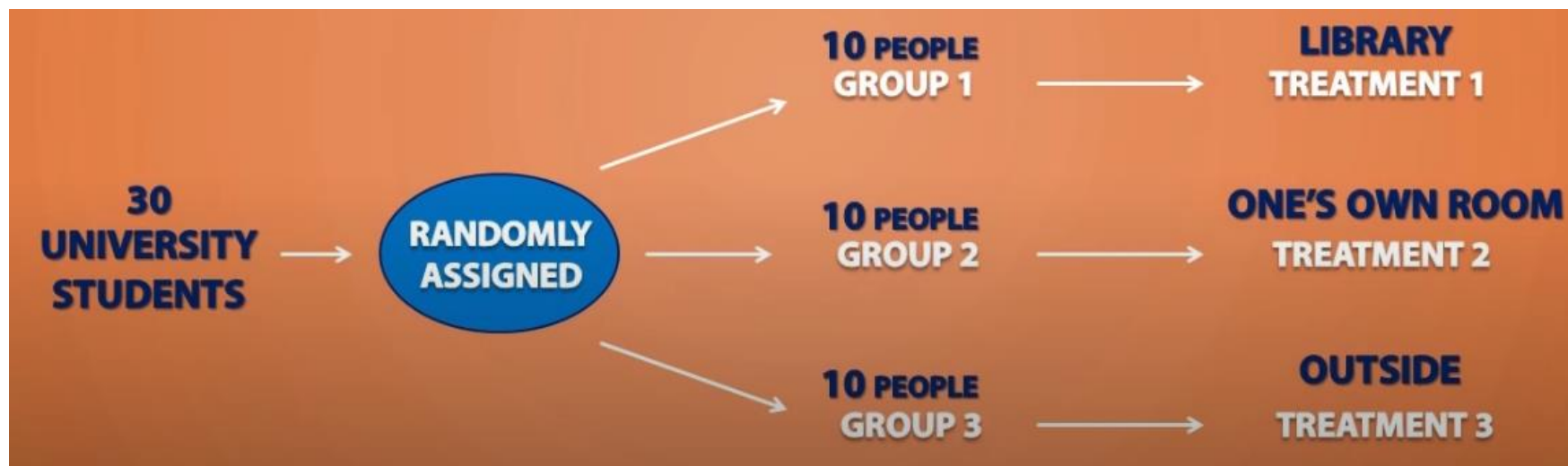


- The randomization applies on the allocation of the **objects**, **subjects** and in which **order** the tests are performed.
- All statistical methods used for analyzing the data, require that the observations be from independent random variables
- Example:
  - The selection of the persons (subjects) will be representative of the designers in the company, by random selection of the available designers.
  - The assignment of subjects to each treatment (object-oriented design or the standard company design principle) is selected randomly.



# Randomized design example

- Research objective is to determine which environment is best for studying: library, room or outside?
  - Sample: 30 university students



*Source: Simple Learning Pro*

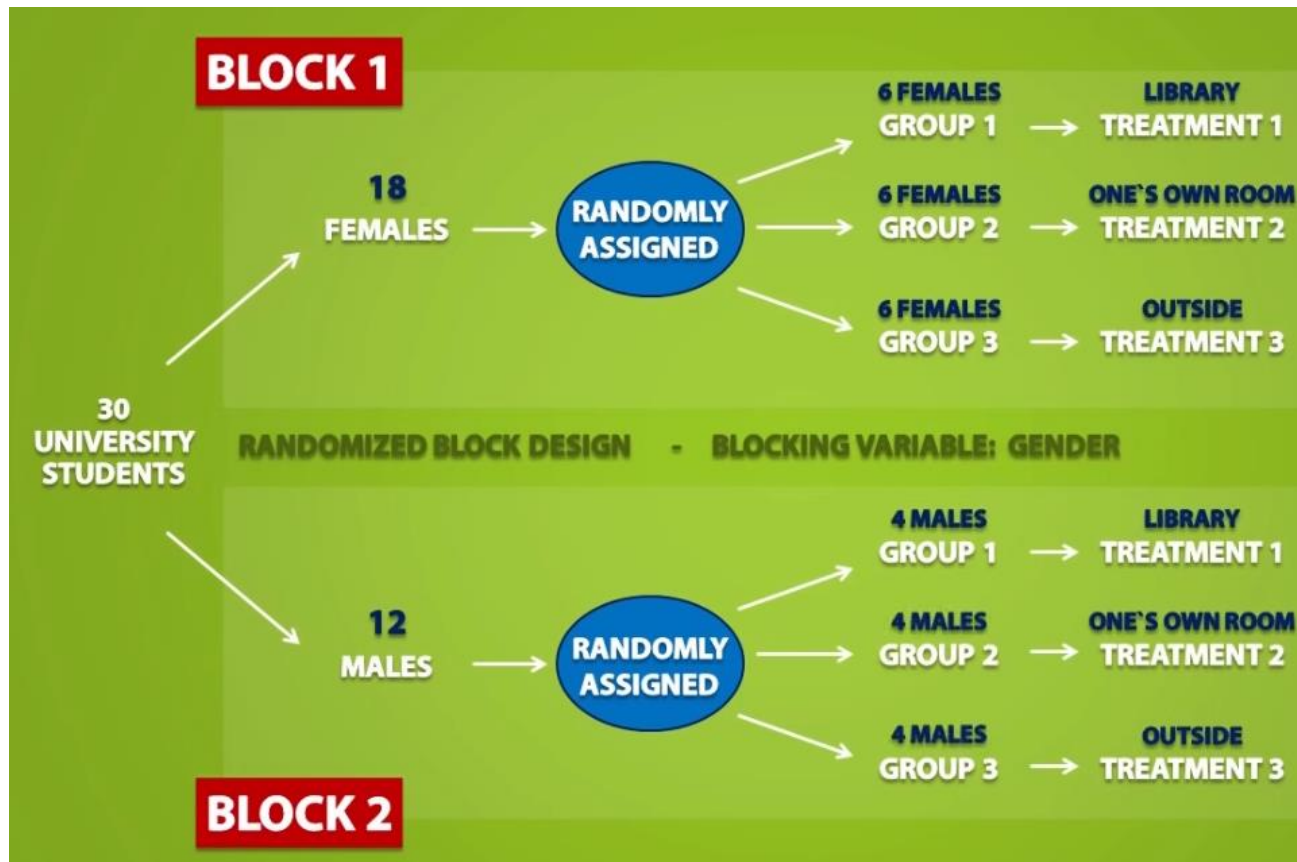
# Blocking



- Sometimes we have a factor that probably has an effect on the response, but we are not interested in that effect
- Blocking is used to systematically eliminate the undesired effect in the comparison among the treatments.
  - Within one block, the undesired effect is the same and we can study the effect of the treatments on that block.
- Example:
  - The persons (subjects) used, for this experiment, have different experience.
  - Some of them have used object-oriented design before and some have not.
  - To minimize the effect of the experience, the persons are grouped into two groups (blocks), one with experience of object-oriented design and one without.

# Blocking design example

- If the researcher believes that gender has an effect on the result, then use “Randomized Block Design”



Source: Simple Learning Pro


# Balancing



- If we assign the treatments (object-oriented design or the standard company design principle) so that each treatment has equal number of subjects, we have a balanced design.
- Balancing is desirable because it both simplifies and strengthens the statistical analysis of the data, but it is not necessary.
- Example:
  - The experiment uses a balanced design, which means that there is the same number of persons in each group (block).

# Standard design types



- One factor with two treatments
  - One factor with more than two treatments
  - Two factors with two treatments
  - More than two factors each with two treatments
- 

# One factor with two treatments

- In this design type, we want to compare the two treatments against each other. The most common is to compare the means of the dependent variable for each treatment. The following notation is used:

$\mu_i$     The mean of the dependent variable for treatment  $i$ .

$y_{ij}$     The  $j$ :th measure of the dependent variable for treatment  $i$ .

- Example of an experiment:
  - The aim is to investigate if a new design method produces software with higher quality than the previously used design method.
  - The factor in this experiment is the design method and the treatments are the new and the old design method.
  - The dependent variable can be the number of faults found in development.

# One factor with two treatments

- **Completely randomized design.** This is a basic experiment design for comparing two treatment means.
  - In this design, the experimenter randomly assigned subjects to one of two treatment conditions (e.g., new and old design method).
  - If we have the same number of subjects per treatment the design is balanced

Example of hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2, \mu_1 < \mu_2 \text{ or } \mu_1 > \mu_2$$

Subjects	Treatment 1	Treatment 2
1	X	
2		X
3		X
4	X	
5		X
6	X	

# One factor with two treatments

- **Paired comparison design.** In this design, each subject uses both treatments on the same object (crossover design).
  - This type of design has some challenges, i.e. to minimize the effect of the order, in which the subjects apply the treatments, the order is assigned randomly to each subject (1 or 2)
  - This design cannot be applied in every case of comparison as the subject can gain too much information from the first treatment to perform the experiment with the second treatment.
  - If we have the same number of subjects starting with the first treatment as with the second, we have a balanced design.

Example of hypothesis:

$d_j = y_{1j} - y_{2j}$  and  $\mu_d$  is the mean of the difference.

$H_0 : \mu_d = 0$

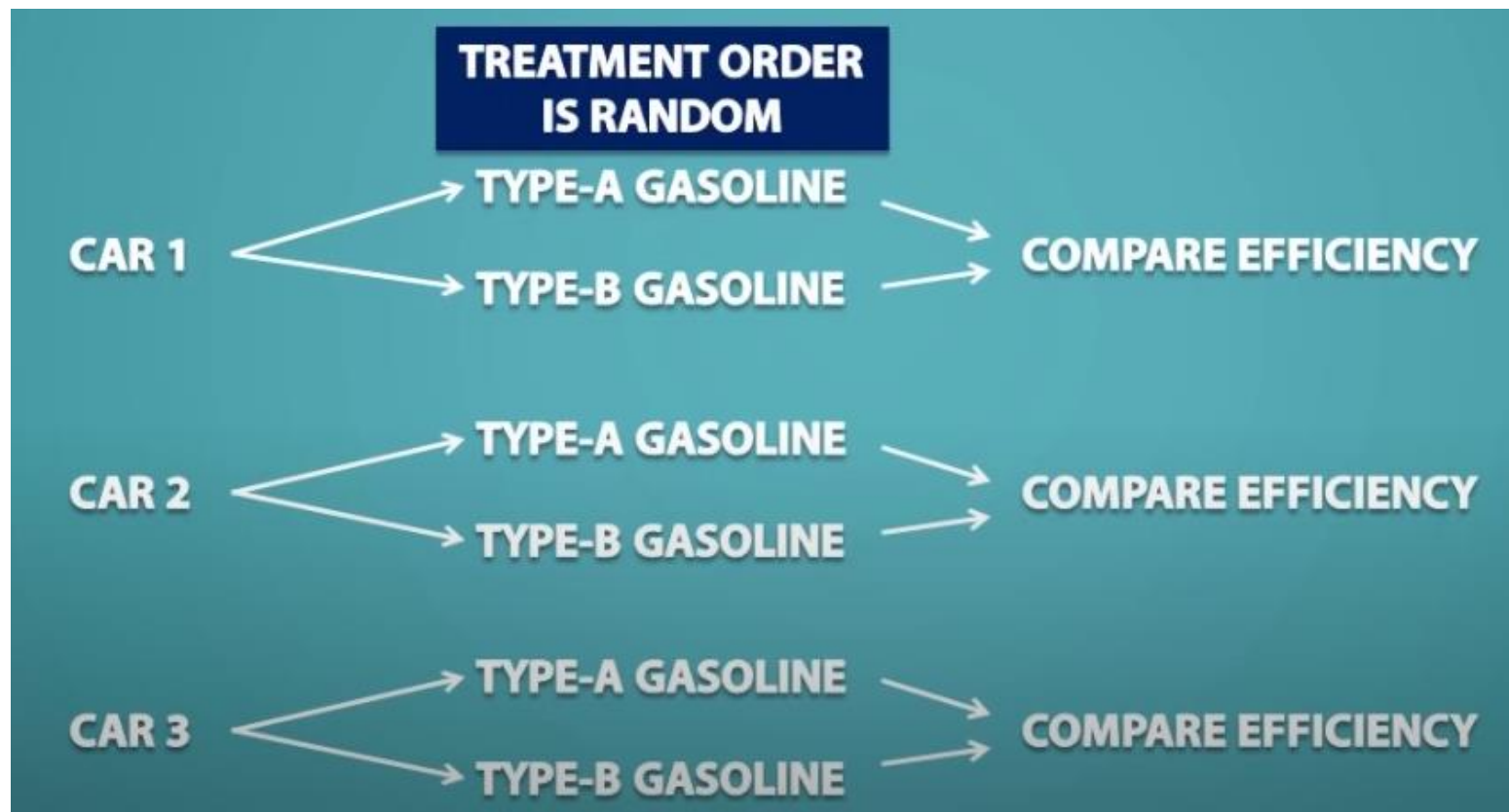
$H_1 : \mu_d \neq 0, \mu_d < 0 \text{ or } \mu_d > 0$

Subjects	Treatment 1	Treatment 2
1	2	1
2	1	2
3	2	1
4	2	1
5	1	2
6	1	2



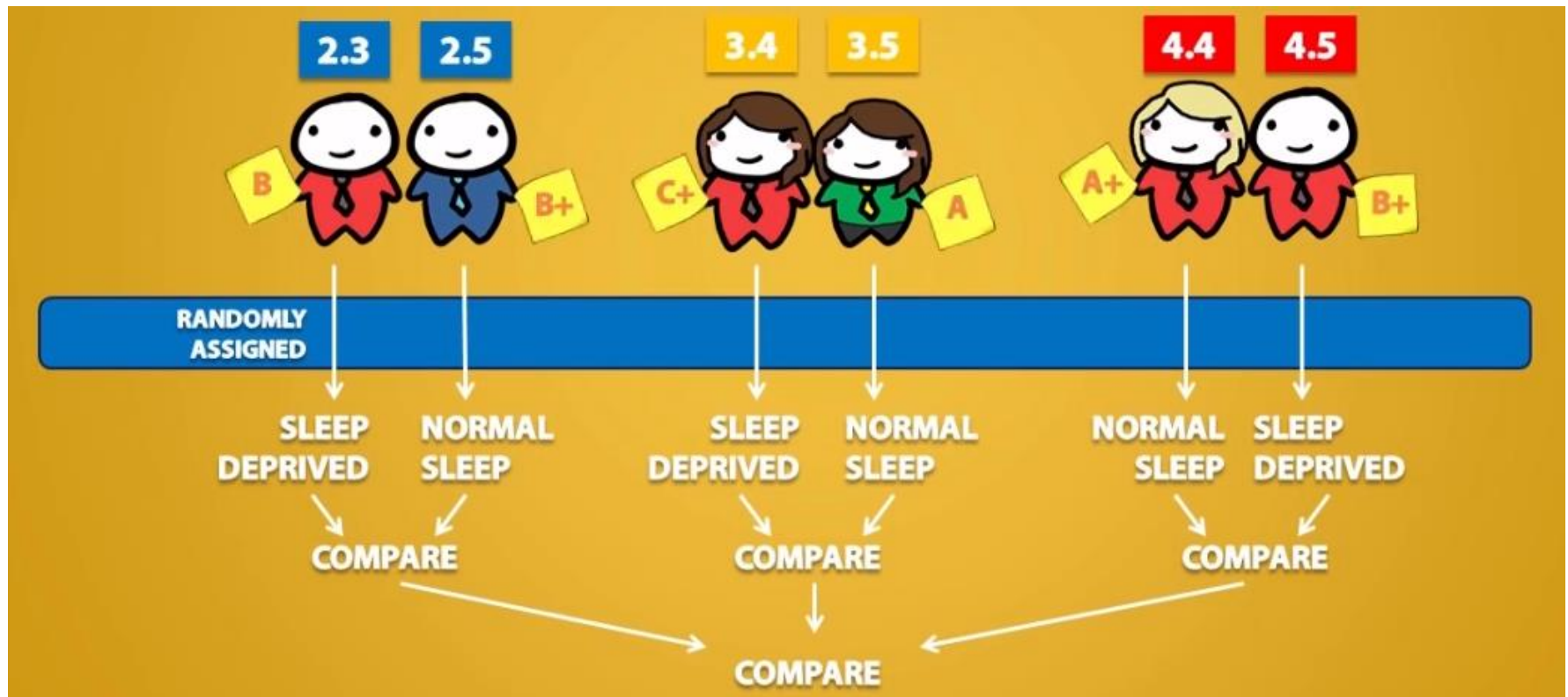
# Pair design example - same experimental unit

- Which type of gasoline is more efficient 91 or 95?
  - Use three cars in the experiment
  - Same experimental unit



# Pair design example - similar experimental unit

- Whether sleep deprivation has effect on test score:



Source: Simple Learning Pro

# One factor with more than two treatments



- Example of an experiment:
  - The experiment investigates the quality of the software when using different programming languages.
  - The factor in the experiment is the programming language and the treatments can be C, C++, and Java.

# One factor with more than two treatments

## ■ Completely randomized design

Example of hypothesis, where  $a$  is the number of subjects:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_a$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j)$$

Subjects	Treatment 1	Treatment 2	Treatment 3
1		X	
2			X
3	X		
4	X		
5		X	
6			X

## ■ Randomized complete block design

Example of hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_a$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j)$$

Subjects	Treatment 1	Treatment 2	Treatment 3
1	1	3	2
2	3	1	2
3	2	3	1
4	2	1	3
5	3	2	1
6	1	2	3

# Two factors with two treatments

- The experiment gets more complex when we increase from one factor to two.
- The single hypothesis for the experiments with one factor will split into three hypotheses:
  - one hypothesis for the effect from one of the factors,
  - one for the other factor
  - and one for the interaction between the two factors.
  - We use the following notations

$\tau_i$	The effect of treatment $i$ on factor A.
$\beta_j$	The effect of treatment $j$ on factor B.
$(\tau\beta)_{ij}$	The effect of the interaction between $\tau_i$ and $\beta_j$ .

# Two factors with two treatments

- 2\*2 factorial design
  - This design has two factors, each with two treatments.
  - In this experiment design, we randomly assign subjects to each combination of the treatments.
- Example

		Factor A	
		Treatment A1	Treatment A2
Factor B	Treatment B1	Subject 4, 6	Subject 1, 7
	Treatment B2	Subject 2, 3	Subject 5, 8

# Two factors with two treatments

## ■ Example of an experiment:

- The experiment investigates the understandability of the design document when using structured or object-oriented design based on one 'good' and one 'bad' requirements documents.
- The first factor, A, is the design method and the second factor, B, is the requirements document.
- The experiment design is a 2\*2 factorial design as both factors have two treatments, and every combination of the treatments are possible.

## Example of hypothesis:

$$H_0 : \tau_1 = \tau_2 = 0$$

$$H_1 : \text{at least one } \tau_i \neq 0$$

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{at least one } \beta_j \neq 0$$

$$H_0 : (\tau\beta)_{ij} = 0 \text{ for all } i, j$$

$$H_1 : \text{at least one } (\tau\beta)_{ij} \neq 0$$

H0 = the effect of treatment 1 and 2 on factor A is zero

H1 = at least one treatment of A is not zero

H0 = The effects of treatment 1 and 2 on factor B is zero

H1 = at least one treatment of B is not zero

H0 = Treatment 1 and 2 on factor A and B is zero

H1 = at least one treatment of either A or B is not zero

# Two factors with two treatments

## ■ Example

		<u>Factor A: Design methods</u>	
		Structured	Object Oriented
<u>Factor B: Req document:</u>	Good document	Subject 4, 6	Subject 1, 7
	Bad document	Subject 2, 3	Subject 5, 8

$H_0$  = The effect of treatment 1 (structured) and 2 (object oriented) on factor A (design methods) is zero

$H_1$  = at least one treatment of A is not zero

$H_0$  = The effect of treatment 1 (good document) and 2 (bad document) on factor B (reg document) is zero

$H_1$  = at least one treatment of B is not zero

$H_0$  = Treatment 1 and 2 on factor A and B is zero

$H_1$  = at least one treatment of either A or B is not zero



# More than two factors each with two treatments

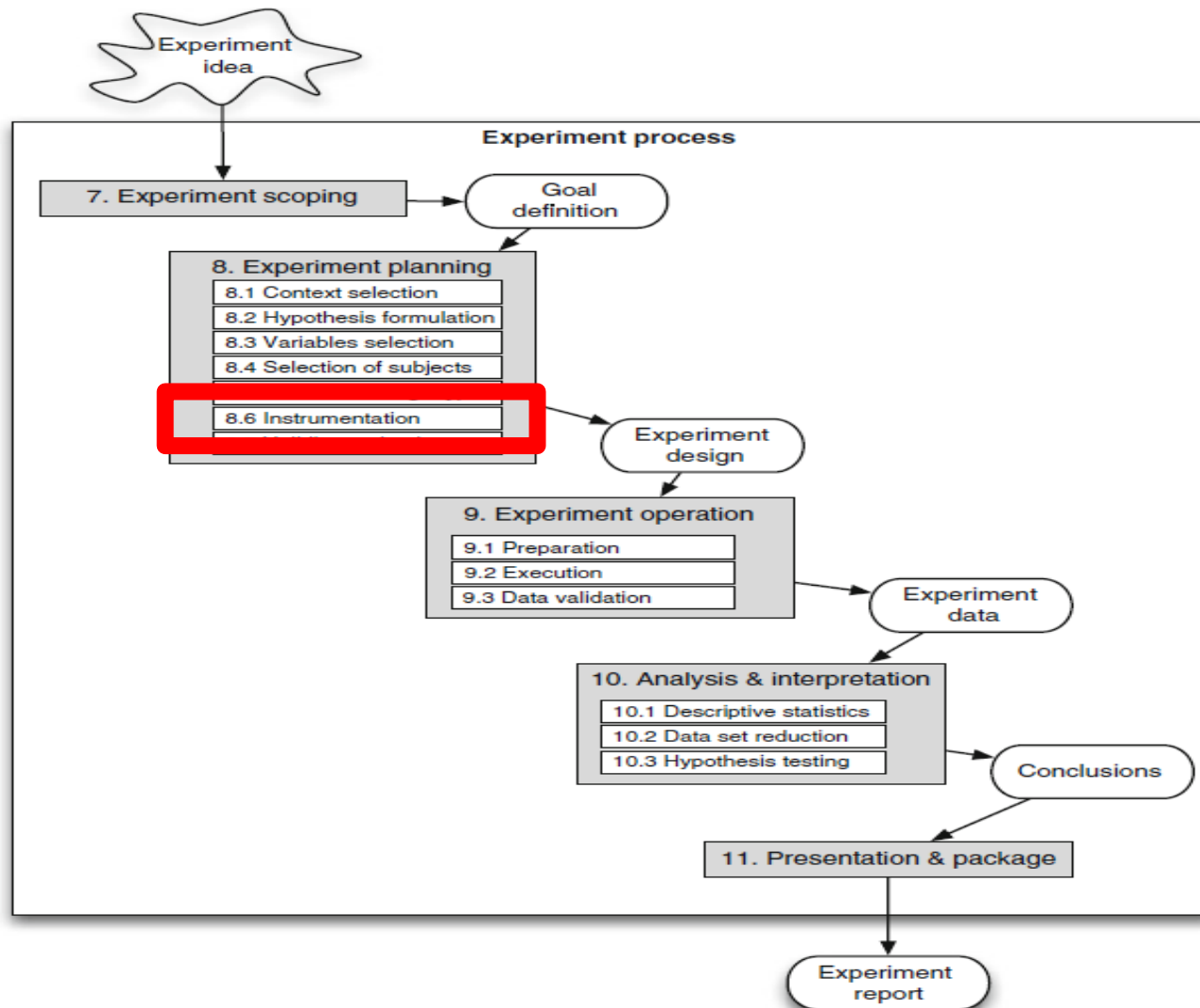


- $2^k$  *factorial design*
  - The  $2^k$  factorial design has k factors where each factor has two treatments.
  - This means that there are  $2^k$  different combinations of the treatments.
  - To evaluate the effects of the k factors, all combinations have to be tested.
  - The subjects are randomly assigned to the different combinations.

# More than two factors each with two treatments

- Example of a  $2^3$  factorial design

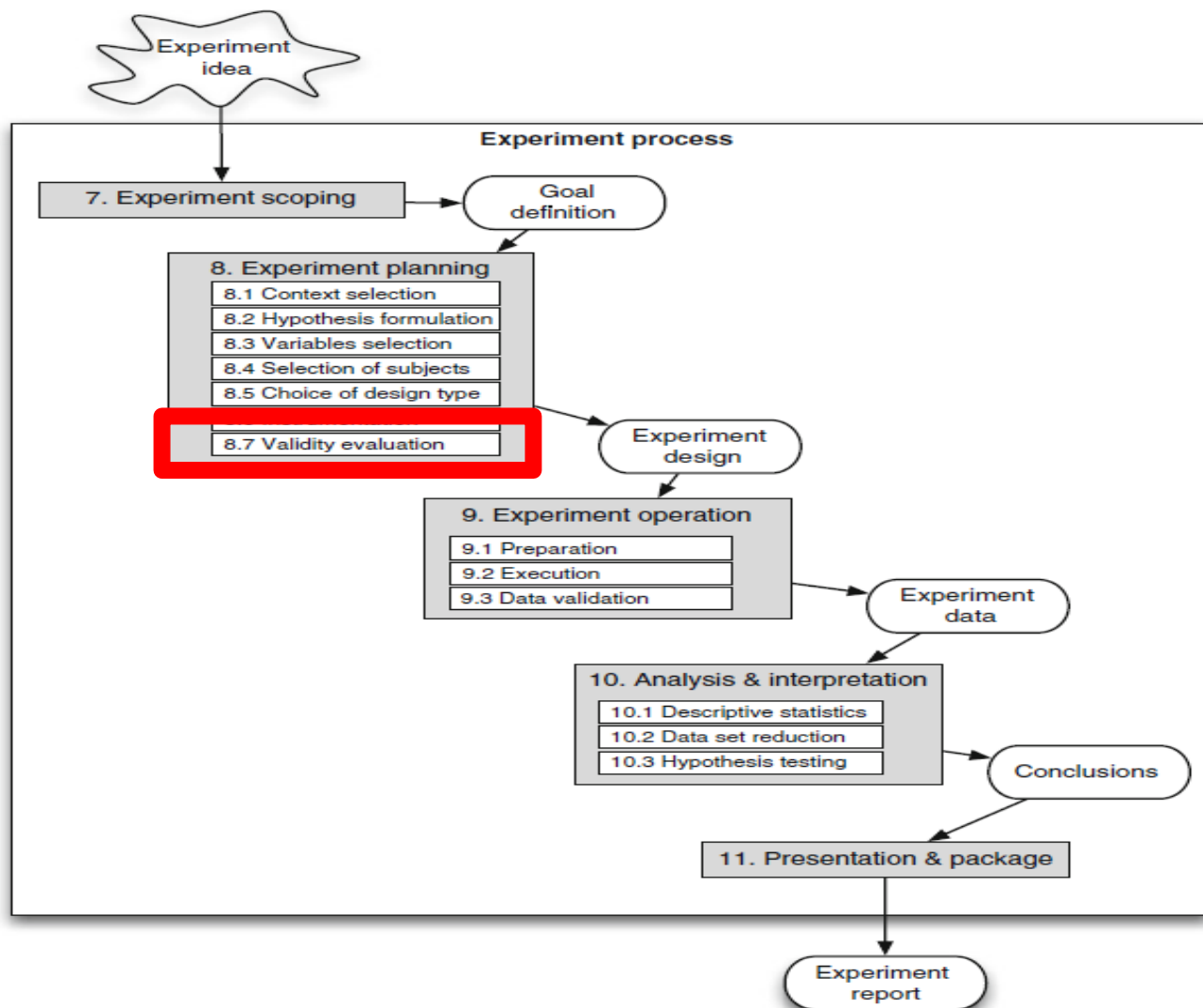
Factor A	Factor B	Factor C	Subjects
A1	B1	C1	2, 3
A2	B1	C1	1, 13
A1	B2	C1	5, 6
A2	B2	C1	10, 16
A1	B1	C2	7, 15
A2	B1	C2	8, 11
A1	B2	C2	4, 9
A2	B2	C2	12, 14



# Instrumentation



- The overall goal of the instrumentation is to provide means for performing the experiment and to monitor it
- If the instrumentation affects the outcome of the experiment, the results are invalid.
- The instruments for an experiment are of three types, namely **objects, guidelines** and **measurement instruments**
  - Experiment objects may be, for example, specification or software or code documents.
  - Guidelines are needed to guide the participants in the experiment., e.g. process descriptions and checklists.
  - Measurements in an experiment are conducted via data collection (interviews, forms)



# Validity evaluation



- A fundamental question concerning results from an experiment is how valid the results are.
- It is important to consider the question of validity already in the planning phase in order to plan for adequate validity of the experiment results.
- Adequate validity refers to that the results should be valid for the population of interest.

# Validity evaluation



- First of all, the results should be valid for the population from which the sample is drawn.
- Secondly, it may be of interest to generalize the results to a broader population.
- The results are said to have adequate validity if they are valid for the population to which we would like to generalize.

# Validity types



- Conclusion validity
- Internal validity
- Construct validity
- External validity



# Conclusion validity



- Conclusion validity is sometimes referred to as statistical conclusion validity.
- This validity is concerned with the relationship between the treatment (the effect of changing one or more independent variables) and the outcome of an experiment.
  - We want to make sure that there is a statistical relationship, i.e., with a given significance.
- Threats to conclusion validity are concerned with issues that affect the ability to draw the correct conclusion about relations between the treatment and the outcome of an experiment (reliability of measures and treatments etc)


# Conclusion validity causes



- Low statistical power: Is a threat when sample sizes are too small.
- Violated assumptions of statistical tests: Is a threat when the assumptions underlying statistical tests (e.g., normality) are not met.
- Failing to account for randomness in simulation studies (e.g., in network protocols or distributed systems simulations) can threaten conclusion validity, as results may not accurately represent the true behavior under different conditions.

# Internal validity



- Threats to internal validity are influences that can affect the independent variable with respect to causality, without the researcher's knowledge.
  - When the researcher is investigating whether one factor affects an investigated factor there is a risk that the investigated factor is also affected by a third factor.
  - Factors that impact on the internal validity are how the subjects are selected and divided into different classes, how the subjects are treated and compensated during the experiment, etc.
- 

# Internal validity



- In research, **internal validity** is the extent to which you are able to say that no other variables except the one you're studying caused the result.
  - For example, if we are studying the variable of pay and the result of hard work, we want to be able to say that no other reason (not personality, not motivation, not competition) causes the hard work. We want to say that pay and pay alone makes people work harder.

# Internal validity causes

- Maturation: Is a threat when an observed effect might be due to the respondents' gets more experienced.
- Testing: Is a threat when participants are tested more than once in the same way
  - The effect might be due to the number of times particular responses are measured, rather than due to the treatment.
- Instrumentation: Is a threat when instrumentation changes between pre-testing and post-testing
  - The effect might be due to a change in the measuring instrument and not to the treatment's differential impact at each time interval.
- Diffusion or imitation of treatment: Is a threat when various experimental (and control) groups can communicate with each other.
- When testing the impact of a new optimization in a compiler, failing to control for other changes in the software environment (e.g., updates to libraries or the operating system) can threaten internal validity, as it's unclear if observed performance improvements are due to the optimization or other changes.


# Construct validity



- **Construct validity** is the degree to which a test measures what it claims, or purports, to be measuring.
- Threats to construct validity refer to the extent to which the experiment setting actually reflects the construct under study.
  - For example, the number of courses taken at the university in computer science may be a poor measure of the **subject's expertise** in a programming language, i.e., has poor construct validity.
  - The number of years of practical use of programming language may be a better measure of **subject's expertise** in a programming language, i.e., has better construct validity.

# Construct validity causes



- Hypothesis-guessing within experimental conditions: Is a threat when participants guess the hypothesis (incorrectly or correctly) and then act on the basis of what they perceive the experimental hypothesis to be.
  - Using lines of code as the sole measure of software complexity in a study investigating the relationship between complexity and bug frequency. This operational definition may not fully capture the multidimensional construct of complexity.
  - In a study on the impact of response time on user satisfaction with a web application, if user satisfaction is measured only through the speed of page loading, this may not fully encapsulate the construct of satisfaction, which could also involve factors like usability and content relevance.
- 

# External validity



- Threats to external validity concern the ability to generalize experiment results outside the experiment setting.
- External validity is affected by the experiment design chosen, but also by the objects in the experiment and the subjects chosen.
- There are three main risks
  - Having wrong participants as subjects
  - Conducting the experiment in the wrong environment
  - Performing it with a timing that affects the results



# External validity causes

- Interaction of selection and treatment: Is a threat when the treatment only works with the people used in the study.
  - To evaluate whether this is the case, you should ask yourself, can the results be generalized to other people beyond social class, race, age, geography, sex, or personality groups?
- Interaction of setting and treatment: Is a threat when the treatment only works in the setting used in the study.
  - To evaluate whether this is the case, you should ask yourself, can the results be generalized to other settings such as military camps, university campuses, offices, factories, malls etc.?
- Interaction of history and treatment: Is a threat when the treatment only works for a group of people with particular experiences.
  - To evaluate whether this is the case, you should ask yourself, can the results be generalized across time? e.g., does it matter if the day that participants were tested was another day?

Population Generalization in Human-Computer Interaction (HCI): An HCI study that recruits participants solely from a computer science undergraduate program may not produce results that are generalizable to a broader range of users with different backgrounds and levels of technical expertise.

# SLR on threats to validity in software engineering secondary studies

**Table 6**

Most Common Threats to Validity.

Threats to validity	Count	Percentage
Study inclusion/exclusion bias	100	17,4%
Construction of the search string	92	16,0%
Data extraction bias	91	15,8%
Selection of DLs	70	12,2%
Researcher bias	40	7,0%
Robustness of initial classification	35	6,1%
Generalizability	27	4,7%
Publication bias	24	4,2%
Repeatability	23	4,0%
Validity of primary studies	13	2,3%
Quality assessment subjectivity	13	2,3%
Coverage of research questions	13	2,3%
Results not applicable to other organizations/domains	12	2,1%
Selection of publication venues	12	2,1%
Search engine inefficiencies	10	1,7%

Ampatzoglou, A., et al., *Identifying, categorizing and mitigating threats to validity in software engineering secondary studies*. Information and Software Technology, 2019. **106**: p. 201-230.

# SLR on threats to validity in software engineering secondary studies

**Table 7**  
Most Common Mitigation Actions.

<p><b>Threat to Validity:</b> Study inclusion/exclusion bias Discussion of marginal cases (44, 6.9%) Definition of inclusion / exclusion criteria in a protocol (22, 3.5%) Revision of inclusion / exclusion criteria (16, 2.5%) Employment of a third opinion for marginal cases (10, 1.6%) Employment of a systematic voting approach (8, 1.3%) Cross-checking of paper selection (6, 0.9%) No mitigation (8, 1.3%)<sup>a</sup> Use of random paper screening (5, 0.8%) Execution of a consensus meetings (2, 0.3%) 32 other actions encountered once (5%)</p> <p><b>Threat to Validity:</b> Data extraction bias Discussion among authors (30, 4.7%) Involvement of more researchers / Work in pairs (15, 2.4%) Use of a data extraction form (12, 1.9%) Cross-checking of data extraction (11, 1.7%) Use of random paper screening (11, 1.7%) Execution of pilot data extraction (10, 1.6%) No mitigation (9, 1.4%) Employment of a third opinion for conflicting data items (8, 1.3%) Definition of a review protocol (5, 0.8%) Use of Codes (3, 0.5%) Ensure the conformance to guidelines (3, 0.5%)</p> <p><b>Threat to Validity:</b> Robustness of initial classification Use an existing classification scheme (15, 2.4%) Discussion among authors (7, 1.1%) Employment of a third opinion for the classification (5, 0.8%) No mitigation (4, 0.6%) Application of keywording of abstracts (4, 0.6%)</p> <p><b>Threat to Validity:</b> Repeatability Development of a review protocol (8, 1.3%) Ensure the conformance to well-established guidelines (7, 1.1%) Documentation of the search process (5, 0.8%) Involvement of more than one researcher in the process (3, 0.5%) Documentation of inclusion/exclusion criteria (2, 0.3%) Documentation of the review process (3, 0.5%) Ensure the public availability of data (2, 0.3%) No mitigation (2, 0.3%)</p> <p><b>Threat to Validity:</b> Generalizability No mitigation (14, 2.2%) Use of broad time and publication coverage (4, 0.6%) Comparison to other studies (3, 0.5%) Use both academic and industrial papers (2, 0.3%)</p>	<p><b>Threat to Validity:</b> Construction of the search string Employment of snowballing (27, 4.3%) Inclusion of synonyms/roots (20, 3.2%) Use of a gold standard (11, 1.7%) Systematic search string construction (13, 2.1%) Constant search string refinement (10, 1.6%) Extension of search scope / Broad terms (10, 1.6%) Execution of pilot searches (9, 1.4%) No mitigation (9, 1.4%) Use from previous studies (4, 0.6%) Use of author and citation analysis (3, 0.5%)</p> <p><b>Threat to Validity:</b> Selection of DLs Inclusion of most known DLs (35, 5.5%) Use search engines and indexes (14, 2.2%) Employment of snowballing (13, 2.1%) Inclusion of specific venues (10, 1.6%) No mitigation (7, 1.1%) Use of expert opinion (2, 0.3%) Consideration of a large time period (2, 0.3%) Inclusion of grey literature (2, 0.3%) Ensure the conformance to guidelines (2, 0.3%)</p> <p><b>Threat to Validity:</b> Researcher bias Discussion among authors (16, 2.5%) No mitigation (13, 2.1%) Execution of pilot data analysis (6, 0.9%) Use of reliability checks (4, 0.6%) Development protocol (3, 0.5%) Comparison with existing studies (3, 0.5%)</p> <p><b>Threat to Validity:</b> Publication bias No mitigation (14, 2.2%) Inclusion of grey literature (5, 0.8%) Use of broad time and publication coverage (3, 0.5%) Scanning of selected venues (2, 0.3%) Use of expert opinion (2, 0.3%) 5 other actions encountered once (0.8%)</p>
---	--

<sup>a</sup> This refers to cases when a threat to validity is reported in a secondary study, but no mitigation action is referenced to resolve it.

Ampatzoglou, A., et al., *Identifying, categorizing and mitigating threats to validity in software engineering secondary studies*. Information and Software Technology, 2019. **106**: p. 201-230.

# SLR on threats to validity in software engineering secondary studies

**Table 8**

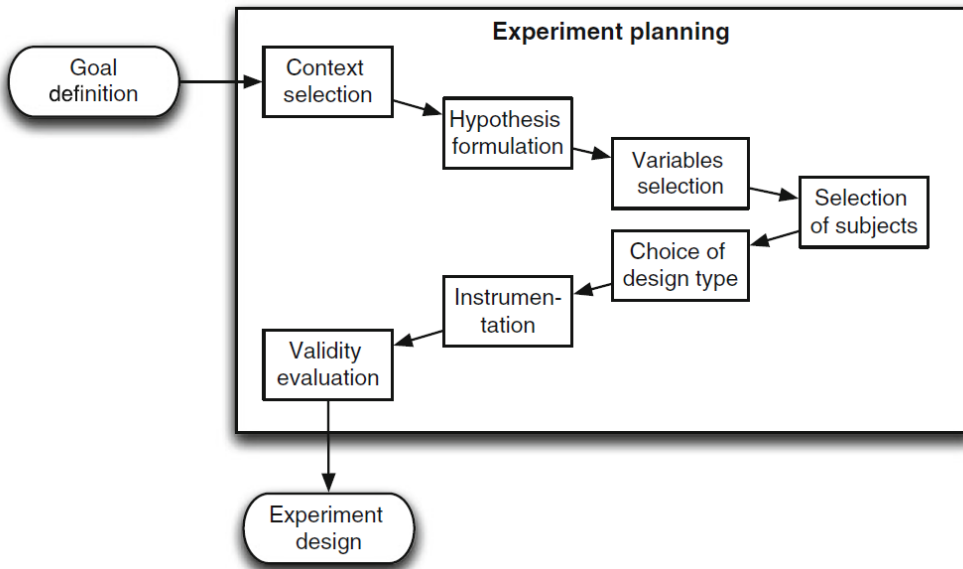
Explicit Categories of Threats to Validity.

Explicit Categories	Count
Not defined	329
Construct	54
Internal	51
External	37
Reliability	24
Conclusion	23
Primary study identification	9
Generalization	7
Data extraction	7
Theoretical	5
Objectivity	5
Publication bias	5
Interpretive Validity	3

Ampatzoglou, A., et al., *Identifying, categorizing and mitigating threats to validity in software engineering secondary studies*. Information and Software Technology, 2019. **106**: p. 201-230.

# Example experiment

- This section is a continuation of the example introduced previously



The goal:

Analyze the **PBR** and **CBR** 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.

# Example experiment




## ■ Context selection

- An off-line experiment will be run with a mixture of M.Sc. and Ph.D. students.
- Two 'toy' requirements documents from a lab package will be used.
- The experiment can be considered as general in the sense that the objective is to compare two reading techniques in general (from a research perspective).
- The main challenge in the specific case is that the participants know the existing technique very well, while a new technique must be taught to them.
  - Thus, the new technique may have a disadvantage since it is not as well known.
  - The potential biases in favor of one or the other technique must be taken into consideration by the researcher.

# Example experiment (*Cont'd*)



- We would like to compare both effectiveness and efficiency when it comes to detecting faults when using two different reading techniques when conducting the inspection.
    - The first method is Perspective-based Reading (PBR) and the second method is Checklist based Reading (CBR).
    - PBR is based on the reviewers having different perspectives when performing the inspection
    - CBR is based on having a checklist for different items that are likely to relate to faults in requirements documents
  
  - Effectiveness refers to the number of faults found out of the total number of faults, while efficiency also includes time, i.e., whether more faults are found per time unit.
    - The number of faults is assumed to be known.
- 

# Example experiment (*Cont'd*)

## ■ Hypothesis formulation

### - Effectiveness:

- $H_0$ : The number of faults found using PBR and CBR are equal.
- $H_1$ : The number of faults found using PBR and CBR are not equal.

$$H_0 : \mu_{NPBR} = \mu_{NCBR}$$
$$H_1 : \mu_{NPBR} <> \mu_{NCBR}$$

### - Efficiency:

- $H_0$ : The number of faults found per time unit using PBR and CBR are equal.
- $H_1$ : The number of faults found per time unit using PBR and CBR are not equal.

$$H_0 : \mu_{NtPBR} = \mu_{NtCBR}$$
$$H_1 : \mu_{NtPBR} <> \mu_{NtCBR}$$



# Example experiment (*Cont'd*)



## ■ Variables selection

- The **independent variable** is the reading technique, and it has two levels (treatments): PBR and CBR, respectively.
- The **dependent variables** are the number of faults found and the number of faults found per time unit.
- This means that we must ensure that the subjects can clearly mark faults found so that the researcher can compare the faults marked with the known set of faults.
- Furthermore, we must ensure that the subjects can keep track of time and fill in the time when a specific fault was found.

# Example experiment (*Cont'd*)



## ■ Selection of subjects

- Preferably it would be possible to find subjects for the experiment by random. However, in most experiments the researcher tends to be forced to use subjects that are available.
- Often students participating in courses at the university become the subjects in experiments run at the university, which is the case in this example experiment.
- It is important to characterize the selected subjects to help assessing the external validity of the study.

# Example experiment (*Cont'd*)



## ■ Choice of **design type**

- A good approach is often to use a pre-test to try to capture the experience of the subjects and based on the outcome of the pre-test divide the subjects into experience groups from which we randomly select subjects to the groups in the experiment (i.e., experienced group and non-experienced group).
- The experiment includes one factor of primary interest (reading technique) with two treatments (PBR and CBR, respectively), and a second factor that is not really of interest in the experiment (requirements document).
- The natural design is a completely **randomized design** where each groups (i.e., experienced and non-experienced) first uses either PBR or CBR on one of the requirements document and then uses the other reading technique on the other requirements document.

# Example experiment (*Cont'd*)

## ■ Choice of **design type**

- However, decisions have to be taken in order too. We have two options:
  - (1) have both groups use different reading techniques on one of the requirements documents first and then switch reading techniques when inspecting the other requirements document
  - (2) have both groups use the same reading technique on different requirements documents.

In either case, there is an ordering issue.

- In the first case, one of the requirements document will be used before the other and in the second case one reading technique will be used before the other.

# Example experiment (Cont'd)

- In both options, there is an ordering issue

## Option-1

	PBR	CBR
Group 1: (Req document 1)	X	
Group 2: (Req document 1)		X
Group 1: (Req document 2)		X
Group 2: (Req document 2)	X	

The primary interest is in the **difference between reading techniques** and not any differences between the two requirements documents

## Option-2

	PBR	CBR
Group 1: (Req document 1)	X	
Group 2: (Req document 2)	X	
Group 1: (Req document 2)		X
Group 2: (Req document 1)		X

## Or... Option-3

Allow one group to use PBR on a requirements document and the other group use CBR on the same document  
*(The advantage would be that a larger requirements document could be used in the same time frame. The downside is that only half as many data points are generated.)*

# Example experiment (*Cont'd*)



## ■ Instrumentation

- Given that the experiment is based on a lab package, the requirements documents are already available and hence also a list of detected faults to be able to determine the effectiveness of the reading technique.
- The **guidelines** for the two reading techniques must be developed or reused from elsewhere.
  - Here it is important to ensure a fair comparison, as mentioned above, by providing comparable support for the two methods.
- **Forms** for filling out faults found must be developed or reused from another experiment.
  - It is crucial to ensure traceability between the requirements document and the form, for example by numbering the faults in the requirements document while capturing the information about the fault in the form.

# Example experiment (*Cont'd*)



- Validity evaluation
  - What are the threats to validity?

# Reference



## Chapter 8