Summary Advanced Research Methods

written by

thomasalf

The Marketplace to Buy and Sell your Study Material

On Stuvia you will find the most extensive lecture summaries written by your fellow students. Avoid resits and get better grades with material written specifically for your studies.

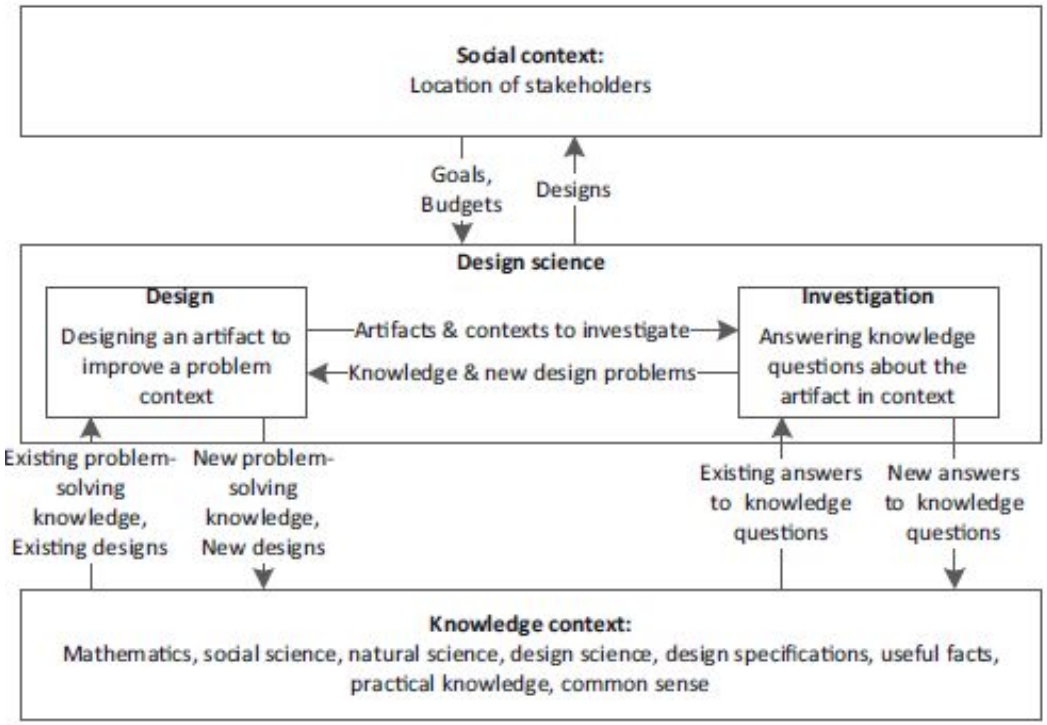
www.stuvia.com

Advanced Research Methods: Literature Summary

Thomas Alflen 2018

Wieringa: Chapter 1 - Design Science

- **Design science** is the design and investigation of artifacts in context. The artifacts we study are designed to interact with a problem context in order to improve something in that context. E.g. the investigation of a software component/system within software (artifact vs. context). Even conceptual tools for the mind can be artifacts. The interaction between the artifact and the context contributes to solving a problem.
- Two kinds of research problems appear in design science: **Design Problems (DP)** and **knowledge questions (KQ).**
    - DP call for a change in the real world and require an analysis of actual or hypothetical stakeholder goals. A solution is a design, and there are usually **many** different solutions. Depends on stakeholder goals.
    - KQ do not call for a change in the world, but ask for knowledge about the world as-is. The answer is a proposition and we assume there is **one** answer. We might give the wrong answer. **Fallibilism** means that we can never be sure that we have actually found the answer to an empirical KQ.
    - DP follow the **design cycle,** while KQ follows the **empirical cycle.**
    - We can start from a DP by asking KQ's about the artifact, context, and the interaction between the two. *Answering a KQ can lead to a new DP.*
- The framework for design science is as follows:



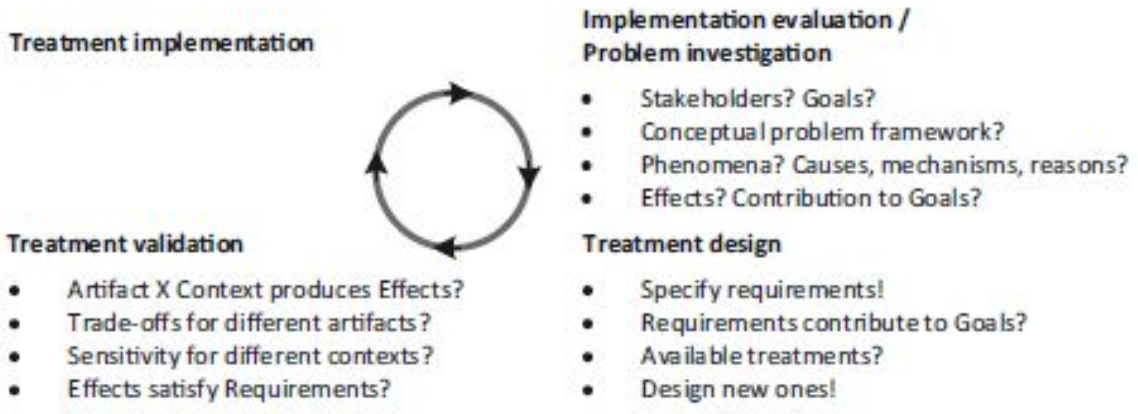
- Knowledge available prior to the project is called **prior knowledge** vs. **posterior knowledge**.
- You also have **basic sciences** (physics, etc.), **special sciences** (about the earth), **applied sciences** (astronomy, geology, etc.), and **case research** (engineering, consultancy, etc.).

Wieringa: Chapter 2 - Research Goals and RQ's

- You have multiple goals in design science:
  - **instrument design goal → knowledge goal → prediction goal → artifact design goal → social context goals (external stakeholder goals)**
- A **design problem** is a problem to (re)design an artifact so that it better contributes to the achievement of some goal. Improve <*a problem context*> by <*(re)designing an artifact*> that satisfies <*some requirements*> in order to <*help stakeholders achieve some goals*>. The requirements are also the interaction as previously mentioned.
- This book focuses on **empirical knowledge questions** (requires the data about the world to answer them), in contrast to **analytical knowledge questions** (answered by conceptual analysis, mathematics).
  - An important classification of KQ's is by their goal: **descriptive** or **explanatory** questions. Also, "what" vs. "why" (not always); both can be open or closed. I'd remember this by simply stating that the first are more easy questions compared to the explanatory questions.
  - There are four important KQ types about designs:
    - **Effect questions** (artifact vs. context) produce effects?
    - **Trade-off questions** (alternative artifact vs. context) produce effects?
    - **Sensitivity questions** (artifact vs. alternative context) produce effects?
    - **Requirements satisfaction questions** Do effects satisfy requirements?

Wieringa: Chapter 3 - The Design Cycle

- The **engineering cycle** is a rational problem-solving process; next figure shows it:



- Problem investigation → Treatment design → Treatment validation → Treatment implementation → Implementation evaluation
  - "?" = Knowledge questions; "!" = Design problems.
- We define an **implementation** of a treatment as the application of the treatment to the original problem context.

- Design science projects are **always** restricted to the first three tasks of the engineering cycle; these are also called the **design cycle** (so this is a shorter version of the E.C.)
- The goal of **validation** is to predict how an artifact will interact with its context, without actually observing an implemented artifact in a real-world context.
- The goal of **evaluation** research, by contrast, is to investigate how implemented artifacts interact with their real-world context.

Wieringa: Chapter 4 - Stakeholder and Goal Analysis

- A **stakeholder** of a problem is a person, group of persons, or institution affected by treating the problem. This chapter falls under the rightmost dot on the picture of the E.C. above. The list of possible stakeholders is listed below:

Table 4.1 List of possible stakeholders of an artifact, based on the list given by Ian Alexander [1]

The system under development (SUD) consists of the artifact and these stakeholders interacting with the artifact

- **Normal operators** give routine commands to the artifact, sometimes called "end users"
- **Maintenance operators** interact with the system to keep it running
- **Operational support** staff support normal operators in their use of the system and help to keep the system operational

Stakeholders in the immediate environment of the SUD, interacting with the SUD

- **Functional beneficiaries** benefit from the output produced by the system, sometimes called "users" of the artifact
- Stakeholders responsible for **interfacing systems** have an interest in the requirements and scope of the artifact

Stakeholders in the wider environment of the SUD

- A **financial beneficiary** benefits from the system financially, such as a shareholder or director of the company that will manufacture the artifact
- A **political beneficiary** benefits from the system in terms of status, power, influence, etc.
- A **negative stakeholder** would be worse off when the artifact is introduced in the problem context
- A **threat agent** is a stakeholder who wants to hurt the system, e.g., by compromising its integrity or stealing confidential information from it

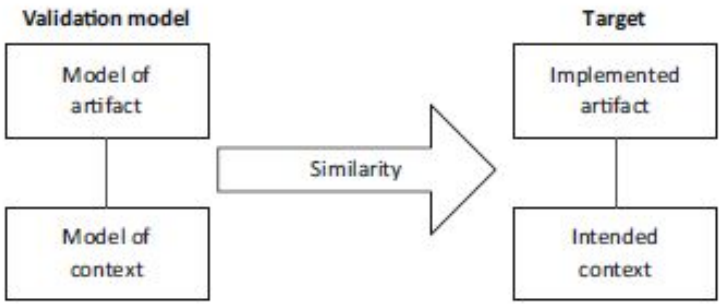
Stakeholders involved in the development of the SUD

- The **sponsor** initiates and provides a budget for developing the artifact. Important source of goals and requirements for the artifact
- The **purchaser** is, in this taxonomy, a stakeholder responsible for terminating development successfully. The purchaser could be a project manager or a product manager responsible for a wider range of projects all related to one product
- **Developers** such as requirements engineers, designers, programmers, and testers build the system. They are not normal operators of the system and do not benefit from its output during normal operation
- **Consultants** support development of the artifact
- **Suppliers** deliver components of the artifact

- We define a stakeholder **goal** as a **desire** for which the stakeholder has committed resources. The stakeholder is willing to achieve this goal and has committed money and/or time to achieve it. **Anything** can be the object of desire; desires can be in conflict, important ones are:
    - Two desires are in **logical conflict** if it is logically impossible to realize them both. E.g. desire to spend money conflicts with the desire to keep it.
    - Two desires are in **physical conflict** if it would violate the laws of nature to satisfy them both. E.g. desire to eat more is in conflict with the desire to lose weight.
    - Two desires are in **technical conflict** if it would be physically possible to realize them both, but we currently have no technical means to achieve this. No example necessary.
    - Two desires are in **economic conflict** if it is technically possible to realize them both, but this exceeds the available budget of the stakeholder. No example necessary.
    - Two desires are in **legal conflict** if it would be illegal to realize them both.
    - Two desires are in **moral conflict** if satisfying them both would be morally wrong.

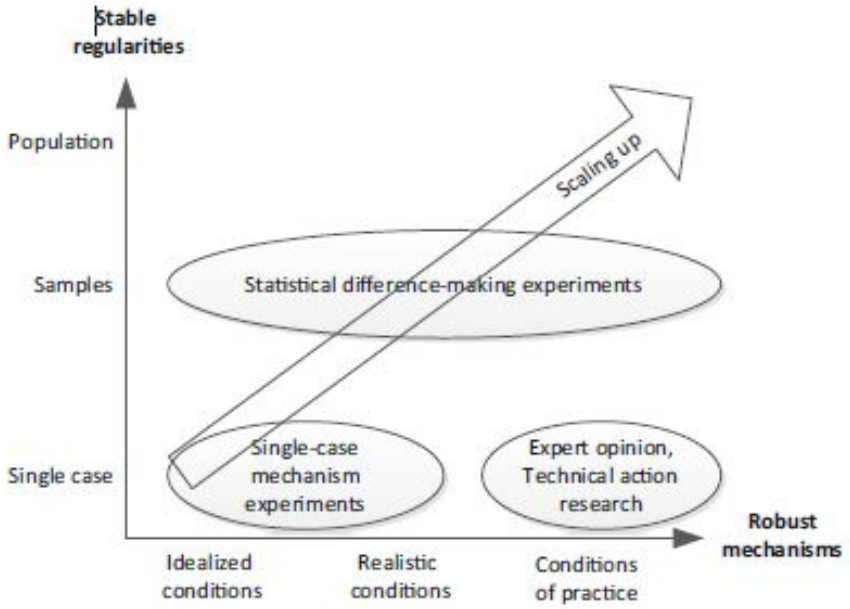
Wieringa: Chapter 7 - Treatment Validation

- This is about the whole left-below corner in the E.C.
- In this book, the word “model” is used as an **analogic model**: an entity that represents entities of interest, called its **targets**, in such a way that questions about the target can be answered by studying the model.
- A **validation model** consists of model of the artifact interacting with a model of the problem context. The targets of a validation model are all possible artifact implementations interacting with real-world problem contexts. This figure represents this notion:



- Three notions about scientific theories:
    - A scientific theory is a belief about a pattern in phenomena that has survived testing against empirical facts and critical reviews by peers. This survival does not imply that the theory is final or even that it is completely true. Any scientific theory is fallible and may be improved in the future.

    - A scientific theory contains a conceptual framework that can be used to frame a research problem, describe and analyze phenomena, and generalize about them.
    - A scientific theory also contains generalizations about patterns in phenomena that may be usable to explain the causes, mechanisms, or reasons of phenomena. This in turn may be useful to predict phenomena or to justify artifact designs. Not each generalization may be usable for each of these purposes.
- In validation research, we develop **design theories**, which are theories of the interaction between an artifact and its intended problem context.
- There are multiple research methods concerning the validation models:
    - **Expert Opinion**: Simple, panel of experts that are used to “observe” (as instruments!).
    - **Single-Case Mechanism Experiments**: A single-case mechanism experiment in validation research is a test in which the researcher applies stimuli to a validation model and explains the response in terms of mechanisms internal to the model. For example, you build a prototype of a program, build a model of its intended context, and feed its test scenarios to observe its responses.
    - **Technical Action Research (TAR)**: is the use of an artifact prototype in a real-world problem to help a client and to learn from this.
    - **Statistical Difference-Making Experiments:** compare the average outcome of treatments applied to samples. They can be used in validation research by selecting samples of validation models and comparing the average outcome of treatments in different samples.
- New technology is always developed by designing and testing it under idealized laboratory conditions first, incrementally scaling this up to conditions of practice later. In this **scaling up** approach, we follow two lines of reasoning, illustrated in the figure below (also noting the different research methods that were mentioned):



- **Case-based inference** is the reasoning by analogy from the investigated model to real-world cases.
- **Sample-based inference** is observing average outcomes (statistical) in samples and making it plausible that this average effect also exists in the population.

Wieringa: Chapter 10 - The Empirical Cycle

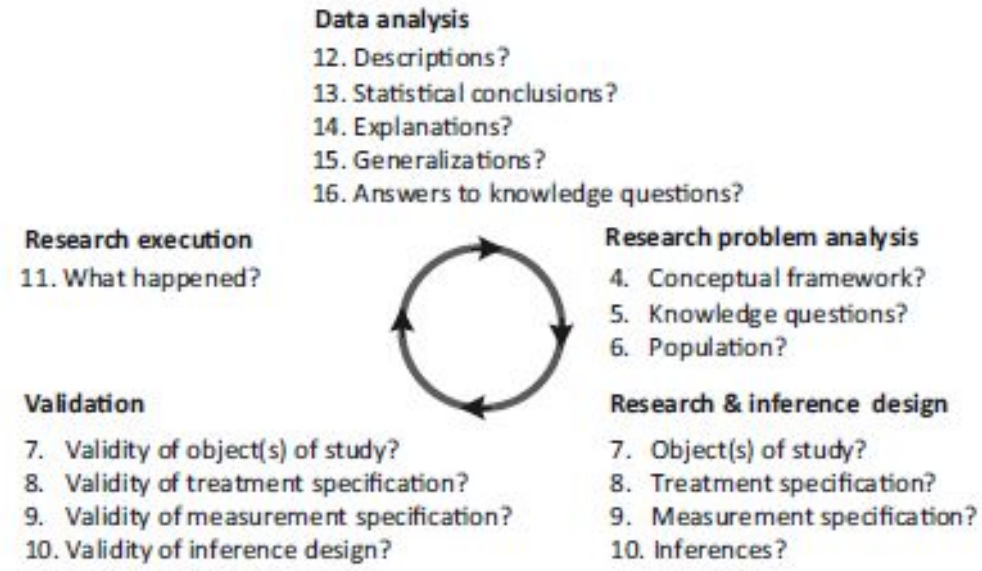
- The empirical cycle (as mentioned before) is a rational way to answer scientific KQ's.
- You should go through this checklist:

Table 10.1 Checklist for the research context

1. Knowledge goal(s)
   - What do you want to know? Is this part of an implementation evaluation, a problem investigation, a survey of existing treatments, or a new technology validation?
2. Improvement goal(s)?
   - If there is a higher-level engineering cycle, what is the goal of that cycle?
   - If this is a curiosity-driven project, are there credible application scenarios for the project results?
3. Current knowledge
   - State of the knowledge in published scientific, technical, and professional literature?
   - Available expert knowledge?
   - Why is your research needed? Do you want to add anything, e.g., confirm or falsify something?
   - Theoretical framework that you will use?

17. Contribution to knowledge goal(s)
    - Refer back to items 1 and 3
18. Contribution to improvement goal(s)?
    - Refer back to item 2
    - If there is no improvement goal, is there a potential contribution to practice?

- The cycle is as below:



The empirical cycle

- **Research problem analysis → Research & inference design → Validation → Research execution → Data analysis**
- There are three kind of validity questions about a research design:
  - **Inference support**: To what extent does the research setup the planned inferences?
  - **Repeatability**: Is the design specified in such a way that competent peers could repeat the research?
  - **Ethics**: Does the treatment of people respect ethical norms?
- For the **research problem**, you need the **conceptual framework**, **KQ's,** and **population**.
- The **research setup** requires the treatment instruments & the measurement of the instruments. In **case-based research** the researcher studies **objects of study (OoS)** separately. In **sample-based research**, the researcher studies a sample of cases.
- In **experimental research**, the researcher applies an experimental treatment to the OoS and measures what happens. In **observational research**, the researcher refrains from intervening and just measures phenomena in the OoS.

Table 10.3 Some different research designs. Between brackets are the numbers of the chapters where these designs are explained

	Observational study (no treatment)	**Experimental study (treatment)**
Case-based research	• Observational case study (17)	• Single-case experiment (18) • Comparative-cases experiment (14) • Technical action research (19)
Sample-based research	• Survey • Quasi-experiment (20)	• Randomized controlled trial (20) • Quasi-experiment (20)

- The process of drawing conclusions from these data is called **inference**. All of the inferences that we discuss are **ampliative**, which means that their conclusions may be false while their premises are true. This is the opposite of a **deductive** inference, of which the conclusions are guaranteed to be true when its premises are true. (You discuss the validity to countermeasure the ampliative inference wrongness).
- There are four kinds of inferences:
    - **Descriptive inference** summarizes the data into descriptions. It is subject to the constraints of **descriptive validity**.
    - **Statistical inference** is the inference of population characteristics from sample statistics. It is subject to the constraints of **conclusion validity**.
    - **Abductive inference** postulates the most plausible explanations for your observations. It is subject to the constraints of **internal validity**.
    - **Analogic inference** is the generalization of your explanations to similar OoS. It is subject to the constraints of **external validity**.

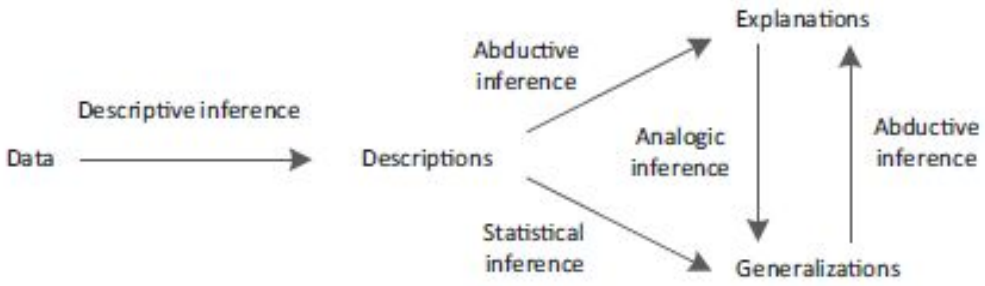


Table 10.4 Case-based and sample-based inference strategies

Case-based inference	Sample-based inference
1. Descriptive inference: Describe the case observations.	1. Descriptive inference: Describe sample statistics.
	2. Statistical inference: Estimate or test a statistical model of the population.
2. Abductive inference: Explain the observations architecturally and/or rationally.	3. Abductive inference: Explain the model causally, architecturally and/or rationally.
3. Analogic inference: Assess whether the explanations would be true of architecturally similar cases too.	4. Analogic inference: Assess whether the statistical model and its explanation would be true of populations of architecturally similar cases too.

- There are three rules in the empirical cycle:
    - **Rule of posterior knowledge:** Knowledge created by the research is present after execution of the research, and it is absent before executing the research.

- **Rule of prior ignorance**: Any knowledge present before doing the research may influence the outcome of the research.
- **Rule of full disclosure:** All events that could have influenced research conclusions must be reported.

Wieringa: Chapter 18 - Single-Case Mechanism Experiments

- A **single-case mechanism experiment** is a test of a mechanism in a single object of study with a known architecture. Goal is to describe the cause-effect behavior of the object of study.
    - Are also causal experiments: The effect of a difference of an independent variable X on dependent variable Y.
    - But not all causal are single-case.
- The **population** of validation research is not the set of similar validation models, but it is the set of all real-world instances of artifact context.
- In the engineering cycle we assess the *validity of a treatment* design with respect to the problem it is designed for,
    - Here, we are interested in the second kind of validity, coming from the empirical cycle: assess *the validity of inferences*.
        - **Descriptive inference**: The meaning of indicators is defined in terms of observable properties of the validation model, i.e., of the artifact prototype and the model of the context. If symbolic data will be produced, then interpretation procedures have to be agreed on too.
        - **Abductive inferences**: If the validation model contains people and you want to do causal inference, you have to assess possible threats to internal validity related to psychological or social mechanisms of people in the validation model or across validation models:
            - *OoS dynamics:* Could there be interaction among validation models?
        - **Architectural inference**: following questions are important:
            - *Analysis*: is there enough information about the architecture of the artifact and context available to do an interesting analysis later?
            - *Variation:* What is the minimal val. model that you can construct to answer the KQ?
            - *Abstraction:* The artifact prototype and context simulation will contain components not specified in the artifact design but required to run the simulation. Influence validation model?
- Threats to validity of rational explanations are: Goals of the actors (differ from explanation), Motivation of actors.
- Population predicate: will the validation model satisfy the population predicate?
- Ambiguity: What class of implemented artifacts in real-world contexts could the validation model represent?

- Relevant consideration to sampling is: **representative sampling**: a case-based research; in what way will the constructed sample of models be representative of the population?

- **Treatments**:
    - **Treatment control**: What other factors than the treatment could influence the validation models?
    - **Treatment instrument validity**: If you use instruments to apply the scenario, do they have the effect on the validation model that you claim they have?
    - **Treatment similarity**:. Is the specified treatment scenario in the experiment similar to treatments in the population? Or are you doing an extreme case study and should it be dissimilar?
    - **Compliance:** Is the treatment scenario implemented as specified?
    - **Treatment contro**l: What other factors than the treatment could influence the validation models? This is the same question as mentioned above for causal
- **Measurement design**
    - *Measurement influence:* will measurement influence the validity model?
    - *Construct validity*: Are the definitions of constructs to be measured valid?
    - Measurement instrument validity: Do the measurement instruments measure what you claim that they measure?
    - *Construct levels:* Will the measured range of values be representative of the population range of values?

- **Descriptive inference** in single-case mechanism experiments is often the presentation of data in digestible form such as graphs or tables with aggregate information.
- **Abductive inference**: If the behavior of the validation model is time independent and if effects are transient, then you can do single-case causal experiments with them.
- **Analogic inference:** Generalization from mechanism experiments is done by architectural analogy: In objects with a similar architecture, similar mechanisms will produce similar phenomena.

Wieringa: Chapter 19 - Technical Action Research (TAR)

- This is the use of an experimental artifact (i.e. still under development and is not yet in the problem context) to help a client and to learn about its effects in practice.
    - Validate the artifact into the field, i.e. the last stage.
    - These are single-case studies; each artifact is studied as a case.
    - With this, the researcher not only applies the artifact, but also helps the client, which makes TAR action research. It is different since it is *artifact driven*.
- With **TAR**, the researcher plays three roles:
    - *Technical researcher*: designs a treatment intended to solve a class of problems (e.g. new effort estimation technique).

- *Empirical researcher:* answers some validation KQ's about the treatment (e.g. wants to know how accurate the effort estimation is).
- *Helper:* researcher applies a client-specific version of the treatment to help a client.
- A lot of bullshit that is similar to chapter 18.

- **Treatment design**:
  - Client cycle: client is treated with an experimental artifact
  - Empirical cycle: experimental artifact is tested by treating it with a real-world context.
  - Researcher & client agree on a treatment plan in which, from one point of view, the client is treated by the artifact and, from another point of view, the artifact is treated to the client context.
  - How you get a client, bla bla bla.

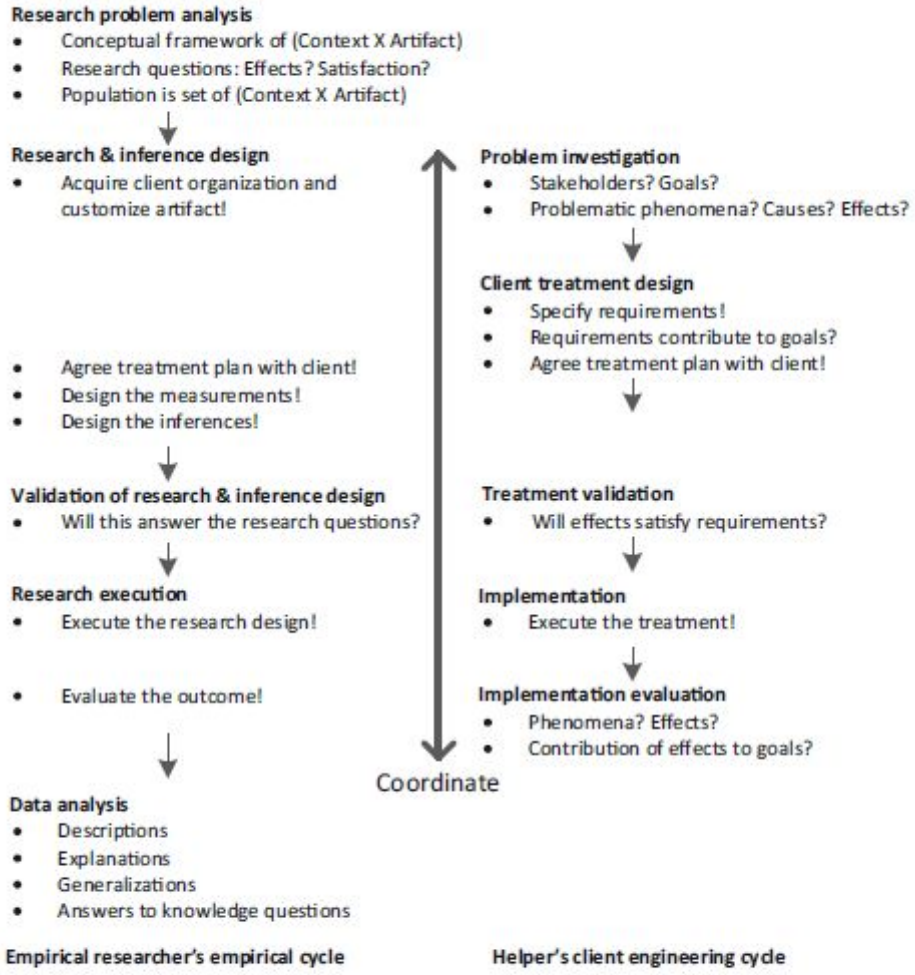
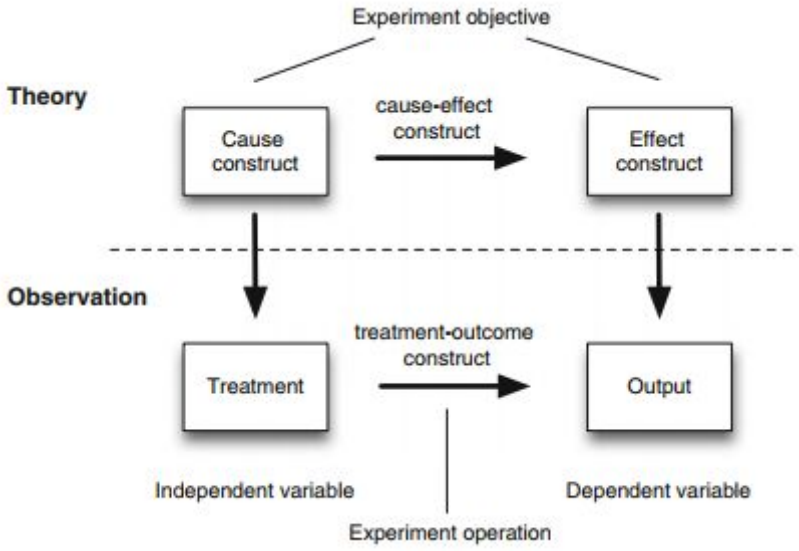


Fig. 19.2 Detailed list of tasks in TAR. Exclamation marks indicate things to do, question marks indicate questions to answer. Coordination with activities in the client cycle starts as soon as a client is acquired and the artifact customized and finishes when the client cycle is evaluated

- The checklist for causal inference is important, but this is the same as in chapter 18: and then the “- Treatments: …”
- Same for measurement design. (very similar).
- Side note: after participating in the ARM exam: this did not occur into the practice exam or real exam.

Wohlin: Chapter 6 - Experiment Process

- The basic principles behind an experiment are illustrated:



- Treatment, Variables:

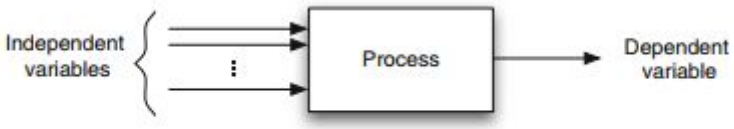


Fig. 6.2 Illustration of independent and dependent variables

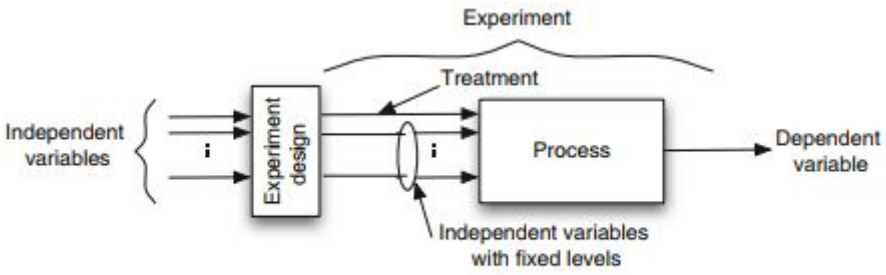
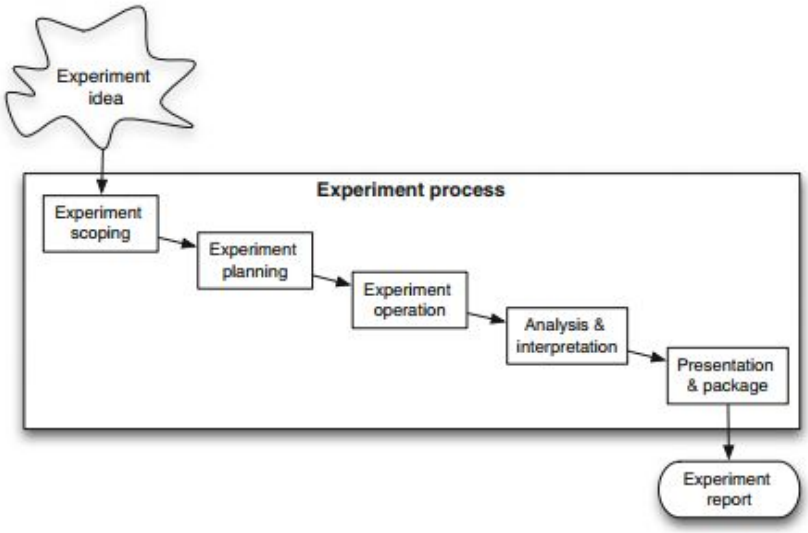
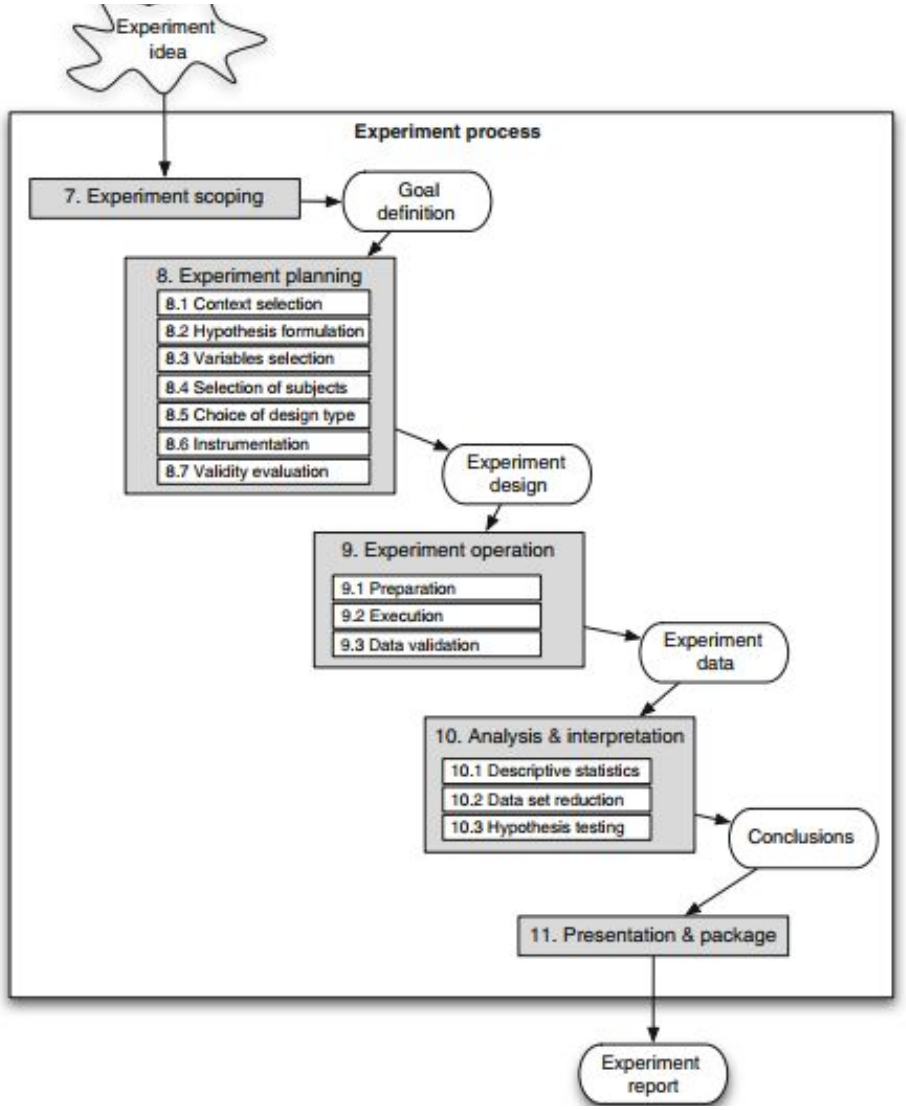


Fig. 6.3 Illustration of an experiment

- Treatments are applied to the combination of objects (documents, programs, etc.) and subjects (people that apply the treatment).
- The following is the overview of the experiment process:



- This process is partly iterative and it may be necessary to go back and refine a previous activity before continuing with the experiment.
- The next picture is the same above picture, but with corresponding substeps:



Wohlin: Chapter 7 - Scoping

- The goal template is:
  Analyze <Object(s) of study>
  for the purpose of <Purpose>
  with respect to their <Quality focus>
  from the point of view of the <Perspective>
  in the context of <Context>
- Output of this process gives you a goal definition.
- You have single & more than one objects and also with subjects:

Table 7.2 Example experiment context classification, from Basili [10]

		# Objects	
		One	More than one
# Subjects per object	One	3. Cleanroom project no. 1 at SEL [14]	4. Cleanroom projects no. 2-4 at SEL [14]
	More than one	2. Cleanroom experiment at University of Maryland [149]	1. Reading versus test [12] 5. Scenario based reading vs. checklist [18]

Table 7.3 Goal definition framework

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		Customer	
			User	
			Researcher	

Wohlin: Chapter 8.1-8.9 - Planning

- After the goal is defined, you need to plan the experiment:
  - **Context selection**
    - Off-line vs. on-line
    - Student vs. professional
    - Toy vs. real problems
    - Specific vs. general
  - **Hypothesis formulation**
    - Null vs alternative H.
    - Type-I-error: False positive (P(reject H0 | H0 true))
    - Type-II-error: False negative (P(not reject H0 | H0 false))

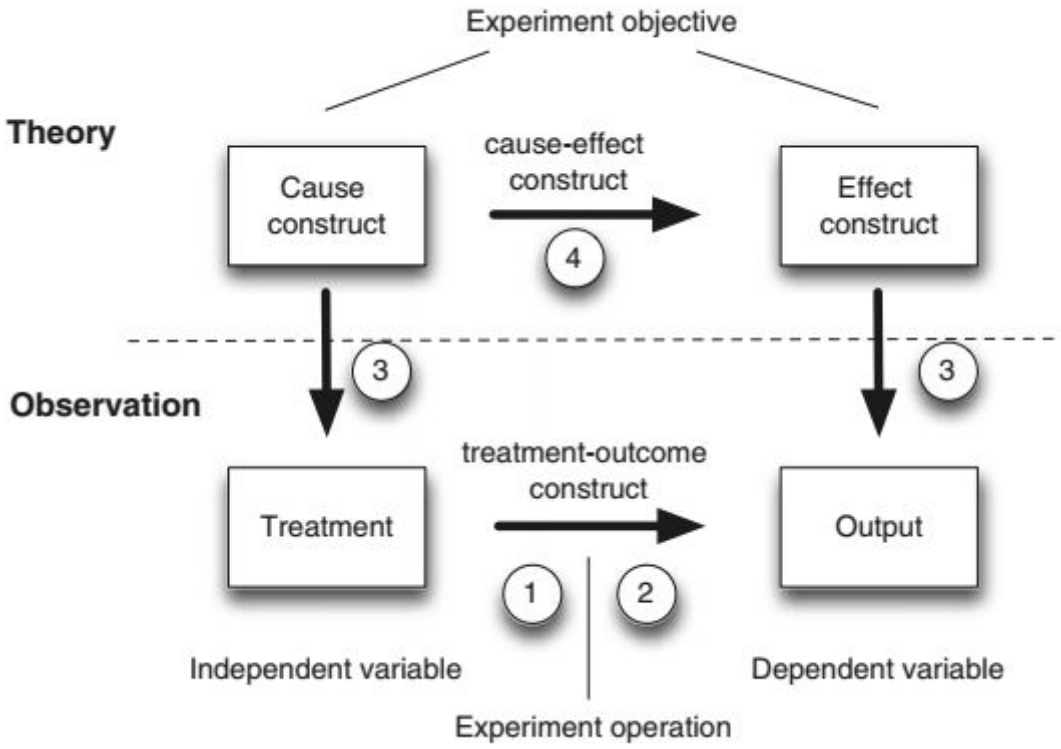
    - Power: the probability (P) that the test will reveal a true pattern if H0 is false. (P(reject H0 | H0 false) = 1-P(type-II-error)
- **Variables selection**
    - Independent variables: that we control and change
    - Dependent variables: the effect of the treatments; often only one, derived from the hypothesis.
- **Selection of subjects**
    - Examples of *probability sampling techniques* are:
        - Simple random sampling: Subjects are selected from a list of the population at random.
        - Systematic sampling: The first subject is selected from the list of the population at random and then every n:th person is selected from the list.
        - 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.
    - Examples of *non-probability sampling* techniques are:
        - Convenience sampling: The nearest and most convenient persons are selected as subjects.
        - Quota sampling: This type of sampling is used to get subjects from various elements of a population. Convenience sampling is normally used for each element.
- **Choice of design type**
    - General design principles:
        - Randomization: All statistical methods used for analyzing the data require that the observations be from independent random variables. It applies on allocation of the objects, subjects and in which order the tests are performed.
        - Blocking: Blocking a factor that has an undesired and known effect on the response. You systematically eliminate the undesired effect in the comparison among the treatments. Increases precision of the experiment. Grouping subjects is a form of blocking.
        - Balancing: Assign treatments to an equal number of subjects.
    - Standard design types:
        - One factor with two treatments: We want to compare the two treatments against each other. Most common is to compare the means of the dependent variable for each treatment.
            - You can have *completely randomized design* or *paired comparison design* (they do both the treatments, also crossover design).
        - One factor with more than two treatments: Same as above, but with more treatments.
            - Here you also have the same, but paired comparison design is now *randomized complete block design*, i.e. they perform every treatment (again at random).

- Two factors: Factor A has two treatments and Factor B has two treatments, so 2 hypotheses (one per treatments) and +1 hypothesis about the interaction between the two.
    - *Two-stage nested design*: one of the factors is similar but not identical for different treatments of the other factor (B vs. A for example). Factor B is then nested under factor A.
- More than two factors: The effect in the dependent variable can be dependent not only on each factor separately but also on the interactions between the factors.
    - *2k factorial design*: 2k different combinations of the treatments; all combinations have to be tested.
    - *2k fractional factorial design*: Because of the rapid growing number of factors, you limit the research to a fraction, because of negligible interactions. It is based on three ideas:
        - *The sparsity of effect principle*: It is likely that the system is primarily driven by some of the main and low-order interaction effects.
        - *The projection property*: A stronger design can be obtained by taking a subset of significant factors from the fractional factorial design.
        - *Sequential experimentation*: A stronger design can be obtained by combining sequential runs of two or more fractional factorial designs.
    - You actually explore and identify the factors that have large effects on the system. Again, here are fractional factorial designs:
        - *One-half fractional factorial design of the 2k factorial design*: Half of the combinations of a full 2k factorial design is chosen. (2^k-1)
        - *One-quarter fractional factorial design of the 2k factorial design*: One quarter of the combinations of the full 2k factorial design is chosen. (2^k-2)

- **Instrumentation**
    - Objects, guidelines, measurements, monitoring...
- **Validity evaluation**
    - (1)Conclusion validity: Concerned with the relationship between the treatment and the outcome; we want to make sure that there is a statistical relationship, i.e. with a given significance.
    - (2)Internal validity: We make sure if the outcome above is a causal relationship, not the result of a factor which has not been measured/controlled, i.e. the treatment causes the outcome/effect.
    - (3)Construct validity: Concerned with the relation between theory and observation. If the relationship between cause and effect is causal, we

must ensure two things: (1) that the treatment reflects the construct of the cause well (see left part of Figure) and (2) that the outcome reflects the construct of the effect well (see right part of Figure)

- (4)External validity: Concerned with generalization; is there a relationship between the treatment and the outcome?



- The complete list of validity threats are on the next page:
  - Side note: in the ARM exam in 2017-2018, specific validity threats were not asked. They did ask about the types of validity threats. I would recommend remembering a few per type.
  - If you do not understand a certain type, I recommend reading the definition in the book. Most of the types should be clear.

Table 8.10 Threats to validity according to Cook and Campbell [37]

Conclusion validity	Internal validity
Low statistical power	History
Violated assumption of statistical tests	Maturation
Fishing and the error rate	Testing
Reliability of measures	Instrumentation
Reliability of treatment implementation	Statistical regression
Random irrelevancies in experimental setting	Selection
Random heterogeneity of subjects	Mortality
	Ambiguity about direction of causal influence
	Interactions with selection
	Diffusion of imitation of treatments
	Compensatory equalization of treatments
	Compensatory rivalry
	Resentful demoralization
Construct validity	**External validity**
Inadequate preoperational explication of constructs	Interaction of selection and treatment
Mono-operation bias	Interaction of setting and treatment
Mono-method bias	Interaction of history and treatment
Confounding constructs and levels of constructs	
Interaction of different treatments	
Interaction of testing and treatment	
Restricted generalizability across constructs	
Hypothesis guessing	
Evaluation apprehension	
Experimenter expectancies	

Table 8.11 Threats to validity according to Campbell and Stanley [32]

Internal validity	External validity
History	Interaction of selection and treatment
Maturation	Interaction of history and treatment
Testing	Interaction of setting and treatment
Instrumentation	Interaction of different treatments
Statistical regression	
Selection	

- There is some conflict between the types of validity threats. Cook & Campbell propose the following priorities for theory testing and applied research (also: the order of the validities!):
    - **Theory testing**: It is most important to show that there is a causal relationship (internal validity) and that the variables in the experiment represent the constructs of the theory (construct validity). Adding to the experiment size can

generally solve the issues of statistical significance (conclusion validity). Lastly comes external validity.

- **Applied research:** Again, the relationships under study are of highest priority (internal validity) since the key goal of the experiment is to study relationships between causes and effects. In applied research, the generalization – from the context in which the experiment is conducted to a wider context – is of high priority (external validity). Third, the applied researcher is relatively less interested in which of the components in a complex treatment that really causes the effect (construct validity). Lastly comes the conclusion validity.

Wohlin: Chapter 9 - Operation

- Experiment design
  - Preparation
  - Execution
  - Data validation
- Experiment data
- Commit participants:
  - Obtain consent
  - Sensitive results
  - Inducements (what I did with beer)
  - Disclosure (reveal as much as possible)

Wohlin: Chapter 10 - Analysis & Interpretation

- **Descriptive Statistics:** deals with the presentation and numerical processing of a data set. It may be used to describe & graphically present interesting aspects of the data set. It may be used before carrying out hypothesis testing → identify abnormal/false data points (**outliers**)
  - The relevant statistics:

Table 10.1 Some relevant statistics for each scale

Scale type	Measure of central tendency	Dispersion	Dependency
Nominal	Mode	Frequency	
Ordinal	Median, percentile	Interval of variation	Spearman corr. coeff. Kendall corr. coeff.
Interval	Mean, variance, and range	Standard deviation	Pearson corr. coeff.
Ratio	Geometric mean	Coefficient of variation	

  - Mean, median, mode are **measures of Central Tendency**.
  - Range, variance, frequency, standard deviation, variation interval, coefficient of variation are **measures of Dispersion.**
  - **Covariance** is dependent on the variance of each variable. Rest of **dependency measures** were not relevant for the course.

    - **Visualization**: You can use a scatter plot, box plot (median in the middle), histogram, cumulative histogram, pie chart.
- When outliers are discovered, you should do something with them: based on the coordinates in the diagram, you should analyse the reasons for the outliers.
    - Strange/rare event, never will happen again? *Exclude it.*
    - Strange/rare event, can happen again? *Not advisable to exclude it.* This could be an undefined variable (unexperienced staff for instance).

- **Hypothesis Testing**
    - The objective of hypothesis testing is to see if it is possible to reject a certain null hypothesis, H0, based on a sample from some statistical distribution.
    - Three important probabilities concerning hypothesis testing are:

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

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

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

    -
    - The **power** of a test increases as alpha decreases, you will likely make less corruptive faults, or Type-I-errors.

- **Parametric tests:** are based on a model that involves a specific distribution. It is assumed in most cases that some of the parameters involved in a parametric test, are **normally distributed**. A normality test is Chi-2 test. PT also require that parameters can be measured at least on an **interval scale**.
- **Nonparametric tests:** Do not make the same type of **assumptions** concerning the distribution of parameters as parametric tests do. These are only very general, an example test is the binomial test. Nonparametric tests can generally be used instead of parametric tests, but not vice versa.
    - Two factors to consider here: **Applicability** (what assumptions are made?) and **Power** (parametric is generally higher power, require fewer data points, smaller experiments, IF assumptions = True)

- Different tests, either parametric or nonparametric:

Parametric	Nonparametric
t-test Most often used, compare two (independent) sample means (one factor with 2 levels/treatments)	**Mann-Whitney** Alternative to the t-test, always possible to use this instead of the t-test.
F-test compare the variance of two (independent) sample distributions	**Wilcoxon** Alternative to the paired t-test, requirements are that it is possible to determine which of the measures in a pair is the greatest and that it is possible to rank the differences.
Paired t-test t-test For a paired comparison design; for example measurements are made with respect to a subject more than once, e.g. two tools are compared.	**Sign test** Alternative to the paired t-test, simpler alternative to the Wilcoxon test, i.e. when it is not possible or necessary to rank the differences.
ANOVA Used for designs with more than two levels of a factor; one factor with more than two levels, one factor & blocking variable, factorial design, and nested design.	**Kruskal-Wallis** Alternative to ANOVA, in the case of one factor with more than two treatments.
	Chi-2 Family of non-parametric tests that can be used when data are in the form of frequencies.

- The first 4 nonparametric tests assume to have a case with few samples.
- Overview for different test **designs**:

Table 10.3 Overview of parametric/non-parametric tests for different designs

Design	Parametric	Non-parametric
One factor, one treatment		Chi-2, Binomial test
One factor, two treatments, completely randomized design	t-test, F-test	Mann-Whitney, Chi-2
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-2
More than one factor	ANOVA[a]	

[a] This test is not described in this book. Refer instead to, for example, Marascuilo and Serlin [119] and Montgomery [125]

- In the book examples are provided of each above test.

- **Chi-2 Goodness of fit test:** can also be carried out in order to check if measurements are taken from a certain distribution, e.g., the normal distribution.
- **Model Adequacy Checking:**
    - *Normality:* If a test assumes that the data is normally distributed, a Chi-2 test can be made to assess to which degree the assumption is fulfilled. The Chi-2 test is described above.
    - *Independence:* If the test assumes that the data is a sample from several independent stochastic variables, it is necessary to check that there is no correlation between the sample sets. This may be checked with scatter plots and by calculating correlation coefficients as discussed in the beginning of this section.
    - *Residuals:* In many statistical models, there is a term that represents the residuals (statistical error). It is often assumed that the residuals are normally distributed. A common way to check this property is to plot the residuals in a scatter plot and see that there is no specific trends in the data (the distribution looks random).

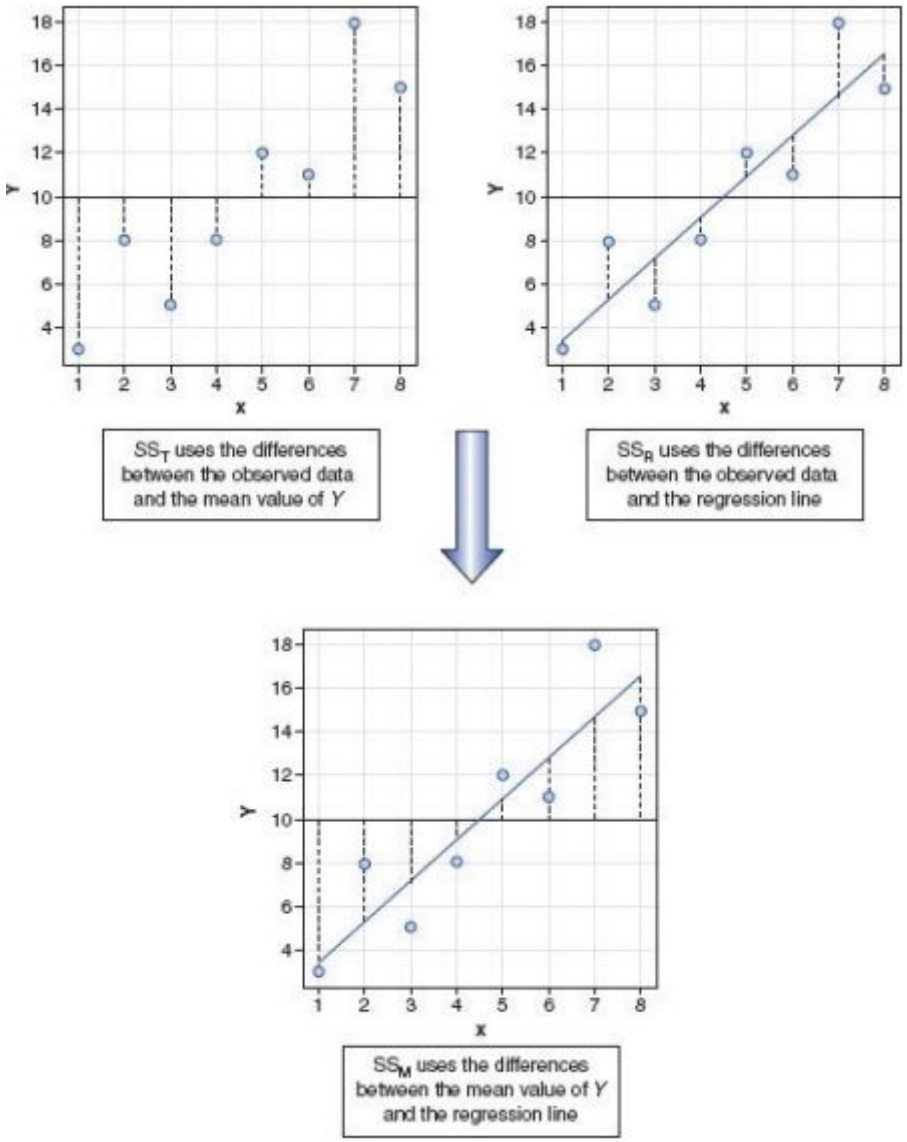
- **Drawing conclusions:**
    - *When the experiment data has been analyzed and interpreted, we need to draw conclusions regarding the outcome of the experiment. If the hypotheses are **rejected** we may draw conclusions regarding the influence of the independent variables on the dependent variables, given that the experiment is valid, see Chap. 8.*
    - *If, on the other hand, the experiment **cannot reject** the null hypothesis, we cannot draw any conclusions on the independent variables influence on the dependent variable. The only thing we have shown, in this case, is that there is no statistically significant difference between the treatments.*

- Some **tactics**:
    - Start with a boxplot, looking for any outliers and maybe exclude them in high exception (which should be well documented! Normally a no-go).
    - Then: is the data normally distributed? Plotting a histogram can help with this, or using a Chi-2 test, Kolmogorov-Smirnov test, Shapiro-Wilks' W test, or the Anderson-Darling test.
        - Minding: with a small sample size it may look normally distributed, without actually being normally distributed.
        - t-test is robust for non-normality, ANOVA is not. (If you do decide to go with ANOVA and there is significance, test extra with Fisher's Protected Least Significant Difference test **(Fisher's PLSD test)**
    - Then we look at what kind of design we got and if there is a parametric/nonparametric test available to reject the H0.

FROM THIS PART ON IT IS EXTRA, I did not finish this book in the course since practicing seemed more relevant. It might help you though, but it is solely an extra.

Field: Chapter 7.1, 7.2

- 7.1 is a cute introduction story.
- You can use one predictor variable (**simple regression**) or several predictor variables (**multiple regression**).
- The method used is the **method of least squares**.
- The straight line is defined by its **slope** (gradient) which is usually b1, and the point at which the line crosses the vertical axis (**intercept**), which is b0. So:
    - *Yi = (b0 + b1*Xi) + error*
    - b1 & b0 are also called the **regression coefficients**.
- You try to fit lines to predict the values from Y from values of the X variable. The differences between the line and the actual data are usually called the **residuals**. Might also be called deviations.
    - This yields positive (underestimates) and negative (overestimates) values.
    - If you add this, they tend to cancel each other out, so, we *square the differences before adding them up.*
        - Large differences, then the line is not representative.
        - So you want low differences in total → a low **MSE**.
        - The lowest sum of squared differences is selected with this model.
- The best fit of the line is called a **regression model**.
- Once the best fit has been found, we need to assess how well this line fits the actual data: the **goodness of fit**.
- The **residual sum of squares** (SSr) represents the degree of inaccuracy when the best model is fitted to the data. You can use this with the **total sum of squares** (SSt) to calculate how much better the regression line (line of best fit) is than just using the mean as a model.
    - Thus, calculating the difference between SSt & SSr → shows the reduction in the inaccuracy of the model resulting from fitting the regression model to the data.
    - This is called the **model sum of squares** (**SSm)**. The picture below summarises this better.
    - If the SSm = large, then the regression model is very different from using the mean to predict the outcome value, thus the regression model has made a big improvement on the prediction.
        - And vice versa with small.
        - ***R2 = SSm/SSt.*** To express it you should multiply it by 100 as a percentage.
        - Take the square root of this value to obtain the **Pearson's correlation coefficient**.
        - *The proportion of variance in the outcome accounted for by the predictor/independent variable or variables.*



-

- A second use of the sum of squares in assessing the model is through the **F-test**. This is usually the amount of systematic variance divided by the amount of unsystematic variance, or: the model compared against the error in the model.
    - You use the average sum of squares → **means squares (MS)**.
    - For SSM the degrees of freedom are simply the number of variables in the model, and for SSR they are the number of observations minus the number of parameters being estimated.
    - The results are the mean squares for the model (**MSm**) and the residual means squares (**MSr**) → ***F = MSm / MSr***
    - A good model should have a large F-ratio (greater than 1 at least), because the top of the equation will be bigger than the bottom.
    - *The F-Ratio is the ratio of variance explained by the model to the error in the model.*

- With a bad model, you would expect the **value of b** to be zero. Think about the model representing the mean line above (value of outcome does not change).

- Important: a bad model (such as the mean) will have regression coefficients of 0 for the predictors → No change in predicted value & gradient is 0.
- The **t-statistic** tests the null hypothesis that the value of b = 0: therefore, if it is significant then the b value is significantly different from 0.
- We use the standard error to look at the expected error of the b-value.
- We use the standard deviation of the distribution of the error samples.
- ***t = (b-observed -- b-expected) / SEb == b-observed / SEb***
- In regression, the degrees of freedom are N - p - 1, where N = the total sample size, p is the number of predictors.
    - A simple regression with one predictor = N-2.
    - If t is very large, it is unlikely to have occured when there is no effect.

Field: Chapter 8 - Logistic Regression

- **Logistic regression** is an extension of regression that allows us to predict *categorical outcomes* based on predictor variables.
    - Outcome = categorical, predictor(s) = continuous or categorical.
    - E.g.: a person has pig-headedness, alcohol consumption, scores high on laziness, has an outcome = male.
    - If we predict the membership of only two categorical outcomes (e.g. male/female) = **binary logistic regression**
    - If we want to predict more than two categories = **multinomial (or polychotomous) logistic regression**
- To keep things simple, the author focuses on the binary logistic regression.
- With multinomial regression, the regression formula is extended with b's and X's as new predictors multiplied by its respective regression coefficient; b1X1 + b2X2, etc.
- In logistic regression, instead of predicting the value of a variable Y from X1, we predict the **probability** of Y occurring given known values of X1 (or Xs).
- The logistic regression (Y) =
    - *P(Y) = 1 / (1 + e^(-(b0+b1*X1)) ;* This has only 1 predictor variable.
    - This equation within the brackets is the same as the linear regression.
    - You can extend this also with several predictors.
    - With the logarithmic you express a non-linear relationship in a linear way.
    - The result, obviously, varies between 0 (unlikely) and 1 (likely).
    - The chosen model has the results in values of Y closest to the observed values; the parameters are estimated using **maximum-likelihood estimation**.
- We can use the observed and predicted values to assess the fit of the model (like the R2 in linear regression), which is the **log-likelihood**:
    - $\sum_{i=1}^{N}\left[Y_i ln\left(P(Y_i)\right)+(1-Y_i)ln\left(1-P(Y_i)\right)\right]$
- The **deviance** is closely related to the log-likelihood: *deviance = -2 * log-likelihood*
    - Or: **-2LL**. It is more convenient, it has a **chi-square distribution**, which makes it easy to calculate the significance of the value.

  - The *baseline* model (like the mean model in linear regression) is just to predict the outcome that occurs most often; logistic regression model when only the constant is included. This is the **likelihood ratio**:

$$\chi^2 = (-2LL(\text{baseline})) - (-2LL(\text{new}))$$
$$= 2LL(\text{new}) - 2LL(\text{baseline})$$
$$df = k_{\text{new}} - k_{\text{baseline}}$$

  -
- You also have an R-statistic:
  - *"This R-statistic is the partial correlation between the outcome variable and each of the predictor variables, and it can vary between −1 and 1. A positive value indicates that as the predictor variable increases, so does the likelihood of the event occurring. A negative value implies that as the predictor variable increases, the likelihood of the outcome occurring decreases. If a variable has a small value of R then it contributes only a small amount to the model."*

$$R = \sqrt{\frac{z^2 - 2df}{-2LL(\text{baseline})}}$$

  - (z2 is the "**Wald statistic**", therefore apparently not a reliable measure)
  - A few analogue ways of the R2 measure are also being mentioned: **Hosmer & Lemeshow's R2L, Cox & Snell's,** and **Nagelkerke's R2n**.
- You can check a model's fit with linear/logistic(?) regression with the Akaike & Bayes Information Criterion:
  - **AIC** = *-2LL + 2k* (k is number of predictors in the model)
  - **BIC** = *-2LL + 2k * Log(n)* (n is the number of cases in the model)
- In linear regression you use the t-test, with logistic regression you use the **z-statistic (Wald statistic)**, which follows a **normal distribution**. It tells if the ***b*** coefficient is significantly different from 0, thus makes a contribution to the prediction Y.
  - *Z = b/SEb* (regression coefficient / its standard error)
  - When b is very large, SEb gets 'inflated', the z-statistic gets underestimated, and therefore the z-statistic should be treated with caution. You reject a predictor while it actually makes a contribution (Type II error).
- Also talked about the **Odds ratio**, which does not seem relevant.
- With logistic regression **methods** you have several:
  - **Forced entry method**: Simply place predictors into one block, and estimate the parameters for each predictor.
  - **Stepwise methods**: Either forward or backward (I already know this). The methods focus on improving the AIC/BIC. Hybrids of forward/backward are better.
- **Assumptions** of logistic regression (share some of those with linear):
  - **Linearity**: Is violated without the log due to categorical outcome, but with the log this is actually fixed.

    - **Independence of errors**: (same with ordinary regression) cases of data should not be related; e.g. you cannot measure the same people on different points in time.
    - **Multicollinearity**: this means that the different predictors should not be too highly correlated with each other. You can check this with VIF statistics, **eigenvalues** of the scaled, uncentered cross-product matrix (8.8.1).
- **Problems with R & Logistic Regression:**
    - Sometimes R makes mistakes with logistic regression: it will not produce an output, or an incorrect output when the predictor values are insufficient/incorrect.
    - Another one is when the outcome variable can be perfectly predicted by one or a combination of variables, which is known as **complete separation**. You get insufficient points in between (think about the burglars & cats weight example).

- What I now skip (it is a lot):
    - 2 examples in R about logistic regression.
    - Multinomial logistic regression in R
    - How to report both

  I can do this in my spare time or see later if this is really necessary to perform.