Chapter 6:

Evaluation: Experimental Design

Overview

1. Variables, scale types and mathematical operations
2. Experimental design
3. Population sampling
4. Validity
5. Ethical considerations
6. Designing the experiment

Variables, scale types and mathematical operations

Variables

The characteristics of every experiment can be described with two different kinds of variables:

Dependent variables (DV)

The observed aspects:
→ e.g. task completion time, words per minute, ...

Independent variables (IV)

The manipulated aspects:
→ e.g. user interface layout, type of keyboard ...

The dependent variables describe the effect of a change in the independent variables. For example, when we change the layout of our interface, the users will probably need a different task completion time. The independent variables, therefore, represent a cause; their value is not dependent on other variables in your experiment. The question is: how can we manipulate a single aspect only? In theory, we need to keep all other factors stable (e.g., mood, weather), but people or situations are never identical!

Causes of variation of DV

The dependent variable in your experiment might vary. These variations can be caused by two different types: systematic and unsystematic variation. Systematic variation is not always unwanted; this variation includes differences we measure for example in two different groups of a study (intervention and control group). It arises from the effects of the independent variables that you want to analyse. However, you should try to minimize unwanted systematic variation through appropriate randomization!

Systematic variation
Due to the experimental conditions

Unsystematic variation
Errors due to random factors

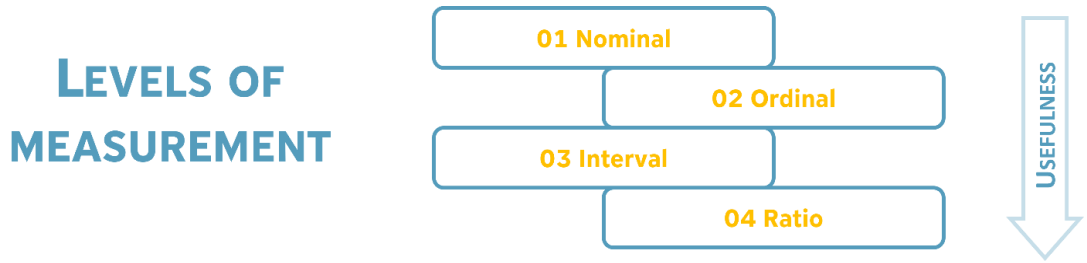
Unsystematic variation is the unintended variation in our results. Sometimes it can be quite difficult to identify causes and eliminate them. Random factors can influence the participants of the study; maybe someone has a bad day and is more negative in his opinion, on the second day of the experiment then his mood changed completely and he might answer the same question differently.

So, to find the unsystematic variations, you need to question your experiment: What could have had an influence? You can find causes, by analysing your design. Is the time of the interview for example too early for people that like to sleep longer? Discuss these questions with peers, supervisors or colleagues and let them take a second look at your design in order to also get a different perspective.

The big role of statistics is to discover how much of the variation in the results we measured is attributed to systematic or unsystematic variation.

Scale types

The variables we measure (DV) can be represented with different types of scales.



01 Nominal: this discrete scale labels the variables in different classes, a value or order is not assigned, and the different values are not related. → Gender, colours, …

02 Ordinal: the ordinal scale represents a natural ranking in a non-mathematical way with discrete values. → education degree, Likert scale, …

03 Interval: the interval scale is a numerical scale measuring variables that exist along a common, ordered scale at equal intervals without a true "zero". → Temperature (there is no "twice as warm")

04 Ratio: this quantitative scale is as well an ordered and continuous scale, but it possesses a true zero. The ratios have to be meaningful (e.g., twice as fast). → Body weight, height

As mentioned above, the nominal and ordinal scales represent discrete scales. This means that we can simply count the different expressions of our variable. Against this, continuous scales like the ratio or interval scale have overcountable many expressions; the scales are steady.

Mathematical operations

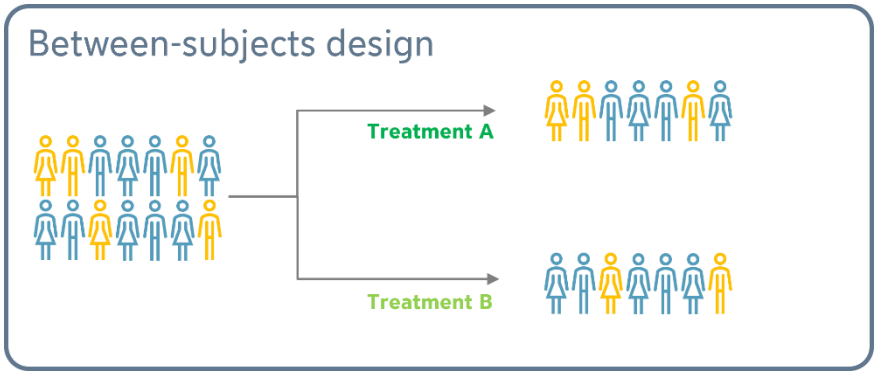
With different scale types, different mathematical operations are allowed. Generally said, all operations that do not change the relations of distances between characteristic expressions are allowed. One example of an inappropriate operation would be calculating the average of the gender. The table below shows you where different mathematical operations with the scale are appropriate.

Scale	Count	Order	Difference	Ratio
01 Nominal	✓	✗	✗	✗
02 Ordinal	✓	✓	✗	✗
03 Interval	✓	✓	✓	✗
04 Ratio	✓	✓	✓	✓

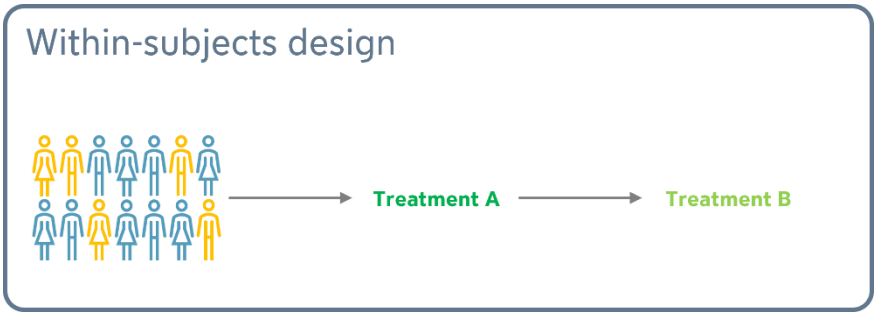
✓ Statistically meaningful operation
✗ Statistically not meaningful operation

Experimental design

The **between-subjects design** is also called independent measures design. Each participant is assigned to one condition only. This design is usually used for clinical trials with medications, where one group gets the new medication while the others get a placebo. This design is easy to conduct and has a low chance of practice or fatigue effects among the participants. It is beneficial when we want to run the experiment only once or when it is impossible for an individual to participate in all possible conditions. However, it is pretty expensive due to the high number of participants required and the time and effort for the recruitment, organisation, and conduction. Furthermore, the individual characteristics of a participant might influence the results of only one of the groups.



The **within-subjects design** can also be called repeated-measures design. In this design, all participants are assigned to all possible conditions. Important to note is that the order must be counter-balanced or randomised. The participants could again be split up into groups, so one group tests, for example, interface A and then interface B. The other group of participants starts with interface B. This design is more economical and features a high level of sensitivity. Since every participant will be assigned to all conditions, individual characteristics have less influence on the results. Disadvantages of this design are that each condition must be reversible and carry-over effects from the previous condition are likely. This could be that fatigue or learning effects among participants.



Besides these two designs, also different hybrid or mixed designs are possible as well. For example, when analysing several independent variables.
However, you should use a **design as simple as possible**, as this also makes **statistical testing easier** in the end.

Quasi-experiments

In a true experiment, the group assignment is performed using randomisation. This way, we have complete control over the independent variable. In a quasi-experimental design, the assignment is not performed randomly! This means not only that we have limited control over the independent variables, but also conclusions we can draw from our study are limited. The control and treatment groups might not even be comparable at baseline.

But still, in some cases, we need to design our study in a quasi-experimental way. Why? Imagine you want to analyse the accidents when driving with and without the lights on. You could think that you just let one group drive always with their lights on, and one group always without lights. After one year, you measure how many accidents occurred. Obviously, this kind of study is ethically problematic and cannot be conducted this way! This is where we use a quasi-experimental design: we search for drivers that already have these habits. In the end, of course, it needs to be discussed whether our conclusions are valid and to what extent!

Population sampling

Choosing participants for your study is called sampling. Since you cannot conduct your study with the whole population you need to take some people from the population as a sample. These samples should reflect the population as good as possible. Otherwise, a selection bias might be present in your results. In 2010 *Henrich et al.* analysed the participants of psychological studies. Their work showed that a vast majority of studies used WEIRD participants (Western, educated, industrialized, rich and democratic). They showed that it is 4000 times more likely that a random American undergraduate student is participating in a study in a psychological journal than an average human being. This obviously leads to a selection bias and should be considered when looking at the study results and drawing conclusions.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi:10.1017/S0140525X0999152X

Validity

Validity is one of the main quality criteria of a study and it describes how suitable a measurement is in relation to its specific objective. Your research should always valid, to provide usable and conclusive results. Validity can be subdivided in two types; internal and external validity.

Internal validity

Internal validity aims at the quality of your experiment design. This includes on the one hand, that you actually measure fit to the question you want to answer. On the other hand, it is defined through the

reproducibility of your experiment. When repeating the measurement, the results you get should match those of the previous run.

To reach internal validity you should ensure that:

- Tasks are clearly defined
- Manipulation is defined
- Measures are defined
- Environment is controlled
- Chosen sample is appropriate
- The experiment is reproducible

The consequence should be that t**he manipulation is the true cause of the outcomes**. You do not only have to have reliable measures for internal validity but also a **strong causality between your independent and dependent variables**. Furthermore, you should be able to rule out other causes for your dependent variables and **unambiguously assign causes to effects**.

External validity

External validity ensures the ability to generalize study results to other people and other situations (outside the laboratory environment).

To ensure this your study should cover:

- Appropriate sample of a population
- Different representative situations
- All relevant time points
- Various locations
- ...

As you can see, the factors influencing the external validity have a wide variety and sometimes not everything can be considered. It is simply impossible to test the whole world! As a result, there will always be a **trade-off between the internal validity (quality of your design) and the external validity (generalizability)**.

Ethical considerations

Ethics

> *'Primum non nocere' – 'First do no harm'*
> Thomas Sydenham

The first thing you should think about before starting your experiment: Am I doing harm to somebody? This is going back to a quite old statement of Thomas Sydenham. Whenever you do studies with real human subjects you should obtain an approval of your study of an ethics committee. Especially, when you are doing research at a university and want to publish your results. If you do studies in the course of your projects or bachelor / master thesis, please discuss the ethical procedure with you supervisor!

> *'One should treat others as one would like others to treat oneself'*
> The "Golden Rule"

Informed consent

One essential part of ethics in a study is the informed consent. With this you get the permission of a person to collect their data before the conduction. This is always necessary when humans participate in a study and personally identifiable information is collected. This includes information that could be used to uniquely identify, contact, or locate a single person directly or with the help of additional sources. The informed consent needs to be obtained regardless of whether they are only interviewed in the study or will take medication for example. The following parts should be included in this informed consent:

Further ethical considerations

Deception:

Sometimes it might be good, if participants do not know about the real context of an experiment. In this case you always need to evaluate, whether it is ethically justifiable. Furthermore, participants must always be informed about the deception afterwards!

Debriefing:

After you finished the conduction of the experiment always answer the participants' open questions.

Confidentiality:

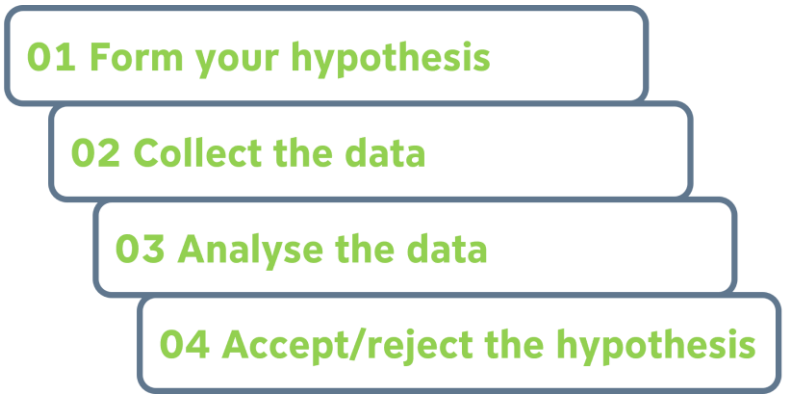
The information and data you collected from the participants should be kept confidential at any time.

Protection from physical and psychological harm:

Never do physical or psychological harm to your participants!

Designing the experiment

When designing an experiment, you can use the following basic scientific method:



In these steps, you need to define

- which system you will test
- which participants will be included
- the hypothesis
- the relevant variables
- the experimental methods that will be used
- the statistical approaches that will be applied

Procedure for user studies

In the beginning you need to set a goal of your experiment. Then you implement the whole design your experiment as described in the previous chapter. As soon as you recruited the participants you schedule them. Typically, the following steps are then completed:

- Inform the participants about the study and let them sign the informed consent form
- Perform a survey on demographics and maybe further questions of interest to the experiment
- Give the participants instructions on the task – do not reveal the hypothesis!
- Depending on the study you can conduct a training run first
- Perform the actual run and measure variables
- (Optional) do a survey on subjective measures
- Be available for questions of participants or their (informal) feedback

When all participants finished the experiment, you can start analysing the data you collected.

Participants

Before the study start, you need to think about the sample size you need for your experiment. This number depends on the specific project and goals you have and also the set up and resources. Sometimes you do not have the time to monitor participants for a long time. Generally, the minimal size should be about 10 participants. Try to be pragmatic!

The participants should be representative for the user group (age, background, skills, experience, ...). In most cases, the other people on your team are **NOT AT ALL** representative!

The recruitment can be performed in various ways: sometimes you can use a customer database, you could use market research devices and apps, or search online or in newspapers for volunteers. Especially the online search is risky, because people actively reaching out to you are often not representative.

Hypotheses

The hypothesis of an experiment is always a prediction of the result. With it, you state how a change in the independent variables will influence the measured dependent variables. To prove your hypothesis, you need to conduct an experiment.

Usually, for every experiment, you state a **null-hypothesis**. This predicts that there is no effect of the change in the independent variable on the measured variable.

As soon as this is defined you can start carrying out the experiment and collect your data. Using statistical measures to disprove the null-hypothesis. Only when a statistical test shows significant difference it is probable that the effect is not random. However, you should always consider the significance in the context of the measured effect size! The larger the sample size, the more likely it is to find statistical significance, but the measured effect size might be extremely small.

Summary

- The experiment should be set up to be **reproducible**!
- Main factors of an experiment: **Your participants, the independent variables, and the hypotheses you stated**
- Always **state the hypotheses**: What do you want to proof?
- **Find the variables**: Which are varied? Which are measured?
- Find the participants: **Representative for the experiment**!
- Fix the **method to use** (between-group / within-groups)

References

1 Carmines, E. and Zeller, R. (1979). Reliability and Validity Assessment. Newbury Park: Sage Publications
2 Colosi, L (1997) The Layman's Guide to Social Research Methods http://www.wiley.com/college/westen/0471387541/instructor/ch02/ar_02.html
3 Field, A. and Hole, G. (2003). How to Design and Report Experiments. Sage Publications
4 Carmines, E. and Zeller, R. (1979). Reliability and Validity Assessment. Newbury Park: Sage Publications
5 Field, A. and Hole, G. (2003). How to Design and Report Experiments. Sage Publications
6 Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale. (1998) Human Computer, Interaction (second edition), Prentice Hall, ISBN 0132398648 (new Edition announced for October 2003)
7 Ben Shneiderman. (1998) Designing the User Interface, 3rd Ed., Addison Wesley; ISBN: 0201694972
8 Meme Obama: https://makeameme.org/meme/your-methods-better