Kasper Hornbæk, Per Ola Kristensson, and Antti Oulasvirta. Upcoming 2025. Introduction to Human-computer Interaction. Oxford University Press.

1 Experiments

Imagine that you have developed a new version of a search feature for an operating system. Users can click a magnifier glass icon and type in a search query, results would be listed underneath. You are certain about the benefits of the new design, but you need something to convince other people. What kind of test could show, convincingly, whether users actually like the feature better than the previous version? You could ask people to use the new feature and inquire if they like it. But that would not help you relate it to the previous search feature; you would not know if it is better. You could ask users of the old feature and users of the new feature to rate how well they liked the respective search features and compare those ratings. But even if a difference was found, an inconvenient alternative explanation would be there: Perhaps the users in the two groups were different, for example, in their experience or age, and that would illustrate the difference. You could also seek expert opinion, consulting colleagues and HCI researchers to learn which search feature is best. However, this may be difficult to assess, even for experts. In the worst case, it might turn into a clash of opinions where everything is trumped by the HIPPO, or the highest-paid person's opinion (see [**?**]). What you need is a method that allows you to firmly attribute an observed difference to the new search feature and nothing else. That method is called an *experiment*.

An experiment is "a study in which an intervention is introduced to observe its effects" [**?** , p. 12]; see also Figure 1.1. An experimenter changes something, or intervenes, while keeping everything else the same, and observes the effect of the change. An experiment is a deliberate change in circumstances: The experimenter imposes some condition or constraint in it. Such intervention may be of a variety of kinds; in HCI it is often a technology, but could be different kinds of training, user groups, use situations, or tasks. In common practice, an intervention is designated as a level of treatment (e.g., comparison of user interface designs), group (e.g., comparison of two age groups), or condition (e.g., comparison of different instructions to users).

The design of experiments boils down to defining an intervention and what is being measured. *An experimental design* associates variables defining the intervention (independent variables) and what is being measured (dependent variables). Something that is systematically varied in an intervention is called an *independent variable*. Consider, for example, changing the color of a button as one independent variable, or the age group of the user. The effects of the intervention are measured as *dependent variables*, those that depend on the intervention. For example, one could measure task completion time or errors. In HCI, measures are often related to usability or experience of the technology.

If the relationship between the dependent and the independent variables was fully under the experimenter's control, the observed changes in the dependent variables could be attributed to the intervention, and to that only. However, experiments with human

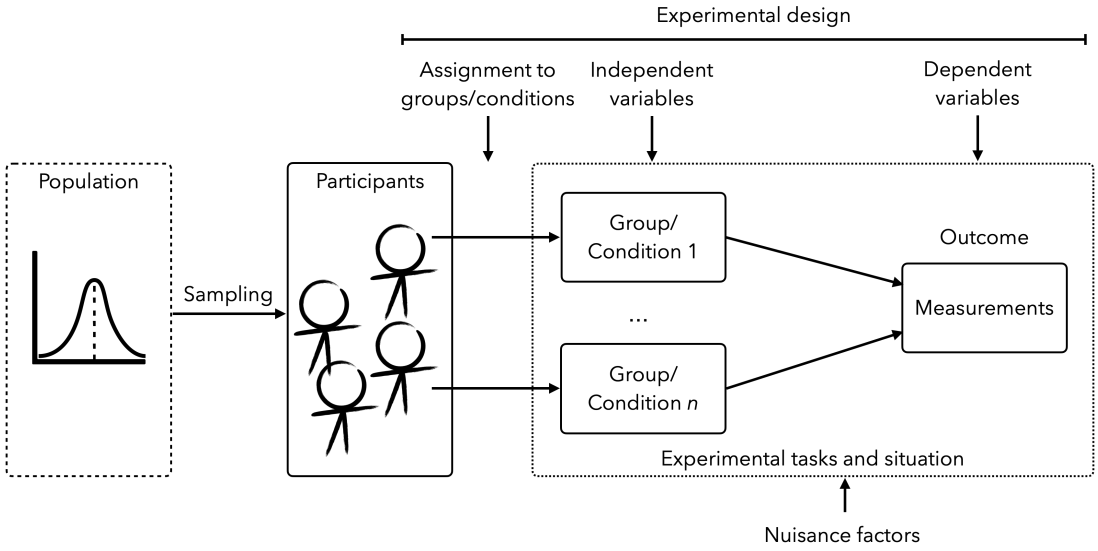


Figure 1.1: The main components of experiments.

participants need to deal with a plethora of *other* factors besides the independent variable. Image a study where users were first asked to carry out a task A with a user interface and then task B. Here, the order of the interventions interacts with users' learning. Any measurements in task B would be affected by what users learned in task A. In general, such factors influence the situation under study, and thus potentially affect the dependent variables. Such factors are called *nuisance factors*: a nuisance in the sense that a factor like this threatens the attribution of cause to the intervention. Was it the task or was it learning that caused the observed difference? Experimental designs have many ways to deal with nuisance factors: controlling such factors, holding them constant, or distributing them randomly across levels of the independent variable. Consider, for example, getting rid of the effect of users in a study having seen the old features in their work. If we would like to get rid of that effect, we talk about controlling it.

Finally, the choice of interventions and measurements must not be arbitrary. *Hypotheses* are statements that connect variation in independent variables with expectations about variation in the dependent variables. Would you expect your new feature to be better than the baseline design in usability; and if so, why? Explicating hypotheses is critical for high quality evaluations. Hypotheses can avoid being fooled by observations subject to noise and error, to avoid being biased by one's own intuitions, and to avoid second-guessing.

Note that the above definition excludes the understanding implied in some common usages of the word experiment, including that of "trying something new" or "an innovative act or procedure". In contrast, experiments in the context of evaluation aim to *establish causal conclusions about which factors influence a situation.* Experiments try to rule out alternative explanations besides the factors being manipulated. This chapter concerns how to enable such conclusions to be drawn, in particular about the use of interactive computing systems. The following subsections explain these components of the experiments in more

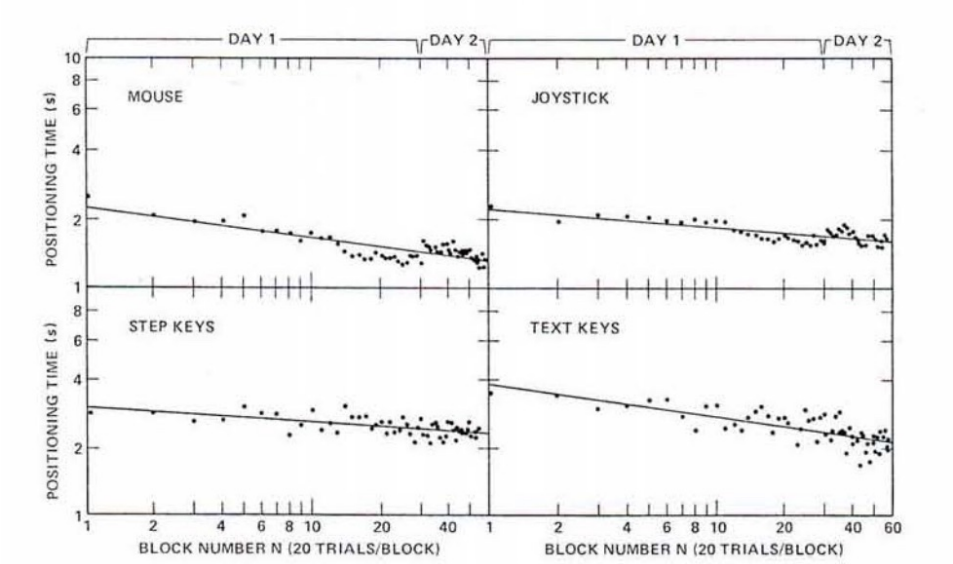
detail; the box below summarizes an early and influential experiment.

Paper Example 1.0.1 : Development of an experimental paradigm for evaluating input devices

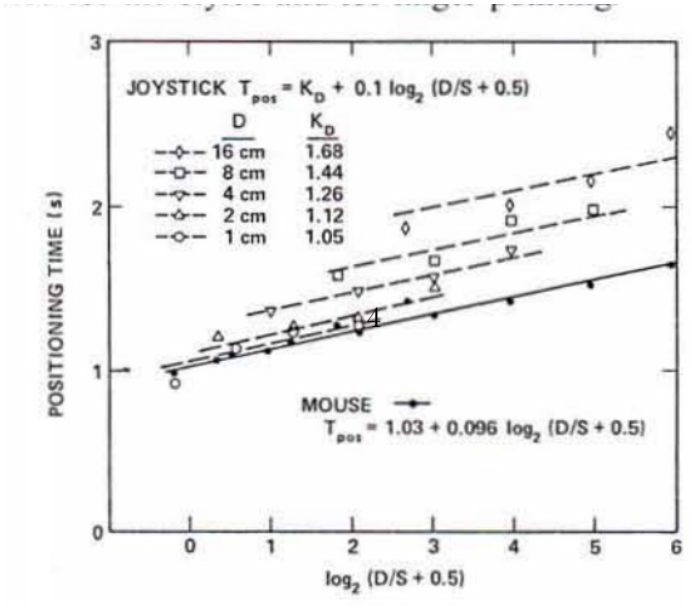
In 1978, **?**] reported a now classic experiment with input devices. The study contrasted the efficiency of input techniques for selecting text. At that time, such comparisons were rare; the paper was among the first in HCI reporting experiments involving a mouse and using Fitts's law [**?**] (see **??**).

The authors had five participants use a mouse, a rate-controlled isometric joystick, and two variants of keys; participants used all four. In the experimental task, the participants had to select highlighted text at varying distances (1 to 16 cm) and sizes (1 to 10 characters). Participants used each device until they did not improve in performance. They did from 1200 to 1800 selections for each device, or four to six hours of pointing.

This experimental design allowed Card and colleagues to plot the development in learning to use the device. They did so in the plot below, which illustrates a power low of practice: $T_n = T_1 - n^{-a}$, where T_n is the time used at trial n and a is a constant. In a log plot this shows as a line whose slope indicates the learning rate.



The authors also use the experiment to plot the positioning time (their main dependent variable) against the index of difficulty – a measure combining the size of targets and the distance between them, two independent variables.



1.1 Research Questions

Planning an experiment, even for evaluative purposes, should always depart from the *research questions* one wishes to address. A research question is a "stated lack of understanding about some phenomenon in human use of computing, or the stated inability to construct interactive technology to address that phenomenon for desired ends" [**?**]. In short, experiments are motivated by *knowledge gaps*. For example, in the example starting this chapter, this knowledge gap concerned the effect of the new search feature on usability.

More broadly, research questions can be divided into three classes: (1) empirical questions: phenomena and effects in human-computer interaction, (2) constructive question: the ability to construct systems and designs with desirable properties, and (3) conceptual questions: relationships between theoretical constructs that represent interaction. Experimental research can serve all three. In addition to empirical questions, we can conduct an experiment with the purpose of setting objectives for design (constructive), or to distinguish between competing theories (conceptual).

But how to come up with a *good* research question? One can think about this by starting form the opposite: a poor research question. Let us entertain two common *objections* to a finished and written up experiment: (1) "so what" and (2) "no surprises" [**?**]. The "so what" objection suggests that the imagined results of an experiment should be interesting and nontrivial; they should matter to theory or practice. Even if running and analyzing the experiment proceeds as imagined, will people find it interesting? Will it add to our understanding of HCI in important ways? Every so often, this objection is voiced by reviewers as "this is not significant", meaning that while the findings are novel and valid, they do not add to the research literature in a substantial and important manner.

The "no surprises" objection suggests that the results should add to or depart from what we already know; they should not be predictable given earlier studies. One should not do an experiment if the results are clear in advance. For instance, if a simple predictive model shows a user interface superior to another or if a technology is without doubt superior to an alternative, then the experiment does not have the possibility to surprise us. Sometimes, of course, new technologies, use situations, or user groups may make it hard to know if earlier findings or theories apply. The "no surprises" objection may be raised both because the experimental setup is biased (we will discuss how to avoid this later) and because the results are easily predictable from the literature (we will also discuss how to avoid this later). Most importantly, both of these objections should be considered *before* deciding to run an experiment.

The experimental method is only one of the many approaches available in evaluation. The decision to choose should be based on careful consideration of the pros and cons. And this boils down to the research questions. Following the discussion of **?**] in **??**, one reason is that experiments maximize precision, but this is achieved at the expense of generalizability and realism. Experiments also allow precise manipulation of tasks and settings, as well as detailed data collection. This allows us to understand *specific mechanisms* involved in interaction. Experiments also allow us to control external factors that may be hard to exclude by other means. For example, if you want to evaluate

user performance of a mobile application, you may want to prevent notifications and multitasking during your measurements, because they would add noise to the data. They also allow us to collect fine-grained data about how users behave with an interface.

Laboratory conditions permit the use of measurement devices such as eye-trackers, motion trackers, and physiological sensing like electromyography (EMG), which annot be easily deployed outside laboratory conditions. If these qualities are important for your research questions, the experimental method is well suited. Another reason is that experiments allow us to investigate research questions about the use of technology without actually deploying it. This is valuable when a prototype does not have all features a real product would need. Experiments are also efficient: they allow us to 'compress time' and study phenomena that occur infrequently. For example, you can arrange a laboratory study that goes through 5-10 tasks within an hour, while the chances of those occurring in real-world use might take days or weeks. Finally, in experiments, we can organize circumstances where we can safely study events that would otherwise be unethical because of causing harm to participants. Research questions along these lines are suitable for experiments.

Another important reason for experiments is the egocentric fallacy [**?**]. According to this fallacy, we tend to overestimate the power of our own intuition of human behavior. However, intuition is weak in discovering latent (unobservable) mechanisms behind our behavior and experience. The mechanisms behind human behavior are beyond intuition. At the same time, we underestimate the extent to which we differ from other people. This is particularly problematic in HCI, where a number of technologies that we propose will have been developed and iteratively refined by ourselves or by close collaborators. Experiments help overcome this fallacy.

In the same spirit, there are also some research questions for which experiments are ill-suited. They rarely work well for studying how people decide to act with technology in real circumstances. See instead the chapters on interviews (**??**) or field studies (**??**).

1.2 Research Hypotheses

Research questions may be further elaborated as *research hypotheses*. Hypotheses are statements that link manipulations of the independent variables to differences in the dependent variables. For example, **?**] hypothesized that "subjects will perceive a computer with dominant characteristics as being dominant" (p. 288). **?**] held the hypothesis that "better support for workspace awareness can improve the usability of these shared computational workspaces" (p. 511). Hypotheses are important also in evaluative studies. However, creating good hypotheses is hard.

Good hypotheses are (a) testable, (b) concise, and (c) name key constructs. The first example given above is testable because one may compare computers with and without dominant characteristics and expect a significant difference in participants' perception of dominance. That example names the key construct dominant, both as something that may be manipulated in computer interfaces (an independent variable) and as something that participants perceive (a dependent variable, assessed, for instance by a questionnaire).

There are many benefits of formulating research questions as hypotheses. First, hypotheses help gain clarity about what one is doing and may help focus a research question. Second, formulating hypotheses helps one think through what earlier work says about the experiment being designed. Third, hypotheses help report an experiment. Fourth, hypotheses are tied to theory. They help think through explanations in advance and allow for genuine surprise about the experimental results.

Not all experiments, however, need hypotheses. The questions that drive the experiments fall in two broad groups, sometimes summarized as "testing theory" and "hunting phenomena" [**?**]. In the former group, there are clear expectations about the outcome of the experiment, typically build on predictions from earlier work. This is often formulated as hypotheses, statements that link the levels of what the experiment is manipulating to outcomes in the measured variables. In the latter group, the experimenter holds less clear expectations and holds more open-minded curiosity. These are also called *explorative* experiments.

1.3 Independent Variables

Independent variables refer to the types of events or factors that we want to draw causal conclusions about. Independent variables can be about anything we can systematically control. They can be about the types of users (e.g., novice, intermittent, expert), types of user interface (e.g., command-line, graphical), form of instruction (offline, online), the type of feedback, and so on.

In selecting independent variables, it is important to remember that the experiments are carried out to *gain information*. Results should not be obvious in advance, and experiments should not be set up to generate winning conditions and losing conditions.

To ensure that we do not pick arbitrary IVs, the choice of independent variables should follow from stated research question. However, that there are many traps in bridging the two. For instance, let us consider again the example from the first part of this chapter (page 1). Recall that we wanted to compare two versions of a search function. In doing so, we need to establish what is considered belonging to the search function: do we want to include the highlighting of search results in our study? Do we want to consider search results opening to a context menu underneath the search button or in a new window, or both? What if the new search function has a case-sensitive option whereas the old does not? For each of those questions, a careless experimenter may invalidate the experiment.

1.3.1 Levels of an independent variable

The *levels of the independent variable* are all possible values that an IV can take in an expeeriment. For example, if your study compares three UI features, the independent variable 'UI feature' would have three levels. If you compare two systems, 'the system' would be one IV with two levels (system A and system B).

1.3.2 Eliminating confounds

One difficulty lies in making levels of the variable (say, two versions of an interface) similar in all essential aspects, except that one is manipulating. If the interfaces one is comparing are not similar, the effects one is studying may be confounded by the dissimilar features. The concerns relating to making conditions similar require an experimenter to ensure that:

- all non-essential aspects are similar. So for instance, if search is supported, it should be across all conditions. If shortcuts are available, it should be in all conditions;
- the screen real estate is similar across interface;
- the training and the users' skill with the variants of interface is similar;
- the setting in which the interfaces is used is similar;
- comparable information available in the interfaces;
- comparable hardware is used (e.g., for input and output);
- the time allotted and the criteria for success are similar across levels of the independent variable.

1.3.3 Selecting meaningful baselines

Another concern is to ensure meaningful baselines. A "strong baseline" refers to a solution that is considered best on the market or literature, the state-of-the-art alternative. This could, for instance, be an interface that implements the typical way of performing a task. Our concern here is to ensure that the baseline (or control) is as strong as possible. **?**] discussed what she termed "Straw Man Comparisons", that is, cases where authors compared their interfaces against outdated work, rather than the state-of-the-art approaches. Although Munzer wrote about information visualization, studies in HCI more generally are sometimes done by comparing novel interfaces with weak or incomplete versions of the current state of the art.

1.4 Participants

The participants in the experiments are the people whose interaction with technology we want to study. They should be seen as a representative sample from the group we want to draw conclusions about. The selection of participants impacts which conclusions can be drawn. It also shapes the practicalities of running the experiment.

The key question is *who* should participate. Representativeness is important. Recruiting a convenience sample – whoever happens to be available – should be done with caution. **?**] found that about half of a sample of studies from the CHI conference used students. Increasingly, participants are also recruited online, for instance, through Amazon Mechanical Turk. The participants hape what happens in the experiment as well

as how well we may generalize findings. We may choose experienced or inexperienced computer users; we may find domain experts or novices. Thus, they should be selected so that we can validly answer the research question we are interested in.

Another important question is *how many* should participate. Typically, the number of participants in studies in HCI is around 12 [**?**]; for controlled experiments where participants are present in person, the number is about 20. However, crucially, this does not mean that 12 is enough for *your study*, only that this number is about the average across most of the published literature.

The principled way of finding an appropriate number of participants is to do *power analysis* [**?**]. Power analysis helps estimate the probability that one detects a difference in dependent variables between the levels of the independent variable if one knows (or can qualify a guess about) the magnitude of the effect one is examining. Power analyses are often a depressing reading. Typically, many participants are required to achieve a reasonable power; say, an 80% probability of finding a difference. To detect medium-sized differences between the two conditions at this probability, one would need 64 participants in each condition in a between-subjects experiment. Medium-sized effects found in the HCI literature include differences between broad and deep menus, or between selection with mouse and keyboards. There are tools to help with such analysis (e.g., G*power).

1.5 Research Ethics

A key consideration in experimental research is the ethical treatment of the participants. They should in no circumstances be harmed. Participating in the study should not have negative consequences on their lives. The principles for the ethical treatment of participants in behavioral research have been established, the Helsinki Declaration on Ethical Principles for Medical Research Involving Human Subjects; recent updates include the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct (http://www.apa.org/ethics/) and the code of conduct for the ACM (https://www.acm.org/code-of-ethics). Research organizations have formal guidelines and requirements in place to ensure respectful, legal, and ethically defensible experiments.

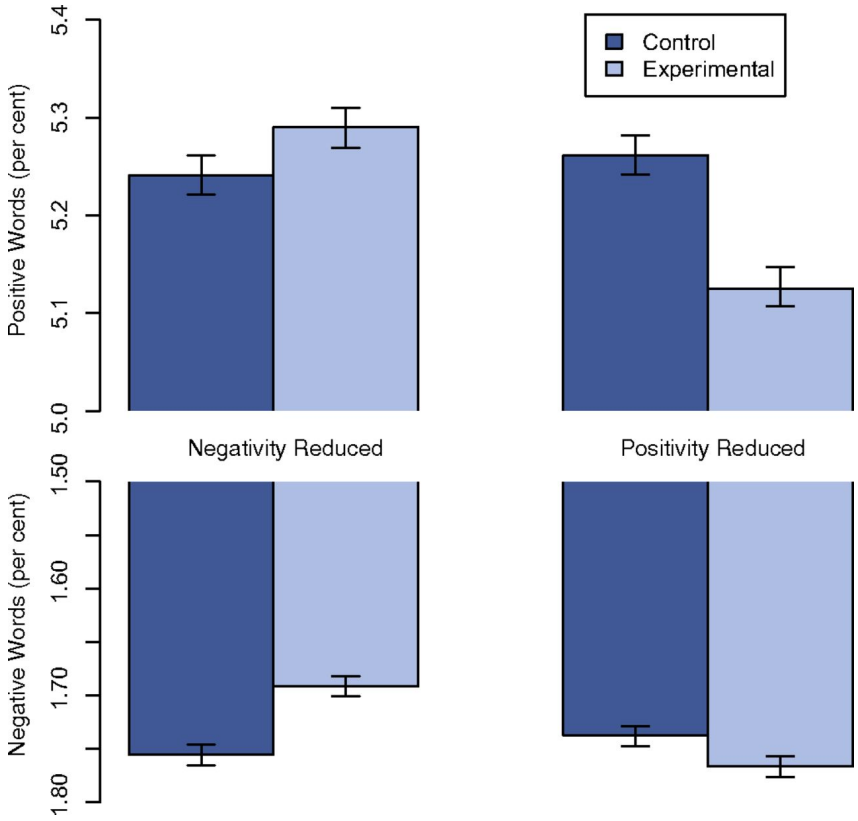
Key goals for ethical experimentation include:

- Treat participants with respect: value their time, honor their opinions, take any criticism seriously.
- Do not expose participants to dangerous or potentially harmful situations; this includes physical, mental, and emotional concerns.
- Make sure participants want to participate; get informed consent (there are templates available online and your institution might also offer one).
- Make sure any reimbursements to participants are adequate: they should not make it necessary for participants to enroll in your study but still compensate for their time and any expenses incurred by participating (e.g., transportation costs).

- Make sure to debrief the participants, explaining them about the purpose of the experiment and taking any questions they might have about your research.

Paper Example 1.5.1 : Controversial Facebook experiments

In 2014, it was reported that a large-scale experiment had been conducted on Facebook which attempted to manipulate the emotions of its users [**?**]. Close to 700,000 Facebook users had their feed of stories manipulated so that they experienced a reduced amount of emotional content. The figure below shows how the manipulation (negativity reduces among one's friends or positivity reduced) change the positive and negative words that users post subsequently (the dependent variable). This figure shows a clear impact of the experimental conditions compared to the control conditions (no change in emotional content. This results are significant because it shows that emotional contagion, having your emotional state changed based on the people that surround you, can happen on social networks and without any direct interaction between people.



However, the paper was controversial. The Facebook users who were part of the study did not give informed consent as would be normally expected in research and did not, in a clear manner, give permission to have their emotions manipulated experimentally.

1.6 Experimental Design

Let us consider again the search functionality example we started with. If you conducted a study where the only functionality being assessed was the new feature, you could not draw a conclusion on whether it improved over the original or not. You would not have a baseline to compare against. This is an example of why we need to introduce *experimental conditions* that allow us to answer our research questions. The variables to which we map those conditions are called independent variables. They are called independent because the observations collected in those conditions are independent of each other. This is ensured by allocating participants either systematically or randomly to those conditions.

Experiments in HCI almost always involve more than one level of an independent variable, for instance two alternative interfaces for a task, a range of different instruction materials, or different approaches to delivering notifications. The assignment of participants to the levels of the independent variable is the main consideration in experimental design. The aim is to ensure internal validity, that is, the ability of an experiment to attribute differences observed in the dependent variables to manipulations of independent variables [**?**]. Furthermore, experimental design need to consider subsequent running and analysis of experiments. Simple designs are typically easier for participants; the statistical analysis and interpretation are easier for experimenters.

One key decision is whether participants experience all or just one level of the independent variable. The former type of design is called *within*-participants, because the independent variable is varied for each participant, the latter type is called *between*-participants.

In within-participants experiments, participants serve as their own control. So even if a person varies in some trait or behavior, a so-called wild-card participant, that variation is cancelled because the participant uses all levels of the independent variable. However, those types of designs are not without problems. For instance, within-participant experiments suffered from learning effects. Participants may learn about the interface or the task, and therefore their experience or performance might change during the experiment.

Between-participant experiments have many benefits. They are easy to analyze and they do not suffer from the possibility of influence across conditions because participants use just one level of the independent variable. The key drawback of between-participant experiments is that they cannot control for individual differences and therefore require more participants.

As suggested in Figure 1.1, other factors in experiments may influence the experimental situation. There are several ways to deal with these. The workhorse of experimental design is *randomization*; this is often cited as a defining characteristic of experiments. Randomization means that participants are assigned to conditions at random. Thereby, the influences of other factors than those being manipulated are randomly distributed over conditions. One of the authors of this book has the motto "when in doubt, randomize". *Control* means simply restricting the variable to one level; one may experiment, for instance, with only left-handed persons. Thereby, the influence of handedness can be ignored in the analysis of the experiment.

Sometimes experiments have more than one independent variable, which complicate

their design. For instance, an experimenter may want to study three interfaces (say, a command line versus a graphical user interface versus a non-computer method). There are easy ways to combine such variation in independent variables, called Latin squares. If we use the first letters of the interfaces (C, G, N) as shorthand, we could have one use do C, then G, then N. Another could do G, N, C. Another, N, C, G. That organization protects against a number of issues in the experimental design by having an equal number of users use each interface first, second, and last (you can see this from Latin squares in that the count of interfaces in each column is the same). More complex ways of setting up experiments. For instance, if we want to compare the interfaces just discussed across three ways of training (hints vs paper manual vs integrated manual), then we can use a Greco-Latin square. It could lead to us organizing the experiment as follows:

```
           1    2    3
User A: C+H  G+P  N+I
User B: G+I  N+H  C+P
User C: N+P  C+I  G+H
```

Again, the sum of types of training is similar across columns and each combination of training and interface occurs just once. Generators for Latin and Greco-Latin square designs may be found online.

1.7 Dependent Variables

In experiments, the aim is to understand how the independent variable influences the interaction. By convention, we call measures of this influence *dependent variables* because they depend or result from our manipulations of the independent variable. Another way to think of dependent variables in HCI is that they indicate the quality of the interaction numerically, say with the duration or accuracy of the interaction.

A crucial question for dependent variables is *conceptualization* [**?**]. Conceptualization concerns making the meaning of concepts in an experiment's research questions clear, defining them precisely, and separating different dimensions of meaning. For instance, while the learnability of a user interface is (superficially) easy as a dependent variable, defining it is much harder. **?**] showed how the literature displays many understandings of learnability. If an experiment does not clearly conceptualize learnability, the validity of any inference from that experiment may be reduced because learnability may mean many different things. Similarly, task completion time is easy to measure in many experiments. But it may not be the best conceptualization of quality of an interface. For instance, studies vary in whether they see low task completion times as good (minimizing resource expenditure) or bad (expressing a lack of engagement); see **?**]. Unthinkingly measuring task completion time therefore reflects too poor a conceptualization of quality. Another example is the notion of user friendliness, which has been used in HCI for decades. On the surface, it may seem like a natural ingredient of a research question and therefore as a dependent variable. But it is difficult to define and it is hard to separate its dimensions. Conceptualization shows that it is a dead end for many experiments.

Several tools of thought help conceptualizing dependent variables. First, one may refer to the models and findings in previous chapters on user experience, usability, performance, and collaboration. Each of these represents important concepts that we may use to make research questions more precise. For instance, the discussion of independent dimensions of usability suggests that we should consider whether we are thinking of effectiveness, efficiency, or satisfaction when we are experimentally trying to find out which of a a set of interfaces that is more usable. Or perhaps we need to consider all of them. Second, Newman and Taylor proposed to think about the critical parameters of a certain situation [**?**]. A critical parameter is a performance indicator that captures aspects of performance that are critical to success, which is domain or application specific, and which is stable over variations of interface. Part of the challenge in applying catalogs of measures is to ensure that at least some measures chosen are critical in the above sense (and not just generic time or error measures).

A second crucial question for dependent variables is *operationalization* [**?**]. Operationalization is about turning the concepts that our research question name into something we may measure. The main consideration concerns the extent to which the actual measures collected reflect what the experimenter wishes to measure, or whether it is possible to make "inferences from sampling particulars of a study to the higher-order constructs they represent" [**?** , p. 65].

We may ponder the following questions in operationalizing dependent variables. First, how will we actually obtain the measure. Second, use validated measures and questionnnaires. Third, multiple measures of the same construct increase reliability and strengthen the validity of claims about constructs. Using just one operationalization of the construct faces a mono-method threat to validity [**?**]. It means that we are more prone to not measuring what we think we are measuring if we use just one indicator for a construct.

Most measures in research studies in HCI are task completion time, accuracy or error rates, and questionnaire answers. However, a couple of additional types of data are worth collecting in experiments. One important type is about the *interaction process*, for instance, which commands participants activate or how they move their mouse. Such data may help us understand the interaction process (rather than just the outcome) and may help us think about why something happens in an experiment.

Whereas dependent variables need to be numeric, many exemplary experiments also collect *qualitative data*, for instance in the form of interviews and observations. Some experiments also rely solely on qualitative data. For instance, **?**] reported a much-cited experiment on reading from paper and from a computer. While they used an experimental setup—using for instance random assignment of participants to either paper or a computer condition—they only reported qualitative data on reading strategies and activities that differed between paper and computer. Such data is valuable when experiments go well (as in O'Hara and Sellen's study), but it is also useful in understanding why an experiment failed.

1.8 Experimental Situation

Experiments put participants into *experimental situations*: the particular circumstances where they are asked to carry out tasks. The design of these situations matters.

One decision concerns the *activities* that participants will engage in. They are often prescribed as tasks. One may select tasks in many ways. One is to select tasks that are representative of what users would do outside the experiment. **?**] discussed selection of tasks in information visualization and wrote 'A study is not very interesting if it shows a nice result for a task that nobody will ever actually do, or a task much less common or important than some other task. You need to convince the reader that your tasks are a reasonable abstraction of the real-world tasks done by your target users " (p. 147). One way of ensuring representativeness is to use tasks that users have been observed doing.

Another approach to selecting tasks is to use simple tasks that capture the essence of what is being investigated. The idea is to reduce variation and remove non-essential features of a task; this idea is similar to the approach for selecting independent variables that was earlier referred to as essential features. For instance, many studies of pointing techniques use the ISO multidirectional tapping task. This task type requires participants to tap circular or square targets arranged in a circle. It does not represent pointing in the wild, but is widely accepted as a useful task for experiments.

Another decision about the experimental situation concerns whether the experiment take place in the lab or in the field. In lab experiments, the setting is controlled and the effect of external influences minimized. In field experiments, the setting is real, although the experimental manipulations are instigated by the experimenter. The view taken here is that neither choice of setting is better than the other; rather, they have relative benefits and drawbacks. We discuss field experiments in **??**.

1.9 Analysis and Interpretation

What should one do with the collected data? The naive approach would be to take means of data in different conditions and compare them. What could go wrong with this? If users report an experience of 3.5 in condition A and 4.2 in condition B, is this not a sufficient basis to conclude that the one is better than the other?

Perhaps the most critical piece of knowledge in analysing data concerns *variance*. Every observation in an HCI experiment is susceptible to variation. Observations are compromised by variation in repeated attempts of users, our measurement instruments, and various random effects. If variance in the two conditions A and B is large enough, it could be that the means differ because of chance. In other words, if you were to repeat the experiment, the result could be different, even flip.

Statistical analysis gives us rigorous tools to understand what we can conclude from data. Statistics is the science of drawing valid conclusions from datasets. In HCI, we use statistical analysis for different purposes, including:

- Exploring and learning about the distribution of variables or their relationships;
- Describing relationships between independent and dependent variables;

- Testing if relationships describe reliable differences in the population from which the sample was taken;
- Identifying a factor that caused or contributed to an observed effect;
- Testing if a statistical model accurately describes the dataset.

In this book, we do not offer a comprehensive overview of statistical methods for HCI. We here review some of the more popular frequentist methods and thinking behind them. For a more thorough treatment, see the book on the topic by **?**]. Recent research has also looked at Bayesian methods for statistical testing, such as the Bayes factor. An in-depth treatment of frequentist and Bayesian statistics can be found in **?**]. Note that we discuss the analysis of qualitative elsewhere in the book: In chapters on interviews, think-aloud protocols, and conversational analysis.

Statistical analysis is divided into two main classes according to purpose:

- Descriptive statistics, where the goal is to describe relationships between variables in the dataset;
- Inferential statistics, where the goal is to draw a conclusion about the population from which the sample was drawn.

It is a good practice to start analysis by describing the dataset. *Descriptive statistics* refers to the use of summary statistics, like graphs, tables, and models, to describe a set of data. For example, imagine you had collected data on accuracy and speed of pointing with two input methods A and B. What you should do is to plot the distributions of the two dependent variables for the two methods. What could you learn? For example, are the distributions normally distributed, are they skewed, how much variance is there? Based on this information, you could produce *summary statistics*, such as mean, median, min and max of each variable. *Graphs* can then be used to visualize them further, for example histograms, scatter plots, line plots etc. Bivariate graphs show relationships between two variables, trivariate among three, and multi-variate more than that.

Inferential statistics refers to the attempt to generalize observations in a sample. A distinction is made between the set of observations and the population it comes from. For example, in the case of the search functionality example, one may be interested in estimating what task completion time is for regular users (population), not just the one recruited to the study (sample). Obviously, if the sample is not *representative*, conclusions drawn based on it can be flawed. There are many reasons why the sample may be unrepresentative. For example, we often use university students to represent regular users. However, they may differ in many respects: age, socio-economic and background, etcetera. This is why it is important that participants and tasks are sampled such that they represent the population. There are several strategies to ensure that: random sampling, stratified sampling, and systematic random sampling, for example.

Inferential statistics may also start by plotting. *Confidence intervals* provide estimates for the range of values that we think the true population value should fall in. For example, if the 95 % confidence interval of our task completion time variable was [14.5, 17.9], it

would mean that we are 95 % confident that the true population value was between 14.5 s and 17.9.

Statistical testing refers to testing if a difference exists between conditions. Since we are talking about inferential statistics, we are not interested whether the difference exists in the dataset, because there almost always is *some* difference, but whether this represents a true difference in the population. Here, the research hypotheses need to be translated into statistical hypotheses, which can then be tested.

Multiple methods exists for statistical testing. They can be divided into two main groups: parametric and non-parametric. Parametric tests use parameters to describe the population. They make distributional assumptions about the population, which must be first checked in order to proceed to use the corresponding test. The most commonly used parametric tests are t-test and its generalization ANOVA. *Non-parametric tests* make no assumption about the underlying distribution, which makes them more flexible as a class of tests.

Regression models help us understand the nature of relationship among two or more variables. In a regression model, we try to understand the relationship between *predictor* and *response* variables. Typically predictor variables would be our independent variables, and response variables the dependent variables. However, any *covariant*, or uncontrolled but recorded variable, could be included, too. For example, we could use regression to find a relationship between age and task completion time. If the p-value of the regression was below an apriori threshold (often 0.05), we can conclude that there would be a significant trend.

1.10 Hypothesis Testing

A very common approach to analyze experiments in HCI is through *hypothesis testing*.

1.10.1 Statistical significance testing

The high-level logic of a hypothesis testing is as follows. Assume you have a sample X with n measurements, $X = \{x_1, x_2, \ldots, x_n\}$.

We now assume that this sample was drawn from a particular distribution, let us call it N. We call this assumption the *null hypothesis* and it is typically denoted H_0.

We can now ask whether our sample was indeed drawn from a particular distribution N and we do this by asking whether it is possible to *reject the null hypothesis*. This means that we are asking whether we can, with some probability, state that the sample *was not drawn* from the distribution N.

While not a typical significance test in HCI, let us consider the distribution N we are interested in to be a standard Normal distribution $N(x) \sim N(0, 1)$. We can then ask whether an individual measurement x in our sample is drawn from N. In this case, this means x should not deviate much from 0. If a measurement x in our sample is large, say 3, then the probability that it is drawn from a standard Normal distribution is very low. However, if a measure x in our sample is small, say 0.3, then the probability is quite high. This can be readily realized by considering that the range $|x| \leq 1$ occupies 68% of the

probability mass of a standard Normal distribution. In other words, the higher x is, the more confident we can be that x was *not* drawn from a standard Normal distribution.

In hypothesis testing we define a criterion value for determining whether the probability that x is not drawn from N justifies rejecting the null hypothesis. Common criterion values are 0.05, 0.01, and 0.001. These criterion values are called *confidence levels*. For example, if we set the confidence level to 0.046, then we would reject the null hypothesis for a measurement x if x different from the value $x = 0$ by 2. That is, for any value of $x = 2$ or higher would give us cause to reject H_0 at significance level 0.046.

Statistical significance tests perform this type of hypothesis testing. However, they also take into account additional factors, such as several measurements in the sample, and they consider more appropriate distributions than a standard Normal distribution.

1.10.2 Example: between-subjects analysis of variance

Assume there is a difference in the measurements we obtained via the dependent variables when we manipulated the independent variables. This difference can be due to two things.

1. Our manipulation of the independent variable.
2. *Error*, which in this context means there is no true difference. Instead, the difference we measured was just due to random chance.

Significance tests help us decide whether measured differences are statistically significant, meaning that we can be reasonably confident that the difference is not due to chance but due to our manipulation of the independent variables.

Now assume we have sampled two groups from a user population and we exposed each group to a different method, say Method A and Method B. The *method* is then our independent variable with two levels: Method A or Method B.

We now believe a right way to compare these methods is to investigate if the means of the measures we have taken differ between the two groups. The null hypothesis H_0 says that for some predetermined confidence level there is no actual difference between the means and any measured difference is solely due to sampling error. If we reject the null hypothesis H_0 then we have a significant result at said confidence level.

We will demonstrate statistical significance testing in HCI by using a method called one-way analysis of variance (ANOVA). It is a significance test used to determine whether two sample means are significantly different in the statistical sense. The sample means must have been generated from a between-subjects experimental design.

The statistical term *error* is the amount an observation differs from the population mean. Typically the population mean is *unobservable*.

The statistical term *residual* is the amount an observation differs from the sample mean. Unlike the population mean, the sample mean is *observable*.

Assume we have obtained samples from a Normal distribution: $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$. Then the sample mean is:

$$\bar{X} = \frac{X_1, X_2, \ldots, X_n}{n} \tag{1.1}$$

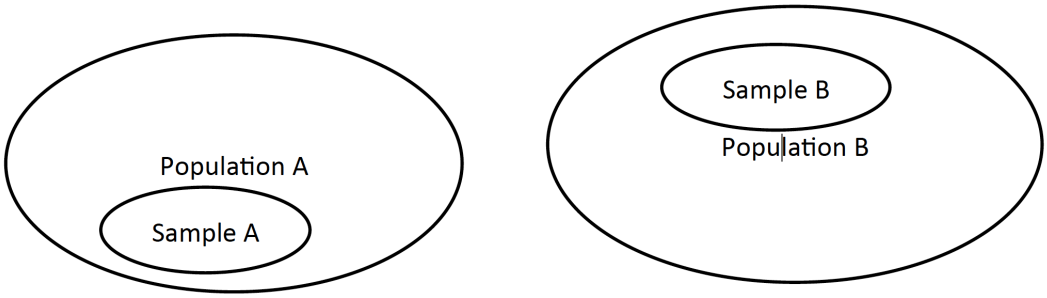


Figure 1.2: An illustration of the acquisition of two samples from two populations.

and the error is $e_i = X_i - \mu$ and the residual is $r_i = X_i - \bar{X}$.

Now, *why* would there be a difference between the sample means of group A and B (Figure 1.2)? There are two possible reasons:

1. *Because of* group membership. This means the difference is due to an effect of the independent variable on the dependent variable.

2. *Not because of* group membership. This means the difference is merely due to sampling error.

The logic of ANOVA is as follows. There are two independent estimates of the population variance that can be obtained: (1) a between-groups estimate, which is the effect of the independent variable *and* error; and (2) a within-groups estimate, which is just error.

Our null hypothesis H_0 is that the two populations A and B have equal means:

$$H_0 : \mu_A = \mu_B \tag{1.2}$$

Given H_0, the between-groups and within-groups variance estimates should be equal. This is because H_0 assumes the effect of the independent variable does not exist. Then both variance estimates reflect error and their ratio is 1. A ratio larger than 1 suggests an effect of the independent variable.

A *sum of squares* (SS) is simply the sum of the squared residuals:

$$SS = \sum_i \left(X_i - \bar{X}\right)^2 \tag{1.3}$$

Now let us consider the sources of variability in a between-subjects analysis of variance. The total variability *within* Sample A is:

$$SS_A = \sum_i \left(X_{Ai} - \bar{X}_A\right)^2 \tag{1.4}$$

The total variability *within* Sample B is:

$$SS_B = \sum_i \left(X_{Bi} - \bar{X}_B\right)^2 \tag{1.5}$$

Finally, the total variability *between* Sample A and Sample B is:

$$SS_A = \sum_i \left(X_{Ai} - \bar{X}_{AB}\right)^2 + SS_B = \sum_i \left(X_{Bi} - \bar{X}_{AB}\right)^2 \tag{1.6}$$

We can now define the variability due to error. This is the total variability *within* Sample A and Sample B—this variability is not due to manipulation of the independent variable and is thus regarded as a source of *error*:

$$SS_{error} = SS_A + SS_B \tag{1.7}$$

The total variability *between* Sample A and B is:

$$SS_{total} = SS_{A+B} \tag{1.8}$$

Finally, The *effect* is the part of the total variability that cannot be explained by the source of error:

$$SS_{effect} = SS_{total} - SS_{error} \tag{1.9}$$

What we have effectively done is we have partitioned the different sources of variability in the samples: (1) variability due to error, that is, the sum of the variability within each group; (2) total variability across both groups; and (3) variability due to an effect of a manipulation of the independent variable—variability that cannot be explained by error. This partitioning is shown graphically in Figure 1.3.

We have found a way to separate out the error from the effect in the data. We first measure the variability of the data within each group (group A and B separately). This gives us the error. Thereafter we measure the total variability (collapsing group A and B into a single group). This gives us the effect + error. We can now obtain the effect by subtracting the error from the total variability.

The sums of squares provide unscaled measures of variability in the data. This can be readily observed because as we keep adding more and more summands the sum becomes larger and larger. Since sums of squares are unscaled they need to be eventually normalized so that we can compare different sums of squares.

Scaled sums of squares are called *mean squares* (MS). Mean squares are obtained by scaling sums of squares by their degrees of freedom (df).

First, we have the degrees of freedom df_{error} within group A and group B, recall this is the error. df_{error} is the number of ways you can arrange the residuals and still have them sum to zero for each group:

$$df_{error} = n - \text{participants} - m \text{ groups} \tag{1.10}$$

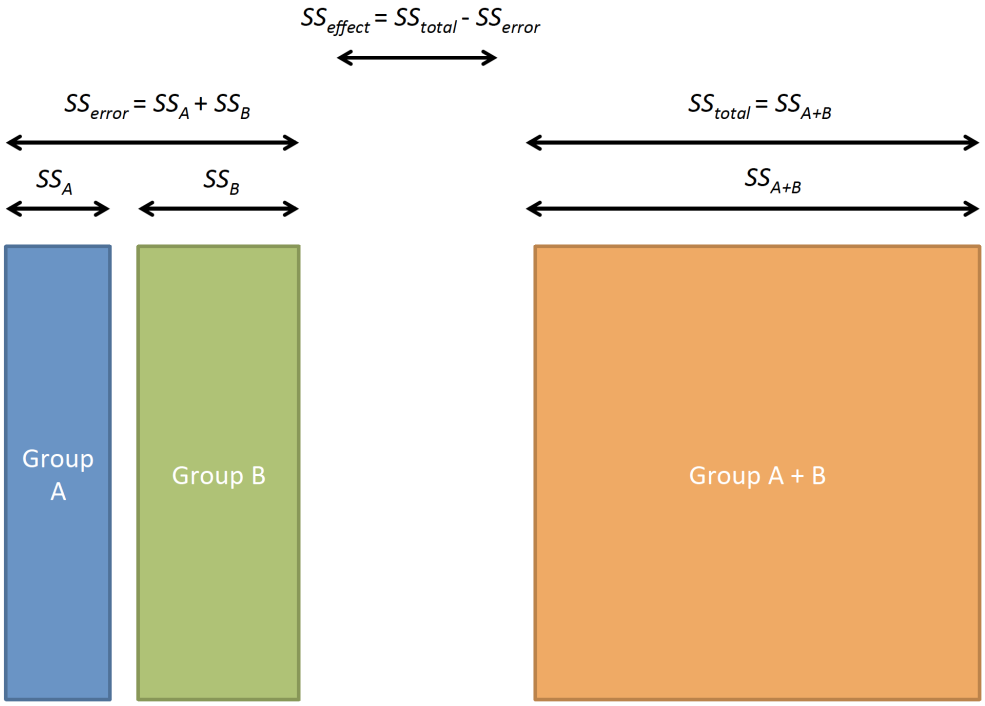


Figure 1.3: A graphical illustration of how one-way ANOVA partitions sums of squares.

Then, we have the degrees of freedom df_{effect} between group A and B, recall this is the effect. df_{effect} is the number of ways you can arrange their deviations away from the mean so that their average always sum to zero:

$$df_{effect} = m \text{ groups} - 1 \tag{1.11}$$

Now we can obtain the mean squares for error and effect:

$$MS_{error} = \frac{SS_{error}}{df_{error}} \tag{1.12}$$

$$MS_{effect} = \frac{SS_{effect}}{df_{effect}} \tag{1.13}$$

Finally, we can calculate the F-ratio, frequently referred to as the F-statistic:

$$F = \frac{MS_{effect}}{MS_{error}} \tag{1.14}$$

The F-ratio will become large if the effect is larger than the error. Vice versa, if the F-ratio will become small if the effect is smaller than the error. This is because the ratio of the between-groups estimate and the within-groups es/mate gives rise to an F-distribution when H_0 is true.

The F-distribution varies as a function of a pair of degrees of freedom, one for each of the variance estimates.

1.11 Experiments need Explanations

The primary goal of evaluation is to estimate the value a design offers to users. Often this goal is better achieved if the obtained result can be explained.

A small thought experiment illustrates this. You run an evaluative study of a prototype design and find that average task completion time is 47.5 seconds, average error rate 0.5, and SUS (system usability scale) 65. What do such results tell you and how confident can you be when taking further actions based on them?

Explanations help understand distributions of dependent variables. Why, for example, had task completion time an average 47.5 s and not, say, 14.4 s? Data collected during the experiment may offer explanations. Verbal protocols, video recordings, interviews etc. can illuminate *explanatory mechanisms* that link independent and dependent variables. Often in HCI experiments, we discover usability problems, conceptual misunderstandings, or issues in perceptual or motor performance.

Without such explanations, results may be 'fragile'. There is a threat that the obtained results do not generalize. Quantitative findings may be underpinned by factors that are contingent (dependent on) experimental conditions. If those conditions change, even slightly, the result may change, too. For example, if your users are not native speakers, perhaps they failed to find an item because they missed the meaning. Now, were you to deploy your system to native speakers, the results might change entirely.

We may also seek explanations outside of data – from theories. For example, theories of cognition may help understand why users have hard time recalling facts or events, and theories of communication may expose why users do not want to engage with a user community. For example, **?**] used decision-making theories from psychology and economics to explain user behavior in using intelligent text entry systems. Because the effect of an intelligent text entry method can be fleetingly small, and contingent on a number of factors, it is important to understand the mechanisms that underpin a user's decision to use a method in a particular way.

To sum up, results of experimental research are more robust and generalizable if they can be explained.

Summary

- Experiments helps give precise measurements and comparisons.
- Explicating research questions and hypotheses is important for high quality evaluations.
- Validity and reliability are key concepts for ensuring trustworthy experiements.
- Data analysis uses methods from descriptive and inferential statistics to draw conclusions about the effects of independent variables on dependent variables.

Exercises

1. Understanding experiments. For each of the questions below, discuss if an experiment is a suitable evaluation approach.
   a) What does it feel like to interact with a chatbot?
   b) Why do people not upgrade software on their devices?
   c) How quickly can people input text on a QWERTY keyboard?
   d) How do we figure out how much people use their mobile phones?
   e) What is the most effective way of organizing email?
2. Formulating testable hypotheses. Please create a hypothesis for an experiment investigating the effects of embodying different avatars in a virtual reality game. Check if you have clear dependent and independent variables. Explain these links with theoretical hypotheses if you can.
3. Experimental design. Please design and run an experiment that will investigate the appropriate mid-air gestures turning on and turning off a television.

4. Comparative studies. You have been asked to run a study that investigates whether a mobile phone application for exercising more during the workday is better than a baseline application. Please consider to think about "better' and how to select dependent variables for the study. One of your collaborators suggested using step count; what are your considerations about validity for this dependent variable? Another colleague considers how to capture the subjective of exercising; how do you engage in discussing this consideration?

5. Measuring user experience. Everybody wants satisfying user interfaces. Imagine that you are conducting an experiment where you are interested in satisfaction. Consider how you would operationalize that in the experimental setup that you imagine. What are the essential and less important aspects of satisfaction? How will you measure them?

6. Improving experimental designs. Read the paper by **?**] and come up with an improved version of the experiment.

7. Reflecting on evaluative practices. Describe how you have previously evaluated designs or software, for example on classes or in your work. Then compare that to the usability testing method as described in this chapter. Give a concrete example and tell how that was tested. Explain whether it was tested from a human-centered perspective. Describe why, or why not, this was done. Next, compare this previous approach to usability testing. Is testing with people relevant or not? Is it your responsibility? What happens if no testing with people is done?