CHAPTER

Experimental design 3

Experiments help us answer questions and identify causal relationships. Well-designed experiments can reveal important scientific findings. By contrast, ill-designed experiments may generate results that are false or misleading. Experiments have been widely used in the human-computer interaction (HCI) field to develop and modify user models or task models, evaluate different design solutions, and answer various other critical questions, such as technology adoption.

Before we discuss specific experimental design methods, we need to differentiate three groups of studies: experiments, quasi-experiments, and nonexperiments (Cooper and Schindler, 2000; Rosenthal and Rosnow, 2008). Figure 3.1 demonstrates the relationship among the three types of studies. If a study involves multiple groups or conditions and the participants are randomly assigned to each condition, it is a true experiment. If a study involves multiple groups or conditions but the participants are not randomly assigned to different conditions, it is a quasi-experiment. Finally, if there is only one observation group or only one condition involved, it is a nonexperiment. True experiments possess the following characteristics:

- A true experiment is based on at least one testable research hypothesis and aims to validate it.
- There are usually at least two conditions (a treatment condition and a control condition) or groups (a treatment group and a control group).
- The dependent variables are normally measured through quantitative measurements.
- The results are analyzed through various statistical significance tests.
- A true experiment should be designed and conducted with the goal of removing potential biases.
- A true experiment should be replicable with different participant samples, at different times, in different locations, and by different experimenters.

In this chapter, we focus on the design of true experiments, which means that all the studies we discuss have multiple conditions or measures and the participants are randomly assigned to different conditions. We start with the issues that need to be considered when designing experiments, followed by discussions of simple experiments that involve only one independent variable. We then examine more complicated experiments that involve two or more independent variables. Three major types of experiment design are discussed: between-group design, within-group design, and split-plot design. Section 3.5 focuses on potential sources of systematic errors

Research Methods in Human-Computer Interaction. http://dx.doi.org/10.1016/B978-0-12-805390-4.00003-0

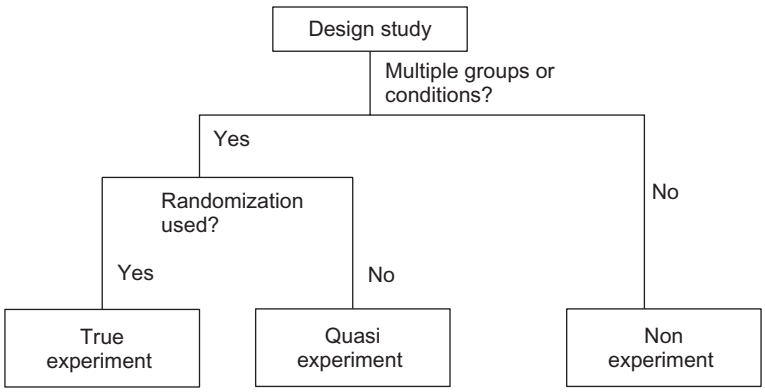

FIGURE 3.1

Defining true experiments, quasi-experiments, and nonexperiments.

(biases) and guidelines for effectively avoiding or controlling those biases. The chapter ends with a discussion of typical procedures for running HCI experiments.

3.1 WHAT NEEDS TO BE CONSIDERED WHEN DESIGNING EXPERIMENTS?

We need to consider several issues when designing an experiment that investigates HCI-related questions. Some of these issues are universal for all scientific experiments, such as research hypotheses, the measurement of the dependent variables, and the control of multiple conditions. Other issues are unique to experiments that involve human subjects, such as the learning effect, participants' knowledge background, and the size of the potential participant pool. Detailed discussions of measurement and generation of research hypotheses are provided in Chapter 2. A complete review on conducting research involving human subjects is provided in Chapter 15.

Most successful experiments start with a clearly defined research hypothesis with a reasonable scope (Oehlert, 2000). The research hypothesis is generated based on results of earlier exploratory studies and provides critical information needed to design an experiment. It specifies the independent and dependent variables of the experiment. The number and values of independent variables directly determine how many conditions the experiment has. For example, consider designing an experiment to investigate the following hypothesis:

> *There is no difference between the target selection speed when using a mouse, a joystick, or a trackball to select icons of different sizes (small, medium, and large).*

There are two independent variables in this hypothesis: the type of pointing device and the size of icon. Three different pointing devices will be examined: a mouse, a joystick, and a trackball, suggesting three conditions under this independent variable. Three different target sizes will be examined: small, medium, and large, suggesting

three conditions under this independent variable as well. Since we need to test each combination of values of the two independent variables, combining the two independent variables results in a total of nine ($3 \times 3 = 9$) conditions in the experiment.

The identification of dependent variables will allow us to further consider the appropriate metric for measuring the dependent variables. In many cases, multiple approaches can be used to measure the dependent variables. For example, typing speed can be measured by the number of words typed per minute, which is equal to the total number of words typed divided by the number of minutes used to generate those words. It may also be measured by number of correct words typed per minute, which is equal to the total number of correct words typed divided by the number of minutes used to generate those words. We need to consider the objective of the experiment to determine which measure is more appropriate.

Another issue to consider when designing experiments is how to control the independent variables to create multiple experimental conditions (Kirk, 1982). In some experiments, control of the independent variable is quite easy and straightforward. For instance, when testing the previously stated hypothesis, we can control the type of pointing device by presenting participants with a mouse, a joystick, or a trackball. In many other cases, the control of the independent variable can be challenging. For instance, if we are developing a speech-based application and need to investigate how recognition errors impact users' interaction behavior, we may want to compare two conditions. Under the control condition, the speech recognizer would be error free and recognize every word that the user says correctly. Under the comparison condition, the speech recognizer would make errors and recognize a percentage of the words incorrectly. This sounds straightforward, theoretically. But in practice, all speech recognizers make errors. There is no way to find a recognizer that would satisfy the requirements of the controlled condition. A possible solution to meet the needs of this experiment is the Wizard-of-Oz approach (Feng and Sears, 2009). That is, we can have a human acting as a speech recognizer, listening to what the user says and entering the user's dictation into the system. The truth would normally not be revealed to the participants until the end of the experiment. Therefore, all participants would believe that they are interacting with the speech recognizer when completing the task. The Wizard-of-Oz approach allows us to test ideal applications that do not exist in the real world. This approach is not without its limitations. Humans also make errors. It is very likely that the human "wizard" would make errors when listening to the dictation or when typing the words. Therefore, it is very difficult to control the independent variable to achieve the desired condition (Feng and Sears, 2009; Li et al., 2006). One approach that addresses this problem is the development of technical tools to assist the human wizard (Li et al., 2006).

3.2 DETERMINING THE BASIC DESIGN STRUCTURE

At the first stage of experimental design, we need to construct the experiment based on the research hypotheses that have been developed. This enables us to draw a big picture of the general scope of the experiment and, accordingly, come up with a

reasonable estimation of the timeline of the experiment and the budget. The basic structure of an experiment can be determined by answering two questions:

- How many independent variables do we want to investigate in the experiment?
- How many different values does each independent variable have?

The answer to the first question determines whether we need a basic design or a factorial design. If there is one independent variable, we need only a basic one-level design. If there are two or more independent variables, factorial design is the way to go. The answer to the second question determines the number of conditions needed in the experiment (see Figure 3.2). In a basic design, the number of conditions in the experiment is an important factor when we consider whether to adopt a between-group or within-group design. In a factorial design, we have a third option: the split-plot design. Again, the number of conditions is a crucial factor when weighing up the three options.

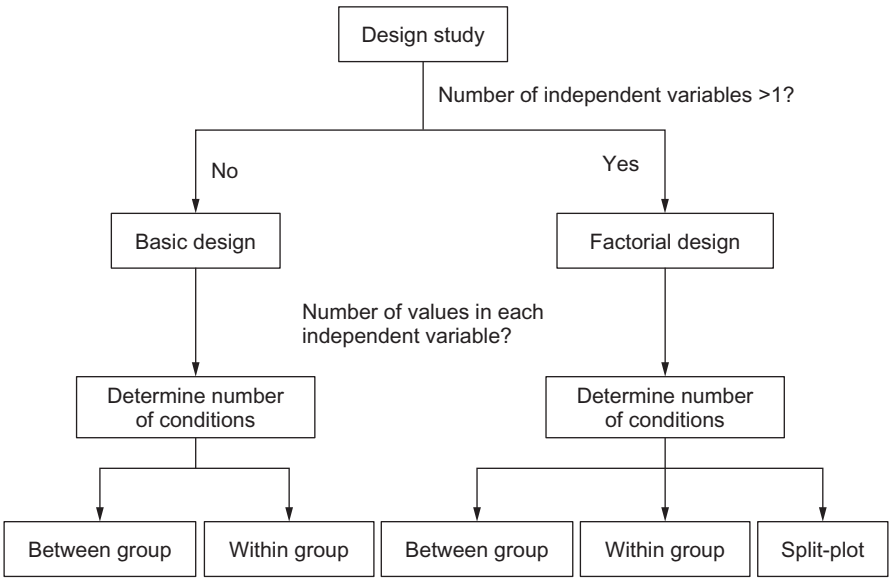

FIGURE 3.2

Determining the experiment structure.

In the following sections, we first consider the basic design scenarios involving one independent variable and focus on the characteristics of between-group design and within-group design. After that, we consider more complicated designs involving multiple independent variables, to which understanding split-plot design is the key.

3.3 INVESTIGATING A SINGLE INDEPENDENT VARIABLE

When we study a single independent variable, the design of the experiment is simpler than cases in which multiple variables are involved. The following hypotheses all lead to experiments that investigate a single independent variable:

- H1: There is no difference in typing speed when using a QWERTY keyboard, a DVORAK keyboard,[1] or an alphabetically ordered keyboard.
- H2: There is no difference in the time required to locate an item in an online store between novice users and experienced users.
- H3: There is no difference in the perceived trust toward an online agent among customers who are from the United States, Russia, China, and Nigeria.

The number of conditions in each experiment is determined by the possible values of the independent variable. The experiment conducted to investigate hypothesis H1 would involve three conditions: the QWERTY keyboard, the DVORAK keyboard, and the alphabetically ordered keyboard. The experiment conducted to investigate hypothesis H2 would involve two conditions: novice users and experienced users. And the experiment conducted to investigate hypothesis H3 would involve four conditions: customers from the United States, Russia, China, and Nigeria.

Once the conditions are set, we need to determine the number of conditions to which we would allow each participant to be exposed by selecting either a between-group design or a within-group design. This is a critical step in experimental design and the decision made has a direct impact on the quality of the data collected as well as the statistical methods that should be used to analyze the data.

3.3.1 BETWEEN-GROUP DESIGN AND WITHIN-GROUP DESIGN

Between-group design is also called "between-subject design." In a between-group design, each participant is only exposed to one experimental condition. The number of participant groups directly corresponds to the number of experimental conditions. Let us use the experiment on types of keyboard as an example. As shown in Figure 3.3, three

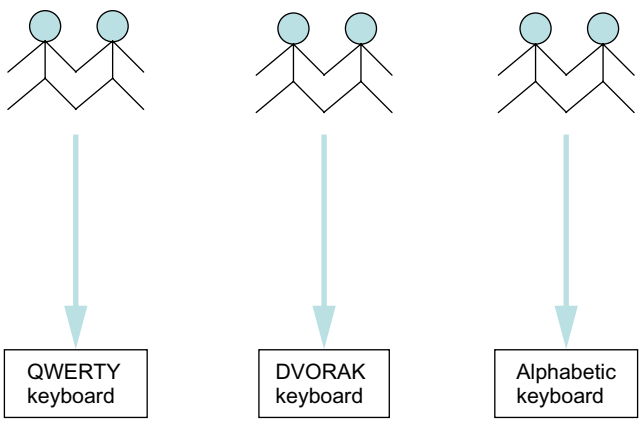

FIGURE 3.3

Between-group design.

[1] Dvorak keyboard is an ergonomic alternative to the commonly used "QWERTY keyboard." The design of the Dvorak keyboard emphasizes typist comfort, high productivity, and ease of learning.

groups of participants take part in the experiment and each group only uses one specific type of keyboard. If the task is to type a document of 500 words, then each participant types one document using one of the keyboards.

In contrast, a within-group design (also called "within-subject design") requires each participant to be exposed to multiple experimental conditions. Only one group of participants is needed for the entire experiment. If we use the keyboard experiment as an example, as shown in Figure 3.4, one group of participants uses all three types of keyboard during the experiment. If the task is to type a document of 500 words, then each participant types three documents, using each of the three keyboards for one document.

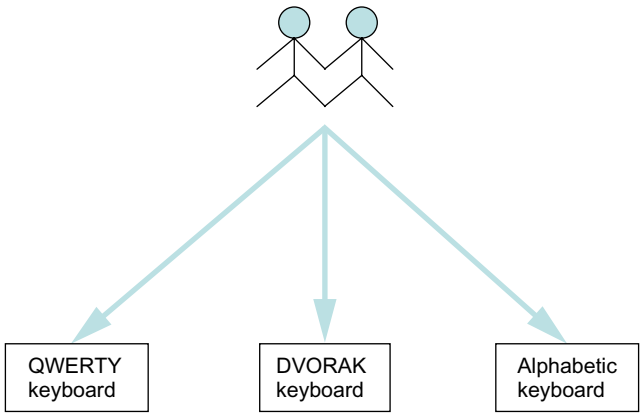

FIGURE 3.4

Within-group design.

Please note that different statistical approaches are needed to analyze data collected from the two different design methods. The details of statistical analysis are discussed in Chapter 4.

3.3.1.1 Advantages and disadvantages of between-group design

From the statistical perspective, between-group design is a cleaner design. Since the participant is only exposed to one condition, the users do not learn from different task conditions. Therefore, it allows us to avoid the learning effect. In addition, since the participants only need to complete tasks under one condition, the time it takes each participant to complete the experiment is much shorter than in a within-group design. As a result, confounding factors such as fatigue and frustration can be effectively controlled.

On the other hand, between-group design also has notable disadvantages. In a between-group experiment, we are comparing the performance of one group of participants against the performance of another group of participants. The results are subject to substantial impacts from individual differences: the difference between the multiple values that we expect to observe can be buried in a high level of "noise" caused by individual differences. Therefore, it is harder to detect significant differences and Type II errors are more likely to occur.

In order to effectively exclude the impact of noise and make significant findings, a comparatively larger number of participants are needed under each condition. This leads to the second major disadvantage of the between-group design: large sample size. Since the number of participants (m) in each condition should be comparatively larger than that in a within group design and approximately the same number of participants are needed for each condition (let n be the number of conditions), the total number of participants needed for the experiment ($m \times n$) is usually quite large. For example, if an experiment has 4 conditions and 16 participants are needed under each condition, the total number of participants needed is 64. Recruiting the number of participants needed for a between-group experiment can be a very challenging task.

3.3.1.2 Advantages and disadvantages of within-group design

Within-group design, in contrast, requires a much smaller sample size. When analyzing the data coming from within-group experiments, we are comparing the performances of the same participants under different conditions. Therefore, the impact of individual differences is effectively isolated and the expected difference can be observed with a relatively smaller sample size. If we change the design of the experiment with 4 conditions and 16 participants from a between-group design into a within-group design, the total number of participants needed would be 16, rather than 64. The benefit of a reduced sample size is an important factor for many studies in the HCI field when qualified participants may be quite difficult to recruit. It may also help reduce the cost of the experiments when financial compensation is provided.

Within-group designs are not free of limitations. The biggest problem with a within-group design is the possible impact of learning effects. Since the participants complete the same types of task under multiple conditions, they are very likely to learn from the experience and may get better in completing the tasks. For instance, suppose we are conducting a within-group experiment that evaluates two types of ATM: one with a button interface and one with a touch-screen interface. The task is to withdraw money from an existing account. If the participant first completes the task using the ATM with the button interface, the participant gains some experience with the ATM interface and its functions. Therefore, the participant may perform better when subsequently completing the same tasks using the ATM with the touch-screen interface. If we do not isolate the learning effect, we might draw a conclusion that the touch-screen interface is better than the button interface when the observed difference is actually due to the learning effect. Normally, the potential bias of the learning effect is the biggest concern of experimenters when considering adopting a within-group design. A Latin Square Design is commonly adopted to control the impact of the learning effect.

Another potential problem with within-group designs is fatigue. Since there are multiple conditions in the experiment, and the participants need to complete one or more tasks under each condition, the time it takes to complete the experiment may be quite long and participants may get tired or bored during the process. Contrary to the learning effect, which favors conditions completed toward the end of the experiment, fatigue negatively impacts on the performance of conditions completed toward the

end of the experiment. For instance, in the ATM experiment, if the touch-screen interface is always tested after the button interface, we might draw a conclusion that the touch-screen interface is not as effective as the button interface when the observed difference is actually due to the participants' fatigue. We might fail to identify that the touch-screen interface is better than the button interface because the impact of fatigue offsets the gain of the touch-screen interface. Similarly, the potential problem of fatigue can also be controlled through the adoption of the Latin Square Design.

3.3.1.3 Comparison of between-group and within-group designs

The pros and cons of the between- and within-group designs are summarized in Table 3.1. You can see from the table that the advantages and limitations of the two design methods are exactly opposite to each other.

Table 3.1 Advantages and Disadvantages of Between-Group Design and Within-Group Design

	Type of Experiment Design	
	Between-Group Design	**Within-Group Design**
Advantages	Cleaner Avoids learning effect Better control of confounding factors, such as fatigue	Smaller sample size Effective isolation of individual differences More powerful tests
Limitations	Larger sample size Large impact of individual differences Harder to get statistically significant results	Hard to control learning effect Large impact of fatigue

3.3.2 CHOOSING THE APPROPRIATE DESIGN APPROACH

It is quite common for experimenters to argue back and forth when deciding which of the two design approaches to adopt. Many times the decision is quite hard to make since the advantages and disadvantages of the between-group design and within-group design are exactly opposite to each other. It should be emphasized that each experiment is unique and the decision should be made on a case-by-case basis with full consideration of the specific context of the experiment. In some cases, a hybrid design may be adopted that involves both between-group factors and within-group factors. The hybrid approach is discussed in detail in Section 3.4.2. This section discusses the general guidelines that help us choose the appropriate approach for a specific user study.

3.3.2.1 Between-group design

Generally speaking, between-group design should be adopted when the experiment investigates: simple tasks with limited individual differences; tasks that would be greatly influenced by the learning effect; or problems that cannot be investigated through a within-group design.

The size of the individual differences is very hard to estimate. However, it is empirically confirmed that individual differences are smaller when the tasks are simple and involve limited cognitive process (Dillon, 1996; Egan, 1988). In contrast, individual differences are larger when the task is complicated or involves significant cognitive functions. For example, when the task mainly involves basic motor skills, such as selecting a target on the screen, the individual differences among participants might be comparatively small.[2] But when the task involves more complicated cognitive or perceptual functions, such as reading, comprehension, information retrieval, and problem solving, the individual differences have a much larger impact. So when the task is simple, the impact of individual differences is limited and a between-group design would be appropriate.

Depending on the types of task, some experiments are more vulnerable to the learning effect than others. For example, in an experiment that compares the navigation effectiveness of two types of menu within a website, a participant who completes the navigation tasks under one condition would have gained a significant amount of knowledge of the website architecture. The knowledge would make a great impact on the participant's performance when completing the tasks under the other condition. Therefore, within-group design is highly inappropriate for this type of task and between-group design would have to be adopted.

There are many circumstances when it is totally impossible to adopt a within-group design. Taking hypotheses H2 and H3, previously stated, as examples:

- H2: There is no difference in the time required to locate an item in an online store between novice users and experienced users.
- H3: There is no difference in the perceived trust toward an online agent among customers who are from the United States, Russia, China, and Nigeria.

You can see that there is no way to compare the performances of novice users and experienced users through a within-group design because an individual cannot be both a novice user and an experienced user of the online store at the same time. For the same reason, a within-group design is not appropriate for H3 since any participant can only represent one of the four cultures. Under those circumstances, a between-group design is obviously the only option we have.

After choosing a between-group design for an experiment, we need to take special caution to control potential confounding factors. Participants should be randomly assigned to different conditions whenever possible.[3] When assigning participants, we need to try our best to counterbalance potential confounding factors, such as gender, age, computing experience, and internet experience,

[2] Note that the individual differences in these types of tasks can be quite substantial when the participants come from different age groups or when individuals with motor disabilities are involved.
[3] We cannot randomly assign participants to different conditions in the cases of H2 and H3, obviously.

across conditions. In other words, we need to make sure that the groups are as similar as possible, except for the personal characteristics that are experimental variables under investigation.

3.3.2.2 Within-group design

Within-group design is more appropriate when the experiment investigates tasks with large individual differences, tasks that are less susceptible to the learning effect, or when the target participant pool is very small. As discussed previously, complicated tasks that involve substantial human cognitive and perceptual capabilities generally encounter much larger individual differences than simple tasks. Therefore, when an experiment investigates complicated tasks such as reading, comprehension, information retrieval, and problem solving, a within-group design might be more appropriate since it effectively isolates individual differences from the main effects.

Most of the tasks that examine complicated or learned skills or knowledge—such as typing, reading, composition, and problem solving—are less susceptible to learning effects. For example, if an experiment investigates the impact of two fonts (i.e., Times New Roman and Arial) on participants' reading speed, the learning effect between the two conditions would be very limited. Reading one text document of several hundred words is unlikely to improve an individual's reading speed. Therefore, a within-group design would be appropriate as long as the text materials presented to the participant under the two conditions are different in content but similar in levels of difficulty.

Difficulty in finding and recruiting qualified participants is a problem frequently faced by many HCI researchers. One typical example is the field of universal usability, which focuses on developing applications usable by diverse user populations. Numerous studies in this field examine how individuals with disabilities interact with computers or computer-related devices. Although the total number of people falling into a specific disability or disease category is quite large, the number of such individuals living in a particular area is very limited. Therefore, the sample sizes are normally smaller than that in studies examining users without disabilities (e.g., Taylor et al., 2016).

Recruiting participants with specific disabilities is always a challenging task. For more detailed discussion on working with participants with disabilities, please refer to Chapter 16. The same problem also occurs when the target population is well trained, highly experienced, professionals, such as business executives or experienced project managers, simply because they are too busy to be bothered. Under those circumstances, it is almost impossible to recruit the number of participants needed for a between-group design, forcing the experimenters to adopt a within-group design.

Having decided to adopt a within-group design, you need to consider how to control the negative impact of learning effects, fatigue, and other potential problems associated with a within-group design. As discussed previously, a general approach to

control these negative impacts is counterbalancing the condition or treatment orders through a Latin Square Design.

When the objective of the study is not initial interaction with the application, an effective approach to reduce the impact of the learning effect is to provide sufficient time for training. Research suggests that, for many types of tasks, the learning curve tends to be steeper during the initial interaction stages and flatter after that stage (see Figure 3.5). People achieve quicker progress in learning during initial stages, followed by gradual lesser improvement with further practice. Therefore, providing sufficient training time for users to get acquainted with the system or the task greatly reduces the learning effect during the actual task sessions. Of course, training cannot completely eliminate the learning effect. It only reduces its impact. This approach, combined with the counterbalancing of task conditions, is widely adopted in HCI studies to control the impact of learning.

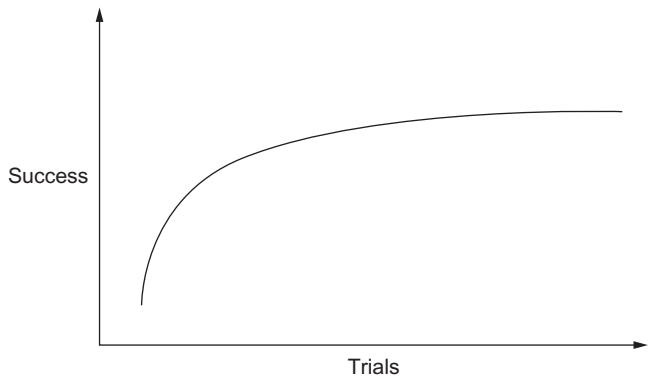

FIGURE 3.5

Typical learning curve.

To address the problem of fatigue caused by multiple experimental tasks, we need to design experiment tasks frugally, reducing the required number of tasks and shortening the experiment time whenever possible. It is generally suggested that the appropriate length of a single experiment session should be 60 to 90 minutes or shorter (Nielsen, 2005). When a session lasts longer than 90 minutes, the participant may get tired or frustrated. It is strongly suggested that a single session should definitely not last longer than 2 hours. During the experiment, the participant should be provided with opportunities to take breaks as needed. Interestingly, even when the experimenter encourages the participants to take breaks, the participants may not realize that they are getting tired and tend to ignore the suggestion to take a break. Therefore, some researchers find it helpful to force the participants to take a break during an experiment. For more discussion regarding the benefit of breaks in HCI studies, please refer to Chapter 15.

3.4 INVESTIGATING MORE THAN ONE INDEPENDENT VARIABLE

3.4.1 FACTORIAL DESIGN

Factorial designs are widely adopted when an experiment investigates more than one independent variable or factor. Using this method, we divide the experiment groups or conditions into multiple subsets according to the independent variables. It allows us to simultaneously investigate the impact of all independent variables as well as the interaction effects between multiple variables.

The number of conditions in a factorial design is determined by the total number of independent variables and the level of each independent variable. The equation for calculating the number of conditions is:

$$C = \prod_{a=1}^{n} Va$$

where C is the number of conditions, V is the number of levels in each variable, and $\prod$ is the product of V_1 through V_n.

The best way to explain a factorial design and this equation is through an example. Consider running an experiment to compare the typing speed when using three types of keyboard (QWERTY, DVORAK, and Alphabetic). We are also interested in examining the effect of different tasks (composition vs transcription) on the typing speed. This suggests that two independent variables are investigated in the experiment: type of keyboards and type of tasks. The variable "type of keyboards" has three levels: QWERTY, DVORAK, and Alphabetic. The variable "type of tasks" has two levels: transcription and composition. Therefore, the total number of conditions in this experiment is calculated according to the following equation:

$$\text{Number of conditions} = 3 \times 2 = 6$$

Table 3.2 illustrates the six conditions in this experiment. In the first three conditions, the participants would all complete composition tasks using different kinds of keyboard. In the other three conditions, the participants would all complete transcription tasks using different keyboards. When analyzing the data, we can compare conditions in the same row to examine the impact of keyboards. The effect of the tasks can be examined through comparing conditions in the same column. As a result, the effect of both independent variables can be examined simultaneously through a single experiment.

Table 3.2 A Factorial Design

	QWERTY	DVORAK	Alphabetic
Composition	1	2	3
Transcription	4	5	6

Either a between-group design or a within-group design may be adopted in this experiment. In a between-group design, each participant completes tasks under only one of the six conditions. As a result, six groups of participants would be required, one group for each condition. In a within-group design, each participant completes tasks under all six conditions. The advantages and disadvantages of between-group design and within-group design that we discussed in Section 3.3.2 also apply to factorial designs. No matter which design is adopted, it is important to counterbalance the orders and conditions in the experiment. In a between-group design, the participants need to be randomly assigned to the conditions. In a within-group design, the order in which the participant completes the six tasks needs to be counterbalanced.

3.4.2 SPLIT-PLOT DESIGN

In experiments that study one independent variable, we can choose to implement the study as a between-group design or a within-group design. In a factorial study, we can also choose a split-plot design. A split-plot design has both between-group components and within-group components. That is, one or more independent variables are investigated through a between-group approach and the other variables are investigated through a within-group approach.

Table 3.3 illustrates an experiment that employs a split-plot design. The experiment investigates two independent variables: age and the use of GPS. The variable "age" has three levels: people who are 20–40 years old, people who are 41–60 years old, and people who are older than 60. The second variable has two levels: driving without GPS and driving with GPS assistance. Therefore, the total number of conditions in this experiment is six.

Table 3.3 A Split-Plot Design

	20–40 Years Old	41–60 Years Old	Above 60
Driving without GPS assistance	1	2	3
Driving with GPS assistance	4	5	6

The impact of age is investigated through a between-group design since three groups of participants from different age ranges are studied. The impact of the use of GPS can be examined through a within-group approach. We can require each participant to complete the same driving task both with and without the assistance of the GPS. This gives us a typical split-plot design that involves both a between-group component (age analysis is based on the columns) and a within-group component (GPS use is analyzed by comparing condition 1 with condition 4, condition 2 with condition 5, and condition 3 with condition 6).

FACTORIAL DESIGN IN HCI RESEARCH

Factorial design has been commonly adopted in user studies in the HCI field. For example, Warr et al. (2016) used a 3×3 factorial design to investigate the differences between three window switching methods in a desktop environment.

The between-group factor of the study was the window switching method: the *Cards* interface, the *Exposé* interface, and the *Mosaic* interface. Three groups of participants took part in the study, each completing tasks under one of the assigned window switching conditions. The within-group factor of the study was the number of open windows on the screen (3, 6, and 9). Under a specific window switching condition, each participant completed the same number of trials with 3 open windows, 6 open windows, and 9 open windows, respectively.

Learning and fatigue might occur during the experiment. In order to address these two factors, participants were given time to practice selecting windows until they were comfortable with the procedure. The order of the 3, 6, and 9 window conditions was counterbalanced through a Latin Square Design.

3.4.3 INTERACTION EFFECTS

One advantage of a factorial design is that it allows us to study the interaction effects between two or more independent variables. According to Cozby (1997), an interaction effect can be described as "the differing effect of one independent variable on the dependent variable, depending on the particular level of another independent variable." When a significant interaction exists between independent variables X and Y, the means of the dependent variable Z would be determined jointly by X and Y.

Let us explain interaction effect through an example. Suppose we are conducting an experiment that investigates how types of device (mouse and touchscreen) and experience impact the effectiveness of target selection tasks. Two types of user are studied: novice users and experienced users. Based on the data collected, we draw a diagram as shown in Figure 3.6. As you can see, novice users can select targets faster with a touchscreen than with a mouse. Experienced users can select targets faster with a mouse than with a touchscreen. The target selection speeds for both the mouse and the touchscreen increase as the user gains more experience with the device. However, the increase in speed is much larger for the mouse than for the touchscreen.

It is critical to study interaction effects in HCI studies since performance may be affected by multiple factors jointly. There are numerous studies that did not identify any significant effect in individual independent variables but found significant results in interaction effects.

Interaction effects may have important implications for design. For example, the interaction effect in Figure 3.6 would suggest that the touchscreen performs better than the mouse during the initial interaction. But users can make greater progress in learning the mouse than the touchscreen and eventually achieve higher efficiency with the mouse. This result may imply that a touchscreen is a more appropriate input

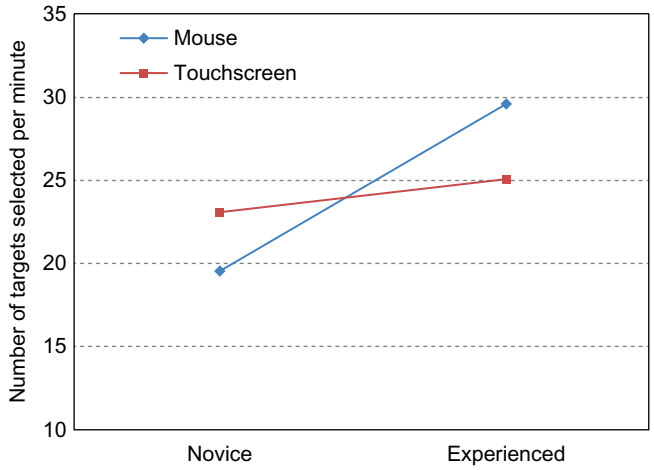

FIGURE 3.6

Interaction effects.

device when the interaction is normally brief and the opportunities for training are limited, such as an ATM interface. In contrast, a mouse might be more appropriate for long-term, frequent tasks, such as interacting with a computer desktop.

3.5 RELIABILITY OF EXPERIMENTAL RESULTS

All experimental research strives for high reliability. Reliable experiments can be replicated by other research teams in other locations and yield results that are consistent, dependable, and stable. One big challenge in HCI studies is that in contrast to the "hard sciences," such as physics, chemistry, and biology, measurements of human behavior and social interaction are normally subject to higher fluctuations and, therefore, are less replicable. The fluctuations in experimental results are referred to as errors.

3.5.1 RANDOM ERRORS

We may observe a participant typing several text documents during five sessions and obtain an actual typing speed of 50 words per minute. It is very unlikely that we would get the same typing speed for all five sessions. Instead, we may end up with data like this:

Session 1: 46 words per minute
Session 2: 52 words per minute
Session 3: 47 words per minute
Session 4: 51 words per minute
Session 5: 53 words per minute

The general relationship between the actual value we are looking for and the observed values can be expressed as follows:

$$\text{Observed values} = \text{Actual value} + \text{Random error}$$

Random errors are also called "chance errors" or "noise." They occur by chance and are not correlated with the actual value. Random errors push the observed values to move up or down around the exact value. There is no way to eliminate or control random errors but we can reduce the impact of random errors by enlarging the observed sample size. When a sample size is small, the random errors may have significant impact on the observed mean and the observed mean may be far from the actual value. When a sample size is large enough, the random errors should offset each other and the observed mean should be very close to the actual value. For example, in the typing task earlier, if we observe only Session 1, the mean would be 46, which is 4 words from the true value of 50 words per minute. If we increase the number of observed sessions to 5, the mean of the observed values is 49.8, very close to the actual value. In reality, we can never claim that we are 100% confident that the observed value is the actual value. But we can be 100% confident that the larger our sample size is, the closer the observed value is to the actual value.

3.5.2 SYSTEMATIC ERRORS

Systematic errors, also called "biases," are completely different in nature from random errors. While random errors cause variations in observed values in both directions around the actual value, systematic errors always push the observed values in the same direction. As a result, systematic errors never offset each other in the way that random errors do and they cause the observed mean to be either too high or too low.

Using the typing task example, the participant might consistently underperform during all five observation sessions, because of tiredness or nervousness, and we may collect the following data:

Session 1: 47 words per minute
Session 2: 44 words per minute
Session 3: 45 words per minute
Session 4: 42 words per minute
Session 5: 46 words per minute

In this case, the mean of the observed values is 44.8, 5 words lower than the actual value. Figure 3.7 shows the performance of the participant in each case. Under the unbiased conditions, the observed values fluctuate due to random errors, but the fluctuations occur in both directions around the actual value and offset each other. However, under the biased condition, the systematic error consistently pushes all values down, causing the mean of the observed values to be significantly below the actual value.

Systematic errors can greatly reduce the reliability of experimental results; therefore, they are the true enemy of experimental research. We can counter systematic errors in two stages: we should try to eliminate or control biases during the experiment

FIGURE 3.7

Comparison of random and systematic errors.

when biases are inevitable, and we need to isolate the impact of them from the main effect when analyzing the data. There are five major sources of systematic error:

- measurement instruments;
- experimental procedures;
- participants;
- experimenter behavior; and
- experimental environment.

3.5.2.1 Bias caused by measurement instruments

When the measurement instruments used are not appropriate, not accurate, or not configured correctly, they may introduce systematic errors. For instance, when observing participants searching for an item on an e-commerce website, we may use a stop watch to measure the time it takes to locate the specific item. If the stop watch is slow and misses 5 minutes in every hour, then we consistently record less time than the actual time used. As a consequence, the observed performance will be better than the actual value. In order to control biases introduced by the measurement instruments, we need to carefully examine the instruments used before experiment sessions. Another approach is to use extensively tested, reliable, and software-driven instruments. A bonus of software-driven instruments is that they can avoid human errors as well.

3.5.2.2 Bias caused by experimental procedures

Inappropriate or unclear experimental procedures may introduce biases. As discussed previously, if the order of task conditions is not randomized in an experiment with a within-group design, the observed results will be subject to the impact of the learning effect and fatigue: conditions tested later may be consistently better than conditions tested earlier due to learning effect; on the other hand, conditions tested earlier may be consistently better than later conditions due to fatigue. The biases caused by the

learning effect and fatigue push the observed value in opposite directions and the combined effect is determined by the specific context of the experiment. If the tasks are simple and less susceptible to the learning effect, but tedious and long, the impact of fatigue and frustration may outweigh the impact of the learning effect, causing participants to consistently underperform in later sessions. If the tasks are complicated and highly susceptible to the learning effect, but short and interesting, the impact of the learning effect may outweigh the impact of fatigue, causing participants to consistently perform better in later sessions.

The instructions that participants receive play a crucial role in an experiment and the wording of the experiment instructions should be carefully scrutinized before a study. Slightly different wording in instructions may lead to different participant responses. In a reported HCI study (Wallace et al., 1993), participants were instructed to complete the task "as quickly as possible" under one condition. Under the other condition, participants were instructed to "take your time, there is no rush." Interestingly, participants working under the no-time-stress condition completed the tasks faster than those under the time-stress condition. This suggests the importance of critical wording in instructions. It also implies that the instructions that participants receive need to be highly consistent. When a study is conducted under the supervision of multiple investigators, it is more likely that the investigators give inconsistent instructions to the participants. Instructions and procedures on a written document or prerecorded instructions are highly recommended to ensure consistency across experimental sessions.

Many times, trivial and unforeseen details introduce biases into the results. For instance, in an experiment that studies data entry on a PDA, the way the PDA is physically positioned may have an impact on the results. If no specification is given, some participants may hold the PDA in one hand and enter data using the other hand, other participants may put the PDA on a table and enter data using both hands. There are notable differences between the two conditions regarding the distance between the PDA screen and the participant's eyes, the angle of the PDA screen, and the number of hands involved for data entry. Any of those factors may introduce biases into the observed results. In order to reduce the biases attributed to experimental procedures, we need to

- randomize the order of conditions, tasks, and task scenarios in experiments that adopt a within-group design or a split-plot design;
- prepare a written document with detailed instructions for participants;
- prepare a written document with detailed procedures for experimenters; and
- run multiple pilot studies before actual data collection to identify potential biases.

A pilot study is not a luxury that we conduct only when we have plenty of time or money to spend. On the contrary, years of experience tells us that pilot studies are critical for all HCI experiments to identify potential biases. No matter how well you think you have planned the study, there are always things that you overlook. A pilot study is the only chance you have to fix your mistakes before you run the main study. Pilot studies should be treated very seriously and conducted in exactly the same way as planned for the actual experiment. Participants of the pilot study should be from

the target population. Having one or two members from the research team completing the designed tasks is not a pilot study in its true sense (Preece et al., 1994).

3.5.2.3 Bias caused by participants

Many characteristics of the participants may introduce systematic errors into the results. Potential contributors may be in a specific age range or have particular computer or internet experience, domain knowledge, education, professional experience and training, or personal interests. For instance, if we are running an experiment to test the user interface of a new mobile phone model, we might recruit participants by posting announcements on a popular blog on https://www.cnet.com. Since this website features highly technical news and reviews related to information technology, its visitors normally have a strong technical background and rich experience in using IT devices. As a consequence, the observed data would tend to outperform what we would observe from the general public. The following guidelines can help us reduce systematic errors from the participants:

- Recruit participants carefully, making sure the participant pool is representative of the target user population (Broome, 1984; Smart, 1966).
- Create an environment or task procedure that causes the least stress to the users.
- Reassure the participants that you are testing the interface, not them, so they are calm and relaxed during the experiment.
- Reschedule a session or give participants some time to recover if they arrive tired, exhausted, or very nervous.

3.5.2.4 Bias due to experimenter behavior

Experimenter behavior is one of the major sources of bias. Experimenters may intentionally or unintentionally influence the experiment results. Any intentional action to influence participants' performance or preference is unethical in research and should be strictly avoided. However, experimenters may unknowingly influence the observed data. Spoken language, body language, and facial expressions frequently serve as triggers for bias. Let us examine the following scenarios:

1. An experimenter is introducing an interface to a participant. The experimenter says, "Now you get to the pull-down menus. I think you will really like them.… I designed them myself!"
2. An experimenter is loading an application for a participant. The response time is a bit long. The experimenter is frustrated and says, "Damn! It's slower than a snail."
3. An experimenter is loading an application for a participant. The response time is a bit long. The experimenter waits uneasily, tapping fingers on the desk and frequently changing body position while staring at the screen impatiently.
4. A participant arrives on time for a study scheduled at 9 a.m. The experimenter does not arrive until 9:10 a.m. After guiding the participant into the lab, the experimenter takes 10 minutes to set up all the equipment. Once the experiment starts, the experimenter finds that the task list is missing and runs out of the lab to print a copy.

In Scenario 1, the experimenter is very demanding and the comment may make the participant reluctant to provide negative feedback about the interface in case it hurts the experimenter's feelings. Therefore, the data collected from the participant, especially the subjective data, are likely to be better than the actual value. In Scenarios 2 and 3, the experimenter's spoken language or body language reveals negative attitude toward the application. Participants would register those cues and would form a negative perspective even before their first encounter with the application and the collected subjective ratings and feedback would be biased against the application. In Scenario 4, the unprofessional and slack style of the experimenter would give a negative impression to the participant, which may impact the participant's performance as well as the subjective ratings and feedback.

When multiple experimenters are involved in the experiment, bias is likely to occur due to inconsistency in instructions and training, as well as individual styles and attitudes. If one of the experimenters is very patient, offers long training sessions, and demonstrates all related commands to the participants before the actual task, while the other experimenter is pushy, offers shorter training sessions, and only demonstrates a subset of the commands, the participants who complete the experiment under the guidance of the first experimenter may systematically outperform the participants who complete the experiment with the second experimenter. In order to control possible biases triggered by experimenters, we need to

- Offer training opportunities to experimenters and teach them to be neutral, calm, and patient when supervising experiments.
- Make sure that the experimenter arrives at least 10 minutes before the scheduled sessions and gets everything ready before the session starts.
- Whenever possible, have two experimenters supervise a session together, one as the lead experimenter and the other as the assistant experimenter. The lead experimenter is responsible for interacting with the participants. The assistant experimenter observes the session closely, fixes errors if noted, and takes notes when necessary.
- Prepare written documents with detailed procedures for experimenters and require all experimenters to follow the same procedure strictly.
- When appropriate, record important instructions before the experiment and play the recording to the participants during the experiment. In this way, we can guarantee that all participants go through the same training process and receive the same instructions.

3.5.2.5 Bias due to environmental factors

Environmental factors play an increasingly important role in HCI research due to the rapid development in mobile computing, universal accessibility, and recognition-based technologies. Environmental factors can be categorized into two groups: physical environmental factors and social environmental factors. Examples of physical environmental factors include noise, temperature, lighting, vibration, and humidity. Examples of social environmental factors include the number of people in the surrounding environment and the relationship between those people and the participant.

Both physical and social environmental factors may introduce systematic errors into the observed data. For instance, a study that examines the performance of a speech-recognition application may yield lower recognition error rates than the actual value if there is a significant level of ambient noise during the experiment session. Even when the study investigates applications other than speech, loud environmental noise may distract the participants or induce fatigue. Regarding social factors, a participant with a person watching over his shoulder may perform differently from a participant who is seated alone. Environmental factors may cause more problems when the experiment is not conducted in a lab, but in locations such as the participant's home or workplace. The following guidelines can help us avoid or control environment-induced biases:

- In a lab setting, make sure the room is quiet, the lighting is appropriate, and the chairs and tables are comfortable. The room should be clean and tidy, without notable distractions.
- Whenever possible, the participant should be seated alone and the experimenter can observe the session from another room via a one-way mirror or monitors.
- In a field study, the experimenters should visit the location before the scheduled time to confirm that the setting meets the requirements of the study.

Finally, it is important to realize that, no matter how hard you try to avoid biases, they can never be completely eliminated. A well-designed experiment with lots of consideration for controlling bias can improve the data, making the observed results closer to the actual values, but still subject to the impact of biases. Therefore, we should be careful when reporting the findings, even when the study results are statistically significant.

3.6 EXPERIMENTAL PROCEDURES

Experiments are conducted in dramatically different fields to answer a myriad of questions. Experiments in the HCI field, similar to many studies in sociology or psychology, typically involve human subjects. Studying human subjects is quite different from studying metal or plant reactions, or other animals, and introduces many interesting issues or challenges. The concerns and practices of working with human subjects are discussed in detail in Chapter 15. In this section, we briefly introduce the procedures for experiments that study human subjects.

In the lifecycle of an HCI experiment, we typically go through the following process:

1. Identify a research hypothesis.
2. Specify the design of the study.
3. Run a pilot study to test the design, the system, and the study instruments.
4. Recruit participants.
5. Run the actual data collection sessions.
6. Analyze the data.
7. Report the results.

Within a specific experiment session, we typically go through the following steps:

1. Ensure that the systems or devices being evaluated are functioning properly, the related instruments are ready for the experiment.
2. Greet the participants.
3. Introduce the purpose of the study and the procedures.
4. Get the consent of the participants.
5. Assign the participants to a specific experimental condition according to the predefined randomization method.
6. Participants complete training tasks.
7. Participants complete actual tasks.
8. Participants answer questionnaires (if any).
9. Debriefing session.
10. Payment (if any).

Some experiments may require more complicated steps or procedures. For example, longitudinal studies involve multiple trials. We need to make sure that the tasks used in each trial are randomized in order to control the impact of the learning effect.

A number of open source platforms have been developed to help researchers design experiments, collect data, and analyze the results. One example is the Touchstone experimental design platform. The Touchtone system includes a "design" platform for examining alternative, controlled experimental designs, a "run" platform for running subjects, and an "analysis" platform that provides advices on statistical analysis (Mackay et al., 2007).

3.7 SUMMARY

Experiment design starts with a clearly defined, testable research hypothesis. During the design process, we need to answer the following questions:

- How many dependent variables are investigated in the experiment and how are they measured?
- How many independent variables are investigated in the experiment and how are they controlled?
- How many conditions are involved in the experiment?
- Which of the three designs will be adopted: between-group, within-group, or split-plot?
- What potential bias may occur and how can we avoid or control those biases?

When an experiment studies only one independent variable, we need to choose between the between-group design and the within-group design. When there is more than one independent variable, we need to select among the between-group design, the within-group design, and the split-plot design.

The between-group design is cleaner, avoids the learning effect, and is less likely to be affected by fatigue and frustration. But this design is weaker due

to the high noise level of individual differences. In addition, larger numbers of participants are usually required for a between-group design. The within-group design, on the other hand, effectively isolates individual differences and, therefore, is a much stronger test than the between-group design. Another bonus is that fewer participants are required. But within-group designs are more vulnerable to learning effects and fatigue. The appropriate design method needs to be selected based on the nature of the application, the participant, and the tasks examined in the experiment.

All experiments strive for clean, accurate, and unbiased results. In reality, experiment results are highly susceptible to bias. Biases can be attributed to five major sources: the measurement instruments, the experiment procedure, the participants, the experimenters, and the physical and social environment. We should try to avoid or control biases through accurate and appropriate measurement devices and scales; clearly defined and detailed experimental procedures; carefully recruited participants; well-trained, professional, and unbiased experimenters; and well-controlled environments.

DISCUSSION QUESTIONS

1. Explain the differences among the three types of study: experiment, quasi-experiment, and nonexperiment.
2. What are the major issues that need to be considered when designing experiments?
3. What is a between-group design? Explain the advantages and disadvantages of a between-group design.
4. What is a within-group design? Explain the advantages and disadvantages of a within-group design.
5. When should a between-group design be considered for an experiment?
6. When should a within-group design be considered for an experiment?
7. What is the benefit of a factorial design compared to experiments that investigate one factor at a time?
8. What is a split-plot design?
9. Explain the differences between random errors and systematic errors.
10. What are the major sources of systematic errors, or biases?
11. What can we do to reduce systematic errors in experiments?
12. Describe the typical procedure of an experiment that involves human subjects.

RESEARCH DESIGN EXERCISES

1. Read the following scenarios. Identify actions or conditions that may induce systematic errors in each scenario and explain the direction of the impact (i.e., whether the observed data will be pulled up or down from the actual value).
Scenario 1: In an experiment that investigates how novice users learn to use the T9 method to enter data into a PDA, a participant has actually used T9 for over a year.
Scenario 2: An experimenter is introducing a website to a participant. The experimenter says, "My team has spent six months on this site. The site is like our baby."
Scenario 3: In an experiment that examines how individuals with severe motor disabilities interact with computers using a brain-computer interface, all participants recruited are healthy individuals without any disability.
Scenario 4: In an experiment that examines speech-based dictation techniques, the experimenter forgets to switch the speech profiles between experiment sessions, so a participant used another person's speech profile to complete the dictation tasks.
Scenario 5: In an experiment that examines the design of an e-commerce website, participants complete multiple tasks to retrieve specific information on the site. However, the network speed is very slow and the participants have to wait significant amounts of time for each page to be loaded.

2. Read the following scenarios. Discuss the type of experiment design (between-group, within-group, or split-plot) that is appropriate for each scenario.
Scenario 1: A study investigates whether people who have attended a security training program generate and use more secure passwords than people who have not received any security training.
Scenario 2: A research team examines the effectiveness of joysticks and trackballs for selecting static targets and moving targets.
Scenario 3: A research team examines whether the gender of an online agent affects the perception of trust for young users, middle-aged users, and older users.
Scenario 4: A research team examines whether virtual teams who use video conferencing are more productive than teams who use phone-based teleconferencing.
Scenario 5: A study examines the effectiveness of three menu structures. The first structure has two levels, with 8 items in the first level and 64 items in the second level. The second structure has three levels, with 4 items in the first level, 16 items in the second level, and 64 items in the third level. The third menu has six levels, with 2 items in the first level and 2^n items in the nth level.

REFERENCES

Broome, J., 1984. Selecting people randomly. Ethics 95 (1), 38–55.
Cooper, D., Schindler, P., 2000. Business Research Methods, seventh ed. McGraw Hill, Boston, MA.

Cozby, P.C., 1997. Methods in Behavioral Research, sixth ed. Mayfield Publishing, Mountain View, CA.

Dillon, A., 1996. User analysis in HCI: the historical lesson from individual differences research. International Journal of Human-Computer Studies 45 (6), 619–637.

Egan, D., 1988. Individual differences in human–computer interaction. In: Helander, M. (Ed.), Handbook of Human–Computer Interaction. Elsevier, North-Holland, pp. 543–568.

Feng, J., Sears, A., 2009. Beyond errors: measuring reliability for error-prone interaction devices. Behaviour & Information Technology, 1–15.

Kirk, R., 1982. Experimental Design: Procedures for the Behavioral Sciences, second ed Brooks/Cole Publishing Company, Pacific Grove, CA.

Li, Y., Welbourne, E., Landay, J., 2006. Novel methods: emotions, gestures, events: design and experimental analysis of continuous location tracking techniques for wizard of Oz testing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1019–1022.

Mackay, W.E., Appert, C., Beaudouin-Lafon, M., Chapuis, O., Du, Y., Fekete, J.-D., et al., 2007. Usability evaluation: Touchstone: exploratory design of experiments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1425–1434.

Nielsen, J., 2005. Time budgets for usability sessions. Alert Box. September 12. Retrieved from http://www.useit.com/alertbox/usability_sessions.html.

Oehlert, G., 2000. A First Course in Design and Analysis of Experiments. Freeman and Company, New York.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., Carey, T., 1994. Human–Computer Interaction. Addison-Wesley Longman Ltd., Essex, UK.

Rosenthal, R., Rosnow, R., 2008. Essentials of Behavioral Research: Methods and Data Analysis, third ed. McGraw Hill, Boston, MA.

Smart, R.G., 1966. Subject selection bias in psychological research. Canadian Psychology 7a, 115–121.

Taylor, B., Dey, A., Siewiorek, D., Smailagic, A., 2016. Customizable 3D printed tactile maps as interactive overlays. In: Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 71–79.

Wallace, D., Anderson, N., Shneiderman, B., 1993. Time stress effects on two menu selection systems. In: Shneiderman, B. (Ed.), Sparks of Innovation in Human–Computer Interaction. Ablex Publishing Corporation, Norwood, NJ.

Warr, A., Chi, E., Harris, H., Kuscher, A., Chen, J., Flack, R., et al., 2016. Window shopping: a study of desktop window switching. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3335–3338.