CHAPTER

Experimental research 2

A variety of laboratory and nonlaboratory research methods are available for human-computer interaction (HCI) researchers or practitioners when studying interfaces or applications. The most frequently used include observations, field studies, surveys, usability studies, interviews, focus groups, and controlled experiments (Shneiderman et al., 2017). In order to study how users enter information into their mobile phones, researchers may choose to observe mobile phone users in a natural setting, such as individuals who are using a cell phone in a company lobby, an airport, or a park. They may develop a survey that addresses questions that they would like to have answered and ask mobile phone users to respond to the survey. They may interview a number of mobile phone users to find out how they enter information into their phones. They may also choose to recruit a number of participants and run a usability test in a lab-based environment. Another option is to specify several conditions and run a strictly controlled lab-based experiment.

We can continue to add more options to the researchers' list: focus groups, field studies, and so on. Each of these options has its own strengths and weaknesses. Unobtrusively observing users in natural settings may allow the researcher to identify the patterns that are most representative of the use of the mobile phone in natural settings, but observation studies can be extremely time consuming. The researchers may wait for hours only to find that none of the individuals being observed has used the functions in which they are most interested. The survey approach may allow the researchers to reach a large number of users, say over a hundred, in a short period of time, but the participants may misunderstand the questions, the data collected may not represent depth in understanding, and the participant sample can be highly biased. Interviews allow the researchers to clarify questions and dig deeper with follow-up questions when a participant provides interesting feedback. However, interviews cost significantly more time and money than surveys. Usability tests provide a quick and comparatively low-cost method of identifying key usability problems in an interface or application, but they cannot guarantee that all critical design problems can be identified.

Choosing which method to use is a highly context-dependent issue related to a variety of factors including the primary purpose of the study, time constraints, funding, the participant pool, and the researchers' experience. We discuss in more detail in Chapter 3 on how to select the best research method. This chapter examines experimental research in general and focuses on the very basics of conducting experimental studies. We discuss how to develop research hypotheses and how to test the validity

Research Methods in Human-Computer Interaction. http://dx.doi.org/10.1016/B978-0-12-805390-4.00002-9

of a hypothesis. Important concepts related to hypothesis testing, such as Type I and Type II errors and their practical implications, are examined in detail.

2.1 TYPES OF BEHAVIORAL RESEARCH

Viewed broadly, all of the methods mentioned above are kinds of empirical investigation that can be categorized into three groups: descriptive investigations, relational investigations, and experimental investigations (Rosenthal and Rosnow, 2008). Descriptive investigations, such as observations, surveys, and focus groups, focus on constructing an accurate description of what is happening. For example, a researcher may observe that 8 out of 10 teenagers in a class who frequently play a specific computer game can touch type while only 2 out of 12 teenagers in the same class who do not play the game can touch type. This raises an interesting observation. But it does not allow the establishment of a relationship between the two factors: playing the game and typing. Neither does it enable the researcher to explain why this happens.

Relational investigations enable the researcher to identify relations between multiple factors. That is, the value of factor X changes as the value of factor Y changes. For example, the researcher may collect data on the number of hours that the teenagers play the computer game per week and measure their typing speed. The researcher can run a correlation analysis[1] between the number of hours and typing speed. If the result is significant, it suggests that there is a relationship between typing speed and the time spent playing the game. The results of relational studies usually carry more weight than what can be learned through descriptive studies. However, relational studies can rarely determine the causal relationship between multiple factors (Cooper and Schindler, 2000; Rosenthal and Rosnow, 2008).

Using the same example, the significant correlation result does not allow the researcher to determine the cause of the observed relationship. It is possible that playing the computer game improves typing speed. It is also possible that teenagers who type well tend to like the game more and spend more time on it. To complicate matters even more, the correlation can be due to hidden factors that the researcher has not considered or studied. For example, it is possible that teenagers who read well tend to type faster and that teenagers who read well tend to like the game more and spend more time on it. In this case, playing the computer game has no impact on the typing speed of the teenagers.

How, then, can the researchers determine the causal effect between two factors? The answer lies in experimental research (Kirk, 1982; Oehlert, 2000). The researchers may recruit teenagers in the same age group and randomly assign the teenagers to two groups. One group will spend a certain amount of time playing the computer game every week and the other group will not. After a period of time (e.g., 3 months or longer), the researchers can measure each teenager's typing speed. If the teenagers who play the computer game type significantly faster than the teenagers who do not

[1] Correlation analysis is a statistical test designed to identify relationships between two or more factors. Details of correlation analysis are discussed in Chapter 4.

play the game, the researchers can confidently draw the conclusion that playing this computer game improves the typing skills of teenagers.

As shown in the above example and summarized in Table 2.1, the most notable difference between experimental research and the other two types of investigation is that experimental research enables the identification of causal relationships. Simply put, it can tell how something happens and, in some cases, why it happens. The ability of experimental research to identify the true cause of a phenomenon allows researchers to manipulate the way we do research and achieve the desired results. To give a few examples, experimental studies are widely adopted in the field of medicine to identify better drugs or treatment methods for diseases. Scientists also use experimental research to investigate various questions originating from both the macro-world, such as the impact of acid rain on plants, and the micro-world, such as how nerves and cells function.

Table 2.1 Relationship Between Descriptive Research, Relational Research, and Experimental Research

Type of Research	Focus	General Claims	Typical Methods
Descriptive	Describe a situation or a set of events	X is happening	Observations, field studies, focus groups, interviews
Relational	Identify relations between multiple variables	X is related to Y	Observations, field studies, surveys
Experimental	Identify causes of a situation or a set of events	X is responsible for Y	Controlled experiments

The three kinds of research methods are not totally independent but highly intertwined. Typical research projects include a combination of two or even three kinds of investigation. Descriptive investigations are often the first step of a research program, enabling researchers to identify interesting phenomena or events that establish the cornerstone of the research and identify future research directions. Relational investigations enable researchers or practitioners to discover connections between multiple events or variables. Ultimately, experimental research provides the opportunity to explore the fundamental causal relations. Each of the three kinds of investigation is of great importance in the process of scientific discovery.

2.2 RESEARCH HYPOTHESES

An experiment normally starts with a research hypothesis. A hypothesis is a precise problem statement that can be directly tested through an empirical investigation. Compared with a theory, a hypothesis is a smaller, more focused statement that can be examined by a single experiment (Rosenthal and Rosnow, 2008). In contrast, a

theory normally covers a larger scope and the establishment of a theory normally requires a sequence of empirical studies. A concrete research hypothesis lays the foundation of an experiment as well as the basis of statistical significance testing.

THEORY VS HYPOTHESIS

The differences between theories and hypotheses can be clearly demonstrated by the extensive HCI research into Fitts' law (Fitts, 1954), one of the most widely accepted theories in the HCI field. It states a general relationship between movement time, navigation distance, and target size for pointing tasks in an interface:

> *In movement tasks, the movement time increases as the movement distance increases and the size of the target decreases. The movement time has a log linear relationship with the movement distance and the width of the target.*

Fitts' law is a general theory that may apply to various kinds of pointing devices. It is impossible to validate Fitts' law in a few experiments. Since Fitts' law was proposed, hundreds of user studies have been conducted on various pointing devices and tasks to validate and modify Fitts' law. The research hypothesis of each of those studies is a much more focused statement covering a small, testable application domain.

For example, Miniotas (2000) examined hypotheses about the performance of two pointing devices: a mouse and an eye tracker. Movement time was shorter for the mouse than for the eye tracker. Fitts' law predicted the navigation time fairly well for both the mouse and the eye tracker, indicating the potential to apply Fitts' law to technologies that do not rely on hand-based control. Accot and Zhai (2003) investigated Fitts' law in the context of two-dimensional targets. More recently, Bi et al. (2013) developed a FFitts law model that expanded Fitts' law to finger touch input.

2.2.1 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

An experiment normally has at least one null hypothesis and one alternative hypothesis. A null hypothesis typically states that there is no difference between experimental treatments. The alternative hypothesis is always a statement that is mutually exclusive with the null hypothesis. The goal of an experiment is to find statistical evidence to refute or nullify the null hypothesis in order to support the alternative hypothesis (Rosenthal and Rosnow, 2008). Some experiments may have several pairs of null hypotheses and alternative hypotheses. The characteristics of null and alternative hypotheses can be better explained through the following hypothetical research case.

Suppose the developers of a website are trying to figure out whether to use a pull-down menu or a pop-up menu in the home page of the website. The developers decide to conduct an experiment to find out which menu design will allow the users

to navigate the site more effectively. For this research case, the null and alternative hypotheses[2] can be stated in classical statistical terms as follows:

- H_0: There is no difference between the pull-down menu and the pop-up menu in the time spent locating pages.
- H_1: There is a difference between the pull-down menu and the pop-up menu in the time spent locating pages.

From this example, we can see that the null hypothesis usually assumes that there is no difference between two or more conditions. The alternative hypothesis and the null hypothesis should be mutually exclusive. That is, if the null hypothesis is true, the alternative hypothesis must be false, and vice versa. The goal of the experiment is to test the null hypothesis against the alternative hypothesis and decide which one should be accepted and which one should be rejected. The results of any significance test tell us whether it is reasonable to reject the null hypothesis and the likelihood of being wrong if rejecting the null hypothesis. We explain this topic in more detail in Section 2.5.

Many experiments examine multiple pairs of null and alternative hypotheses. For example, in the research case above, the researchers may study the following additional hypotheses:

- H_0: There is no difference in user satisfaction rating between the pull-down menu and the pop-up menu.
- H_1: There is a difference in user satisfaction rating between the pull-down menu and the pop-up menu.

There is no limit on the number of hypotheses that can be investigated in one experiment. However, it is generally recommended that researchers should not attempt to study too many hypotheses in a single experiment. Normally, the more hypotheses to be tested, the more factors that need to be controlled and the more variables that need to be measured. This results in very complicated experiments, subject to a higher risk of design flaws.

In order to conduct a successful experiment, it is crucial to start with one or more good hypotheses (Durbin, 2004). A good hypothesis normally satisfies the following criteria:

- is presented in precise, lucid language;
- is focused on a problem that is testable in one experiment;
- clearly states the control groups or conditions of the experiment.

In the early stages of a research project, researchers usually find themselves confronted with a broad and vague task. There are no well-defined research questions. There are no focused, testable research hypotheses. The common way to initiate a research project is to conduct exploratory descriptive investigations such as observations, interviews, or focus groups. Well-conducted descriptive investigations help researchers identify key research issues and come up with appropriate control groups to be manipulated as well as dependent variables to be measured.

[2] Traditionally, H_0 is used to represent the null hypothesis and H_1 to represent the alternative hypothesis.

2.2.2 DEPENDENT AND INDEPENDENT VARIABLES

A well-defined hypothesis clearly states the dependent and independent variables of the study. Independent variables refer to the factors that the researchers are interested in studying or the possible "cause" of the change in the dependent variable. The term "independent" is used to suggest that the variable is independent of a participant's behavior. Dependent variables refer to the outcome or effect that the researchers are interested in. The term "dependent" is used to suggest that the variable is dependent on a participant's behavior or the changes in the independent variables. In experiments, the primary interest of researchers is to study the relationship between dependent variables and independent variables. More specifically, the researcher wants to find out whether and how changes in independent variables induce changes in dependent variables.

A useful rule of thumb to differentiate dependent variables from independent variables is that independent variables are usually the treatments or conditions that the researchers can control while dependent variables are usually the outcomes that the researchers need to measure (Oehlert, 2000). For example, consider the null hypothesis proposed in the research case in Section 2.2.1:

> *There is no difference between the pull-down menu and the pop-up menu in the time spent locating pages.*

The independent variable is the type of menu (pull-down or pop-up). The dependent variable is the time spent in locating web pages. During the experiment, the researchers have full control over the types of menu with which each participant interacts by randomly assigning each participant to an experimental condition. In contrast, "time" is highly dependent on individual behavioral factors that the researchers cannot fully control. Some participants will be faster than others due to a number of factors, such as the type of menu, previous computer experience, physical capabilities, reading speed, and so on. The researchers need to accurately measure the time that each participant spends in locating pages and to relate the results to the independent variable in order to make a direct comparison between the two types of menu design.

2.2.3 TYPICAL INDEPENDENT VARIABLES IN HCI RESEARCH

Independent variables are closely related to the specific research field. It is obvious that the factors frequently investigated in medical science are drastically different from those examined in physics or astronomy. In the HCI field, independent variables are usually related to technologies, users, and the context in which the technology is used. Typical independent variables that relate to technology include:

- different types of technology or devices, such as typing versus speech-based dictation, mouse versus joystick, touch pad, and other pointing devices;
- different types of design, such as pull-down menu versus pop-up menu, font sizes, contrast, background colors, and website architecture.

Typical independent variables related to users include age, gender, computer experience, professional domain, education, culture, motivation, mood, and disabilities. Using age as an example, we know that human capabilities change during their life span. Children have a physically smaller build and shorter attention span. Their reading skills, typing skills, and cognitive capabilities are all limited compared to typical computer users between ages 20 and 55. At the other end of the scale, senior citizens experience deterioration in cognitive, physical, and sensory capabilities. As a result, users in different age groups interact differently with computers and computer-related devices. Most computer applications are designed by people between 20 and 50 years of age who have little or no knowledge or experience in the interaction style or challenges faced by the younger and older user groups (Chisnell, 2007). In order to understand the gap created by age differences, a number of studies have been conducted to compare the interaction styles of users in different age groups (Zajicek, 2006; Zajicek and Jonsson, 2006).

Typical independent variables related to the context of use of technologies include both physical factors, such as environmental noise, lighting, temperature, vibration, users' status (e.g., seated, walking or jogging) (Price et al., 2006), and social factors, such as the number of people surrounding the user and their relation to the user.

2.2.4 TYPICAL DEPENDENT VARIABLES IN HCI RESEARCH

Dependent variables frequently measured can be categorized into five groups: efficiency, accuracy, subjective satisfaction, ease of learning and retention rate, and physical or cognitive demand.

Efficiency describes how fast a task can be completed. Typical measures include time to complete a task and speed (e.g., words per minute, number of targets selected per minute)

Accuracy describes the states in which the system or the user makes errors. The most frequently used accuracy measure is error rate. Numerous metrics to measure error rate have been proposed for various interaction tasks, such as the “minimum string distance” proposed for text entry tasks (Soukoreff and Mackenzie, 2003). In HCI studies, efficiency and accuracy are not isolated but are highly related factors. There is usually a trade-off between efficiency and accuracy, meaning that, when the other factors are the same, achieving a higher speed will result in more errors and ensuring fewer errors will lower the speed. Consequently, any investigation that only measures one of the two factors misses a critical side of the picture.

Subjective satisfaction describes the user's perceived satisfaction with the interaction experience. The data is normally collected using Likert scale ratings (e.g., numeric scales from 1 to 5) through questionnaires.

Ease of learning and retention rate describe how quickly and how easily an individual can learn to use a new application or complete a new task and how long they retain the learned skills (Feng et al., 2005). This category is less studied than the previous three categories but is highly important for the adoption of information technology.

Variables in the fifth category describe the cognitive and physical demand that an application or a task exerts on an individual or how long an individual can interact with an application without significant fatigue. This category of measures is less studied but they play an important role in technology adoption.

2.3 BASICS OF EXPERIMENTAL RESEARCH

In order to understand why experimental research can allow causal inference while descriptive and relational investigations do not, we need to discuss the characteristics of experimental research. In a true experimental design, the investigator can fully control or manipulate the experimental conditions so that a direct comparison can be made between two or more conditions while other factors are, ideally, kept the same. One aspect of the full control of factors is complete randomization, which means that the investigator can randomly assign participants to different conditions. The capability to effectively control for variables not of interest, therefore limiting the effects to the variables being studied, is the feature that most differentiates experimental research from quasi-experimental research, descriptive investigations, and relational investigations.

2.3.1 COMPONENTS OF AN EXPERIMENT

After a research hypothesis is identified, the design of an experiment consists of three components: treatments, units, and assignment method (Oehlert, 2000). Treatments, or conditions, refer to the different techniques, devices, or procedures that we want to compare. Units are the objects to which we apply the experiment treatments. In the field of HCI research, the units are normally human subjects with specific characteristics, such as gender, age, or computing experience. Assignment method refers to the way in which the experimental units are assigned different treatments.

We can further explain these three terms through an example. Suppose a researcher is running an experiment to compare typing speed using a traditional QWERTY keyboard and a DVORAK keyboard.[3] The treatment of this experiment is the type of keyboard: QWERTY or DVORAK. The experiment units are the participants recruited to join the study. To achieve the goal of fair comparison, the researchers would have to require that the participants have no previous experience using either keyboard. If most participants can touch type using the QWERTY keyboard but have never used a DVORAK keyboard before, it is obvious that the results will be highly biased towards the QWERTY keyboard. The researcher can employ different methods to randomly assign the participants into each of the two conditions. One well-known traditional method is to toss a coin. If a head is tossed, the participant is assigned to the QWERTY condition. If a tail is tossed, the participant is assigned to the DVORAK condition. Obviously, researchers are not busy tossing coins in their lab; more convenient randomization methods are used today. We discuss those methods in Section 2.3.2.

[3] 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.

The keyboard comparison case illustrates a simple between-subject[4] design with two conditions. There are much more complicated designs involving multiple treatments and both between-subject and within-subject[5] comparisons. No matter how complicated the design is, all experiments consist of these three major components: treatments, units, and assignment methods.

2.3.2 RANDOMIZATION

The power of experimental research lies in its ability to uncover causal relations. The major reason why experimental research can achieve this goal is because of complete randomization. Randomization refers to the random assignment of treatments to the experimental units or participants (Oehlert, 2000).

In a totally randomized experiment, no one, including the investigators themselves, is able to predict the condition to which a participant is going to be assigned. For example, in the QWERTY vs. DVORAK experiment, when a participant comes in, the researchers do not know whether the participant will be using the QWERTY keyboard or the DVORAK keyboard until they toss a coin and find out whether it settles as heads or tails. Since the outcome of tossing the coin is totally random and out of the control of the researchers, the researchers have no influence, whether intentionally or subconsciously, on the assignment of the treatment to the participant. This effectively controls the influence of hidden factors and allows a clean comparison between the experiment conditions.

Traditional randomization methods include tossing a coin, throwing dice, spinning a roulette wheel, or drawing capsules out of an urn. However, these types of randomization are rarely used in behavioral research and HCI studies nowadays. One method to randomize the selection of experimental conditions or other factors is the use of a random digit table. Table 2.2 is an abbreviated random digit table taken from the large random digit table generated by RAND (1955). The original table consisted of a million random digits.

Table 2.2 An Abbreviated Random Digit Table

Line	Random Digits				
000	10097	32533	76520	13586	34673
001	37542	04805	64894	74296	24805
002	08422	68953	19645	09303	23209
003	99019	02529	09376	70715	38311
004	12807	99970	80157	36147	64032
005	66065	74717	34072	76850	36697

There are several ways to use this table. Suppose we are running a study that compares three types of navigation schemes for a website: topical, audience split,

[4] A between-subject design means each participant only experiences one task condition. The details of between-subject design are discussed in Chapter 3.

[5] A within-subject design means each participant experiences multiple task conditions. The details of within-subject design is discussed in Chapter 3.

and organizational. We recruit 45 participants and need to assign each of them to one of the three conditions. We can start anywhere in the random digit table and count in either direction. For example, if we start from the third number on the first row and count to the right for three numbers, we get 76520, 13586, and 34673. We can assign the first three participants to the conditions according to the order of the three random numbers. In this case, 76520 is the largest, corresponding to condition 3; 13586 is the smallest, corresponding to condition 1; and 34673 corresponds to condition 2. This means that the first participant is assigned to the design with the organizational navigation scheme, the second participant to the topical scheme, and the third participant to the audience split scheme. We can continue counting the numbers and repeating the process until all 45 participants are assigned to specific conditions.

Nowadays, software-driven randomization is also commonly used among researchers and practitioners. A large number of randomization software resources are available online, some of them free of charge, such as the services offered at http://www.randomization.com. Randomization functions are also available in most of the commercial statistical software packages, such as SAS, SPSS, and SYSTAT.

In a well-designed experiment, you will frequently find that you not only need to randomize the assignment of experiment conditions, but other factors as well. In a longitudinal study[6] reported by Sears et al. (2001, 2003), the researchers investigated the use of recognition software to generate text documents. Each of the 15 participants completed a total of nine tasks on different days. During each task, the participant composed a text document of approximately 500 words in response to one of nine predefined scenarios. The researchers found it necessary to randomize the order of the scenarios being used in the nine tasks. If the order of the scenarios were not randomized, it is likely that the characteristics of the scenarios would become a factor that influences the results. Randomizing the order of the scenarios cancels out the potential errors introduced by differences in scenarios.[7]

Counter balancing is commonly used in experiments to address the problem of systematic differences between successive conditions. In this case, researchers usually rotate the sequences of treatments or conditions through a "Latin Square Design" illustrated in Table 2.3 (Rosenthal and Rosnow, 2008). In this table, letters A, B, C,

Table 2.3 Latin Square Design

	Order of Administration			
	1	2	3	4
Sequence 1	A	B	C	D
Sequence 2	B	C	D	A
Sequence 3	C	D	A	B
Sequence 4	D	A	B	C

[6] A study in which data is gathered for the same participants repeatedly over a period of time.

[7] Special attention was paid during the development of the scenarios so that they are similar to each other in the degree of difficulty in responding, which was confirmed by the reported results. However, it is good practice to randomize the order of the scenarios in case there are unanticipated differences between them.

D each represents a condition. Each row represents a sequence of four conditions to which one participant can be randomly assigned. Note that each condition only appears once in each row and column, suggesting that the order of the conditions is completely counter balanced for these four participants.

2.4 SIGNIFICANCE TESTS

2.4.1 WHY DO WE NEED THEM?

Almost all experimental investigations are analyzed and reported through significance tests. If you randomly pick up an HCI-related journal article or a conference paper, it is very likely that you will encounter statements similar to the following:

> *On average, participants performed significantly better ($F(1,25) = 20.83$, $p < 0.01$) … in the dynamic peephole condition … rather than the static peephole condition. (Mehra et al., 2006)*
>
> *A t test showed that there was a significant difference in the number of lines of text entered ($t(11) = 6.28$, $p < 0.001$) with more entered in the tactile condition. (Brewster et al., 2007)*

Why do you need to run significance tests on your data? What is wrong with the approach of comparing two mean values of error rate and then claiming that the application with the lower mean value is more accurate than the other application? Here we encounter a fundamental issue in statistics that has to be clarified in order to understand the numerous concepts, terms, and methods that will be discussed in the rest of this chapter and in Chapters 4 and 5. Let us consider the following two statements:

1. Mike's height is 6′2″. Mary's height is 5′8″. So Mike is taller than Mary.
2. The average height of three males (Mike, John, and Ted) is 5′5″. The average height of three females (Mary, Rose, and Jessica) is 5′10″. So females are taller than males.

It should not be difficult for you to tell that the first statement is correct while the second one is not. In the first statement, the targets being compared are the heights of two individuals, both known numbers. Based on the two numbers, we know that Mike is taller than Mary. This is simple to understand, even for a child. When the values of the members of the comparison groups are all known, you can directly compare them and draw a conclusion. No significance test is needed since there is no uncertainty involved.

What is wrong with the second statement? People may give various responses to this question, such as:

- Well, by common sense, I know males are generally taller than females.
- I can easily find three other males and three other females, in which the average height of the three males is higher than that of the three females.
- There are only three individuals in each group. The sizes of the comparison groups are too small.
- The individuals in both the male group and the female group are not representative of the general population.

All of the above responses are well grounded, though the last two responses have deeper statistical roots. The claim that females are taller than males is wrong due to inappropriate sampling. The distribution of the heights of the human population (and many other things in our life) follows a pattern called "normal distribution." Data sets that follow normal distribution can be illustrated by a bell-shaped curve (see Figure 2.1), with the majority of the data points falling in the central area surrounding the mean of the population (μ). The further a value is from the population mean, the fewer data points would fall in the area around that value.

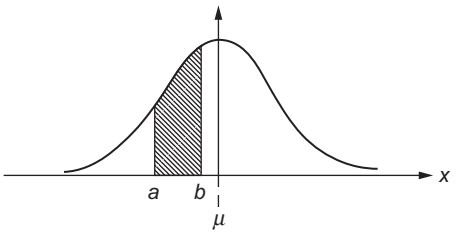

FIGURE 2.1

Normal distribution curve.

When you compare two large populations, such as males and females, there is no way to collect the data from every individual in the population. Therefore, you select a smaller group from the large population and use that smaller group to represent the entire population. This process is called sampling. In the situation described in statement 2 above, the three males selected as the sample population happened to be shorter than average males, while the three females selected as samples happened to be taller than average females, thus resulting in a misleading conclusion. Randomization methods and large sample sizes can greatly reduce the possibility of making this kind of error in research.

Since we are not able to measure the heights of all males and females, we can only sample a subgroup of people from the entire population. Significance tests allow us to determine how confident we are that the results observed from the sampling population can be generalized to the entire population. For example, a *t* test that is significant at $P < 0.05$ suggests that we are confident that 95% of the time the test result correctly applies to the entire population. We further explore the concept of significance tests in the next section.

2.4.2 TYPE I AND TYPE II ERRORS

In technical terms, significance testing is a process in which a null hypothesis (H_0) is contrasted with an alternative hypothesis (H_1) to determine the likelihood that the null hypothesis is true. All significance tests are subject to the risk of Type I and Type II errors.

A Type I error (also called an α error or a "false positive") refers to the mistake of rejecting the null hypothesis when it is true and should not be rejected. A Type II error (also called a β error or a "false negative") refers to the mistake of not rejecting

the null hypothesis when it is false and should be rejected (Rosenthal and Rosnow, 2008). A widely used example to demonstrate Type I and Type II errors is the judicial case. In the US justice system, a defendant is presumed innocent. This presumption leads to the following null and alternative hypotheses:

- H_0: The defendant is innocent.
- H_1: The defendant is guilty.

A Type I error occurs when the jury decides that the defendant is guilty when he is actually innocent, meaning that the null hypothesis is rejected when it is true. A Type II error occurs when the jury decides that the defendant is innocent when he is actually guilty, meaning that the null hypothesis is accepted when it is false. Table 2.4 illustrates these errors. In the ideal case, the jury should always reach the decision that the defendant is guilty when he is actually guilty and vice versa. But in reality, the jury makes mistakes occasionally. Each type of error has costs. When a Type I error occurs, an innocent person would be sent to prison or may even lose his life; when a Type II error occurs, a criminal is set free and may commit another crime.

Table 2.4 Type I and Type II Errors in the Judicial Case

	Jury Decision: Not Guilty	**Jury Decision : Guilty**
Reality: Not Guilty	No error	Type I error
Reality: Guilty	Type II error	No error

Let us further examine Type I and Type II errors through a study in the HCI domain. Suppose a bank hires several HCI researchers to evaluate whether ATMs with a touch-screen interface are easier to use than the ATMs with buttons that the bank branches are currently using. In this case, the null hypothesis and the alternative hypothesis are:

- H_0: There is no difference between the ease of use of ATMs with touch screens and ATMs with buttons.
- H_1: ATMs with touch screens are easier to use than ATMs with buttons.

The possible Type I and Type II errors in this study are illustrated in Table 2.5. A Type I error occurs when the research team decides that touch-screen ATMs are easier to use than ATMs with buttons, when they are actually not. A Type II error occurs when the research team decides that touch-screen ATMs are no better than ATMs with buttons, when they are. Again, each type of error can induce negative consequences. When a Type I error occurs, the bank may spend money to switch to touch-screen ATMs that do not provide better service to the customers. When a Type II error occurs, the bank chooses to stay with ATMs with buttons and loses the opportunity to improve the service that it provides to its customers.

Table 2.5 Type I and Type II Errors in a Hypothetical HCI Experiment

	Study Conclusion: No Difference	Study Conclusion: Touchscreen ATM is Easier to Use
Reality: No Difference	No error	Type I error
Reality: Touchscreen ATM is Easier to Use	Type II error	No error

It is generally believed that Type I errors are worse than Type II errors. Statisticians call Type I errors a mistake that involves "gullibility." A Type I error may result in a condition worse than the current state. For example, if a new medication is mistakenly found to be more effective than the medication that patients are currently taking, the patients may switch to new medication that is less effective than their current treatment. Type II errors are mistakes that involve "blindness" and can cost the opportunity to improve the current state. In the medication example, a Type II error means the test does not reveal that the new medication is more effective than the existing treatment; the patients stick with the existing treatment and miss the opportunity of a better treatment.

2.4.3 CONTROLLING THE RISKS OF TYPE I AND TYPE II ERRORS

When designing experiments and analyzing data, you have to evaluate the risk of making Type I and Type II errors. In statistics, the probability of making a Type I error is called alpha (or significance level, *P* value). The probability of making a Type II error is called beta. The statistical power of a test, defined as $1-\beta$, refers to the probability of successfully rejecting a null hypothesis when it is false and should be rejected (Cohen, 1988).[8]

It should be noted that alpha and beta are interrelated. Under the same conditions, decreasing alpha reduces the chance of making Type I errors but increases the chance of making Type II errors. Simply put, if you want to reduce the chance of making Type I errors with all other factors being the same, you can do so by being less gullible. However, in doing so, you increase the odds that you miss something that is in fact true, meaning that your research is more vulnerable to Type II errors.

In experimental research, it is generally believed that Type I errors are worse than Type II errors. So a very low *P* value (0.05) is widely adopted to control the occurrence of Type I errors. If a significance test returns a value that is significant at $P<0.05$, it means that the probability of making a Type I error is below 0.05. In other words, the probability of mistakenly rejecting a null hypothesis is below 0.05. In order to reduce Type II errors, it is generally suggested that you use a relatively

[8] How alpha and beta are calculated is beyond the scope of this book. For detailed discussion of the calculation, please refer to Rosenthal and Rosnow (2008).

large sample size so that the difference can be observed even when the effect size is relatively small. If interested, you can find more detailed discussions on statistical power in Rosenthal and Rosnow (2008).

2.5 LIMITATIONS OF EXPERIMENTAL RESEARCH

Experimental research methods originated from behavioral research and are largely rooted in the field of psychology. Experimental research has been a highly effective research approach and has led to many groundbreaking findings in behavioral science in the 20th century. Experimental research certainly plays an important role in the field of HCI. A large number of studies that explored fundamental interaction theories and models, such as Fitts' law, employed the approach of experimental research. To date, experimental research remains one of the most effective approaches to making findings that can be generalized to larger populations.

On the other hand, experimental research also has notable limitations. It requires well-defined, testable hypotheses that consist of a limited number of dependent and independent variables. However, many problems that HCI researchers or practitioners face are not clearly defined or involve a large number of potentially influential factors. As a result, it is often very hard to construct a well-defined and testable hypothesis. This is especially true when studying an innovative interaction technique or a new user population and in the early development stage of a product.

Experimental research also requires strict control of factors that may influence the dependent variables. That is, except the independent variables, any factor that may have an impact on the dependent variables, often called potential confounding variables, needs to be kept the same under different experiment conditions. This requirement can hardly be satisfied in many HCI studies. For example, when studying how older users and young users interact with computer-related devices, there are many factors besides age that are different between the two age groups, such as educational and knowledge background, computer experience, frequency of use, living conditions, and so on. If an experiment is conducted to study the two age groups, those factors will become confounding factors and may have a significant impact on the observed results. This problem can be partially addressed in the data collection and data analysis stages. In the data collection stage, extra caution should be taken when there are known confounding factors. Increasing the sample size may reduce the impact of the confounding factors. When recruiting participants, prescreening should be conducted to make the participants in different groups as homogeneous as possible. When confounding factors are inevitable, specific data analysis methods can be applied so that the impact of the confounding factors can be filtered out. A common method for this purpose is the analysis of covariables.

Lab-based experiments may not be a good representation of users' typical interaction behavior. It has been reported that participants may behave differently in lab-based experiments due to the stress of being observed, the different environment,

or the rewards offered for participation. This phenomenon, called the "Hawthorne effect," was documented around 60 years ago (Landsberger, 1958). In many cases, being observed can cause users to make short-term improvements that typically do not last once the observation is over.

However, it should be noted that the context of the Hawthorne studies and HCI-related experiments is significantly different (Macefield, 2007). First, the Hawthorne studies were all longitudinal while most HCI experiments are not. Secondly, all the participants in the Hawthorne studies were experts in the tasks being observed while most HCI experiments observe novice users. Thirdly, the Hawthorne studies primarily focused on efficiency while HCI experiments value other important measures, such as error rates. Finally, the participants in the Hawthorne study had a vested interest in a successful outcome for the study since it was a point of contact between them and their senior management. In contrast, most HCI studies do not carry this motivation. Based on those reasons, we believe that the difference between the observed results of HCI experiments and the actual performance is not as big as that observed in the Hawthorne studies. But still, we should keep this potential risk in mind and take precautions to avoid or alleviate the impact of the possible Hawthorne effect.

EMPIRICAL EVALUATION IN HCI

The validity of empirical experiments and quantitative evaluation in HCI research has been doubted by some researchers. They argue that the nature of research in HCI is very different from traditional scientific fields, such as physics or chemistry, and, therefore, the results of experimental studies that suggest one interface is better than another may not be truly valid.

The major concern with the use of empirical experiments in HCI is the control of all possible related factors (Lieberman, 2007). In experiments in physics or chemistry, it is possible to strictly control all major related factors so that multiple experimental conditions are only different in the states of the independent variables. However, in HCI experiments, it is very difficult to control all potential factors and create experimental conditions that are exactly the same with the only exception of the independent variable. For instance, it is almost impossible to recruit two or more groups of participants with exactly the same age, educational background, and computer experience. All three factors may impact the interaction experience as well as the performance. It is argued that the use of significance tests in the data analysis stage only provides a veneer of validity when the potentially influential factors are not fully controlled (Lieberman, 2007).

We agree that experimental research has its limitations and deficiencies, just as any other research method does. But we believe that the overall validity of experimental research in the field of HCI is well-grounded. Simply observing a few users trying two interfaces does not provide convincing results on the

performance and preference of the target population. Controlled experiments have allowed us to make critical and generalizable findings that other methods would not be able to provide. The truth is, experimental research and significance testing is the only approach that enables us to make judgments with systematically measured confidence and reliability. The control of confounding factors is challenging but the impact of those factors can be reduced to acceptable levels through well-designed and implemented experiments, which we discuss in detail in Chapter 3.

2.6 SUMMARY

Research in HCI examines human behavior in relation to computers or computer-related devices. There are three major types of research methods for studying human behavior: descriptive, relational, and experimental. The major strength of experimental research, compared to the other two types, is that it allows the identification of causal relationships between entities or events.

After a hypothesis is constructed, the design of an experiment consists of three components: treatments, units, and the assignment method. In an experiment, the process of sample selection needs to be randomized or counter-balanced, as does the assignment of treatments, or experiment conditions. Many methods can be used to randomly select samples or assign experiment conditions, including, but not limited to, the random digit table and software-generated randomization schemes.

Successful experimental research depends on well-defined research hypotheses that specify the dependent variables to be observed and the independent variables to be controlled. Usually a pair of null and alternative hypotheses is proposed and the goal of the experiment is to test whether the null hypothesis can be rejected or the alternative hypothesis can be accepted. Good research hypotheses should have a reasonable scope that can be tested within an experiment; clearly defined independent variables that can be strictly controlled; and clearly defined dependent variables that can be accurately measured.

Significance testing allows us to judge whether the observed group means are truly different. All significance tests are subject to two types of error. Type I errors refer to the situation in which the null hypothesis is mistakenly rejected when it is actually true. Type II errors refer to the situation of not rejecting the null hypothesis when it is actually false. It is generally believed that Type I errors are worse than Type II errors, therefore the alpha threshold that determines the probability of making Type I errors should be kept low. The widely accepted alpha threshold is 0.05. With its notable strengths, experimental research also has notable limitations when applied in the field of HCI: difficulty in identifying a testable hypothesis, difficulty in controlling potential confounding factors, and changes in observed behavior as compared to behavior in a more realistic setting. Therefore, experimental research methods should only be adopted when appropriate.

DISCUSSION QUESTIONS

1. What is descriptive research?
2. What is relational research?
3. What is experimental research?
4. What is randomization in experimental research? Discuss several examples of randomization methods.
5. What is a research hypothesis? What are the characteristics of a good research hypothesis?
6. What is a dependent variable?
7. What is an independent variable?
8. What is a significance test? Why do we need to run significance tests?
9. What is a Type I error? What is a Type II error?
10. Discuss the practical implications of Type I errors and Type II errors.

RESEARCH DESIGN EXERCISES

1. A research team is investigating three possible navigation architectures for an e-commerce website. Thirty participants are recruited to test the website, with 10 participants testing each architecture. How should the participants be assigned to the three conditions?
2. Read the following hypotheses and identify the dependent variables and independent variables in each hypothesis.
    1. There is no difference in users' reading speed and retention rate when they view news on a desktop computer or a PDA.
    2. There is no difference in the target selection speed and error rate between joystick, touch screen, and gesture recognition.
    3. There is no difference in the technology adoption rate between two speech-based applications with different dialog designs.
    4. There is no difference in the reading skills of children who used educational software for 6 months compared to those who have never used the software.

3. A spam filter assigns ratings to all incoming emails. If the rating of an email is higher than a specific threshold, the email is deleted before it reaches the inbox. Answer the following questions based on this scenario:
   a. What is a Type I error in this scenario?
   b. What is a Type II error in this scenario?
   c. If the rating is assigned using a scale of 1–10, with 1 representing "definitely not spam" and 10 representing "definitely spam," what happens if the threshold is set to 1, 2, 3, …, 10?
   d. What do you think the appropriate threshold should be? Why?

REFERENCES

Accot, J., Zhai, S., 2003. Refining Fitts' law models for bivariate pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 193–200.

Bi, X., Li, Y., Zhai, S., 2013. FFitts law: modeling finger touch with Fitts' Law. In: Proceeding of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1363–1372.

Brewster, S., Chohan, F., Brown, L., 2007. Mobile interaction: tactile feedback for mobile interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–162.

Chisnell, D., 2007. Where technology meets green bananas. Interactions 14 (2), 10–11.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Academic Press, New York.

Cooper, D., Schindler, P., 2000. Business Research Methods, seventh ed. McGraw Hill, Boston, MA.

Durbin, C., 2004. How to come up with a good research question: framing the hypothesis. Respiratory Care 49 (10), 1195–1198.

Feng, J., Karat, C.-M., Sears, A., 2005. How productivity improves in hands-free continuous dictation tasks: lessons learned from a longitudinal study. Interacting with Computers 17 (3), 265–289.

Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47 (6), 381–391.

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

Landsberger, H., 1958. Hawthorne Revisited. Cornell University, Ithaca, NY.

Lieberman, H., 2007. The tyranny of evaluation. http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html (retrieved 16.11.07.).

Macefield, R., 2007. Usability studies and the Hawthorne Effect. Journal of Usability Studies 2 (3), 145–154.

Mehra, S., Werkhoven, P., Worring, M., 2006. Navigating on handheld displays: dynamic versus static peephole navigation. ACM Transactions on Computer-Human Interaction 13 (4), 448–457.

Miniotas, D., 2000. Application of Fitts' Law to eye gaze interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 339–340.

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

Price, K.J., Lin, M., Feng, J., Goldman, R., Sears, A., Jacko, J.A., 2006. Motion does matter: an examination of speech-based-text entry on the move. Universal Access in the Information Society 4 (3), 246–257.

RAND Corporation, 1955. A Million Random Digits with 100,000 Normal Deviates. Free Press, New York.

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

Sears, A., Karat, C.-M., Oseitutu, K., Karimullah, A., Feng, J., 2001. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. Universal Access in the Information Society 1 (1), 4–15.

Sears, A., Feng, J., Oseitutu, K., Karat, C.-M., 2003. Speech-based navigation during dictation: difficulties, consequences, and solutions. Human–Computer Interaction 18 (3), 229–257.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., 2017. Designing the User Interface: Strategies for Effective Human–Computer Interaction, sixth ed. Addison-Wesley, Boston, MA.

Soukoreff, W., MacKenzie, S., 2003. Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 113–120.

Zajicek, M., 2006. Aspects of HCI research for older people. Universal Access in the Information Society 5 (3), 279–386.

Zajicek, M., Jonsson, I., 2006. In-car speech systems for older adults: can they help and does the voice matter? International Journal of Technology, Knowledge, and Society 2 (6), 55–64.