15.3.3 Statistics: *t*-tests

There are many types of statistics that can be used to test the probability of a result occurring by chance, but *t*-tests are the most widely used statistical test in HCI and related fields, such as psychology. The scores, for example, time taken for each participant to select items from a menu in each condition (that is, context and cascading menus), are used to compute the means ($\bar{x}$) and standard deviations (SDs). The *standard deviation* is a statistical measure of the spread or variability around the mean. The *t*-test uses a simple equation to test the significance of the difference between the means for the two conditions. If they are significantly different from each other, we can reject the null hypothesis and in so doing infer that the alternative hypothesis holds. A typical *t*-test result that compared menu selection times for two groups with 9 and 12 participants each might be as follows:

$$\mathrm{t} = 4.53, \mathrm{p} < 0.05, \mathrm{df} = 19$$

The *t*-value of 4.53 is the score derived from applying the *t*-test; df stands for degrees of freedom, which represents the number of values in the conditions that are free to vary. This is a complex concept that we will not explain here other than to mention how it is derived and that it is always written as part of the result of a *t*-test. The df values are calculated by summing the number of participants in one condition minus 1 and the number of participants in the other condition minus 1. It is calculated as $\mathrm{df} = (N_a - 1) + (N_b - 1)$, where N_a is the number of participants in one condition and N_b is the number of participants in the other condition. In our example, $\mathrm{df} = (9 - 1) + (12 - 1) = 19$, p is the probability that the effect found did not occur by chance. So, when $p < 0.05$, it means that the effect found is probably not due to chance and that there is only a 5 percent possibility that it could be by chance. In other words, there most likely is a difference between the two conditions. Typically, a value of $p < 0.05$ is considered good enough to reject the null hypothesis, although lower levels of p are more convincing, for instance, $p < 0.01$ where the effect found is even less likely to be due to chance, there being only a 1 percent chance of that being the case.

15.4 Field Studies

Increasingly, more evaluation studies are being done in natural settings with either little or no control imposed on participants' activities. This change is largely a response to technologies being developed for use outside office settings. For example, mobile, ambient, IoT, and other technologies are now available for use in the home, outdoors, and in public places. Typically, field studies are conducted to evaluate these user experiences.

As mentioned in Chapter 14, evaluations conducted in natural settings are very different from those conducted in controlled environments, where tasks are set and completed in an orderly way. In contrast, studies in natural settings tend to be messy in the sense that activities often overlap and are constantly interrupted by events that are not predicted or controlled such as phone calls, texts, rain if the study is outside, and people coming and going. This follows the way that people interact with products in their everyday messy worlds, which is generally different from how they perform on fixed tasks in a laboratory setting. Evaluating how people think about, interact with, and integrate products within the settings in which they will ultimately be used, gives a better sense of how successful the products will be in the real world. The trade-off is that it is harder to test specific hypotheses about an interface because many environmental factors that influence the interaction cannot be controlled. Therefore, it is not possible to account, with the same degree of certainty, for how people react to or use a product as can be done in controlled settings like laboratories. This makes it more difficult to determine what causes a particular type of behavior or what is problematic about the usability of a product. Instead, qualitative accounts and descriptions of people's behavior and activities are obtained that reveal how they used the product and reacted to its design.

Field studies can range in time from just a few minutes to a period of several months or even years. Data is collected primarily by observing and interviewing people, such as by collecting video, audio, field notes, and photos to record what occurs in the chosen setting. In addition, participants may be asked to fill out paper-based or electronic diaries, which run on smartphones, tablets, or other handheld devices, at particular points during the day. The kinds of reports that can be of interest include being interrupted during an ongoing activity or when they encounter a problem when interacting with a product or when they are in a particular location, as well as how, when, and if they return to the task that was interrupted. This technique is based on the experience sampling method (ESM), discussed in Chapter 8, which is often used in healthcare (Price et al., 2018). Data on the frequency and patterns of certain daily activities, such as the monitoring of eating and drinking habits, or social interactions like phone and face-to-face conversations, are often recorded. Software running on the smartphones triggers messages to study participants at certain intervals, requesting them to answer questions or fill out dynamic forms and checklists. These might include recording what they are doing, what they are feeling like at a particular time, where they are, or how many conversations they have had in the last hour.

As in any kind of evaluation, when conducting a field study, deciding whether to tell the people being observed, or asked to record information, that they are being studied and how long the study or session will last is more difficult than in a laboratory situation. For example, when studying people's interactions with an ambient display, or the displays in a shopping mall described earlier (Dalton et al. 2016), telling them that they are part of a study will likely change the way they behave. Similarly, if people are using an online street map while walking in a city, their interactions may take only a few seconds, so informing them that they are being studied would disrupt their behavior. It is also important to ensure the privacy of participants in field studies. For example, participants in field studies that run over a period of weeks or months should be informed about the study and asked to sign an informed consent form in the usual way, as mentioned in Chapter 14. In studies that last for a long time, such as those in people's homes, the designers will need to work out and agree with the participants what part of the activity is to be recorded and how. For example, if the designers want to set up cameras, they need to be situated unobtrusively, and participants need to be informed in advance about where the cameras will be and when they will be recording their activities. The designers will also need to work out in advance what to do if the prototype or product breaks down. Can the participants be instructed to fix the problem themselves, or will the designers need to be called in? Security arrangements will also need to be made if expensive or precious equipment is being evaluated in a public place. Other practical issues may also need to be considered depending on the location, product being evaluated, and the participants in the study.

The study in which the Ethnobot (Tallyn et al., 2018) was used to collect information about what users did and how they felt while walking around at the Royal Highland Show in Scotland (discussed in Chapter 14) was an example of a field study. A wide range of other studies have explored how new technologies have been used and adopted by people in their own cultures and settings. By adopted, we mean how the participants use, integrate, and adapt the technology to suit their needs, desires, and ways of living. The findings from studies in natural settings are typically reported in the form of vignettes, excerpts, critical incidents, patterns of behavior, and narratives to show how the products are being used, adopted, and integrated into their surroundings.

15.4.1 In-the-Wild Studies

For several years now, it has become increasingly popular to conduct in-the-wild studies to determine how people use and persist in using a range of new technologies or prototypes *in situ*. The term *in-the-wild* reflects the context of the study, in which new technologies are deployed and evaluated in natural settings (Rogers, 2011). Instead of developing solutions that fit in with existing practices and settings, researchers often explore new technological possibilities that can change and even disrupt participants' behavior. Opportunities are created, interventions are installed, and different ways of behaving are encouraged. A key concern is how people react, change, and integrate the technology into their everyday lives. The outcome of conducting in-the-wild studies for different periods and at different intervals can be revealing, demonstrating quite different results from those arising out of lab studies. Comparisons of findings from lab studies and in-the-wild studies have revealed that while many usability issues can be uncovered in a lab study, the way the technology is actually used can be difficult to discern. These aspects include how users approach the new technology, the kinds of benefits that they can derive from it, how they use it in everyday contexts, and its sustained use over time (Rogers et al, 2013; Kjeldskov and Skov, 2014; Harjuniemi and Häkkila, 2018). The next case study describes a field study in which the researchers evaluated a pain-monitoring device with patients who had just had surgery.

CASE STUDY:

A field study of a pain monitoring device

Monitoring patients' pain and ensuring that the amount of pain experienced by them after surgery is tolerable is an important part of helping patients to recover. However, accurate pain monitoring is a known problem among physicians, nurses, and caregivers. Collecting scheduled pain readings takes time, and it can be difficult because patients may be asleep or may not want to be bothered. Typically, pain is managed in hospitals by nurses asking patients to rate their pain on a 1–10 scale, which is then recorded by the nurse in the patients' records.

Before launching on the field study that is the focus of our case study, Blaine Price and his colleagues (Price et al., 2018) had already spent a considerable amount of time observing patients in hospitals and talking with nurses. They had also carried out usability tests to ensure that the design of Painpad, a pain-monitoring tangible device for patients to report their pain levels, was functioning properly. For example, they checked the usability of the display and appropriateness of the device covering for the hospital environment and whether the LED display was working and was readable. In other words, they ensured that they had a well-functioning prototype for the field study that they planned to carry out.

The goal of the field study was to evaluate the use of Painpad by patients recovering from ambulatory surgery (total hip or knee replacement) in the natural environments of two UK hospitals. Painpad (see Figure 15.4) enables patients to monitor their own pain levels by pressing the keys on the pad to record their pain rating. The researchers were interested in many aspects related to how patients interacted with Painpad, particularly on how robust and easy it was to use in the hospital environments. They also wanted to see whether the patients rated their pain every two hours as they should do and how the patients' ratings using Painpad compared with the ratings that the nurses collected. They also looked for insights about the preferences and needs of the older patients who used Painpad and for design insights around visibility, customizability, ease of operation, and the contextual factors that affected its usability in hospital environments.

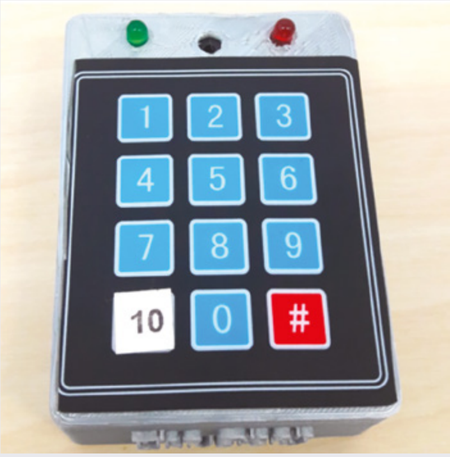

Figure 15.4 Painpad, a tangible device for inpatient self-logging of pain

Source: Price et al. (2018). Reproduced with permission of ACM Publications

Data Collection and Participants

Two studies were conducted that involved 54 people (31 in one study and 23 in another). Data screening excluded participants who did not provide data using Painpad or for whom the nurses did not collect data that could be compared with the Painpad data. Because of the confidential nature of the study, ethical considerations were carefully applied to ensure that the data was stored securely and that the patients' privacy was assured. Thirteen of the patients were male, and 41 were female. They ranged in age from 32–88, with mean and median ages of 64.6 and 64.5. The time they spent in the hospital ranged from 1–7 days, with an average stay of 2–3 days.

After returning from surgery, the patients were each given a Painpad that stayed by the side of their bed. Patients were encouraged to use it at their earliest convenience. The Painpad was programmed to prompt the patients to report their pain levels every two hours. This two-hour interval was based on the hospital's desired clinical target for collecting pain data. Each time a pain rating was due, alternating red and green lights flashed on the Painpad for up to five minutes, and an audio notification of a few seconds sounded. The patients' pain rating was automatically time-stamped by the Painpad and stored in a secure database. In addition to the pain scores collected using Painpad, the nurses also collected verbal pain scores from the patients every two hours. These scores were entered into the patients' charts and later entered into a database by a senior staff nurse and made available to the researchers for comparison with the Painpad data.

When the patients were ready to leave the second hospital mentioned, they were given a short questionnaire that asked whether Painpad was easy to use, how often they made mistakes using it, and whether they noticed the flashing light and sound notifications. They were also asked to rate how satisfied they were with Painpad on a 1–5 Likert rating scale and to make any other comments that they wanted to share about their experience in a free text field.

Data Analysis and Presentation

Three types of data analysis were used by the researchers. They examined how satisfied the patients were with Painpad based on the questionnaire responses, how the patients complied with the bi-hourly requests to rate their pain on Painpad, and how the data collected with Painpad compared with the data collected by the nurses.

Nineteen fully completed satisfaction questionnaires were collected that indicated that Painpad was well received and easy to use (mean rating 4.63 on a scale 1–5, where 5 was the highest rating) and that it was easy to remember to use it. Sixteen of the respondents commented that they never made an error entering their pain ratings, the aesthetics of Painpad were rated as "good," and participants were "mostly satisfied" with it. Responses to the flashing lights to draw patients' attention to Painpad were polarized. Most patients noticed the lights most of the time, while others only noticed the lights sometimes, and three patients said they did not notice them at all. The effectiveness of the sound alert received a middle rating; some patients thought it was "too loud and annoying," and others thought it was too soft. More nuanced reactions and ideas were collected from the free-text response box on the questionnaire. For example, one patient (P49) wrote, "I think it is useful for monitoring the pattern of pain over the day which can be changeable" Patient P52 commented, "A day-to-day chart might be helpful." Some patients, who had limited dexterity or other challenges, reported how their ability to use Painpad was compromised because Painpad was sometimes hard to reach or to hear.

After removing duplicate entries, there were 824 pain scores provided by the patients using Painpad compared with 645 scores collected by the nurses. This indicated that the patients recorded more pain scores than would typically be collected in the hospital by nurses. To examine how the patients complied with using Painpad every two hours compared with the scores collected by the nurses, the researchers had to define acceptable time ranges of compliance. For example, they accepted all of the time scores that were submitted 15 minutes before and 15 minutes after the bi-hourly time schedule for reporting time scores. This analysis showed that the Painpad scores indicated stronger compliance with the two-hour schedule than with scores collected by the nurses. ■

Overall, the evaluation of Painpad indicated that it was a successful device for collecting patients' pain scores in hospitals. Of course, there are still more questions for Blaine Price and his team to investigate. An obvious one is this: "Why did the patients give more pain scores and adhere more strongly to the scheduled pain recording times with Painpad than with the nurses?"

ACTIVITY 15.3

1. Why do you think Painpad was evaluated in the field rather than in a controlled laboratory setting?
2. Two types of data were collected in the field study: pain ratings and user satisfaction questionnaires. What does each type contribute to our understanding of the design of Painpad?

Comment

1. The researchers wanted to find out how Painpad would be used by patients who had just had ambulatory surgery. They wanted to know whether the patients liked using Painpad and whether they liked its design and what problems they experienced when using it over a period of several days within hospital settings. During the early development of Painpad, the researchers carried out several usability evaluations to check that it was suitable for testing in real hospital environments. It is not possible to do a similar evaluation in a laboratory because it would be difficult, if not impossible, to create realistic and often unpredictable events that happen in hospitals (for example, visitors coming into the ward, conversations with doctors and nurses, and so forth). Furthermore, the kind of pain that patients experience after surgery does not occur, nor can it be simulated, in participants in lab studies. The researchers had already evaluated Painpad's usability, and now they wanted to see how it was used in hospitals.
2. Two kinds of data were collected. Pain data was logged on Painpad and recorded independently by the nurses every two hours. This data enabled the researchers to compare the pain data recorded using Painpad with the data collected by the nurses. A user satisfaction questionnaire was also given to some of the patients. The patients answered questions by selecting a rating from a Likert scale. The patients were also invited to give comments and suggestions in a free text box. These comments helped the researchers to get a more nuanced view of the patients' needs, likes, and dislikes. For example, they learned that some patients were hampered from taking full advantage of Painpad because of other problems, such as poor hearing and restricted movement. ■

15.4.2 Other Perspectives

Field studies may also be conducted where a behavior of interest to the researchers reveals itself only after using a particular type of software for a long time, such as a complex design program or data visualization tool. For example, the expected changes in user problem-solving strategies using a sophisticated visualization tool for knowledge discovery may emerge only after days or weeks of active use because it takes time for users to become familiar, confident, and competent with the tool (Shneiderman and Plaisant, 2006). To evaluate the efficacy of such tools, users are best studied in realistic settings in their own workplaces so they can deal with their own data and set their own agenda for extracting insights relevant to their professional goals.

These long evaluations of how experts learn and interact with tools for complex tasks typically starts with an initial interview in which the researchers check that the participant has a problem to work on, available data, and a schedule for completion. These are fundamental attributes that have to be present for the evaluation to proceed. Then the participant will get an introductory training session with the tool, followed by 2–4 weeks of novice usage, followed by 2–4 weeks of mature usage, leading to a semistructured exit interview. Additional assistance may be provided by the researcher as needed, thereby reducing the traditional separation between researcher and participant, but this close connection enables the researcher to develop a deeper understanding of the users' struggles and successes with the tools. More data, such as daily diaries, automated logs of usage, structured questionnaires, and interviews can also be used to provide a multidimensional understanding of the weaknesses and strengths of the tool.

Sometimes, a particular conceptual or theoretical framework is adopted to guide how an evaluation is performed or how the data collected from the evaluation is analyzed (see Chapter 9, "Data Analysis"). This enables the data to be explained at a more general level in terms of specific cognitive processes, social practices such as learning, or conversational or linguistic interactions.

BOX 15.1

How Many Participants Are Needed When Carrying Out An Evaluation Study?

The answer to this question depends on the goal of the study, the type of study (such as usability, experiment, field, or another type), and the constraints encountered (for instance, schedules, budgets, recruiting representative participants, and the facilities available). Chapter 8 "Data Gathering," discussed this question more broadly. The focus here is on the types of evaluation studies discussed in this chapter: usability studies, experiments, and field studies.

Usability studies

Many professional usability consultants use to recommend 5–12 participants for studies conducted in controlled or partially controlled settings. However, as the study of the iPad illustrates, six participants generated a lot of useful data. While more participants might have been preferable, Radiu Budiu and Jakob Nielsen (2010) were constrained in that they needed to complete their study and release their results quickly. Since then, Radiu Budiu and Jakob Nielsen (2012) has said, "If you want a single number, the answer is simple: test five users in a usability study. Testing with five people lets you find almost as many usability problems as you'd find using many more test participants." Others say that as soon as the same kinds of problems start being revealed and there is nothing new, it is time to stop.

Experiments

Knowing how many participants are needed in an experiment depends on the type of experimental design, the number of dependent variables being examined, and the kinds of statistical tests that will be used. For example, if different participants are being used to test two conditions, more participants will be needed than if the same participants test both conditions. These kinds of differences in experimental design influence the type of statistics used and the number of participants needed. Therefore, consulting with a statistician or referring to books and articles such as those by Caine (2016) and Cairns (2019) is advisable. Fifteen participants is suggested as the minimum for many experiments (Cairns, 2019).

Field studies

The number of participants in a field study will vary, depending on what is of interest: it may be a family at home, a software team in an engineering firm, children in a playground, a whole community in a living lab, or even tens of thousands of people online. Although field studies may not be representative of how other groups would act, the detailed findings gleaned from these studies about how participants learn to use a technology and adapt to it over time can be very revealing. ■

In-Depth Activity

This in-depth activity continues work on the online booking facility introduced at the end of Chapter 11 and continued in Chapter 12. Using any of the prototypes that you have developed to represent the basic structure of your product, follow these instructions to evaluate it:

1. Based on your knowledge of the requirements for this system, develop a standard task (for instance, booking two seats for a particular performance).
2. Consider the relationship between yourself and your participants. Do you need to use an informed consent form? If so, prepare a suitable informed consent form. Justify your decision.
3. Select three typical users, who can be friends or colleagues, and ask them to do the task using your prototype.
4. Note the problems that each user encounters. If possible, time their performance. (If you happen to have a camera or a smartphone with a camera, you could film each participant.)
5. Since the system is not actually implemented, you cannot study it in typical settings of use. However, imagine that you are planning a controlled usability study and a field study. How would you do it? What kinds of things would you need to take into account? What sort of data would you collect, and how would you analyze it?
6. What are the main benefits and problems in this case with doing a controlled study versus studying the product in a natural setting?