Kasper Hornbæk, Per Ola Kristensson, and Antti Oulasvirta. Upcoming 2024. Introduction to Human-computer Interaction. Oxford University Press.

Part III.

User Research

10. Introduction to User Research

Many products, systems, and services are ultimately unsuccessful because their designs do not meet the needs, wants, and motivations of users. This part concerns how to obtain actionable knowledge about users, including their activities, their needs and wants, and their contexts of use. *User research* comprises the principles and methods of empirical research used to obtain, analyze, and represent this knowledge. For instance, we might observe users as they do their work (see Figure 10.1) to learn what it is about and what concerns them. Or we might conduct a large-scale survey study to learn about practices and experiences related to a piece of technology that we are interested in. A key aspect of user research is the synthesis of observational data in a form that can inform design.

A main motivation for user research is the realization that, in general, *you are not the user*. Basing critical design decisions on our opinions is risky, as developers and designers often represent very different points of view from the users for whom they design. Moreover, neglecting end-user viewpoints can often be outright unethical because it may result in systems, products, or services that cause harm or distress. Thus, the data we collect as part of user research should be about actual users engaged in specific activities with specific needs and capabilities. We should also obtain insights about other (non-using) stakeholders. For example, the parents of a child playing a mobile game are non-using stakeholders, which user researchers would aim to produce knowledge about.

User research complements the part on Understanding People (Part II). The theories and principles discussed in that part aim to be *general*. They relate to the psychological and social factors that pertain to most, if not all, cases of human-computer interaction. However, this understanding is necessarily limited. The current theoretical view of humans is patchy. As a consequence, almost every practical HCI project will face questions that are not covered by existing theories. Any interactive system also involves *particular* users performing particular activities in particular contexts. We need to know these particulars when we do not have a general understanding that covers them or when it is not clear which general ideas from Part II we may apply. General theories in psychology are at the population level and may not be good predictors of behavior at the individual level [237]. Therefore, the general understanding of people in Part II must be complemented by empirical insights on the particulars of the case to inform design and decision making. For instance, the selection of icons for an interactive system depends on general features of human visual perception (see Chapter 3). However, selecting the actual icons to integrate in a system requires designers to know the particulars of the work that the particular users do and the representations that would match the actual work.

The main aim of user research is to *obtain concrete, empirical knowledge about users*. This goal has been central to human-computer interaction since the early days of the field. Its importance is captured in slogans such as "know the user" [306], "early focus on users"

[280], and "start with the needs of the user" [582]. User research is where "know the user" happens. This requires that we, as researchers and designers, be in direct contact with users, rather than relying on second-hand descriptions or assumptions about them. We cannot indirectly know the particulars of users. As mentioned in Part I, this is a central tenet of being human centered. We need concrete empirical knowledge about users to design interactive systems that match what they do and want.

A key commitment in user-centered design is to *understand first and design later*. The idea is simple: Do not jump to solutions or try to come up with interactive systems that will solve the problem that you see. Instead, try to determine what obstacles and challenges users are facing and identify their needs and wants.

User research aims at design-neutral descriptions. This means that we wish to understand users without tying ourselves to specific design decisions. Due to this aim, user research is rooted in social science, psychology, and other behavioral sciences. We discuss design separately in Part VI. In that part, we will treat methods that break assumptions behind thse methods. For example, in co-design, we do not study users' activities and context to establish a basis for design. Instead, users are included in the design process to break the barrier that separates user research and design. When using design probes, we do not aim for accurate descriptions of people, but to generate knowledge that inspires

Figure 10.1.: A researcher observing a user [2]. User research is about empirical description of interaction in the present, particularly the users, their activities, the roles of any existing interactive system, and their contexts of use, including the social, organizational, and physical context. It also concerns the users' needs and wants. These descriptions are collected and represented to inform design and decision-making.

new ideas. More generally, in many design-oriented methods, designers who immerse themselves in the world and data may not need to articulate their understandings to others. In contrast, this part approaches user research as an applied empirical science, where such articulation is central. This is a key difference between user research and design.

Unfortunately, obtaining knowledge about users is complicated. Many factors contribute to this complexity, and understanding those factors helps avoid pitfalls and put the care into user research that it requires. These factors include the following:

- The say-do problem. This problem comes from the frequent observation that *what users say they do* might differ from *what they actually do*. For instance, when people are asked how much they use the Internet, their self-reports have only a low correlation with the data of the log file [719]. In this case, directly asking the users is not a valid method.
- In their activities, users rely on knowledge that is difficult to articulate [652]. Many factors driving our behavior are *latent* in such a way that they are not accessible to conscious inspection and are difficult or impossible to describe verbally. *Tacit knowledge* refers to knowledge that we are largely unaware of, but can be easily put into use in the right situation. High-repetition motor skills, such as riding a bike or typing in a familiar password, are good examples. You cannot ask people how they ride a bike and get the precise and full picture.
- Users' needs may only be recognized in the future [707]. For example, no user of a batch computer system were likely to be able to articulate the need for a graphical user interface. This means that identifying what users need requires them to imagine the impact of new technology on their activities, which is very difficult.
- Social reasons also make it difficult to get insight into users. Suchman [783] discussed how the ways work is done are often kept secret. Making work visible is fraught with difficulties. For instance, making work practices visible can challenge the power between those doing the work and other members of an organization.

Due to these factors, user research is both required and difficult to do right. This is the reason why it occupies an entire part of this book.

If user research is that difficult, why attempt to do it at all? Perhaps the most important reason is ethical: user research is a commitment to listening to your users and taking them seriously. Rooting decisions in empirical data can help avoid biases in decision-making. However, even the most rigorous user research is no guarantee that the result will not be harmful to end users. Another motivation is that user research can help build better interactive systems. Gathering insights about users that can inform decisions about whether to build an interactive system and how to build it. Properly conducted, empirical research outweighs empathy and opinion. The insights help to make decisions about how to have the desired impact on the future situation, possibly using interactive computer systems. User research can also challenge assumptions about products. For example, Kumar and Whitney [436] described user research to uncover patterns of daily life in a culture, including its value systems and social structures. The results of such user research can challenge the one-design-fits-all idea. Discovering drastic differences in a user population that was thought to be homogeneous can help rescoping and diversifying a product offering.

Finally, user research can pay off. Most technology companies engage in some sort of user research, for example, through log data, surveys, or qualitative research. Some academics and companies have tried to gauge the value of user research. Harper [307] described several cases of user-centered design at Hewlett-Packard. Estimated ROIs ranged from 6 million to 125 million dollars per case. However, the return-of-investment (ROI) is difficult to estimate, for the obvious reason that the effect of design decisions on revenue is difficult to measure [72].

The rest of this chapter explains the key parts of user research, and the rest of this part details the specific methods for doing user research.

10.1. Goals of User Research

The primary goal of user research is to obtain, analyze, and represent knowledge about current or prospective users. We call this goal *gathering knowledge for design*. What is it that we seek to gather knowledge about? In some cases, the user researcher will have a brief design, a focus, or a relatively clear idea of the scope of the interactive system under consideration or the use-related issues that need to be studied. In such cases, this will shape the user research. In other cases, the outcomes of user research will determine what the focus of the interactive system will be, or whether something needs to be designed at all. In both situations, the following list of types of insights to collect may be of inspiration. They may be remembered by the acronym used by Benyon [63], PACT: People, Activities, Contexts, and Technologies.

People People-related insights include user skills, personalities, socioeconomic status, abilities, beliefs, habits, motivations, needs, and wants. These types of insight involve the experiences and attitudes of users and their first-person views of themselves and their activities.

Activities The tasks users perform and the practices they engage in give us important information for design. What do users do? What do they strive to achieve, and in which contexts are they finding themselves in doing this?

Contexts of use The different contexts people find themselves in when interacting with systems provide a rich source of potential insights. There are several contexts to consider. The physical context is related to the physical environment and may include weather, lighting, or the built environment. The social context covers the social environment of interaction, such as social relationships and activities with others. The organizational context relates to how interaction is situated within organizations, for example, power structures, division of labor, and formal and informal work hierarchies. The historical context may be related to previous exposure to practices and systems. Finally, the cultural context captures cultural beliefs and norms that affect the use of the system.

Technologies Existing interactive systems and tools with which users interact serve as another important type of insight. This may include perceived pros and cons and any opinions on alternative solutions. Nevertheless, it is important to remember that the point of user research is not to primarily focus on technology or to begin imagining technological solutions, but to focus on people.

A primary goal of user research is to *inform* design and make practical decisions, such as whether to launch a product or if it contains biases. However, because user research is often open ended, it may not speak directly to concrete decisions. More severely, we face the *is–ought problem*, that is, from descriptions of how things are (*is*), we cannot conclude how they should be (*ought*). For example, imagine that you have collected vast amounts of data on what your users do on mobile devices, say, when they are cooking. What can you say about your need for a new service? The difficulty is that the essential characteristics of that new service may not be present in that user data. Hence, user research results are rarely strongly prescriptive of designs alone. That said, a significant value of user research is that it can convince stakeholders about the quality of the decisions made. Compare a designer who refers to personal experience with one who presents statistics, interview data, and a model that describes users. Which one will be more persuasive?

Generating insights for design is not the only goal of user research. Another prominent goal is to *involve users* in the design of interactive systems. One motivation is that participation in design is an argument that users have a political or moral right to be part of shaping the interactive systems with which they might work in the future. This argument comes from the Scandinavian tradition of system development called *participatory design* [403]. Another motivation to involve users is pragmatic: user involvement enables the creation of better systems that users are more committed to using [432, 320]. Part VI discusses techniques for involving users not only as a source to obtain insight about the current situation but also as co-designers of the future situation.

Another goal of user research is to *create empathy* between the user researcher (who could also be the designer) and the user. This goal transforms the activity of knowing the user into "understanding what it feels like to be that person, what their situation is like from their own perspective. In short, it involves empathy" [896]. This view permeates much work on design. Hansen [306], for instance, argued that one important function of getting to know the user is to "remind the designer that the user is a human". Rams argues that "indifference towards people and the reality in which they live is actually the one and only cardinal sin in design". Thus, direct and continuous contact between user researchers and the people they seek to understand might help establish a good connection among them. However, empathy is not a substitute for empirical inquiry. As we will discuss in Part VI, it tends to fail. Just as it is difficult to guess—even after years—what a significant other may want, it is difficult to emphasize with the lived world of someone else.

Finally, user research is valuable as is, even with no product in mind. User research can lead to discoveries and contribute to *basic research findings*. Many examples of such findings are provided in Part II on understanding people. This is the reason some HCI researchers have argued that not all papers describing studies of people should contain implications for design [206]. Some may primarily help understand people.

10.2. Who is the 'User' in 'User Research'?

If you are not the user, then who is? A central step in all user research is to identify who the user is and then select which users to investigate. It is essentially a three-step process, but allsteps are crucial to a successful user research.

Target audience The first step is to specify the *target audience*, which is a set of profiles that describe the user groups that the product, system, or service is intended to reach.

Other stakeholders The second step is to specify other people directly affected by or involved in the use of the interactive system.

Sampling Having settled on a set of profiles that accurately describe the target audience and other stakeholders, it is now possible to select representative users to research, this is called sampling.

These steps are critical because failure at either step risks introducing incorrect, misleading, or low quality into the design process. For example, failure to describe the target audience means that there is a high risk of conducting research with users who will ultimately not be affected or interested in the product, system, or service that will be designed. Although the findings of this user research may still be relevant by sheer luck, there is a possibility of *undetected bias* in the user research data incorporated into the design process. Such undetected bias is dangerous, as the designers are unaware of it and its consequences only discovered at deployment.

10.2.1. Target Audience

The *target audience* is a description of the intended customers or users affected by your system, service, or product. It is usually created by constructing a set of profiles that are intended to capture various subgroups of the target audience.

There are many ways to form a profile for a target audience. One process that is frequently used in user research for product development is the following.

- First, consider the inclusion criteria for the profile. The *behavioral criterion* means including people who actually do or want to do the things you intend to enable with your product. The *technological criterion* involves considering the background and interest of people in relating technology, such as how frequently they use social media or interact with a system based on spoken dialogue. Finally, the *demographical criterion* means capturing people's anticipated age, geographical location, household income, and so on.

- Having considered behavioral, technological, and demographical criteria, it is possible to develop an initial profile. For example, consider a hypothetical wearable health monitor coupled with a smartphone that is meant to use persuasive gamification methods to encourage a healthy lifestyle. An initial profile may then have a *Demographics* heading that says the intended age range is 25–55, occupation can be anything, and it is specifically aimed at single people who are educated at the university level and have a post-tax income of USD 60,000. A *Behaviors* heading may state that this typical user walks or uses public transportation to go to work. Finally, a *Technology-use and Experience* heading may read that a person within the profile may use their mobile phone to browse the web and Facebook several times a day.

- The next step is to review this initial profile in an attempt to make it suitable for user research. One trick is to try to identify additional criteria for ideal users that would provide the most valuable feedback. Several prompts can be used to aid in this exercise, such as on which *segments* within your target audience should you focus on; how much experience should they have with *your product*; how much experience should they have with *competing products*; and finally, are there any undesirable characteristics to avoid, for instance, fitness power users. The revision of the profile will then concentrate on removing factors that are not going to affect users' behaviors when interacting with your product or otherwise add any additional information and instead focus on isolating factors that are more likely to give rise to an ideal target audience. In the previous example, such a revision may, for example, involve updating the *Behaviors* heading to include additional criteria, such as a person in the target audience: (1) prefers not to exercise; (2) currently does not wear any form of a watch; and (3) talks about health issues with their friends.

In practice, the process of creating such a set of profiles will vary. For example, a company developing a new variant of a product in a market in which the company has been active for a long time will likely have acquired more knowledge about its customers and other stakeholders, such as, for instance, members of procurement teams. In contrast, early-stage research may be required to discover potential product development experiences, and this may require a process of iteration in which the target audience is gradually refined until a set of users is discovered that find the new potential product appealing.

In general, a number of ways of thinking about users have emerged in human-computer interaction. Perhaps the most prominent is that the users that we focus on in user research—interviewing, observing, surveying—should be *representative*, that is, a group of users who vary enough to be typical of the major types of users. What is representative is largely an empirical question. User research can chart both existing and potential user base, for example via methods like customer surveys. However, defining who a user is often a political decision. Are we designing for wealthy young professionals or managers, or are we designing while considering the real and diverse needs of all users? User-centered design should strive to cater to all audiences equally while considering their diverse needs.

This means striving away from the urge to design for managers and oneself, and instead empirically studying who will, or might be, a user.

Another approach to selecting users is to focus on extreme or *extra-ordinary users* [663]. This idea originally emerged in accessibility research, where it was argued that we should focus on users whose abilities are outside of what we may consider ordinary. The point here is to learn from users at the extremes of ability distributions since they may be the ones posing the hardest requirements for design. Pullin and Newell [663] listed a number of products that originate in these user groups, including the tape recorder (developed for blind people) and large grip kitchen products (for people with tremors).

Another way is to think of users in terms of *early adopters*. The rationale here is that early adopters of technology have needs and wants that will be general for other users months or years later. One popular approach is *lead users*. The idea there is that some users who have more experience and stronger needs with respect to an emerging technology than users in general. A lead user may be a social media influencer, a journalist, or a member of a community of practice. In this way, they can give earlier and deeper information.

Another way to think about users is to deliberately look for *variety* in user groups. The motivation is that we want to understand users and their activities with as much nuance, complexity, and diversity as possible, because these may challenge us to create better designs. Identified groups should be different in terms of their needs, beliefs, contexts, purchasing power, and so on.

10.2.2. Other Stakeholders

In addition to direct users of interactive systems, other people are often affected by the systems, indirectly involved in their use, or benefiting from the system. For instance, they may be parents of children who use a kid's web page or coworkers of users of an accounting system. We call these people *stakeholders*. We may conduct research with stakeholders in the same way as we do with users. However, methods that focus on observing and analyzing human-computer interaction are less relevant. Furthermore, stakeholders do not directly need to interact with an interactive system, and therefore may have different views and varying knowledge of the benefits, drawbacks, and activities relevant to the system. Therefore, questions of power, interest, and relationships with users must be answered and may require the participation of stakeholders.

How do we identify a stakeholder? *Stakeholder analysis* refers to a systematic attempt at identifying stakeholders. Its goal is to identify groups and organizations that affect or are affected by the system.

Mitchell et al. [536] proposed three relationship attributes to start defining a stakeholder group: power, legitimacy, and urgency. First, *power* refers to a relationship in which one group of users can get another group to do something. For example, election monitors may have certain powers over local organizations, who in turn may affect the layout of a voting interface, which in turn affects interaction. Management at a construction company has power over contractors, who in turn may have power over subcontractors. To the extent that a system is involved in the control of this power, such groups become stakeholders in that system. Second, *legitimacy* refers to the legal or normative basis of a stakeholder group. For example, hackers are illegitimate stakeholders in (too) many networked systems. Third, *urgency* refers to how time-sensitive a relationship is. Some groups are perceived as ''calling for attention" or ''in need of immediate attention" or ''driving" development. Urgency can also be thought about by considering its criticality: What is the opportunity loss of *not* addressing this group versus the gain from addressing it? In short, urgency is about the timeliness and criticality of a group.

Stakeholder analysis is hard and has not gained such popularity as it maybe should outside of academia. The reason lies in the complexity of relationships that a system can have with people. Some stakeholder groups are latent; they are only affected not directly but through some mediating mechanisms. For example, excessive use of gamification affects not only children playing mobile games, but also their parents. Some stakeholders are dormant: they are silent most of the time and are triggered only in special circumstances. For instance, legal experts are not stakeholders in the regular use of a system, but in the case of an accident or crime, this relationship changes. The authors also note that managers have a tendency to be biased. They focus not necessarily on those stakeholders who are the most urgent, but those that are most important for them, for example, because of commercial or political reasons. Stakeholder analysis is further complicated by the fact that some stakeholders are not humans. Animal welfare, for example, is foregrounded in attempts to design ''the Internet of dogs". There is increasing interest in including the natural environment as a stakeholder.

10.2.3. Sampling

To do user research, specific participants should be chosen. One consideration is *representativeness*: Is this user group representing the target audience or a valid segment of other stakeholders? This is an important consideration, as failing to sample representatively can provide a very skewed dataset later on, and this skewness may be difficult to detect.

Another consideration is *variability in the sample*: Will we learn about the breadth and variety of the target audience's attitudes and work? In interview studies and observations we are often interested in covering as varied situations as possible to get the full complexity of users' worlds.

Yet another consideration is *statistical power*: Will this sample size be large enough to draw reliable conclusions about it? Often, a larger sample is always better—assuming it is representative. However, in practice, it will often be necessary to sample only as many people as absolutely necessary to draw reliable conclusions. In situations where our aim is to get statistically reliable results, we can perform power analysis; see, for instance, Cohen [162].

There are a few practical considerations that are worth dwelling on. First, *access and availability*: are these users available to study? Second, *cost-efficiency*: Is this user group providing information for design that warrants the cost of studying them? Third, *costs for users in participating*: Does being part of user research present any cost or danger to participants? These considerations may lead to different choices regarding who and how many users to include. The individual chapters that follow this introduction contain

Research method	Explanation
Interview	Ask users questions about their attitudes, experiences, and activities.
Contextual inquiry	Observe and speak to users as they do their work.
Observation	Observe users without affecting them.
Ethnography	Explore the viewpoint of the user through observations, interviews, and participation.
Surveys	Collect a large sample of structured self-report data.
Diaries	Have users keep a diary about their use of interactive systems.
Log file analysis	Automatically analyze what users actually do with interactive systems.
Analysis of archival data	Analyze the activities and products users create in or with interactive systems.

Table 10.1.: Overview of popular user research methods learned in this part.

more detailed advice.

10.3. User Research Methods

Research methods in user research are means to obtain empirical knowledge about users, including research-based instruments, techniques, and procedures for collecting, analyzing, and reporting insights. As in any empirical study, the goal of user research is to obtain accurate and defensible knowledge about the subject under study.

Many research methods help to conduct user research. In this part, we cover a number of qualitative and quantitative methods used extensively in both universities and companies for user research. As an overview, Table 10.1 shows some of the most commonly used. Some methods are open-ended, that is, they make few assumptions about the categories of observations, but instead "let the data speak". Log analyses and observations, for example, can be carried out in an open-ended way. By contrast, some methods are closed-ended. They assume that we know in advance what the interesting phenomena and factors are and mainly want to know their statistical tendencies. The survey, for example, is a closed-ended method. Another aspect of user research methods is whether it provides a first-person or third-person perspective. Ethnography, for example, attempts to provide a perspective to the lived, experienced world of people. Log file analysis, in contrast, leads to a more distant third-person account.

10.3.1. Research Strategy

Research strategy concerns how to select one or more research methods for gathering insights about users. The selection of research methods depends on what the user researchers think is important to obtain insights about. On the one hand, the anticipated outcomes obtained with a particular method should be useful for the problem at hand. On the other hand, the resources needed for the method are also important to consider. However, the choice of method is not just about cost efficiency.

McGrath [518] introduced three general principles for thinking about research strategy. The first principle states that *research methods bound what we can empirically learn*. According to McGrath, “all methods have inherent flaws, although each has certain potential advantages”. From this principle follows that there is no correct method; the selection on methods needs to follow from careful consideration of the goals of the user research. Only then can you determine which method to use. Thus, the user researcher should never state that interviews are better, or that field studies are superior. Only that each has certain limitations and certain strengths.

The second principle of McGrath [518] states that the research strategy is about *trading off conflicting criteria*. The criteria are as follows, in no particular order:

Realism concerns how similar the situation being studied is to the situations that the researcher wants to gather insights about. Studies that look at user-initiated behavior in its naturally occurring surroundings, such as field studies, are high in realism.

Precision concerns how much accuracy and detail one obtains about the users’ behavior and attitudes, and how much control there is over variables and circumstances of no interest. User studies that track every detail of people’s hand movements with motion-tracking equipment, for instance, are high in precision.

Generalizability refers to how well the findings generalize to other people or situations. Studies that are able to get input from a lot of users, such as surveys, are often high in generalizability.

Each of these criteria works against any of the others and, in general, it is impossible to maximize all of them. For instance, let us say that we seek to maximize realism and change the situation being studied as little as possible. This has the consequence that any feature of the context, even those we are not interested in, influences our study. It also means that we obtain evidence for just about one situation and face challenges when we try to generalize it to others. Those principles can also be used to consider the previous discussion on how to select users. Part of that discussion was about whether one should see insights about users in general (maximizing generalizability) or about particular users (maximizing realism).

The third principle of research strategy follows from the inherent limitations of methods—*triangulation*. McGrath [518] explained the idea as follows: “If you only use one method, there is no way to separate out the part that is the ’true’ measure of the concept in question from the part that mainly reflects the method itself. If you use multiple methods, carefully selected to have different strengths and weaknesses, the methods can add strength to each other by compensating for each other's weaknesses". The meaning of triangulation in ordinary language is to establish the location of something by taking bearings to it from two or more positions. In research, triangulation means the combination of multiple research methods to study the same phenomenon. The rationale is that if all research methods are limited in different ways (as argued above), then their combination should be able to mitigate some of those limitations. Thus, triangulation is a characteristic of the user research process, rather than of a single method.

How to do triangulation in practice? The following checklist helps to understand the requirements for triangulation Munafò and Smith [550, p.400]:

- The approaches chosen to the study must be able to address the same underlying question. For example, users' beliefs about interaction can be studied with interviews, surveys, and possibly observations, but may not be – easily at least – studied via log analysis.
- For each approach, the inherent biases should be analyzed and made explicit. Every method has tendencies to focus attention to particular facets of the world, ignoring others; thus, bias. Being aware of such observational biases, we can better select methods that complement each other.
- Mechanisms that produce bias should be understood and acknowledged. For example, researchers have a tendency to use terminology for interviews and surveys that is better approachable by user groups that are similar to them. Each of the methods we discuss in this Part is susceptible to biases, which can be mitigated via careful triangulation.
- Critical shortcomings or biases should be compensated for somehow, for example, by adding a method that compensates for them. If, for example, one method is more susceptible to the researcher's own biases, it should be compensated by adding a method that is less so.
- The results of using the two methods are compared and assessed to understand how the methods differ. When reporting the results of a triangulated study, we should be systematically aware of the limitations of the methods that inspired the observations.

In triangulation, several things might happen. The different methods may result in the same overall results, corroborating the findings and offering converging evidence. Methods can also offer complementary findings, highlighting different phenomena or relations. Finally, the findings of each method may, of course, contradict each other. Pettersson et al. [636] showed that about half of a set of user experience studies used methodology triangulation. The other half, in contrast, relied on just one method, which carried the risks outlined above.

10.4. Methodological Quality

Research methods for user research have different strengths and weaknesses. For instance, we may characterize those in terms of their realism, generalizability, or precision. In addition, the research methods may be used more or less well. We call the criteria by which that is judged *methodological quality*. Next, we will discuss the four dimensions of methodological quality.

Validity: Validity concerns whether the conclusions drawn from a study are justified. *Threats to validity* are anything that could go wrong and threatens the validity of the conclusions drawn. A classic taxonomy of validity covers four types [739]. *Internal validity*: whether a variable under the control of the researcher, for example, user groups, devices, or tasks, has an effect on observations. *Construct validity*: if a measurement is supposed to measure something, such as user experience, it actually measures it. *Statistical conclusion validity*: if the conclusions drawn based on the data are statistically reliable. *External validity*: do the conclusions hold for other participants and settings?

Reliability: The reliability of user research and empirical research in general concerns that user research gives consistent results. That is, we would like our methods or measures to give the same result, for instance, if they are applied immediately again to the same person. McDonald et al. [516] offered guidelines on how to assess and use reliability in HCI research.

Transparency: The transparency of user research refers to the idea that researchers should make the design, data, analysis approach, and derivation of conclusions accessible and inspectable. Transparency is often associated with open scholarship and a desire to increase the replicability of research. To many, this ideal of research quality applies to all forms of research in HCI, including those that draw on social science methods [528]. However, current practice in HCI is not transparent, and many studies do not share artifacts, protocols for research methods, and so on [844]. There are several practices associated with transparency, such as sharing data, deciding on plans before seeing data; care in analysis and reporting, and maintaining a chain of evidence throughout the study.

Ethics Doing user research requires that the person conducting the research carefully considers what is right and wrong in collecting, analyzing, and reporting data. Concerns about ethics also include allegiances to the various stakeholders, most importantly the prospective users who participate in the research, the client for which the research is done, the professional standards of the field of HCI, and the responsibility toward society at large. We also discuss ethics in Chapters 43 and 45.

Validity, reliability, and transparency are all key issues for analysis and frequently ethics becomes relevant in analysis as well. For instance, the transcription of interviews impacts the reliability of the analysis. If transcription is sloppy, the resulting analysis may suffer. Similarly, the care by which we record observations of users might impact the validity of our findings; if we miss writing down certain observations, our data might be invalid. The specific techniques differ depending on the research method. They include thematic analysis, affinity diagramming, statistical analysis, and machine-learning classifiers. The rest of this part will detail the techniques.

10.5. Data Analysis and Representation

A distinct challenge in user research is to transform the results obtained into a useful and actionable form. Thus, the analysis and representation of the results of user research is important.

For representation, the key issue is to synthesize the results of user research in a way that is helpful for design. A representation is helpful if it brings forward critical insights from user research clearly and transparently. It may also be helpful to have an impact on people who are making decisions about the design. In the rest of this part, we discuss different techniques for representing user research. They include the following techniques. *Personas* are profiles of fictional characters that describe representative users within relevant segments of the target audience. *Scenarios* are narrative accounts of what happens when users try to use a system to achieve their goals. *Artefact and context models* are descriptions of key objects and their contextual relationships within a situation. *Quotes* are rich representations of users' attitudes and needs. Finally, *journey maps* are temporal accounts of encounters that users have with a system. Chapter 15 discusses representation in depth.

10.6. Does User Research Work?

But what is the evidence that user research works? This is a difficult question to answer. One answer is that user research is akin to an axiom: It is considered valuable as a basic commitment to do user-centered research. As pointed out above, one can go as far as saying that *not* doing user research is unethical; it may lead to outcomes that harm end-users in ways not predicted by designers. The time lost to frustrations with computers is one example. Another, even more severe, is algorithmic bias: AI algorithms that discriminate users based on race or gender. But there is often also an economic benefit in engaging in user research. As mentioned in Part I, many HCI activities have been linked to a positive return on investment. Another answer is to point to the many success stories of work in HCI that have departed from users, like the ”tire track diagrams” of HCI's history that we discussed in the introduction to this book (Chapter 16). In the history of HCI, there are examples of discoveries made in user research that have helped shape a product.

However, the evidence is by no means unequivocally in favor of user research. Before modern user-centered design, many interactive systems were developed without involving users and systematically the systems before launch. Many of the tools we use, or musical instruments, are the result of several, even hundreds, rounds of iteration.

The HCI pioneer Donald Norman [585] argued that human-centered design can sometimes lead to incoherent systems. Such systems may strive to fulfill user requests, but end up as a jungle of disentangled and confusing functionalities. *Feature creep* is the unavoidable consequence of too closely adhering to even the weakest signal in user research. Moreover, as Norman argued, users adapt. They appear willing to adapt to systems that offer value, even if such systems are not necessarily adapted well to human abilities.

In summary, user research has not been proven to be necessary or sufficient for building interactive systems. However, one can argue that despite the lack of conclusive evidence, user research is the best available approach for understanding people in design and development. Its core value may lie more in informing good designs than driving radical innovations, as discussed by Norman [585].

Summary

- You are not the user: empirical research is required to obtain knowledge on the practices and experiences of other people.

- User research uses empirical methods to understand users, their activities, the contexts in which they act, and the interactive systems they presently use, if any.

- The goal of user research is to produce actionable knowledge about users; insights that may subsequently help design an interactive system or make a decision about its use.

- All methods for user research are limited, for instance, with respect to their realism, generalizability, and precision.

- The key qualities of research methods concern validity, reliability, transparency, and ethics.

11. Interviews

One way to learn about users is to speak with them. However, just talking to people is not research. From a research point of view, the *way* we talk to people in everyday encounters is fraught with problems. Everyday conversations are often not planned and, therefore, may provide a scattered view of a topic. Everyday conversations may not aim to understand the conversational partner's viewpoint. Moreover, we rarely record our conversations, and it is practically impossible to later form a systematic account of what happened.

Interviewing, in user research, concerns learning about the subjective experience of users. The purpose is to learn how *they* view their tasks, values, experiences, practices, and what their wishes are for interactive systems. The goal of such learning is primarily to understand users and secondarily to inform design and decision-making related to their technology. aiming to establish insights about users rather than input for a particular design.

An interview is where the researcher meets the user. An interview aims to elicit first-person views of a particular user on a selected topic while avoiding omissions and biases. In interviewing, an interviewer talks with an interviewee following some planned script or schema. The session itself can be organized as a structured series of fixed questions or as an open-ended, but thematically defined conversation. It is the systematics of interviewing that distinguishes it from everyday talk.

The interview method has helped establish many important findings in HCI and is routinely used in both user research and interactive system evaluation. For example, from interviews we have learned about the motivations behind the use of social networks (see Paper Example 11.0.1). Other examples of uses of an interview in user research:

- Ljungblad et al. [477] studied human-drone interactions by interviewing professional drone pilots. Their data show how drone pilots are inventing new work and business enabled by drones, such as new types of photogrammetry, such as imaging potential rockfall areas. Pilots are particularly pleased with improved safety over earlier methods. For example, drones reduce the risk to workers in hazardous jobs, such as inspecting power lines or in forestry work.

- Kumar et al. [435] studied the experience of 21 Uber drivers in Dhaka, Bangladesh. Their data exposed a surprising finding: The design of the ridesharing service amplified feelings of oppression. One reason for these experiences was the use of English in the map and in the application. Many drivers, some of whom were illiterate, had to rote-learn to use the application, while others had to ''rely on luck" in using it. Drivers had to invent workarounds in a creative way.

- Simpson and Semaan [757] studied the experience of TikTok algorithms in LGBTQ+ communities. Their interviews exposed that TikTok's *For You Page*, on the one hand, supports their identity work and, on the other, can violate them. The authors describe the practices that the users had adopted in response to these issues.

Paper Example 11.0.1 : Why We Tag?

In their paper *Why We Tag: Motivations for Annotation in Mobile and Online Media*, Ames and Naaman [21] studied motivations for the use of tags, or annotation in photo-sharing. At the time when the study was conducted in 2007, photo sharing was becoming popular as a type of social media, and there was a need to understand why tagging was gaining popularity. Tags, at that point, were used in some social media services but not all, and the question arose of what motivated people to use them. Why would users put effort into using tags? Ames and Naaman approached this question by interviewing 13 Flickr users.

The interviews revealed a surprising diversity of motivations, which Ames and Naaman categorized using two dimensions: Sociality and Function. While we usually think about tagging as something we do to attract attention and do self-promotion, they showed that tags are also used to help memory and remind oneself to do things. According to the first dimension, people tag for social and self-related purposes. According to the second dimension, they either organize the information or communicate it. For example, Participant 2 in their study told:

> If I have the time, the neighborhood, or the event, I have enough information to look at my own collection and know where this came from. I don't have the bandwidth to tag for the benefit of the Flickr system. ... I want at least one hook of association in there that can help me reconstruct what I was thinking. I don't have time to put all the hooks in but I can put one in.

This example illustrates the strength of interviews: They can help shed light on phenomena for which there is no systematic understanding yet. They complement other methods by an open-ended, non-committal approach to data collection that does not need to commit to predefined categories of observation, nor assume there is a 'true' or 'valid' point of view. Instead, they seek to expose the individual, subjective viewpoints to technology.

In HCI, interviews can be lamented as 'easy' or, in the most ignorant, as 'not scientific'. Nothing could be further from the truth. While we all know how to talk to other people, the interview method requires not only understanding of the underlying scientific principles, but also years of practice to master. Interviewing is hard. Numerous pitfalls threaten an uneducated interviewer. Learning to ask questions and listen to their answers to uncover the depths that experienced interviewers uncover is a craft. However, many guidelines and ways to organize interviews to help you talk to users in a valid and useful way. This chapter covers some of those guidelines and types of interviews; more details are found in chapters and books on the topic, such as Rubin and Rubin [696].

The fundamental idea of using interviews for research is that they allow interviewers to develop an understanding of the lives of *others*: people who are very different from themselves. The strength of the interview is to get the attitudes and interpretation of the interviewee of the world in which *they* live and work. Interviews thus allow a unique perspective to felt experience that is accessible exclusively to such structured approaches. To this effect, interviewers must not impose their views or beliefs on those they interview: "don't argue, never give advice" [514]. More than that, interviews can give voice to people, empowering them.

It is important to understand when to use an interview and when not. Interviews are not particularly well suited for enumerative or statistical information. Instead, interviews are typically about gaining an understanding of beliefs, attitudes, and experiences. For example, it is not necessary or advisable to use interviews to figure out how often people use a particular software or if they are using a particular website.

There are many forms of interviews that have been developed for a range of different purposes in many research fields. At a high-level these forms of interview can be divided into structured, semistructured, and unstructured interviews.

A *structured interview* has two fundamental variants. The purpose of a quantitative structured interview is to collect statistical survey data, and because of this objective, the format of the interview and the sequence of the questions is predetermined. Such interviews are common for various demographic or market surveys. A qualitative structured interview consists of a set schedule and a set sequence of questions.

An *unstructured interview*, sometimes also referred to as nondirected interview, is as its name implies an interview where there is no fixed schedule or sequence of questions. Questions may or may not be prepared in advance. Unstructured interviews are an important research methodology in some fields, such as sociology, however, they are not that common in HCI and will not be treated in detail in this chapter.

Often, a compromise between these two variants is used. The *semi-structured* interview is on a continuum between a structured interview and an unstructured interview. It does not have a rigorous schedule or a fixed sequence of questions, but it is also not completely loose in its structure. Semi-structured interviews tend to follow a loose structure and are frequently referred to as *open-ended interviews* in HCI. Open-ended interviews will be treated in detail in this chapter.

Paper Example 11.0.2 : Death on social media

People on social media sometimes pass away. Brubaker et al. [101] were interested in how their digital profiles were involved in grief and how such profiles were used after death.

The interviews were open-ended, allowing participants to guide the discussion to those topics that most interested them, but with a general focus on feelings about and approaches to death, experiences with social media and other communication technologies, and interactions and experiences with death on Facebook in particular. The interviews showed critical views of postmortem Walls in Facebook:

”To be honest, I just don’t think death on Facebook is ever appropriate... I feel like all that’s doing is attention calling . . . maybe you want to share that you are in pain and in grief, but you probably just want people to know that you knew somebody who died and it makes you sad . . . there’s a reason people put that crap on their Facebook profile, and I don’t think it’s for the benefit of the dead person.”

More surprising was the finding that encounters with the deceased allowed users to participate in the mourning in new ways, such as in this quote:

”Maybe about a year and a half ago, he contacted me on Facebook and he wanted to know what I was up to. And we had a long conversation on instant messenger...that’s the last time I was in contact with him. . . . I went on his Facebook to wish him a happy birthday and saw that he had died. . . . It had been nine months or so ago he was in a car accident.”

This example shows how the interview method can help get access to personally deep experiences that would be very hard to gauge by other means. For example, how would one form a questionnaire on mourning on social media?

11.1. Open-Ended Interviews

Open-ended interviews are the most common form of interviews in HCI. They are called open-ended because the questions asked during the interview are not fully planned in advance and, at least in part, developed and adapted depending on their answers. In this way, the interviewer can pursue unexpected but interesting answers or adapt the questions so that they help the interviewees. Open-ended interviews try to minimize the influence of the person asking the questions. Sometimes such interviews are also called *responsive* interviews [696], highlighting that the structure and the questions develop as the interview progresses. Nevertheless, the interviews retain a certain theme, structure, and focus.

Open-ended interviews can be thought of as a development of conversation. In an open-ended interview, a conversation takes place – typically face-to-face – between two conversational partners, the interviewer and the interviewee. The interview shares the following characteristics with a conversation:

- The interview is flexible in content and structure, just as a conversation may freely develop. Thus, the interviewer adapts their questions based on what is being said. They follow up on answers and may pursue new information with additional questions.

- The interview has a certain continuity; the conversation partners understand where the conversation is going.
- The interview is about understanding what the conversational partner says. If you do not understand what the interviewee says, you follow up and probe with additional questions. If you are not being understood, for instance, if your interview questions are unclear, you explain them.
- The interview needs the full attention of the interviewer and, ideally, of the interviewee. Minimize the attention paid to notes and note-taking, other people, and the environment.
- The interview requires conversational partners to treat each other with respect and protect opinions if shared in confidence.

However, not all aspects of ordinary conversations are reflected in interviews. First, the interview is initiated by the interviewer for the purpose of doing user research. Thus, there is a basic asymmetry in the conversation that needs to be carefully managed so that the interviewee feels valued and can follow the purpose and content of the conversation.

Second, even if an interview may appear as a casual conversation, the interviewer is carefully prepared and the questions are highly designed. This is the result of preparation to identify a theme for the interview, design the interview situation, and carefully word the questions.

Third, during the interview, an interviewer often spends more time listening than in an average conversation; in short, "listen, don't talk" [514, p. 65]. Listening closely, noting the relation of what is said to other interviews as well as what is left unsaid, is a craft far beyond most of the conversations that most of us have most of the time.

Fourth, an interview may often be recorded to be analyzed in depth later; indeed, some follow-up questions may be posed in anticipation of that later analysis.

User research with open-ended interviews typically consists of four phases: (1) developing the themes of the interview, (2) planning and wording questions, (3) conducting the interview, and (4) analyzing the ensuing conversations. Often these phases are done iteratively; for instance, conducting the interview and analyzing it may lead to different questions and wordings thereof. Next we discuss each of these phases.

11.1.1. Themes

The basis of any interview is the development of the themes of the interview. The themes refer to the topics the interviewer wants to learn about. In user research, themes may concern practices of using an existing system, patterns of collaboration in an organization, perceptions of implantable devices, or exceptionally thrilling experiences with computers. Themes can be specific or broad. They may focus on a specific system, event, or use situation—or they may be broad and cover the totality of the interviewee's life and aspirations. The former might be appropriate when trying to the breakdowns of an existing system, whereas the latter might be appropriate to understand the effects of remote work.

Existing interview studies in HCI contain examples of the outcomes of developing themes. For instance, Grudin [298] was interested in the (at that time) increasing use of multiple monitors. He discussed the literature on multiple monitors and found that a number of simple questions around this topic had not been answered including ”How do multimonitor users make use of the extra space?”, and ”Is the second (or third) monitor an extension of their workspace, or do they use it for different purposes?”. Those themes would then form the basis for designing questions for interviewing 18 multidisplay users.

In addition to technology-centered themes, an interview can also be centered on research questions [696] or more closely formulated hypotheses [440]. In these cases, themes depart from theory, attempt to build theory, or on purpose challenge a known theory. However, when developing such themes, the intent is the same as in the general case: to spell out why the interview is done and what it should be about.

11.1.2. Questions

Given a set of themes for an interview, the next phase consists of finding effective ways to ask about those themes. In open-ended interviewing, you want to ask simple questions that allow interviewees to give complex answers. Three types of questions may be distinguished: main questions, probes, and follow-up questions [696].

Main questions directly corresponds to the themes of the interview; typically, there are a few of these questions for any study. Each of those questions can start a conversation. Main questions are often broad and allow the interviewee to cover the basis. For instance, Kirk et al. [408] was interested in the work that people do on their digital photos after capture but before sharing them. They argued that previous work had not investigated this aspect of digital photography but mainly focused on sharing photos. Based on an initial analysis of what such photo-work might involve, they developed a series of themes topics that were currently unclear and that an interview study could investigate. These themes are represented in the paper as questions such as “When and why do you delete your pictures?”, “How do you file your photos away (using file structure, folder labels, and so on)?”, and “When do you look at your pictures?”.

A few types of main questions may be differentiated.

- In the *tour* question, the interviewee is asked to give a tour of topic of the interview, be it a workplace, an interactive system, a work task, or a location. For instance, one may ask "Step me through what happens when X", "show me Y", or "tell me about Z". For instance, Odom et al. [594] organized their interviews with tweens and teens on virtual possessions as a tour: "We asked participants to give us a tour of their material possessions both stored and on display in their bedroom, and to describe their relationships with these artifacts. This was typically followed by a tour of participants’ virtual possessions, where we observed virtual artifacts on their personal computer, phone, media player, etc".

- In a grounded question, interviewers are instead asked to explain a specific example that grounds their answer. For instance, the most recent file they downloaded or the most extreme mistake they have made in sending out a social-media message.
- In the your-experience question, the interviewee is asked to explain their beliefs or attitudes on a certain topic or event.
- In key concepts questions, interviewees are asked to explain a certain topic as to how they understand it.
- In stages questions, interviewees are asked to account for the stages or chronology.
- In comparison questions, interviewees are asked to compare or contrast events, technologies, or anything else of interest to the interviewer.

Main questions in specific interviews may, of course, differ completely from these types. What is advisable is that there is a clear link from each theme to one or more main questions. Together, these questions should allow for a good exploration of the theme.

The second type of question, *probes*, is designed to make the interviewee comfortable and continue to talk. Thus, such questions have no explicit link to the themes of a study but are simply tactics for moving the conversation forward, asking for elaboration and clarification, or for coming back to the indented point of a main question after a digression. For instance, the interviewer may say ”mmhmm”, ”then what...”, ”please go on”, ”what hapened next?”, or ”tell me more about that”. Non-verbal acknowledgment, like nodding and smiling, may work similarly. Sometimes interviewers may simply echo the interviewee’s words to indicate that they find them interesting and worthy of further explanation. If the interviewee says ”but the system does not work as intended”, the interviewer may echo ”does not work as intended”.

The third type, *follow-up questions*, aims to further extend the answer to a main question. The intent is to ensure that the conversation of a main question is thorough and deep and to explore new ideas and concepts that the interviewee brings up. Some of those questions can be worked out in advance, in anticipation of likely answers. They may also be improvised during an interview, when new information or interesting stories are shared. Sometimes they may come up in analysis of earlier interviews, where the interviewer realizes that they should have followed up.

There may be many things to follow up on in an interview. Below we offer a brief list of possible lines of follow-up.

- Clarify answers. Follow up on undefined terms or unclear explanations. For instance, ”you said that the system was X; what do you mean by X?” or ”what you said there seemed very important and I want to be sure that I understand it; please explain it again”.
- Getting answers concrete. Follow up on answers that oversimplify, lack depth, or are abstract. For instance, ”could you give me an example?”.

- Compare or contrast answers. You may follow up on the answers by asking the interviewee to compare or contrast them to previous answers. For instance, ”you said earlier that X; how does that situation to compare to what you just told me”.
- Understanding variations. Here, you follow up to understand how responses vary over circumstances. For instance, ”do you always do it this way?” or ”how has this changed over time?”.
- Ensure thoroughness. You may follow up when answers does not cover all aspects of a main question or when you feel that information is missing. For instance, ”I think we have not covered X much; please tell me about that”.
- Personal attitude. Follow up on questions where it is unclear how the interviewee thought or felt. For instance, ”why was this important to you” or "how did you feel about that”.

Across main and follow-up questions, the wording of questions in interviews is crucial. The right wording keeps the interview on track and questions do not bias the interviewee. Questions should be as neutral as possible.

Avoid wording that could influence the answers in a particular direction, biasing the questions, or giving cues to their answer. For instance, instead of asking ”How much did like the system?” – highly biasing, you can ask ”Can you tell me about your experiences while using the system?”. Questions should be asked one at a time; avoid the double-barrelled question (e.g., ”Let me know how you normally do the same task when at home and what kinds of features you use in your toaster.”). The questions should be clearly expressed. Make them short and be sure to use the appropriate terms, jargon, and culturally-appropriate terms; if in doubt, ask in an open-ended manner and learn the clearest terms and ways of questionning.

A key issue is to avoid *leading questions*. A leading question is a question that implies that there is a specific set of “right” answers to a question. Consciously or unconsciously, asking leading questions is a frequent source of error in user research. Non-directed interviewing is about setting up an interview situation that is completely focused on the interviewee and leading questions prevent this by injecting conscious or unconscious prejudices from the interviewer into the interview.

While avoiding leading questions sounds obvious, it is in practice very difficult to avoid direct questioning. It requires a constant awareness of the risk of suddenly injecting bias or steering the conversation in a way that contaminates the interviewee’s own thoughts. A central idea is to be a *neutral interviewer* to fully understand any positive and negative information provided and to be able to relate back what the *user*—not the interviewer—wants or needs. This requires the interviewer to try to step out of everything they know and feel about the HCI idea or system in question. Depending on the interviewer this may easy or difficult. More prior personal investment into an HCI idea or system will make it more difficult to forget all previous hard work and creativity may have been involved previously. It is important to remember this and be honest with yourself about whether you are really suitable as an interviewer if the aim is to try to become a neutral interviewer on the matter. In practice, it demands self-imposed distancing to the issues being discussed, a comprehensive self-evaluation and introspective examination of assumptions, values, etc. relating to the interview topic.

So how do we create non-directed questions? In general, it is best to focus each non-directed question on a single topic to avoid ambiguity. For example, a question such as "How would you use your camera on a holiday or at work?" is best broken down into two separate questions. Non-directed questions should ideally also focus on immediate experience and impressions. People are poor at predicting the future—including their own future. Therefore, asking users whether they will use something in the future is not meaningful. A better strategy is to try to find out what they find compelling and usable right now in the moment. Related, it is frequently important to avoid having interviewees extrapolate from their own understanding of their behavior as people's understanding of their behavior often does not correspond to their actual behavior. For instance, a question such as "Is handwriting recognition useful?" may be understood by the interviewee as a question on whether handwriting could be useful for *someone* in *some* hypothetical situation. However, the interviewee is unlikely to know the answer to such a question and is likely to simply speculate. A better way to phrase such a question is to simply ask "Is handwriting recognition useful to you right now?". Questions should also not use judgmental language, in other words, the interviewee should not get the impression that they are requested to answer a question a specific way or that their answer to a question was wrong. For example, a question such as "Don't you think it is a great idea to enable autocorrect by default on all smartphones?" implies the interviewer holds this opinion and is looking for affirmation.

One word of warning about questions that appear in the methodology literature [684, 696]. Question are often of the **5W1H-form**: who, what, when, where, why, and how (or variants thereof). Although the question "why?" is simple and in many seems a good interview question, use it carefully. It may prompt participants to give simple answers where they do not know or to offer justifications when they do not know. This may spill over into nervousness on their part and may affect the rest of the interview. So be careful.

11.1.3. Conducting Interview Conversations

An interview conversation consists of several parts [696]. Although these may vary depending on the theme and circumstances of the interview, they typically include the following six parts.

Introduction In the first phase, both the interviewer and the participant introduce themselves, and the interviewer explains the purpose of the interview. Typically, the interviewer would want to establish themselves as neutral but sympathetic. It is often a good idea to start interviews with a bit of casual chat to create a relaxed, conversational atmosphere. Show people that they are competent/noteworthy. Introduce yourself and your focus: Why do you want to speak to the interviewer?

Warm-up In the second phase, the objective is to make the participant step away from their normal life and change their focus to consider the objective of the interview, typically a particular system, product, or service. This part typically includes some easy questions to start the conversation and make the interviewee comfortable. They may be about background information, for example about the theme of the interview.

General issues In the third phase, the interview is centered around direct experiences and related attitudes and expectations with the product, system or service in question. Asking these types of questions early in the interview prevents adverse effects from the interviewer inadvertently skewing the interviewee's perception later in the interview.

Deep focus In the fourth phase, the central focus of the interview is introduced. This part is the key part of the interview, where the main questions are asked, prompts are given, and follow-up questions are used to deepen the interview.

Retrospective In the fifth phase, the interviewee has time to reflect on the broader issues of the HCI idea or system that have been discussed. This is also an opportunity for the interviewee to relate the *general issues* identified earlier with thoughts that arose in the deep focus phase. Another purpose of this part may be to minimize any intellectual or emotional tension in the interview and ensure that it ends on a relaxed note.

Wrap-up In the sixth phase, the interviewer completes the interview to avoid interviewees feeling left hanging at the end of the interview. It includes thanking the interviewee for their time and answers. Typically, the interviewer also explains what will happen next, for instance, if they expect to send the whole or parts of the interview, or the analysis of the interviewee, to the interviewee for comments or approval. It may also cover other plans for staying in touch, such as sending a final report to the interviewee.

An important objective for an interviewer is to be pleasant, polite, and at least appear to be enjoying the conversation.

Interviews often last about one hour. Robson [684] offered the following advice: "Anything under half an hour is unlikely to be valuable; anything going much over an hour may be making unreasonable demands on busy interviewees, and could have the effect of reducing the number of persons willing to participate, which may in turn lead to biases in the sample that you achieve". In practice, depending on the setting and the themes of the interview, interviews in HCI ia sometimesmuch shorter.

During the interview, the interviewer captures the answers in some way or another. There are several options available. The interviewed can pay full attention to the conversation if they do not take any notes; needless to say, the interviewer may not at a later stage analyze what participants actually said, cannot easily discuss the content of the interview with persons that were not present, and is generally not recommended. The interviewer may record the conversation so that it may later be transcribed and revisited;

usually, an audio-recording device is sufficient, video does not add much unless the talk is about artifacts or practices that are better shown.

Interviews are often supported by an *interview guide*. Such a guide contains the main questions and a list of anticipated prompts and follow-up questions for each of them. It can also contain a procedure for the entire interview, written out in full or only outlined. Writing it down in full is recommended for novice interviewers.

11.1.4. Analysis of Open-ended Interviews

The analysis of interviews transforms the interview conversation into insights. The recommendation for open-ended interviews is to do this as soon as possible after the interview. The point is to keep as much of the conversation and the accompanying hunches and associations clear when the interview is analyzed. In this way, insights from the analysis of the conversation may also inform later interviews, for instance, by changing the questions, improving the follow-up questions, or improving the tone set in the introduction to the interview.

The details of how to do the analysis will be covered in section 11.5, as they are similar in several forms of interviewing. Ideas on how to construct representations of users that may inform future design work are covered in Chapter 15.

11.2. Structured Interviews

Structured interviews are a variant of interviews that are more rigid than open-ended interviews; their questions and the way in which their structures are predefined. This allows scripted interviews that are easier to conduct because they involve less improvisation. It also simplifies the analysis and ensures a minimum level of overlap between interviewees. In this sense, they come closer to surveys than interviews.

Paper Example 11.2.1 : A mixed methods study of phishing

Have you received a suspicious email asking you to fill in information on a Web site that mimics a legitimate one? *Phishing* refers to fraudulent websites or emails to acquire personal information on users. Dhamija et al. [189] conducted structured interviews with 22 participants.

Their method combined a structured interview with questionnaire-based data collection. They showed websites and asked participants to determine which ones are fraudulent: ”Imagine that you receive an email message asking you to click on one of the following links. Imagine that you decide to click on the link to see if it is a legitimate website or a spoof”. Their data collection involved a structured part, where they asked participants to state if the shown site was legitimate or not and assess their confidence. They also asked for their reasoning and to think aloud when assessing. Finally, in a debriefing session, the experimenter told about incorrect answers the participant had given and entered an open discussion about the mistakes.

The paper found that they did not look at cues such as the address bar, which would actually expose the fraud. This led to 40 % incorrect choices. The post-session interviews shed more light on how this happens. For example, participants did not know what the SSL padlock icon means and could not explain its presence.

Structured interviews can give both general but also detailed information. For example, they can be used to ask participants to describe their experiences using adjectives or nouns. For instance, [593] studied the experiences associated with the stimulation of two classes of mechanoreceptors on participants' hands. Participants were subjected to different simulations and asked detailed questions about what they felt on their hand, including ”what words would you use to describe how it felt on your hand, if at all?” and ”was there anything else you felt or thought or are there any sensations or pictures that come in your mind?”. From the resulting descriptions, it was possible to build 14 categories of tactile experiences.

11.3. Micro-Phenomenological Interviews

Micro-phenomenological interview has recently gained interest as a method for understanding the lived experience of users. *Phenomenology* here refers to the felt experience of interviewees, and *micro* to the idea that the interview consists of sequences of smaller questions and responses that iteratively elaborate an account of an experience:

Interviewee: I was frustrated when installing an update to the software.
Interviewer: What happened during the update?
Interviewee: The screen just blanked in the middle. The computer was running but the display was blank.
Interviewer: When did you see it go blank?
Interviewee: I left the computer to do the update and went to lunch. I saw it when I came back.

Interviewer: When you saw the display, what did you feel?

Three aspects are important here.

- First, the interview focuses on an evocative experience and keeps the focus there.
- Second, the interviewer follows up on a response asking to elaborate a particular moment or facet of what has been said earlier. This allows iteratively elaborating the account of the experience.
- Third, the interviewer's questions should be ''empty of content". Content-free here means that the interviewer asks questions that do not suggest an experience. A question should not lead the interviewee to misconstrue the experience. Moreover, Why-questions, and other questions that trigger reasoning as opposed to remembance, should be avoided.

For example, Petitmengin et al. [634, p.55] describe how they applied this method to learn about people's experience of meditation:

Interviewee: When I realized that I was gone, the thought vanished.
Interviewer: How did it vanish? Was it instantaneous or gradual?
Interviewee: It was very quick, but it nevertheless took a moment.
Interviewer: And what happened during this moment?

When asking how the experience vanished, the interviewer gave a spectrum options but invited the interviewee to describe. By contrast, a leading question would have asked ''Was it quick?".

As an interview method micro-phenomenological interview is a semi-structured method. It sits between open-ended and structured interviews. On the one hand, the interviewer picks the themes and follow-up questions. On the other hand, the interviewee has degrees of freedom in picking aspects that are personally important. In this interview method, it is important to allow the interviewee enough time to recall and construct the experience. Hence, even if the interview consists of short backs–and–forths, there are many pauses and hesitations present. Moreover, sometimes the same experience is iterated a few times. The interviewer's task is to keep the discussion focused on the event and guide the interviewee's attention to the relevant aspects of time (moments) and experience, and slow down the conversation to create room for thinking.

Prpa et al. [659] guide on applying the method in the context of HCI. After establishing a communication contract (agreement on the content and timing of interviews), the interviewer sets the theme of the interview: ''I would like to go back the moment [when something was experienced] and talk about it. Take your time and tell me when you are ready." To start interviewing on an experience, the interviewer should first induce an evocative state of remembrance. The goal is to get the interviewee to relive what happened. This can be tried by first asking about context of the experience (''Where did this take place? Who was there?") and then asking about the experience. After this, different techniques can be used to iteratively elaborate the accounts:

- Shifting attention between what to how: What-questions ask the interviewee to focus on what something felt like, while How-questions unfold the dynamics of how that happened (‘’How did you know the installation was still running even if the display was blank?”). The authors warn against asking Why-questions, because they ask for analytical explanations and may break the flow of the interview.
- Redirecting attention to the evocative experience: If the discussion departs too far away from the seed experience, the interviewer should refocus the discussion. ‘’Could we go back to the moment when you were seeing the blank display?”.
- Asking to describe the temporal order of the event (‘’diachronic”): ‘’What did you do then?”, or ‘’How did you start the installation?”.
- Deepening to a desired level of detail: By referring to an element in the interviewee’s response, the interviewer can ask the interviewee to elaborate an aspect of interest.
- Reflecting about the experience: Asking the interviewee to come back to the present and reflect on the experience from the present moment.

Learning to master these techniques in the thick of an interview requires practice.

11.4. Contextual Inquiry

Contextual inquiry is a type of interview developed for HCI by Holtzblatt et al. [343]. It has been influential in HCI because of its intertwining of observing users’ activities with interviewing and because of the many hands-on descriptions of how to conduct such interviews (see also [342]).

The key idea in contextual inquiry is to watch users do activities that are important to them and talk to them about those activities. Activities should be actual ongoing work or leisure, whenever possible, or concrete summaries thereof. The interviewing of users should focus on understanding the activities and the reasons why users engage in them. Holtzblatt and Beyer argue that the resulting understanding of users’ world and their desires form an important basis for the design of technology. This approach to interviewing emphasizes *realism*, in that it attempts to come close to users’ activities and their perception thereof. It emphasizes concrete data about a few users over generalizability and a full understanding of activity over precise details about a single task. A typical contextual inquiry session lasts between one-and-a-half and two hours: just enough to engage deeply on a matter but not too long to disrupt.

11.4.1. Principles of Contextual Inquiry

Inquiry is guided by four principles. These principles guide the interviewer’s attitude toward the interview situation and the subsequent analysis of the data.

Context The first principle is called context. By being close to the activity and the interviewee's perception of it we gain insights into all of those aspects of context that shape what people do. In other words, we prefer to be with the interviewee during the activities: watching them cook, listening in on them making a phone call, watching them finish their tax returns, or whatever is the focus of our inquiry. Presence allows us to see how the activity are performed, how coordination happens, which things and devices are part of the activity, why it is important. It helps us notice all of those elements of the activity that only becomes apparent in context; a normal interview, for instance, would risk missing many of them.

Sometimes being present during the activity is impossible. For instance, the topic of interest may be confidential, occur very rarely, and last so long that it cannot be followed, or distributed. In those cases, *retrospective interviews* may be an alternative. This will lack the cues usually present in context, such as particular devices, physical layout, or colleagues.

Contextual inquiry pursues concreteness. The inquiry should be about the ongoing activity because that is what reveal the details of what it is about and why it matters. The role of the interviewer is to help the interviewee focus on the activity; it is important to avoid summaries, abstractions, and generalizations that detach from actual evidence. When interviewees use such descriptions, the interviewer should use probes to direct attention to the concrete activity. This may be done in many ways but could be "could you show me how" (in response to "I typically do it some way"), "tell me about when you last did that" (after "I do this often"), "may I see one" (after "often people file complaints").

Partnership The principle of partnership suggests a collaboration between the interviewer and the interviewee to understand the activity and the users' perception of it. This contrasts with a number of other models of the relationship. For instance, it is best to avoid the role of the interviewer with a long list of question, the host visiting the users' workplace, or the expert in computer systems.

The ideal partnership in contextual inquiry is described as one between a master and an apprentice. This has two implications. First, the user needs to do the talking. Second, the interviewer must take on the role of the apprentice, attempting to inquire about the activity and immerse themselves in it. Thus, in the words of Holtzblatt and Beyer, the interviewer needs to be "nosy".

One way of supporting the partnership is that the interviewer move between withdrawal and return. During withdrawal, the interviewer discusses with the user, probing about what just happened while trying to understand what the user thinks. During return, the interviewer looks, listens, and try to learn about the activity.

Interpretation The principle of interpretation suggests that users and interviewer together attempts to create meaning of the activity. The interviewer should be listening for "no", "but", "maybe". The interviewer should also strive to confirm views and be sensitive to cues about the interviewee being hesitant about the interpretation.

Focus Finally, to make the most out of the interview, the interviewer should go for depth. Focus. This does not mean avoiding an open, explorative mindset, but rather trying to electively engage in topics that provide depth and insight in to the interaction. This is the most potent ground for surprises. It is also the one where most of the energy and emotion can be found.

11.5. Analysis of Interviews

Analysis of interviews concerns how we go from data collected in structured or open-ended conversations to insights about people, their activities, and contexts of use. As in other user research, the goal is to make sense of the participants' world, from their perspective. Any considerations of possible technological interventions like clever designs is secondary.

One common approach to analysis is to separate the four phases as follows.

- During *transcription*, the audio or video of the interview are typed so that a later analysis may be performed with the text. The detail with which that is done differs, from cursory notes on the process to detailed annotations of rythm, pauses, and intonations in speech. This phase concerns returning to the participants in interviews and verifying open questions or tentative conclusions with them. Reliability is a key concern here.

- During *the analysis*, the interviewer does rich readings of transcripts while trying also to organize and condensate the interviews. In particular, the focus is on the meanings that interviewees assign to their behavior and the world around them.

- During *verification*, the interviewer may return to participants for clarification and to check if interpretations are correct. While this is not always possible, it may give valuable input. As such, it is a tactic to increase the validity of the findings from the interview.

- During *reporting*, the interviewers write up the outcomes of the interviewing. This may be as a paper or it may be in the form of representations of users, such as models of work sequences, personas, or something third (Chapter 15 concerns these and other examples). The focus on reporting is, or should be, on the transparency of the findings.

The rest of the chapter focuses on how an analysts creates meaning; the details of the other three phases are well covered in guides on how to analyze interviews.

11.5.1. Expanding and Condensing Meaning

The analysis of interviews encompasses two processes with different intent. In *the expansion of meaning*, analysts aim to expand the interpretation and meaning found in particular parts of interviews. For beginning interviewers, this process may seem daunting. How can one sentence of an interview lead to so many interpretations and so many discussions among a team of researchers? The simple answer is that if one pays intense interest, a multitude of interpretations and implications *do* follow from one sentence. For example, imagine doing an interview on Instagram filters. If an interviewee remarks that "I use them to liven up my feed", this may be interpreted in different ways. The question that analysts should be asking is what "liven up" means. Among others, it can mean making images more energetic or aesthetic, or inviting comments via exaggeration. How can an analyst resolve which interpretation is the best?

Interpretation demands comprehensive yet sufficiently detailed readings of relevant interview data. Such readings might be supplemented by asking questions of data, in particular comparing within and across participants, seeking to understand interviews relative to the themes, and comparing interviews to what is known. The methods for analyzing interviews each prescribe different techniques for expanding the meanings and interpretations of interviews. Because it is difficult to identify and expand meanings in interviews, this phase is often done in collaboration. The benefit is also to increase the reliability of the process: the knowledge, skills, and peculiarities of a single analyst thereby have less of an influence on the final result.

The other process in interview analysis is concerned with *condensation* of insights. Rather than to expand the possible meanings and interpretations of the interview, the aim here is to reduce what is being said to patterns or concepts into an organized and manageable structure. Those will form answers to the themes of the interview and its research questions. This process may be based on previously held beliefs or theory, which is used to condensate data (sometimes called deductive or top down). Or it may be guided by the data (sometimes called inductive or bottom-up).

The main strategy for condensation interviews is referred to as *coding*. Coding implies that parts of interview transcripts are assigned one or more codes that describe some important aspect of them (e.g., the interviewer describes a particular activity). Such codes may be identified from the data, so-called bottom-up coding, or may be driven by a set of a-priori codes, for instance from an existing theory.

Analysis may be done at two different times: intertwined with conducting interviews or after all interviews are done. On the one hand, some researchers find that "analysis is not a one-time task, but an ongoing process" [696, p. 16]. Intertwining the conduct of interviews with the analysis brings many benefits. It allows researchers to change their questions based on insights from analysis, and it allows them to test conclusions from earlier interviews in later ones. This approach is often combined with open-ended interviews (see section 11.1). In that way, it supports a series of expansions and condensations of meanings. The alternative approach is to do the analysis after all interviews are done. In any case, insights about the themes and content of interviews may occur at any time during user research with interviews. A good practice is, therefore, to capture those insights in notes (sometimes called memos, field notes, or field diaries).

11.5.2. Affinity Diagramming

Affinity diagramming is a simple technique to identify themes among interviews, as well as other data. It was created and initially called the KJ method [398], but was made popular in HCI, in particular, through its use in contextual design [342]. The basic idea is to do a collaborative, bottom-up coding of parts of interviews, or other data, based on any affinity (or similarity) those parts bear on each other.

Affinity diagramming can be done in many ways; here we separate four phases based on the empirical work by Lucero [482]. These phases are typically done by a group.

1. When *creating notes*, the interviews become manageable distinct notes. These notes may come from transcriptions, subdividing interviews into sentences, question-answer pairs, or other segments that contain an understandable point. The notes may also come from people who were present at the interviews and contain interpretations or insights of what was said in the interviews. If affinity diagramming is used on other sources of data, these are similarly segmented into small, meaningful units. Notes may be put on post-it notes or small pieces of papers.

2. When *clustering notes*, the group doing the clustering either read all the notes or read each note together, one at a time. Then, the group will begin to form rough of clusters of related notes, for instance at a wall or a table (see Figure 11.1). Once clusters begin to form, the group will explain the rationales for placing notes in a particulate way.

3. When *walking the wall*, the group looks over the entire organization of notes (called the wall because they may be stuck on one). At this stage, clusters may be merged or split up, if that makes the organization of notes clearer. The group may also discuss the content of the wall to add notes, split notes (because they contain multiple themes), or clone notes because they contain multiple ideas.

4. When *documenting* the analysis, the group transfers the organization of the wall to some digital form for further use.

11.5.3. Thematic Analysis

Themetic analysis is a systematic approach to data analysis that helps identify and organize insights into the meaning across a set of qualitative data. That data may come from interviews, but could also come from surveys, online content (such as blog posts or social media activity), or physical documents. The classic description of thematic analysis is Braun and Clarke [94].

As all other methodology, thematic analysis is driven by the researchers' question. It can explicitly be inductive, going from data to themes (as affinity diagramming), or deductive, going from theoretical constructs to themes, or any combination thereof. Thematic analysis may also be concerned with obvious meanings in data or more latent insights or any combination thereof. In that way, it provides a flexible framework for analysis.

Thematic analysis consists of six steps:

Figure 11.1.: An example of activities in affinity diagramming from [482]. (a) Shows pre-clustered notes. (b) Talking through clusters and notes. (c) Reclustering notes. (d) Verbally explaining to others reasons for moving notes between clusters. (e) Bringing in interview data in the form of written quotes (on yellow post-its here).

.

1. Familiarize yourself with the data. Read the transcripts of interviews or whatever form the data is in. From a first opinion about interesting parts of the data based on active and critical thinking about the data. What does what is said or written or otherwise conveyed mean?

2. Generating initial codes. Codes are the building blocks of analysis and label a feature of data which might be important in analysing the data. Some of the codes may be based on what participants are saying, some may be based on the researchers' conceptual framework; codes may concern factual aspects of the data or deep interpretations. Coding may happen at any level, including individual sentences or words up to long segments of an interview. Creation of coding happens as one go over the data. Therefore, be generous in creating codes; at any stage of reading, it is difficult to know if something is relevant. Moreover, be sure to make several passes over data because your codes develop and you learn.

3. Searching for themes. This step requires the analyst to generate themes from the codes. A theme captures a pattern of meaning in data which is important to the research question. According to Braun and Clarke [94], themes are generated or constructed relative to the research question; unfortunately, they cannot just be found or identified. Thus, searching for themes is an active, interpretive process. In practice, this happens through collapsing or clustering codes based on similar features.

4. Reviewing themes. This step begins to consolidate and quality check the themes developed relative to the entire data set. For instance, such checks may concern

overlaps among themes, if a theme is interesting given the research question, or if a theme has sufficient and substantial data to back it up. As a result of these questions, themes may be merged, split, discarded, or developed in another way.

5. Defining and naming themes. This step is about ensuring that themes have a clear focus and address your research question. Part of ensuring that is to define and name themes. Often, participants' words and phrases are used to define and name a theme, letting them speak clearly through the themes.

6. Producing a report. This step is linked with the previous one and aims to combine a coherent argument with a narrative of the thematic literature, excerpts of the data, and the existing literature. On the one hand, this is about describing, and staying close to participants' words and what they said. On the other hand, this step is crucially about explaining how and why themes are interesting—about interpreting the interviews and qualifying them as answers to the research question.

11.6. The say/do problem

The interview method is in a privileged position as a method to understand the experienced life of users. It gives special access to the concepts and structures that people use to think about computer use. However, it is not a panacea. Its main limitation boils down to one question: Is it possible to obtain accurate information about users' practices by asking them?

The say/do problem poses a practical obstacle to this aim. The problem refers to the discrepancy between what an interviewee says and what they does. For example, when faced with the question, the vast majority of people would say that they wash their hands after using a toilet. However, according to one study that directly observed hand washing behavior [86], 15% of men and 7% of women do not wash their hands after using a toilet. In computer use, there are events that users may be embarrassed to disclose or simply do not remember.

More generally, there are many causes for the say/do problem. One cause is the inability to imagine a future. This is important for user research. In user research, we are often tempted to ask about alternative and possible technologies. However, as interviewees lack first-hand experience with them, how accurate can their estimates be? Even professionals have difficulties in predicting the development of their fields. For example, most IT professionals were unable to predict how pervasive impact smartphones would have on our use of computers, and we are witnessing the same with emerging forms of AI. Interviews are not ideal for obtaining insights into these kinds of questions.

Another cause of the say/do discrepancy is that not all knowledge can be articulated. Some knowledge is tacit. However, tacit knowledge can be exposed by acting on it out. If you ask regular users about their use of smart features on smartphones, such as autocorrection or password managers, they may not be able to tell what they do. Using methods such as contextual inquiry, such tacit behaviors can be observed, recorded, and discussed about.

A third cause is social. Users may deliberately exaggerate or omit aspects of their behavior. They may want to give a positive impression of themselves or avoid negative social consequences. Recognizing such situations is important for interviewers. For example, forming a trust position with vulnerable groups can take months or years.

A skilled interviewer is aware of the problem and these causes. Part of the expertise of interviewers is to recognize issues like this, avoid misinterpreting their data, and design the study method accordingly.

Summary

- The interview method is an indispensable method in user research. It allows access to the first-person view, to learn about how users experience and perceive interaction.
- Interviews can activate people, empower them, and get them narrate their experiences.
- Interview methods are divided into structured, semi-structured, and open-ended interviews.
- Contextual inquiry is a mixed-methods approach that combines interviews with field observations.
- Methods for analyzing interview data aim at comprehensive overviews of data on theone hand and interpretations that are justified by data on the other.

Exercises

1. Open-ended interview. Pick an application you do not know really well but your friend (the interviewee) does. Pick a theme for an interview and design a few lead-in questions following the guidelines given in this chapter. Interview your friend and record the interview (about 10-15 minutes). Listen to the interview afterwards. Instead of analyzing the content, focus on how you lead the interview.
2. Designing an interview structure. Think about a future product, something you believe might come out in 3-4 years. Think about the potential user needs. Write an interview structure for this product. The goal is to find out early adopters' experience of the product.
3. Analyzing transcripts. There are many data sets available with interview transcripts (of varying quality) available online. Pick any of your liking that has more than one person's data and analyze it. 1) Identify a theme relevant for the interview. 1) Develop a code (of minimum 2-3 categories) for the theme. 3) Verify your code against the transcripts; iterate if needed. 4) State your conclusion about the theme.

4. Contextual inquiry. Conduct a contextual inquiry on an everyday activity with a computer. Plan, run and analyze a contextual inquiry with one person (other than you) that focuses on an everyday activity, for instance, checking latest posts in social media.

5. Microphenomenological interview. Following the guidance given this chapter, design an interview structure to find out how your friend experiences awareness cues (see Chapter 9 in social media. Mastering this type of interview takes practice. Pay attention to how you respond and lead discussion without biasing the interviewee.

12. Field Research

This chapter concerns user research methods for learning about users *in situ*; that is, in their natural environments. *Field research*, in the context of user research, refers to the collection of data on users in their real contexts. In much field research, users are observed doing things that would also occur without the researcher being present, and – as much as possible – with the minimum bias of the researcher being present (for reactivity; see Chapter 14).

Observation is a key method of field research. In observation, we pay close attention to what users do, aiming to understand what happens and make inferences about why activity unfolds as it does. Observations may be supplemented with both informal conversations and open-ended interviews. In this chapter, we focus on observation. Interviews and contextual inquiry were covered in Chapter 11.

The strength of field research is *realism*. Field research provides insights into how users act, collaborate, and communicate in their natural environments. For this reason, it is sometimes called naturalistic research. In the palette of user research methods, field research is vital in capturing aspects of use that other methods might miss. Due to its openness, it sheds light to the complexity of the social and organizational setting in which interaction is accomplished. By making a few commitments to predefined notions of the phenomenon under study, field research can uncover factors that might not be discovered by other means. Field research also produces knowledge about the practices of stakeholders and how they understand that practice. It also teaches us about the relation between prescribed procedures (e.g., manuals, organizational procedures) and how work actually gets carried out.

Such insights can be put to multiple uses. Field research has helped describe many important phenomena in HCI, for example,

- how interactive televisions are used in the home [592]
- how people develop romantic relations in computer games [250]
- how new information systems create unexpected *new* work for end-users (e.g., [647]).

Field research may also be used to inform the design of new interactive systems, serving as a way to emphasize with prospective users and to provide background knowledge about them and their practices. It may also help inform decisions about whether to adopt a system or not. The side box describes the use of field research to uncover safety issues that a planned automation of a paper-based system in air traffic control would pose.

An example of an observation site is given in Figure 12.1. It may seem deceptively simple to observe what is happening. However, one cannot simply 'go and watch' users.

Figure 12.1.: Field research is uniquely positioned to understand how technology use is contingent on its context, including the material (think: setup of the room), social (relations between the workers), and technical (digital and other artefacts) aspects. Observing technology use in its context is not trivial; it requires planning and skills in deciding what to observe and how. Example recording from a field study of a Rally control center [846].

It is not obvious who to observe, what to look for, or how to make sense of it all. This chapter provides an introduction to these questions. A key challenge in field research is understanding *contextual contingencies of human-computer interaction*; that is, the different ways interactions are produced and shaped by their social, physical, and technical circumstances.

Every field research must answer three formative questions [479]:

1. What to focus on? In any complex real-life situation, there are several things to observe. How to decide what to pay attention to and what to ignore?

2. How to gather data? In particular, how to plan and conduct observations. The way this is done affects everything from recorded data to access to sites and ethics.

3. How to analyze the data? Field research often produces an overabundance of data.

For example, a fieldwork on Facebook use in rural Kenya collected "24 transcribed interviews, 174 pages of field notes, and 1,375 digital photographs" [897]. The authors' intention was to understand "what is the Facebook experience in contexts where the social, economic, and technical contexts affecting use differ from the American college campuses where the site was first popularized?" (p. 33).

In the rest of this chapter, we first discuss observation in general and then ethnography, a particular form of observation where the observer engages in and participates in what is observed. The last section deals with analysis of data from field research.

Paper Example 12.0.1 : Understanding automation and work practices

Many interactive systems digitalize or automate human practices. However, such attempts may fail if the social, cognitive, and material aspects of practices are misunderstood or ignored.

On a case in point, in the late 1990s, MacKay [490] conducted a four-month ethnographic study of air traffic controllers and their use of paper. Her study was motivated by the objective of mimicking and replacing paper with digital systems. However, her research has shown the benefits of paper and its role in coordinating work flexibly.

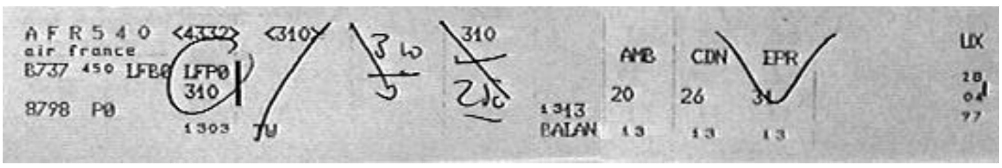
The author observed how a team consisting of senior controllers, qualified controllers, and students. They worked together as a team. Their task was to organize planes at an airport and to provide instructions to pilots on safe and efficient routes. To respond to changing situations during traffic hours, the team had to collaborate and maintain awareness.

A paper flight strip is a paper-based representation of a single flight; see the figure below). It contains a few basic parameters of a flight in a structured format. As it is in paper format, it can be directly edited with a pen. Paper strips can also be spatially organized. At first, it may seem trivial to replace such simple representations digitally; after all, there is no information that could not be presented on a display or interacted on. Mackay's study found out, however, that this not the case.

She found that an essential aspect of a controller's work routine is sequentially checking each aircraft, first on radar and then on the paper strips. She found that regardless of how many aircraft are in the air, controllers execute this cyclical checking routine because it allows them to stay vigilant and ensure a certain level of awareness at all times. This, some controllers said, is also important during " boring" times.

Paper strips could be organized on a control table in such a way that a quick overview of a situation could be obtained. This spatial organization also helped the team maintain awareness of each others' situations. But paper strips also served an anticipatory function. Mackay observed that controllers make markings "for the future". For example, a controller might insert information that would serve as a reminder to call a pilot a few minutes from now. Having that information on a paper strip helps the controller "offload" information to the strip so that it does not have to be kept in memory. A computerized representation that would not help to do free-form annotations, some controllers argued, might increase workload.

Mackay concluded that the change to a computerized system could endanger air traffic due to unforeseeable consequences. She proposed a radical alternative: maintain the physical paper strips, but turn them into an the interface to the comptuer. This, she argued, would help controllers leverage their existing, safe practices with paper strips. Alas, this vision was difficult to realize using the technology of that era.



12.1. Observation

Observation is a method of data collection where a researcher is positioned to perceive and make note of activity related to the use of interactive systems. The goal in collecting such notes is to account for relevant types of events, their frequencies, and patterns occurring over time. This section deals with with the basic idea of observation and how to use it to describe what users do.

One ideal in observation is to interfere as little as possible in what is being observed. The researcher should be positioned in such a way that there is little interference or bias due to the presence. This type of observation is sometimes called an outsider's or *etic* view. For instance, etic observation could be relevant to study

- how much time do people spend doing something, for instance looking at interactive exhibits in shopping windows
- how users of a virtual reality world engage with strangers
- how does a team of air traffic controllers respond when unplanned events occur.

All of these may be observed, enumerated, and verified by a researcher describing the events from an outside point of view. By contrast, in *emic* research, which we discuss in section 12.2, the goal is to immerse oneself into a community to describe it from their perspective.

In terms of the three purposes for field research presented at the beginning of this chapter, the focus of data collection is about what the goal of the observation is; gathering data is mainly about finding an appropriate site of observation and good ways of capturing data; analysis of data mainly about the reliability of data collection and analysis of types and frequency of observations. Next, we discuss these focusing, gathering, and analysing activities.

12.1.1. Focus

What to look for in observation is largely prescribed by the goals of the user research. For instance, to design an interactive system for groups it may be useful to know how it is used by single users and when users are in a group—an existing system may be observed to figure this out. What to look for in observation is also influenced by how much is already known. For instance, observation in the field may be used to confirm usage-patterns identified in interviews (confirmatory) or it may be used to figure out what the most frequent activities are altogether (exploratory or open-ended).

The literature offers an assortment of frameworks that offer researchers inspiration for what to focus on in inspiration. Table 12.1 shows nine dimensions that may be the subject of open-ended or confirmatory observation. For instance, in observing space a researcher would pay particular attention to the use of space in collaboration around a shared technology. Their observations may then concern how different individuals use different parts of the space, how they place and retrieve artifacts in that space, how they physically position themselves relative to the space, and so on. Observations focusing

Dimension	Description
Space	Physical layout and organization of the observation site
Actors	Names, roles, and other characteristics of the people at the site
Activities	The doings of the actors
Objects	The physical elements of the site
Acts	Individual things that people do and say
Events	Particular important things that happens
Time	The sequence of events and activities
Goals	The things actors are trying to accomplish
Feelings	Actors' emotions and moods

Table 12.1.: Observations in the field can be structured. This helps the researcher focus on aspects of the field directly relevant for the research question. The table shows dimensions of observation according to [768].

exclusively on space has for instance been done in research on tabletops, large vertical displays that invite users to collaborate. However, observational research on tabletops have shown that users' behaviors are shaped by a sense of territoriality [731]. Users treat different parts of the display space differently, depending on whether it is a personal or a shared territory.

Other frameworks offer other dimensions for thinking about. Goetz and LeCompte [273] suggested looking at who is present, what their role is, what is happening, when something happens, where it happens, why it happens, and how the happening is organized. Others suggest to look at the physical setting, the activities, the social environment, the formal interactions among people, the informal and unplanned activities, the nonverbal communication, and the things that does not happen (but might somehow be expected) [273].

12.1.2. Gathering Data

Finding a site

An immediate question facing field researchers concerns *where* to observe users, including how to obtain access to the site of observation and whether observations should be known to participants or not. Picking a site of observation again relates to the purpose of doing the user research in the first place. The site should reflect that purpose. Every observation site permits certain types of observations to be made. A poorly selected site will result in limited or skewed observations.

Another concern is the effect of the selected site on the observed individuals. *Reactivity* refers to the impact of the observer on the observed individuals. Consider, for example, having a field researcher hired by your employer observing you, how might that affect your behavior?

The first task in coming to a site of observation is to orient. The observer should understand the structure of the site, including the positions of the individuals being observed, as well as those of relevant physical and digital artefacts. The following kinds of questions should be answered:

- Who are the actors in this space and what roles do they have?
- What are the relevant material or digital artefacts in this space?
- What are some typical or probable locations or routes of the relevant actors?
- How am I and my recording instruments positioned in this space? In particular, if the use of computers need to be recorded, how to position in a way that that can be achieved.

To this effect, an annotated floorplan can be used. The site of observation, including the positions of cameras, if any, should be noted. **??** shows an example from the field study of a control center shown in Figure 12.1.

Shadowing

Rather than settling on a fixed site of observation, particular people may be followed. This is known as *shadowing*. González and Mark [278] were interested in how users switched between different activities at work. To obtain information on this, they shadowed information workers at an investment management company. The researchers "sat with the informant at her cubicle and followed her, whenever possible, to meetings or other activities" (p. 114). Observations were recorded on an "activity tracking log where we transcribed the observation notes collected during the day". These observational data was complemented with other data collection, including interviews. The observations exposed a fragmentation of work. The authors found that "people spend an average of less than two and a half minutes reading email before they switch to another event, or are interrupted" (p. 116).

Observation may also be *covert* so that users do not know that they are observed. In virtual environments, for example, covert observation is possible since the observer may not need to have an embodied presence in the space that is being observed. In public and semi-public spaces, video cameras may be used for collecting data. Researchers can also ask study participants to collect data using wearable cameras or other sensors that they carry with them [115]. This may, however, alter the behavior of the users who carry such cameras as well as people around them. However, outside of institutionally cleared circumstances, covert observation can be against local regulations. It is also ethically precarious because the individuals being observed may not be informed nor consented to observation.

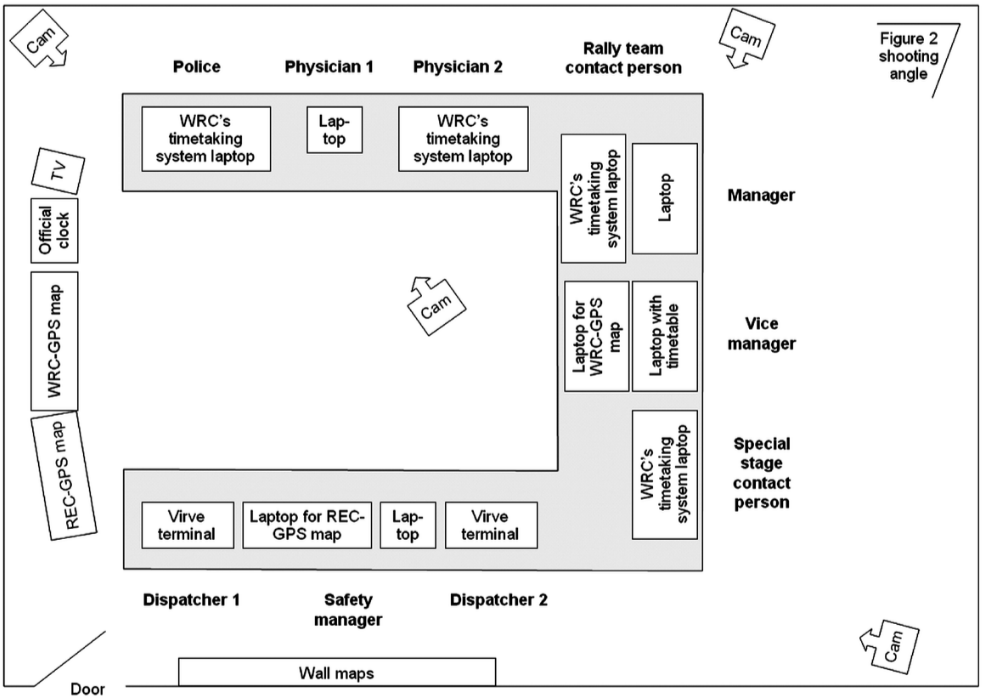


Figure 12.2.: Entering a site of observation requires the researcher to orient with the actors, materials, and artefacts in the setting. Example from a study of large-scale sports event control center [846].

Data Capture and Note Taking

Observations in the field should be recorded for later analysis and verification. Observations may include audio, video, photos, as well as textual descriptions. There are a few principled ways to do this.

Field notes are recordings of remarks made in the field. They are the "raw data" of field research. One method for field notes is to keep a *field diary* A diary acts as a log of observations, typically taken with pen and paper or digitally, for example, with a mobile device. A diary entry relates an observation made to a time of occurrence. One issue with diaries is that they do not necessarily guide observations to be consistent over time. Diary entries can be sporadic and focus on wildly different points.

Structured notes are field notes recorded into a previously defined observation structure [535]. The difference to the diary method is that this structure guides observations and allows the discovery of patterns that occur over time. Moreover, quantitative analyses can be done based on structured notes. The structure can, for example, concern chronology (what happened and when), artifacts, participant roles, or particular actions. Structured notes are often made with pen and paper or a mobile device. For example, an observer could note all events where a user sent an email, marking further observations on how that was done. Every log entry would entail a description of one such event.

We provide ideas on possible categories of observation in Table 12.1. Often, however, the structure evolves as the study progresses. As it is not clear during the early part of a study what will prove important, *non-structured field notes* may be recorded, which may then later lead to a coding structure.

12.1.3. Analysis of Observations

From video clips to diaries, field notes can be of radically different types. However, there are a few shared principles to consider when analyzing them.

Immediate recall: When in the field and "in the thick of it", events pass quickly. A field researcher often has to choose whether to immerse themselves in the ongoing event versus writing notes about them. Since even the most comprehensive recording offers just a glimpse into what transpired, field research almost always relies on the observer's memory. It is therefore critical that observations are analyzed quickly after being recorded. Knowing when to observe, when to take notes, and what to remember from an event, are some of the skills that one develops as a field researcher.

Thick description: Important events should get more attention in analysis. To this end, a field researcher should try to recall as much as possible of an event and write them down, answering questions like: What was observed? What does it mean? Why does it matter? This can lead to a rich account of an event called *a thick description.* A researcher can also jog memory, for example by revisiting audio/video recordings or using representations of events like floorplans.

Coding data: Any post hoc data coding should strive to be reliable and reproducible. Categorizations of observations should be based on as unambiguous definitions as possible. Those definitions, with examples and counter-examples, should be rigorously written down into a *coding manual*. A definition can tell, for example, that

> ”Map reading”: This event occurs when a participant manipulates or glances a map (physical or digital) for more than 5 seconds. Keeping a map open on the side does not count as map reading.

Coding manuals can be attached as appendices to field reports or scientific papers.

When statistical reliability is important, *inter-rater reliability* should be measured and reported. They are statistical metrics that quantify how reliably two raters, working independently of each other, end up with the same codings given the same materials and the same coding manual. Since coding the whole dataset is time-consuming, the raters are usually asked to code an overlapping sample. When conclusions drawn from a study are quantitative, low reliability is a serious concern and calls for rethinking the coding manual.

Validation with participants: Conclusions that are drawn based on field research are statements made about users. From this point of view, users are stakeholders in field research. For them, these statements can be wrong: irrelevant, biased, partial, or incorrect. Even in the case of covert observation, it is their lifes, opinions, and practices that are being described. Validation can seek to verify conclusions drawn from data. It can also provide some agency to users on what is being said about them. This is particularly important for vulnerable user groups. Engaging with them may help correct misinterpretations. For example, task analysis explicitly requests task analyses to be validated with stakeholders [25] (see Chapter 15).

12.2. Ethnography

Ethnography is a method developed in anthropology to describe activities from the perspective of an individual or group. Although the data collection methods are similar to other field research – be those observations, interviews, analysis of documents, etc. – the purpose is different. The goal of ethnography is *emic*; in other words, to describe activity from the users' point of view, often focusing on the experiences or social organization of the activity.

In HCI, ethnographic research contributes to understanding of how work that involves interactive computing systems is carried out and experienced. As surprising as it can sound, emic research on computer use can provide insights. Computer use often involves meanings, practices, expressions, and activities that may require an ethnographic approach to be understood. This insists the researcher to minutely and over extended periods of time to observe the participants in naturally occurring settings. For instance, in the side box, a study of system administrators is described.

Ethnography requires long immersion and commitment into field work; Instead of days, rigorous ethnographic research can take months or even years. As a consequence, an ethnographer may become a participant in the community he or she is studying. If the researcher becomes accepted as a member – even if peripheral – of the community, there may be less of an issue of reactivity. Individuals may be less prone to change their behavior when the observer is familiar. In some cases, this may be a requirement to access a community as a legitimate member; think, for example, studies of Internet criminals.

In practice, ethnography often comprises multiple data collection methods, including observation, interviews, and collection of artefacts. In the present chapter, however, we focus on the high-level principles

Paper Example 12.2.1 : An ethnography of sysadmins

Thanks to deeper immersion to the experienced lives of the participants, ethnographic research can illuminate technically specialized user groups. One such group is system administrators. System administrators, or sysadmins for short, are professionals that set up, maintain, and repair computer systems and infrastructure.

To understand their practices, Barrett et al. [48], then researchers at IBM Almaden Research Center, observed web and database administrators for 25 days, a total of over 100 hours of recorded materials. They followed a mixed-methods approach consisting of diaries, interviews, and observations. What their data uncovered would have been difficult to uncover with other than an open-ended method.

In one of the episodes, a sysadmin team was creating a new web server with a tight deadline. Configuring the server produced a vague error message "Error: Could not connect to server." The team went into intense troubleshooting mode, which included phone calls, emails, messaging, and discussions with several people in the organization. Some came to the terminal of one of the sysadmins, called "George", and gave ideas on how to solve the issue.

> The problem was eventually found to be a network misconfiguration. George misunderstood the meaning of a certain configuration parameter for the new web server (ambiguously labeled as 'port') to be for communication from the webserver to the authentication server, when in fact it was the opposite. The former would have been permitted by the firewall, but the latter was not. George's misunderstanding affected the remote collaborators significantly throughout the troubleshooting session. We witnessed several instances in which he ignored or misinterpreted evidence of the real problem, filtering what he communicated by his incorrect understanding of the system configuration, which in turn greatly limited his collaborators? ability to understand George's error propagated to his collaborators. The solution was finally found by the one collaborator who had independent access to the systems, which meant that his view of the systems was not contaminated by George's incorrect understanding.

The paper showed that the work of system administrators includes hours of troubleshooting complex information, tools, and organizational knowledge. Yet, existing tools that sysadmins use do a poor job supporting that. Most importantly, they are opaque, poor at communicating their state, which means that misunderstandings can propagate also in a collaborative setting like this. This has led sysadmins to build their own scripts and GUI tools.

12.2.1. Principles of Ethnography

Ethnography is centered on three principles [80].

Members' point of view: Ethnography aims to provide the member's point of view. Instead of describing what a person does, the goal is to describe its meaning as an experience. The term 'member' here refers to a member of a group or culture that is being studied. This means understanding meanings, concepts, values, and other factors that define that perspective. An ethnographer starts with a minimal set of assumptions. This is important when seeking appreciate the stance of the member's. The side box provides an example in a case where print operators' practices were studied. Sometimes getting the participants' viewpoint requires long groundwork. One first needs to gain acceptance and credibility. After a sufficient level of trust and rapport has been established, the participants may be more willing to disclose information and engage with a researcher.

The strive for a member's point of view differs from the *etic* descriptions we discussed in the previous section. Etic research routinely utilizes extraneous concepts to describe events. For example, in task analysis, we use a predefined ontology to describe an activity. An ethnographic researcher would seek to find out the member's viewpoint instead.

Descriptive: Ethnographic research does not seek to impose extraneous values to the group being studied. In other words, it describes and does not prescribe. This position allows an ethnographer to approach groups that might be reluctant to value systems or concepts that they consider threatening.

Holistic: Ethnography strives away from predefined narrow categories of observation, such as we need to use in surveys, and instead aims to describe the totality of events. 'Holistic' is often criticized as an empty, vague goal. In ethnographic research, it does not refer to the impossible feat of describing every possible detail of an event, rather to the idea of including all of those factors that members themselves consider relevant. If something is relevant to them, it should be included in the description.

Paper Example 12.2.2 : Print operators' practices

Button and Sharrock [112] provide advise for field observations. To help gauge the participants' perspective, what people do should be described in its own terms. For example, managers and consultants may describe work in formal terms that do not meet the experience of the workers. What does, for example, 'business process re-engineering' mean? To a worker, it may mean very concrete changes in who sits where or what buttons need to be clicked. Ethnographers should aim to "tell it like it is"; that is, tell how people do their jobs without masking these descriptions with jargon.

The authors illustrate this point with a description of experienced print operators (original study published by Bowers et al. [90]). The company had adopted a new workflow and a management system proposed by consultants, who however were not in touch with the operators' work on the ground. Via field observations, the researchers wanted to learn how operators actually organize their work. One key lesson was that operators engage in a number of actions to manage contingencies in a print room, and develop sensory expertise to "know" and coordinate work with others in a room. A print room, they explain, for an unexperienced bystander, may appear cacophonic, is however experienced differently by the operators [p.71]:

> In addition to the resource provided by direct line of sight, operators have other ways of monitoring each other's work. First, the machines emit warning noises that alert operators to completed processes or machine problems such as paper jams or depleted paper trays. In addition to 'designed-in' sounds, the machines also make regular noises when engaged in particular operations and skilled operators can monitor the printing process just by the sound that a machine is making at any one time. The neophyte to the print room encounters a cacophony of undifferentiated sounds, merged together into a booming, sometimes overwhelming noise. To experienced print workers, however, the noise can be disassembled into its component sound parts, and they can draw out a sound from the noise. Thus, what is mere noise to a neophyte is a rich resource of meaningful sounds that can be used to monitor the overall work of the print room and the state of particular machines by the experienced operator.

First-person perspectives to practices can be illuminating and offer a vastly different view than formal descriptions of work. The authors argued that, in this case, the (consultants') new management system failed to achieve its goals because of dismissing the workers' practices.

12.2.2. Rapid Ethnography

While these attributes necessarily make ethnographic studies longer than non-ethnographic field research, there is no intrinsic value in prolonged exposure per se. *Rapid ethnography* is the idea that immersion should only be as long as it has to (see [531]). Sometimes an actionable understanding of the subject matter may be achieved in a matter of few days, or even hours. This may be necessary to align ethnographic work with rapid product development cycles.

Rapid ethnography requires relaxing the ideal of holistic description and, instead purposefully limiting the scope of the study. Several observers may be used instead of one, field data can be augmented with digital data collection (e.g., logs, documents), and the focus may be on preselected "key informants". Generally, such "quick and dirty" ethnography is more feasible in the special case when the observer already shares essential aspects of the culture or where the researcher is approved within the community. For example, a user researcher coming from an IT company may be more easily accepted by special groups like system administrators.

12.2.3. Ethnography in Practice

In HCI, ethnography can be used for at least three main purposes. First, ethnography can be used to obtain a holistic view of how end-users view some technology. The method is indispensable when the user group in question is very different from the lifeworld of the researcher. Second, ethnography can also be used to evaluate some technology from the perspective of members. Third, ethnography can also be used concurrently with system or service development. In this case, user researchers may need to engage with a user group for a longer time, even years.

In practice, ethographic studies involve four distinct phases:

Getting access and getting along: Finding and accessing informants is a type of work that is decisive for the success of a study, yet rarely appreciated by outsiders. (How informants were found is rarely mentioned in HCI papers reporting ethnographic studies.) Approaching the members with a humble, unthreatening, and polite mindset is necessary for establishing trust.

Developing membership: Unlike in experimental research, where observations are supposed to be independent of each other, in ethnographic research the observational capability of researcher developers over time. This is due to epistemological development on the one hand. In short, the researcher grows to understand the topic. On the other hand, it is due to changing membership characteristics. As the researcher becomes more familiar and trusted, there is a better chance of discovering knowledge that would be inaccessible to outsiders.

Collecting Data: From a methodological perspective, an ethnographer can flexibly combine data collection methods, both quantitative and qualitative. However, a main aspect of ethnographic data collection is the ability to talk to and watch people and make recordings and field notes. This never ends, it only deepens. Ironically, for an outsider an ethnographer may appear passive.

Button and Sharrock [113] outline advise for making observations in the field. One advise is to follow people with open eyes and adopt the mindset of an apprentice: bear as little predisposition as possible. Another is to follow leads: keep asking people to describe what they do as they are doing it, such observations can provide leads on what to look at next. For example, one should learn about where one's work's outputs go and how others use those outcomes. All the time, a field researcher should make notes of everything that occurs as richly as possible. The authors remind of the importance of investing time into thinking what the data means as opposed to spending all time just collecting data.

Analyzing Data Data analysis differs between ethnographic research and laboratory research in a few vital points. First, often in lab research extraneous variables, those that are not under the control of the researcher, are considered a nuisance, something to get rid of. By contrast, in ethnographic research, such factors are of essence in offering a holistic description of the member's perspective. This unnecessarily results in appreciation of anecdotal evidence, like "war stories" that an IT worker can recount. However, such accounts are not antithetical to the aim of understanding "the big picture". Second, in laboratory research the focus is often on understanding causal effects of some variables on some other variables. By contrast, in ethnographic research it is assumed that the world is constructed in via activities of the informant. That is, it should not be pre-fixated on some variables but "let the data speak". Third, while laboratory studies often strive for statistically reliable findings, in ethnographic studies the goal is to offer accounts that are consistent with the data that have been obtained. The burden is on the ethnographer to show that the interpretations can be traced back to obtained data. Fourth, while in laboratory studies, every observation is equally important, ethnographic research can be selective. It can focus on events and people deemed "important" by some account.

12.2.4. Ethics

The deep connection that an ethnographer can achieve with the individuals under study is both an opportunity and a source of risk. On the one hand, there is the positive side: ethnographic research is sometimes the only viable means of understanding the viewpoints of groups that are vulnerable or neglected. For example, economically or otherwise disadvantaged people's experience of computer use is completely different from that of others. How else could one seek to report what they experience than via ethnography?

On the other hand, ethnographic research can be harmful. It can harm participants by releasing information about them. It can also harm them by being present in their lives. The following suggestions are essential to avoid harm [84, p.176]:

1. To be extremely careful with field notes, being sure not to leave them in a place where anyone might pick them up.
2. Not to discuss the study or any of the people in it with others outside the class.
3. To use pseudonyms in field notes as well as in final reports.
4. Not to report any of the information they find to people who might use it in a way that might embarrass or hurt the subjects.

5. To strike a clear bargain with the subjects as to what they can expect to get out of the study, and to fulfill that obligation.

12.3. Can Field Observations Inform Design?

Realism is the strength of field research. Its results offer a level of fidelity and open-endedness that is out of reach of other approaches in user research. Field research conveys in detail how actions and experiences take place in context. However, there is a catch. Realism also makes it hard to draw implications to design. When every observation is detailed and contingent – tied to a unique circumstances, how can one draw conclusions that are general enough to inform design? One might be tempted to summarize field observations in a 'takeaway bullets' for design. However, the threat is that this would blunt the work and discount the richness of the underlying data.

The 'So What?' question riddles both academics and practitioners. In addition to the issue that the value of field research is hard to quantify, it is also costly when compared to other methods such as surveys. To make field research more cost-efficient, *rapid ethnography* has been proposed, as we discussed. While this involves less immersion in the domain, 'quick-and-dirty' methods can more efficiently gather information that designers may seek.

However, there is long-term value to field research. It is hard to quantify but valuable. While field research cannot offer results that define designs, it offers knowledge of strategic value that complements other methods. Instead of asking "Should we do field research?" one should ask "Can we succeed without it?".

First, field research is particularly strong in identifying non-obvious problems affecting computer use. While for a designer something may appear straightforward to use, there are many factors that can impede use in real life. Some of such factors are mundane (e.g., weather) while others are subtle and hard to expose, such as power structures. For example, field research at workplaces has reported 'discrepancy problems': problems where managers' expectations of an information system do not match what happens on the ground [647]. New practices or workflows assumed by a system may end up producing more work to end-users, at times without managers or developers knowing about this. To draw implications like this, the data may need to be triangulated with other data. This can help an analyst to construe models of data – 'theories' of user practices – that posit observed factors in some systematic form. We discuss methods for representing user data in Chapter 15.

Second, field research is important when changing or designing new socio-technical practices. Many computing systems are not just 'adopted', but they need to be domesticated and appropriated (see Chapter 19). Users need to reorganize their practices. Understanding the fabric of such practices is where field research excels.

Third, sometimes field observations are instrumental to radical new concepts in design. For example, early field studies of wireless connectivity led to the concept of *seamfulness* [142]. The lack of wireless connectivity can normally be considered a disruption, something to avoid exposing users to. The concept of seamfulness, on the contrary, suggests that to make connectivity more visible and understandable, users should better learn to manage it in their practices. A seamful connectivity map would show where connectivity is strong vs. poor.

Fourth, field research also has narrative power. Well recounted accounts of users' lived world are something that stakeholders can easily relate to. This can get buy-in within an organization. It can also sensitize developers. It can make them realize that system requirements are harder to achieve 'in the wild' than imagined.

Finally, one can argue that deep understanding of situated use of computers is valuable on its own. If field work is considered as basic research, there is no need to justify it via its instrumental value in the design.

Summary

- Field research relies on observation of users in their natural contexts.
- Field observation is not just about 'keeping the eyes open' but a complex skill that balances between 'being there' and taking notes.
- Observation methods are divided into etic and emic methods. Ethnographic methods, which belong to the latter, insist on longer-term immersion in a culture as a key to obtain first-person accounts of practices.
- The results of field research are valuable in identifying problems, deriving models of contextual use, and generally for sensitizing with factors relevant in a project.

Exercises

1. Organizing field studies. Imagine that you are tasked with the goal of running a field study of a help desk in a large IT company. Following the guidance given in this chapter, how would you organize it? Consider aspects like choosing a site of observation in your answer.
2. The value of field observations. Field research is expensive, but there are many cases where it is more expensive not to do that. Come up with three examples of an HCI project where field research is necessary.
3. Membership. Why is membership in the community that is being observed important?
4. Try to do a practical observation. Decide on a setting where you can observe people buying tickets. Also decide on the focus of the observation and the practicalities of observing. Then try to do it three times, 20 minutes each time. Then analyze your observations, possibly supported by some of the material in the book.

13. Survey Research

In *survey research*, a researcher designs *a questionnaire* and distributes it to respondents who fill it in, typically without the researcher's presence. A questionnaire consists of a series of questions presented and answered in a structured way. For example, you have probably filled in questionnaires that use the Likert scale. Here, a question is delivered in the form of a claim where the response must be given on a symmetrical 5 or 7 point ordinal scale, ranging from example 'Strongly Disagree' to 'Strongly Agree':

```
''I enjoy reading textbooks in an e-book format on a digital device''
1 - Strongly Disagree
2 - Somewhat Disagree
3 - Neutral
4 - Somewhat Agree
5 - Strongly Agree
```

Figure 13.1 shows another example of a questionnaire used in HCI.

Surveys are widely usable in user research. They can be used to understand users' (1) behaviors, for example based on their reports on their activities, routines, or uses of interactive systems; (2) experiences, such as their reports of positive and negative experience during computer use; (3) needs, desires, and wants; (4) attitudes, preferences, and beliefs. For example, surveys have taught us the following insights:

- Hiniker et al. [337] used surveys to investigate how parents and children establish rules around digital technologies in their families and how effective they perceive such rules to be. Parents and children each reported two rules in free text along with ratings (between disagree and agree) of the rules. The results suggest that rules that constrain activities (e.g., no Snapchat) are more likely to be followed than rules relating to the context (e.g., no phone at dinner table).

- Ceaparu et al. [141] used surveys to study the frustrations that users experience with computer systems. They showed that users frequently experience frustration with error messages, download times, and features that are difficult to find. About half of the time spent on a computer is perceived to be lost due to such experiences. These are alarming results.

- Vaniea and Rashidi [827] used surveys to investigate when and why users update their software. They learned that users go through six distinct phases when deciding to upgrade software and that several of these prevent barriers to users upgrading, therefore potentially compromising security. Some users choose not to upgrade if the current version works fine or if upgrades look like they will be disruptive.

> Respondents frequently mentioned troubleshooting the installer and the installation process as barriers to successfully updating software.

Surveys may also be used for understanding psychological factors behind user behavior. Pre-validated questionnaires can be used to measure latent psychological constructs such as attitudes. To this end, several specialized questionnaires have been developed. Examples include the visual aesthetics of websites [544], as well as the frustrations [785] and engagement that people experience when using computers [591]. Chapter 7 provides examples of user satisfaction and experience.

Generalizability of findings is the distinct benefit of survey research over other user research methods. A carefully designed survey allows a researcher to statistically estimate distributions of phenomena in a target population. This is thanks to two characteristics of surveys. First, response data is structured and defined in advance. Every respondent answers the same questions using the same response options. This reduces heterogeneity in data and lends them statistical analysis. Second, sample sizes in survey research can be large. Assuming some efficient way to distribute the survey, cost per respondent can be moderate or low. If the sample that is collected is representative of prospective users, the generalizability of findings from surveys can be high.

However, designing a good survey is hard. A number of things can go wrong. For example, questions used in survey research must be understandable to respondents. Poorly formulated questions are interpreted by the respondents in dissimilar ways. The researcher cannot be present to clarify, follow up, and quality answers. When compared to the interview method, this is a significant price for generalizability. Also the answer options need to be comprehensive. If an answer option is missing, or the options are biasing the respondents, the resulting data will be biased, too. Survey results are also easily misinterpreted as representing the ”true” behaviors or experiences of users. However, self-reports may not be truthful. For example, if a questionnaire asked about your salary, use of pornographic web sites, or tendency to procrastinate, would you answer truthfully? Moreover, distributing questionnaires to the actual population of interest may be hard. Surveys often see a low response rate: some users self-select to respond and some not, which may cause a bias in the data. The data will have those respondents overrepresented who, for some reason, chose to answer the survey. We will discuss these and other biases in detail throughout the chapter. Developing a good survey is hard work including multiple iterations and stages.

Survey research may be done in many ways, depending on purposes and constraints. However, typically, the workflow consists of the following steps:

1. Planning survey research
    - Establishing research focus
    - Selecting survey type
    - Deciding the sampling strategy
2. Selecting a questionnaire
    - Using an established questionnaire or

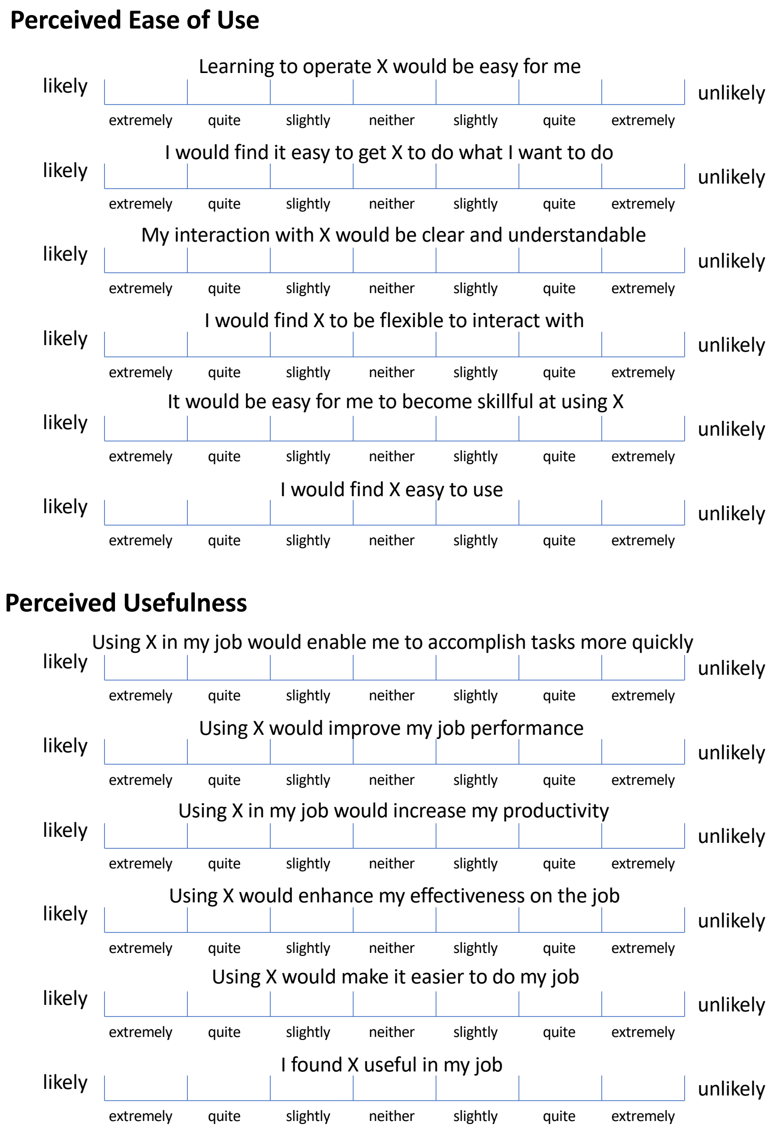


Figure 13.1.: An example of a questionnaire used to measure the extent to which people accept software (denoted by X in the table). The questionnaire is based on the Technology Acceptance Model [179] (see Chapter 19). The questionnaire contains two measurement constructs: (i) perceived ease of use and (ii) perceived usefulness, each measured with six questions. These two constructs are used to predict the acceptance of the system.

   - Designing a new questionnaire, including questions, their types, answer options, and order
3. Piloting
4. Collecting data
5. Analyzing data
   - Processing the answers, describing answers, and carrying out qualitative and statistical analyses
   - Drawing conclusions and reporting.

Next we discuss these phases. Further instructions may be found in general textbooks on surveys (e.g.,[549, 603]).

Paper Example 13.0.1 : How do people use Facebook?

The uptake of social media over the past decades has been fast and widespread. Many studies have set out to describe that uptake, either as a way of understanding our use of social media or to improve such media. Lampe et al. [442] used questionnaires to study the development in Facebook use at its early stages 2006–2008.

The study focused on three research questions: 1) How has communication with others on Facebook changed over time? 2) How has perception of the audience on Facebook changed over time? 3) How have attitudes towards Facebook changed over time? The study was *cross-sectional*. That is, the authors sent advertisements to undergraduates at the campus to fill in the questionnaire. They repeated this for three consecutive years. While this does not allow researchers to track changes in individual use, it will enable them to infer it by comparing across years.

To gauge attitudes towards Facebook, the authors included questions on "Facebook is part of my everyday activity" and "I would be sorry if Facebook shut down"; participants rated this on a five-point Likert scale. Such scales typically have labelled endpoints (e.g., "strongly disagree" and "strongly agree") but all five scale points may also have labels (e.g., "neutral" for the middle point).

The table below summarize a few changes found. It shows ratings (in rows) of attitudes, answered on a scale ranging from 1 to 5, across three years. Low numbers mean low agreement, high numbers more agreement. (The data were also subjected to analysis of variance, reported in the column named F. The subscript letters indicate significant post-hoc tests.)

Year of survey		2006		2007		2008	
	F	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Facebook is part of my everyday activity	36.12	3.12	1.26	3.75[1]	1.11	3.85[2]	1.12
Facebook has become part of my daily routine	35.82	2.96	1.32	3.70[1]	1.16	3.66[2]	1.19
I am proud to tell people I am on Facebook	2.90	3.24	0.89	3.40[1]	0.87	3.34	0.85
Facebook is just a fad	12.15	3.14	1.03	2.96	1.09	2.75[1,2]	1.00
I would be sorry if Facebook shut down	5.21	3.45	1.14	3.69[1]	1.19	3.72[2]	1.34
I use Facebook to get useful information	78.51	2.55	1.10	3.39[1]	1.02	3.54[2]	1.00
I use Facebook to find out about things going on at MSU	56.59	2.59	1.08	3.34[1]	1.18	3.51[2]	1.10
My Facebook use has caused me problems	22.51	1.67	0.89	2.14[1]	1.10	2.20[2]	1.12
I spend time on Facebook when I should be doing other things	9.44	3.16	1.15	3.52[1]	1.23	3.54[2]	1.18

Table 6: Ratings of attitudes towards Facebook. A mean reported with a "1" superscript indicates a significant (p<.05 or better) difference with the year before. A "2" indicates a significant difference between 2006 and 2008.

This table suggests a couple of noteworthy findings about the development of attitudes about Facebook. For instance, over the years the site has become increasingly part of users' daily life, as suggested by answers to the first two questions. This is perceived by the respondents as useful (question six), although the site also appears to increasingly cause problems for respondents.

13.1. Design of Survey Research

The design of survey research concerns establishing the focus of the research, selecting an appropriate type of survey, and figuring out whom to give the survey and how.

13.1.1. Research Focus

The focus of the research needs to be clear and spelled out before a questionnaire is selected. Survey research is often appropriate for research goals like the ones listed below. The first set of items are about behaviors or what would later be called *factual questions*; the last set of items is about users' subjective experiences or what will later be called *attitude questions.*

- Research about users' characteristics (e.g., what would often be called marketing research, demographics)
- Research about users' activities, including use of existing interactive systems, how they spend their time.
- Research about users' perceptions of their activities, the technology that surrounds them, and the context they work in.
- Research about users' attitudes towards future technology or future activities.

A second principled way to think about survey research is to consider questionnaires as a *measurement tool* [603]. The idea is that well-designed questionnaires can measure behavior or attitudes precisely and at scale. For instance, we can consider questionnaires a measurement tool for assessing the loyalty of visits to a website. For example, the net promoter score is the percentage of visitors to a website that rate the likelihood of recommending a site 9 or 10 (on a 0 to 10 scale) minus the percentage that scores it 6 or lower [676].

A third principled way to think about survey research is that it can help collect *qualitative data at scale.* . The popularity of this purpose has increased with the use of crowdsourcing. As we will see, there are distinct limitations to the data that may be collected in this way, but such surveys have been used to study attitudes in depth.

13.1.2. Survey Types

Two main types of surveys may be distinguished: (1) descriptive and (2) analytic [603]. *Descriptive surveys* are known from census, election surveys, and surveys on how people spend their time. The main use of these surveys is to describe and characterize attitudes or behavior. Often such questionnaires focus on answering questions on "how many" – for instance, how many times do you experience frustration when using a computer – or "what" – for instance, what content do you enjoy on social media. For example, the GVU Center at Georgia Tech began administering surveys in 1994 on people's behavior on the World Wide Web [645]. The survey documented which browsers the respondents were using, how much time they spent in front of their computer, and their age.

In *analytic surveys*, a deeper attempt is made to explain why a certain situation exist. This type of surveys borrows some of the logic from experiments (see Chapter 43) to answer questions about "why" – for instance, why do people like an app. The idea is to analyze "what goes with what"; that is, to see how different variables are related to each other. Compared to experiments, which can do interventions to expose causal factors, surveys are poorer at exposing causality but better at generating initial ideas about mechanisms and factors. For example, Tuch et al. [811] used surveys to collect narratives about positive and negative experiences with technology. The comparison of those experiences highlighted what differs in users' experiences.

A central characteristic of analytic surveys is the attempt to relate answers to each other. Sometimes the relation is done by giving the same questionnaire to different people, a so-called *cross-sectional design* (see side box on 278). Cross-sectional designs may deal with multiple factors at the same time (similar to a factorial design for an experiment, see 43). For instance, Tuch et al. [811] had four groups of participants. These four groups were obtained by crossing two factors of interest. One factor concerned the difference between positive experiences with technology to negative experiences. The other factor concerned positive wordings of questions about psychological needs versus negative wordings. Together, these factors gives the four groups of respondents (positive experience and positive wording, negative experience and positive wording, negative experience and positive wording, and negative experience and negative wording).

In *longitudinal designs*, respondents fill out the same questionnaire on multiple occasions. This may be before or after an intervention, or it may be done to study the development of attitudes, learning, or expertise over time. For instance, Karapanos et al. [396] administered a questionnaire to six mobile phone users over four weeks. It required that participants at the end of each day listed all the activities that had somehow involved their phone. Based on that, and other data collected, Karapanos et al. [396] formulated a model of their user experience.

13.1.3. Sampling

We can rarely ask all users in a population of interest to fill in a questionnaire. We therefore need select a subset, or *sample*, and base our conclusions on that. To ensure we can generalize our findings to the target population, the sample has to selected carefully.

Sampling refers to decisions relating to how people are approached to respond to a survey (see also the chapter on experimental studies, Chapter 43). This can be done in multiple ways: via personal contact, email lists, advertisements, social networking groups, and so on.

The goal of sampling is to acquire a sample (of respondents) that allows the researcher draw valid conclusions about the target population. All actual or potential users are called the *population*. We are able to contact a portion of the population, the *sampling frame*. We then get our questionnaire to all, most or some of the sampling frame, which is the actual *sample*. And among those users, a portion chose to fill out the questionnaire, the *respondents*. Each of these steps require care on part of the researcher, because excluding users unwittingly impact the validity of the results.

We may distinguish two principled ways of thinking about the sampling frame: randomization and stratification. First, sampling may be done at random from the population; this may be facilitated by panels of representative users from the population or by having your questionnaire distributed by opinion pollsters. Sampling may also be designed to ensure that certain types of users are represented equally often – often referred to as quota sampling, cluster sampling, or stratification.

In *snowball sampling*, each respondent shares the invitation to other participants in their network or with similar characteristics; that way, the sample of respondents grows as a snowball rolling downhill. However, this may introduce unwanted (out-of-quota) characteristics. The reality of survey research in HCI is that many samples frames are informal, being established through word-of-mouth, through one's network, and posting to online fora. Such so-called convenience sampling does not ensure that every member of the population is equally likely to be part of a study and is therefore making valid inferences difficult.

The question of which *sample size* to use in questionnaire research is difficult to answer. The practices in HCI research vary greatly. Published survey research in HCI can have as few respondents as six [396] or as many as 50,000 [226] (see the side box on page 284). A review of sample sizes concluded that the mean number of respondents in surveys in HCI papers is 371 [118].

The choice of sample size may also be informed by statistics. In descriptive surveys, the appropriate sample size depends on *the margin of error* you are willing to accept in quantitative measures. For instance, if you accept a 10% margin of error in estimates of which browser people use, you need ask more people, for example hypothetically 100 people; however, if you accept only a 1% margin of error you need ask much more, for example 700. In analytic surveys, the sample size may be determined by *power analysis*. Power analysis is a class of methods that can be used to estimate the number of respondents needed given the analyses that will be carried out (see also Chapter 43).

13.2. Selecting a Questionnaire

Survey research may be done using existing questionnaires or by developing one's own. Many questionnaires have been carefully validated and are known to be reliable. Designing ones own questionnaire from scratch is time-consuming and difficult, even for experiences researchers. However, uch research in HCI does not use existing questionnaires. Hornbæk and Law [349] surveyed usability studies for which the full data were available. For the 36 studies where questionnaire data were available, 16 (44%) used existing questionnaires. The remained developed their own. Those questionnaires had, on average, lower reliability than existing ones, and six of them failed to reach commonly accepted criteria for minimum reliability (Cronbach's α of .70 or above). This survey confirms the value of using existing questionnaires.

Questionnaire	Purpose
Nasa Task Load Index (TLX)	Measure the experienced workload when using an interactive system using six items, including mental demands, physical demand, and effort [311].
Godspeed questionnaire	Measures five distinct components of users' perception of social robots, including their anthropomorphy, animacy, and perceived intelligence [49].
User Engagement	O'Brien and Toms [591] presented a questionnaire that assess six factors in being engaged with interactive systems, including focused attention, felt involvement, and novelty.
Player Experience	The Player Experience inventory helps measure payer expereince as functional consequences (e.g., audiovisual control) and psychosocial consequences (e.g., mastery) [4].
Single-item post-task usability questionnaire	Sauro and Dumas [712] showed that questionnaires can be short by offering a single-item questionnaire for usability after a task: "Overall this task was", answered on a seven-item scale going from "very easy" to "very difficult".

Table 13.1.: Examples of questionnaires that are often used in user research.

13.2.1. Finding Appropriate Questionnaires

How to find existing questionnaires so that they can be used in user research? The key step is to identify if the constructs of interest (e.g., perceptions of privacy, feelings of satisfaction, awareness of a group) have been developed into questionnaires. Table 13.1 shows examples of existing questionnaires that are often used within HCI. Previous research on a particular topic and HCI handbooks are also great sources of existing questionnaires. For example, Sauro and Lewis [714] lists 24 standard questionnaires to assess the perception of usability and user experience. Outside of HCI, Such questionnaires may be found in handbooks [e.g.,], on dedicated search engines for questionnaires [448], and by searching scholarly databases (e.g., ERIC, Google Scholar, arXiv, PsychAbstracts) for the construct of interest.

13.3. Developing Questions

Our advise is to use validated questionnaires when possible; the situations in which there is nothing available are few. Developing a questionnaire involves careful creation, validation and testing of questions. This section covers some basic ideas for the cases where no existing questionnaire is available. It also illustrates some of the reasons why developing valid and reliable questions are difficult.

Below we distinguish two kinds of questions [718]. First, some questions are about *events and behaviors.* They concern how often something is done, which types of activities that were done yesterday, frequency of use of a system, and so on. For instance, we might be interested in how frequently people use a particular website. These questions are formulated such that they in principle have a correct answer – even if the respondent may not be able to recall it. Second, some questions are about *subjective phenomena*, including attitudes, intentions, wants, and so. For instance, we might be interested in how somebody values virtual possessions, like skins in a game, relative to physical possessions. Those questions include another set of difficulties and particular biases in answering.

13.3.1. Wording of Questions

The wording of questions may, if done sloppily, lead to answers that are invalid and unreliable. For example, let us assume that we wish to design a question that probes which piece of software people use the most. Let us start with considering the question "Which software do you use and engage with the most". What do we mean by "software"? There is software in cars, hearing aids, and washing machines—do we want answers all of those devices? Would it be better to spell this out (e.g., "software such as apps on a phone or tablet, or programs on a computer"), or limit the question to be about a subset of devices, as in "computer (e.g., laptop, tablet, smartphone)"? If we specify it as "your computer", one limitation there is that most frequent use might not occur on a computer that one owns, but rather on a work computer or a borrowed computer. And what about "use"? And what if respondents consider it different from "engage with"? And what about "the most"? It may be interpreted as the most times or the longest duration; which one should respondents answer? As another example, consider the study in the box on how to ask about how often people use social media.

These examples suggest some general principles for thinking about how to craft valid and reliable questions.

Use language that respondents understand. Questions should be user centered, that is, use terminology that users are familiar with and that they find clear and straightforward. We should avoid hard-to-parse questions, including those using difficult words (e.g., ""), double negatives (e.g., ""), and excess words.

Ask one question at a time. "How satisfied are you with the system and support for it" does not work because it is possible for respondents to be simultaneously satisfied with the system but not with the support. Such questions are called double-barreled and should be avoided.

Ask neutral questions. Questions should be formulated so as to not impose the researcher's views or ideas about answers on the respondent. Avoid leading questions such as "do you agree that the web site is usable"; instead use something like "do you agree or disagree that the web site is usable". Also, avoid making make assumptions about the respondents: "What is the brand of your laptop?" assumes that the respondent own a laptop which might not be the case.

Ask specific questions. Questions should be specific, so that the respondent is clear about what is being asked. For instance, "How does this product compare to others" does not specify on what aspects the respondent is to compare products (e.g., prize, functionality, usability). Specificity is particularly important for objects and events. For instance, we already saw some difficulties around the specificity of the word ”computer”. Further, specificity may also be about words and phrases that are assessed relative to a personal standard. For instance, ”frequently”, ”rarely”, ”expensive”, and so on may vary much for different respondents. Even something like "last week" may be considered unclear: is it the last seven days or the week before the current one?

Be aware of response biases. It is well known that participants have biases in answering questions. Effective questions are worded so that they mitigate those biases to the greatest possible extent. For instance, social desirability bias means that respondents may provide answers that they think are more socially acceptable or prestigious than others. We report more bias types below.

Paper Example 13.3.1 : Formulating questions to learn about time usage

Designing questions that participants understand and answer validly is hard. Ernala et al. [226] wanted to find questions that would help them get valid and reliable answers on how long time people spend on social media. They administered 10 ways of asking about this to about 50.000 respondents. For instance, one question asked“In the past week, on average, approximately how much time have you spent actively using Facebook”, another “How many hours a day, if any, do you typically spend using Facebook?”. The research additionally collected data on how much time participants *actually* used Facebook. In that way, they could account for differences between recorded and actual time usage, as well as investigate how the demographics of their respondents influenced these differences.

They found that the formulation with least error was the following:

```
In the past week, on average, approximately how much time PER DAY
have you spent actively using Facebook?
    Less than 10 minutes per day
    10-30 minutes per day
    31-60 minutes per day
    1-2 hours per day
    2-3 hours per day
    More than 3 hours per day
```

The study also showed that self-reports were actually only moderately correlated by the actual time people spent on Facebook.

13.3.2. Question Types

Two main types of question types exist. With *closed questions*, respondents are offered a choice between different answers. Closed questions may have many different forms, including the following:

- Single-choice questions offer a list of alternatives, of which one may be selected.
- Multiple-choice offer a list of alternatives, of which one one or more be selected.
- Rankings allow respondents to rank order alternatives.
- Ratings offer a textual or graphical scale that the answer is given on. Those scales may have different number of steps and different anchors (i.e., description of the end points).

As homework, we recommend trying to formulate the same question in these four different ways.

The benefits of closed questions are that they are easy to analyze: Because options for answering are restricted, statistics are straightforward to apply to the answers. The drawback is that such questions may be too crude and not capture all the relevant options for respondents. While generally recommended, the availability of options like "I don't know", "non-applicable", or "other" do not remove this issue. This may work if the options are derived from earlier research or pilot work; Oppenheim [603] suggested that all closed questions should start their lives as open ones.

The other main type of question is *open questions*, also called free-response questions or free-text questions. Open-ended questions are useful when it is not clear what respondents want to answer. This is for instance the case in exploratory work where no closed questions are yet known. Also, open questions allow respondents to answer freely and spontaneously. Often questions about sensitive or socially disapproved behaviours and frequency or forms of behaviour are better reported with open questions. However, open questions are "easy to ask, difficult to answer, and still more difficult to analyse" [603, p.113].

The decision between open-ended and closed questions is (again) a tradeoff between the objectives of the research and the pros and cons of different types of questions just outlined.

13.3.3. Presentation and Order of Questions; Instructions

Questionnaires were traditionally presented over a telephone or on a paper, but most surveys are now done online. Online surveys are easier to deliver to a large number of respondents. However, there are two main drawbacks. First, online surveys require a digital device. Those who do not have access to a digital device are excluded. This can lead to bias in sample. Second, computer terminals differ in the way they present the survey, creating variability in responses.

A questionnaire should start with overall instructions to respondents. This should involve a broad context of the study. It should also instruct the respondents on how to interpret and respond to the questions. Instructions are sometimes shown one by one to ensure they are not glossed over.

The order in which questions are presented matters. *Funneling* means that questions start with broad questions and move toward specific. The argument is that broad questions help jog respondents' memory on the topic of the questionnaire.

Surveys must be designed for visual clarity and ease-of-use. The response options should be quick to choose with the terminals respondents use. For example, dropdown lists are more time-consuming on touchscreen devices, while some touchscreen devices do not support sliding.

13.3.4. Sets of Questions and Scales

One of the key issues in survey research is reliability. Unfortunately, single item questions can be unreliable, in particular when asked about subjective experiences. More generally, using single questions to uncover an underlying attitude or experience is generally risky: A particular question may be vague for a respondent. They may also misclick the answer option.

The response to this challenge is to use sets of questions, sometimes called scales. A scale consists of items, which are intended to be about the same underlying construct (or constructs, if the scale is made up of multiple items). Some popular scales used in HCI research include:

- Likert scale, which consists of a claim and a symmetric negative–to–positive range with the option ”Neutral” in the center (1-5, 1-7, or 1-9).
- Semantic differentials, which posits a scale between two adjectives (e.g., ”exhausting” and ”effortless”). Semantic differentials are often used for measuring attitudes and opinions.
- Graphical ratings, in which several options are presented in a symbolic fashion (e.g.,the pain measurement scale).

13.3.5. Biases in Answering Questions

A question should be thought of as a measurement. It probes the respondent and records a response. The validity of this measure depends on a number of things that we have discussed so far. There are many biases that threaten validity, such as

Acquiescence bias Tendency to respond in a positive manner.

Social desirability Tendency to mirror socially desirable behaviors and views. For example, if a question asks to report if the respondent takes back ups, some proportion of responses might overstate this.

Response order bias First and last parts of a questionnaire see different behaviors. Especially long questionnaires see effects of fatiguing.

Neutral item bias Items in the middle of the scale (e.g,. "Neutral", "I do not know") tend to be overused. By contrast, the extrema (lowest and highest values) tend to be underused.

Demand bias Users who know about the purpose of the study may start behaving according to its expectations. For example, if you tell participants that you are going to measure their knowledge of phishing attacks, they may pay more attention to those during the study.

There are many ways these biases can be mitigated. We look into these tactics in the exercises.

13.4. Questionnaire Piloting

All questionnaires should be piloted, in particular if they are modified from existing questionnaires. This includes not only the questions but also the instructions for the questionnaire.

One overarching principle for creating good questions is to consider the experience of answering them; this principle has direct implications for how to pilot questionnaires. In Chapter 42, we discuss think-aloud studies. One adaptation of such studies is to use them to get insight into how respondents of questionnaires experience the questions and the instructions. With such insights, the questionnaires may be adapted based on discrepancies between the idea of the question and participants' experience.

13.5. Data Collection

Data collection in survey research is related to how questionnaires are distributed to respondents and how and when respondents complete the questionnaire.

13.5.1. Distribution of Surveys

Traditionally, questionnaires were distributed to respondents on paper or by telephone, with a researcher reading the questions aloud to the respondents. Presently, questionnaires are most commonly answered online using a browser or mobile app. These questionnaires are typically answered once, say, after using a new interactive system or at the time when people elect to answer the survey.

HCI researchers have modified these forms to suit the purpose of their studies. Palen and Salzman [619], for instance, allowed participants to leave open-ended and structured answers on voice-mail. This allowed them to collect data from participants on the go. Schwind et al. [728] investigated if it impact survey research whether participants answer questionnaires in virtual reality rather than in the real world. Their data suggested that it does not and that researchers can validly use questionnaires in virtual reality.

Within the past decade, *crowdsourcing* has allowed tasks that humans are good at but computers bad at to be distributed to people at scale [666, 411]. The idea is that tasks are broken down to be small in scale and distributed over the internet to people. These people may do the tasks for payment. These are often small fees, so-called micropayments, because the tasks are short. Crowdsourcing has, in particular, impacted how survey research is conducted.

With crowdwork, including surveys, a key concern is that respondents pay attention and produce high-quality answers. If participants are just racing through a survey – sometimes called satisficing – the data will not be valid and such answers may dilute the overall data set. There are several ways of avoiding this. Researchers have included *attention checks* in surveys, that is, questions with obvious answers—if respondents fail to answer them correctly, their data may be excluded. Similarly, in *instructional manipulation checks* [604], respondents are instructed in text to do something while an seemingly obvious next step is presented elsewhere (e.g., a continue button or a rating scale). Thus, participants who do not follow the instructions may be excluded.

13.5.2. Compensation and Treatment of Participants

Survey participants must be treated ethically, as should all participants in user research, see Part III. The purposes of the study should be told, as well as the way the data is used. Respondents should be offered a non-punitive way to stop participating in the study if they want to stop. In the cases respondents are compensated, the compensation should be done in an ethical manner. That means that participants should be payed at least minimum wage for their work, or otherwise appropriately compensated for their effort.

13.5.3. Improving response rates during study

If the response rate is low during study, it may be improved with different techniques: 1) doing advertisements in not one but several campaigns, 2) improving the usability of the questionnaire, 3) asking a known person to approach members of the community personally, and 4) sending personal reminders.

However, the caveat of such techniques is that they may bias the sample away from the intended sample. Such biases can be subtle. For example, if you call people personally to fill in your survey, it may skew respondents toward those whose contact number you have available.

13.6. Analysis

Before starting to analyze the obtained dataset, the original research questions should be revisited. They should be formulated in a way that can be answered via computations done on the dataset. A typical process from there on is that the response data are cleaned, it is checked that they are reliable and valid, and they are analyzed statistically or qualitatively in relation to the aims of the research. In this book, we cover five main steps with a focus on conceptual understanding:

1. Preprocessing the dataset for subsequent computations: cleaning, imputing missing data if neeeded, and recoding.

2. Assessing reliability: computing response rates and internal validity
3. Descriptive analysis: Describing tendencies and relations of interest among variables
4. Explorative analysis: Exploring emergent tendencies not written up as research questions
5. Testing hypotheses (covered in Chapter 43)
6. Drawing valid conclusions and reporting.

13.6.1. Preprocessing

To be used for further analyses, response data need to be cleaned and and variables prepared for analyses. *Cleaning* means removing those answers that are invalid. For example, some may have failed attention checks, or they may have answered using the same response to all questions, or they have copy-pasted or otherwise produced irrelevant data. Participants who are inconsistent with their answers may also be removed. When some answers are removed, the analyst faces a decision. Should one remove a participant entirely – so-called list-wise deletion, or just the problematic answers – so-called pair-wise deletion? One should be careful with cleaning, though. If a large proportion of the data needs to be removed, the questionnaire is probably poorly designed. There is no clear-cut criterion for deciding what is 'too large' a proportion. A good rule of thumb is that if the removed proportion is larger than 5%, cleaning may start to adversely affect the validity of conclusions.

Some data may be missing, for example due to technical reasons (browser not storing user responses), or due to inattentive respondents skipping questions. Such cases may be removed, however, again, the proportion of removed data should not be too large. *Data imputation* refers to statistical techniques of replacing missing data with estimates computed based on the rest of the data. For example, averages (of all respondents) can be used, in cases where there is not a lot of data missing. If there is, more sophisticated statistical techniques (e.g., MICE [822]) can be used. They try to form a better model of what a missing value might have been, given the rest of the data matrix.

After cleaning and imputation, some response data may need to be *recoded*. Items of a scale may be summed to compute an index score. For example, responses to the NASA-TLX questionnaire, which consists of several components of workload, are summed to compute an aggregate index of workload. The type of recoding depends on the analyses that are done later on. However, one must remember that recoding may lose information.

Open-ended questions in questionnaires may provide extensive qualitative data. The techniques for analyzing data from open-ended questionnaires are the same as those covered earlier for interview data, for instance thematic analysis (see Chapter 11), and for archival data, for instance content analysis (see Chapter 14).

13.6.2. Assessing reliability

Response rates

In data collection, *the response rate* refers to the number of participants in our sampling frame that completes the questionnaire. It is a good practice to compute response rates, when it is possible to know how many have seen the survey invitation. For example, if you send an email to a list with 1,000 employees, and 100 take the survey, the rate is 10 %.

In larger-scale surveys, it is not uncommon for response rates to be low. However, as pointed out by Oppenheim [603], the key issues is not the rate of responses or proportion of non-reponses but "the possiblity of bias" (p. 106). The question one should ask is *how* do non-respondents differ from respondents. For example, respondents can be more educated (or less) or younger (or older) than the population of interest. This will unavoidably affect the generalizability of results. To address such issues, it is a good practice to gauge how well the respondents' characteristics match that of the population of interest. For example, one can compute descriptive statistics (e.g., mean, median, standard deviation) of background factors like age, education etc. These can be compared to known characteristics of the target population.

Internal reliability

Testing the reliability of questionnaires is relevant for scales meant to measure some broader construct using multiple questions. For such scales, we would like the answers to vary in the same way. For example, if a user answers 'low' in response to an item on "is the software useful", they should also answer 'low' in response to an item on "does the software help you do things that are important to you". If so, we say that the answers are internally reliable.

There are many ways to check the reliability of scales. One is the simple correlation coefficient between items: If they are reliable, we expect them to correlate strongly. Another is the commonly used Cronbach's alpha:

$$\alpha = \frac{N * \bar{c}}{\bar{v} + (N-1) * \bar{c}} \tag{13.1}$$

where N is the number of items, $\bar{c}$ is average inter-item covariance among the items, and $\bar{v}$ is average variance. It quantifies internal consistency of a test between zero and one. A high value of Cronbach's alpha is expected for scales. One rule-of-thumb for an value is .7.

Cronbach's alpha can tell us whether to trust the result of a survey. For instance, when Hirzle et al. [338] explored eye strain with virtual headsets they used several earlier questionnaires, including one on digital eye strain, comprising four subscales. They then analyzed Cronbach's alpha to understand the consistency of each of the symptoms. "We found acceptable values for Mall ($\alpha = 0.77$) and M_{ex} ($\alpha = 0.77$), but a slightly low values for M_{in} ($\alpha = 0.67$), and a poor value for M_{vr} ($\alpha = 0.51$). Therefore, the results for $M_v r$

must be considered with caution". Thus, in this case, the items in the subscale M_vr did not correlate as strongly as expected with each other.

O'Brien and Toms [591] developed a new questionnaire for measuring user engagement. They initially developed 11 subscales of engagement and checked that alpha was satisfactory. They further used alpha to identify questions that were not relevant in the further development of the questionnaire because they, if deleted, did not change alpha out of its optimal level (between .7 and .9). Another example (QUIS) is given in the side box.

If internal reliability is low, one can try to remove items that have low reliability. If this is not possible, then the questionnaire may need to be redesigned. Note that reliability is only relevant for questions that purport to measure the same thing. If items are about different things, we should not expect them to score similarly.

Paper Example 13.6.1 : Evaluating a user satisfaction questionnaire

Chin et al. [150] reported an evaluation of a questionnaire called QUIS to measure user satisfaction called QUIS. The questionnaire, at the time of the evaluation in its fifth version, comprised 27 rating scales from 0 to 9. It included questions like "Overall Reactions to the software" on a scale going from "terrible" (0) to "wonderful" (9), and "Learning to operate the system" on a scale going from "difficult" (0) to "easy" (9). To assess the validity of the scale, the authors asked respondents to fill in the questionnaire about a product they liked and a product they did not like. The idea here is that the questionnaire should show differences among those two to be considered valid.

The results show that there was indeed significant differences between the liked and disliked products. This serves as a first, rough indicator of validity because the questionnaire can identify the difference between a liked and a disliked product. Moreover, the reliability of the questionnaire was found adequate. Cronbach's α was .94 suggesting that questions that were about the same aspect of satisfaction were answered in the same way.

QUIS may be used by researchers that want to assess the satisfaction of a user interface. There would be no need for them to construct a questionnaire on their own because it is laborious to ensure validity and reliability at the level of the study [150].

13.6.3. Descriptive analysis

Descriptive analysis refers to the study of distributions of variables in the dataset. This can be done using any of the widely available means in statistical tools: histograms, scatterplots, bar charts, line charts etc. It is a good practice to start by visualizing univariate distributions (i.e., how single items are responded to), for example using histograms. This helps spot potential issues like skewness (tendency for a distribution to 'lean' toward one or other end and multimodality (several peaks). Another good practice is *crosstabulation*. It means computing tables that report how two or more response variables are related. Another good practice is to plot *error bars*, such as *confidence levels* to charts. For example, the 95 % confidence interval depict the confidence with which the true mean lies within the interval. It helps understand degree of variability at different levels of responses. This protects against overinterpreting a plot when in fact the underlying data is highly varied.

13.6.4. Exploratory analysis

Besides visualizing data in different ways, *factor analysis* is a form of statistical analysis that helps reveal the underlying dimensions of a set of variables. Those underlying dimensions are called factors, or sometimes latent variables because they cannot be directly obtained by respondents' answers but must be inferred from them. Factor analysis is helpful when a questionnaire has many items; in such situations, it helps determine which items that are most related to each other. For instance, if items are about how frequent you use an interactive system and how long time a session with the system typically last would expect to be able to identify a factor related to frequency (which likely consists of several items, i.e., questions about frequency) and one related to session time. If a factor analysis do not identify those two factors, then something seems off in our questions or how respondents have answered. Factor analysis is mathematically complex and several variants exists;] contains an introduction to the principles, calculations, and software packages involved.

Factor analysis has several uses in survey research. Unlike reliability analysis, factor analysis typically concerns the relationship among all variables of a survey. The side box shows a detailed example.

Paper Example 13.6.2 : Factor analysis in surveys

Law et al. [452] report a study of the Game Experience Questionnaire (GEQ). This questionnaire is widely used in evaluating games. After a review of existing work using the GEQ, the authors did their own survey study using the GEQ to validate the scale.

A total of 633 participants described a recent experience happened recently. They then answered the GEQ relative to that experience, answering the 33 questions on a five-point Likert scale, with anchors Not at all (0) and Extremely (4).

These data were subjected to a so-called confirmatory factor analysis. This is a variant of factor analysis described in text. Instead of trying to find out what the factors are, the goal here is to assess how good some *given* factors are. The authors' analysis showed that two of the seven scales of the GEQ were not strong enough (i.e., below .7). This finding means that the proposed scales of the GEQ are not independent, as expected in the literature.

	Component	MR2	MR1	MR5	MR3	MR4	MR6	MR7	h2
15 I was good at it	Competence	-.030	.060	-.026	**.798**	-.006	-.111	.030	.739
02 I felt skillful	Competence	-.043	.058	.063	**.701**	-.019	.240	-.051	.451
17 I felt successful	Competence	-.073	.096	.035	**.604**	.102	.011	-.092	.697
21 I was fast at reaching the game's targets	Competence	.114	.042	.122	**.593**	.005	-.090	.067	.399
10 I felt competent	Competence	.016	.170	.105	**.491**	.076	-.042	.115	.650
19 I felt that I could explore things	Immersion	-.025	-.074	**.751**	.004	.030	-.065	.032	.697
03 I was interested in the game's story	Immersion	.002	.050	**.722**	-.037	-.045	-.049	-.017	.668
18 I felt imaginative	Immersion	-.037	-.027	**.680**	.023	.068	.080	-.107	.426
27 I found it impressive	Immersion	.017	.177	**.613**	.025	.015	.111	.005	.546
30 It felt like a rich experience	Immersion	.037	.152	**.489**	.070	.136	.088	.115	.518
12 It was aesthetically pleasing	Immersion	-.050	.269	**.337**	.094	.009	.018	.071	.521
31 I lost connection with the outside world	Flow	.069	.047	-.010	-.070	**.850**	-.077	-.023	.529
13 I forgot everything around me	Flow	-.086	-.128	.065	.110	**.724**	.090	-.042	.631
25 I lost track of time	Flow	.068	.086	.033	-.037	**.696**	-.023	.004	.598
05 I was fully occupied with the game	Flow	.018	.069	.101	.150	**.451**	.140	**.315**	.685
28 I was deeply concentrated in the game	Flow	-.008	.066	.159	.171	**.345**	.240	.282	.639
24 I felt irritable	Tension	**.806**	.013	-.012	-.014	.054	-.011	-.009	.595
22 I felt annoyed	Tension	**.800**	-.004	-.007	-.071	.004	.048	-.015	.693
29 I felt frustrated	Tension	**.656**	.089	-.126	-.131	.035	.275	-.013	.780
23 I felt pressured	Tension	**.393**	-.111	-.044	.104	.088	**.339**	.046	.597
32 I felt time pressure	Challenge	**.338**	-.040	-.007	.119	.000	.253	.041	.283
11 I thought it was hard	Challenge	.116	.005	.005	-.151	.001	**.679**	-.088	.583
26 I felt challenged	Challenge	-.032	.134	.049	.081	.059	**.661**	.085	.658
33 I had to put a lot of effort in to it	Challenge	.107	-.066	.137	.109	.034	**.563**	.094	.253
07 It gave me a bad mood	Negative Affect	**.782**	-.074	.009	.059	.012	-.078	.075	.470
09 I found it tiresome	Negative Affect	**.505**	-.240	.113	.063	-.043	.013	-.114	.692
16 I felt bored	Negative Affect	**.455**	-.213	.015	.089	-.043	-.208	-.212	.732
08 I thought about other things	Negative Affect	.278	.011	.102	.066	-.241	-.118	**-.327**	.467
06 I felt happy	Positive Affect	-.011	**.734**	.096	.097	.049	.049	-.184	.401
04 I thought it was fun	Positive Affect	-.041	**.703**	.111	- .006	-.012	.003	.273	.517
20 I enjoyed it	Positive Affect	-.073	**.653**	.095	.073	.020	.026	.184	.499
14 I felt good	Positive Affect	-.090	**.607**	.018	.213	.110	.049	-.097	.407
01 I felt content	Positive Affect	-.088	**.575**	.003	.217	.089	-.017	-.115	.596
After rotation Sums of Squares		3.17	3.70	3.03	2.90	2.61	2.07	1.44	
% of variance explained		9.7	11.2	9.2	8.8	7.9	6.3	4.4	

Table 5. Rotated pattern matrix of the EFA with 33 items loading on seven factors.

	ω	95% CI	Cronbach's α	AVE	MSV
Immersion	.85	[.83,.87]	.85	.48	.54
Flow	.86	[.84,.88]	.86	.54	.51
Competence	.86	[.83,.86]	.85	.54	.60
Tension	.82	[.79,.85]	.82	.55	.75
Challenge	.71	[.67,.74]	.57	.38	.42
Positive affect	.91	[.89,.92]	.91	.66	.60
Negative affect	.69	[.64,.74]	.68	.31	.75

Table 7. GEQ-33 reliability analysis. Reliability coefficient ω with bias-corrected and accelerated bootstrap (1000 iterations) 95%-confidence intervals as implemented in [35].

13.6.5. Drawing conclusions

It is easy to report plots and statistical results based on a collected dataset. It is much harder to draw solid conclusions about the population of interest. This requires consideration of different types of validity:

- *Construct validity* refers to the selection of the questionnaire: does it capture what we intended to capture? For example, if we want to draw conclusions about user engagement but used a usability questionnaire, we should sustain from making conclusions about engagement.

- *Internal validity* refers to the strength of statistical relationship between variables. Even in nominally strong relationships, there can be confounding factors we have failed to account for. For example, if one finds a relationship between the size of a mobile device and typing speed, is that because large displays are better for typing or because users who are wealthier and buy larger phones are also older and therefore more experienced?

- *Statistical conclusion validity*: All data comes with noise and variance. One should take into account the statistical confidence with which the claimed conclusions can be drawn.

- *External validity* refers to ability to extend conclusions from a sample to the population of interest. If the sample poorly represents the population, external validity is threatened.

Besides statistical analysis, it is important to seek to explain the findings. What do extant theories say about them? A compelling explanation increases confidence on results. Results should also be compared against those from previous studies on the same topic or group: are they similar or dissimilar? What might explain such discrepancies?

13.6.6. Reporting

Finally, the analyses that were carried out should be reported. It is important to report all analyses. They are as much part of survey method as the questionnaire. A good practice is to store the scripts or formula that were used and release them as part of the study report.

There is an important ethical dimension to conclusions, too. The publication of findings should not harm respondents, for example by disclosing facts that can endanger their jobs or cause discrimination.

13.7. Situating Questionnaires

Survey studies are inexpensive to conduct; one only needs a computer or a paper handout and a pen. After a survey has been designed, the researcher is not needed for administering it, only for distributing it. This stands in contrast to methods that depend on the presence of the researcher, such as the interview method. Sample sizes tend to be much larger in survey studies than interview studies.

However, the de-situatedness of surveys poses a challenge for user research. The moment when a respondent fills in a questionnaire is detached, in time or space, from the moments that these surveys are about. Respondents need to rely on memory – or their imagination, to close this gap. The question is how to re-situate questionnaires?

With *diaries*, this gap can be decreased albeit not closed. A diary consists of questionnaires that are filled in periodically, for example every day or week. This can be done in many ways: for example on a mobile device or via phone calls. The study by Palen and Salzman [619] discussed above used phone diaries. They had participants call a phone service, answering structured questions like "Please tell us how you have used your phone service since the last time you called us, as well as any problems you have had. Also, describe any changes you may have made to your phone or service". Respondents could call when they wanted and were payed per call. While diaries allow responding more flexibly, it still depends on the user *when* they want to make entries. Some may want to fill them in close to the pertinent events, others may not. This approach does not therefore solve the issue of de-situatedness.

Experience sampling is an attempt to bring surveys even closer to the contexts they refer to. Experience sampling refers to computer-controlled cuing of surveys. A notification pops up on the mobile device, or a message is sent in a messaging application. These invite the respondent to fill in the survey within some time window. For example, an app can follow when the user enters office, and trigger the survey at that moment, asking the user to fill it in within, say, 5 minutes. These surveys, however, need to be very brief. They could consist of just one question. More details can be found in [200].

Summary

- Surveys help collect information from many respondents; their strength is generalizability.
- Since respondents of surveys cannot ask for clarifications of questions, they need to be clear, unambigious, and relavant.
- Questionnaires should be reliable and have a validated factor structure.
- Diaries and the experience sampling method attempt to bring surveys closer to the moments of interaction.

Exercises

1. Alternative formulations. Re-formulate the following question in four different ways (single-choice, multiple choice, rankings, ratings): ”If I suspect that an email is a phishing email, I do not even open it.”
2. Learning from surveys. Go to a crowdsourcing website, such as Amazon Mechanical Turk or Prolific. Therein, find a questionnaire and complete it. Assess the questions: 1) Does it use easily understood language? 2) Does it ask one question at a time? 3) Are the questions neutral? 4) Are the questions specific? 5) Are there potentials for response biases.
3. Survey as a research method. Consider the following situations. For which of them would surveys be an appropriate research method? Why? (a) Understanding how workers in a microbrewery use a mobile application for coordinating their work; (b) Charting general expectations toward augmented reality applications; (c) Charting user needs for a new version of a product.
4. Getting rid of biases. Consider the following questionnaire question: ”How good is the present version of the application you use? Rate: 1 - Not good at all; 2 - Bad; 3 - Neutral; 4 - Good; 5 - The best”. What are some potential biases in this formulation? How would you rephrase the question and the scale?

5. Designing a question. This exercise requires you to design one question that asks the respondent to ask about events or behaviors in computer use. You may come up with your own target event or behavior, or you may attempt to identify how many computers a person uses (imagine a context where we want to characterize how people link or do not link their different computers). Design the question. Go through it based on the chapter. Think about sources of bias, issues of validity, concerns about reliability.

6. Data analysis. Following the guidance given in this chapter, analyze qualitative data from this survey. Data may be found at `https://osf.io/m3fbk/`.

14. Unobtrusive Research

Typical user research affects, or changes, the activities users engage in. In other words, it *intervenes*. For example, consider carrying out user research on how people use passwords. During the course of the study, without you knowing, your participants may changed their behavior or opinions. The fact that they selected themselves to the study, or the way they were informed about the purposes of their study, may have made them more sensitive to it. They may also want to act according to what they believe is socially desirable. To sum up, taking part in a study may affect the behavior of participants. *Reactivity*—or the impact of research on what is being studied—may fundamentally change what is discovered and therefore potentially undermining the validity of the research.

Unobtrusive research is a form of non-interventional user research. It aims to not affect or change the activities that users engage in. Sometimes unobtrusive research is called non-reactive [858]. This term emphasizes that the measures that are obtained are not affected in the same manner as data from observations or interviews. Contrary to reactive methods in user research, in unobtrusive research, the data may be generated even before the research activity has begun. However, it may still be used to characterize users, their activities, and their opinions.

Unobtrusive research uses *traces* of users' behavior or archival records to make inferences about users and their activities. In HCI, such data may come from, for example:

- application logs of activities in a system, such as what users click on a web site (click data). For example, Huberman et al. [355] analyzed the distributions of clicks on web sites, finding statistical tendencies that were explained using the information foraging theory.

- records of posts and comments written in a social media application. For example, Chandrasekharan et al. [143] studied messages in the online community Reddit, using both computational and qualitative methods to analyze the posts. They reported on norms that users uphold. They distinguished between "macro norms" that are shared by most Reddit users, meso norms that are shared among some groups, and micro norms that are specific to subreddits (discussion groups).

- video recordings of a site of interest. For example, Brown and Laurier [100] analyzed publicly available YouTube videos showing incidents or near-by-incidents involving self-driving cars. The authors analyzed how drivers interpret the intentions of other drivers and the car.

Realism is the defining aim of unobtrusive research. It allows us to study interaction with interactive systems as it happens in the wild, without reactivity to data collection and without sensitization of users to what they perceive to be the goal of data collection. Thereby, such research in general helps access people's actual behavior, rather than what they chose to report (see also the say/do problem discussed in Chapter 11).

Unobtrusive research has other benefits. Often it is inexpensive, allowing researchers to collect large amounts of data with little effort. For instance, in the boxed example of web revisitation patterns, data were collected from 612 000 users [11]. Besides not affecting users' attitudes and behavior, unobtrusive research does not disturb users or require them to take time out to be part of interviews or have researchers shadow their work. For instance, [26] used YouTube videos to research how people with motor impairments interacting with touch-enabled mobile phones. That way, they could collect much more intensive data spanning many more situations of use compared to previous work.

In a classic text on unobtrusive measures, Webb et al. [858], identified four sources of nonreactive data:

Traces are either obtained by logfile analysis, for instance from web logs, or from instrumenting people, things or places. Traces can be divided into two classes: direct and indirect.

Direct traces are recordings that are caused by users' actions. For example, mouse movements and clicks can be logged by an operating system and stored to a file.

Indirect traces are caused only indirectly by users via some intermediary mechanism. For example, think about a numpad used to enter passwords. Wear and tear on a physical object tell something about the users and their patterns, but are not direct recordings of any instance of an action. There is no need to elicit user needs about routes in a park through interviews if a path has been trodden that reveals how users want to cross.

Archive data may be of many kinds, including social media content. Next, we discuss these forms.

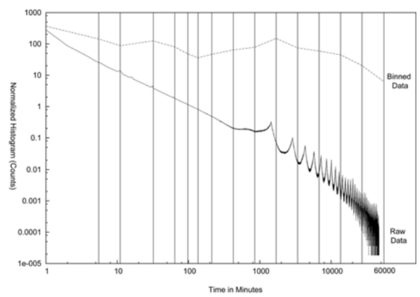
We continue next by discussing the analysis of unobtrusively collected data.

Paper Example 14.0.1 :

Revisitation patterns in web browsing

One of the most prominent applications of unobtrusive research in HCI has been using automatically logged information from computer systems to make inferences about what people do or want to do. Behavior on the WWW is one example of data that may be inexpensively logged at scale to allow researchers understand how users browse the www and how to support that activity. One example of such research focused on how users return to web pages [11]. It is well known that users often return to the same pages, but at that time this was mainly known from studies with few participants.

In contrast, Adar et al. [11] used data from users that had opted in to using a tool bar for web search. Each tool bar was associated with a unique identifier and all www-pages visited through the tool bar was collected for five weeks. This data were thoroughly cleaned, extreme users were removed, and the data grouped according to the duration between revisits to individual www-pages and types of pages. Below is show one resulting insight: Revisitation patterns vary between types of page with distinct patterns. The graphs show the plot of times between revisits. Thus, shopping and reference web pages are revisited with less than an our between visits.



Cluster Group	Name	Shape	Description
Fast Revisits (< hour) 23611 pages	F1		Pornography & Spam, Hub & Spoke, Shopping & Reference Web sites, Auto refresh, Fast monitoring
	F2		
	F3		
	F4		
	F5		
Medium (hour to day) 9421 pages	M1		Popular homepages, Communication, .edu domain, Browser homepages
	M2		
Slow Revisits (> day) 18422 pages	S1		Entry pages, Weekend activity, Search engines used for revisitation, Child-oriented content, Software updates
	S2		
	S3		
	S4		
Hybrid 3334 pages	H1		Popular but infrequently used, Entertainment & Hobbies, Combined Fast & Slow

The study gives a realistic picture of what people do on the web, matching one of the key strengths of unobtrusive research. It also gave data on an unprecedented scale. The study by Adar et al. [11] faced several challenges typical of unobtrusive research. First, it needed to also collect survey data to be able to interpret the logfiles; in themselves, they gave insufficient information to allow concluding. The study also beg the question if searches using the tool bar are indeed representative of general www search. Finally, the researchers had to make sure that users had opted into the study and also needed to anonymize data; this represents frequent ethical concerns in unobtrusive research.

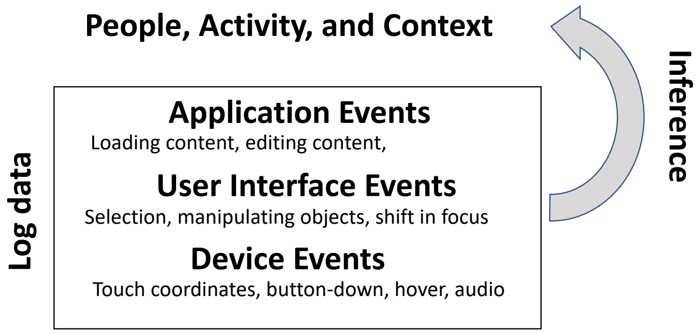


Figure 14.1.: Hierarchy of log file recordings: application, UI, and device events. From these, the researcher intends to draw inferences about the people, activities, and context in which the log files were produced.

14.1. Log Files

Logfile analysis is a term derived from automatic capture of interaction with interactive systems that are captured in a computer file, a *log*. Logs may be obtained from a variety of systems, including standalone programs, logs of interaction with web pages, logs of interaction in applications, and so on. In some cases, these logs are generated automatically, in other cases programs or servers may be modified so as to generate the logs.

Information of interest in log file analysis spans many time frames or levels of abstraction [335]; Figure 14.1 shows some of these. Log files contain data that are produce by people or systems; here we consider only data that are produced as a result of interaction. Such interaction may be movement, gestures, speech, and so forth: we have access to those in log files in as far as the input devices sense them: often this data may be at the milliseconds level (e.g., button presses) and happen synchronously. This is called *device events*. At a slightly higher level, perhaps spanning seconds to minutes, is *user interface events*. They are about objects and actions in the user interface. Slightly higher in abstraction is *application events* which is related to the functions available in the interactive system. All of these may be logged. The hierarchy reflects the fact that nowadays almost every aspect of hardware and software in a computing system can be instrumented to register events.

Next we discuss how to do logging, how to clean and transform data, and how to describe what may be learned from the log files. A crucial part of interaction may not be logged by applications. That part concerns people, their activity, and context. We discuss how to use logged data to make inferences about that.

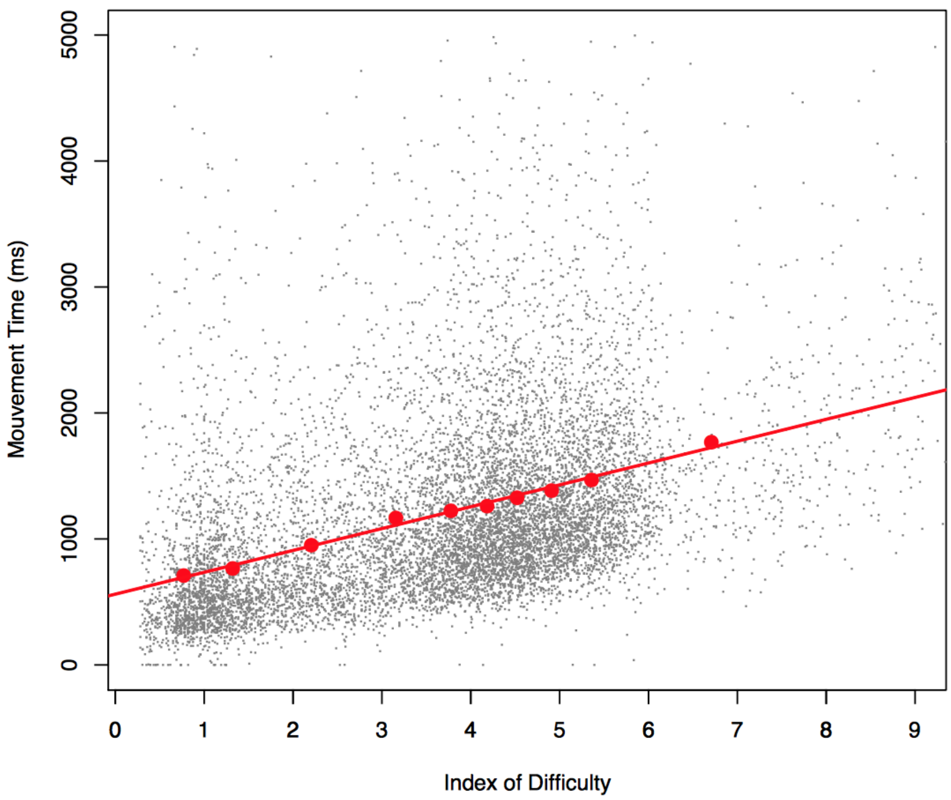
With the advent of the www, tool kits were developed that supported researchers and practitioners log web events. For instance, WebQuilt is a system that support web design teams in logging interaction with web pages, analyse the logged data, and visualize common path taken by users through a site [345]. Many other similar systems now exists, including Google Analytics and other software for web analytics. These may be tailored to particular interaction styles. W3touch [568], for instance, unobtrusively collect performance data related to touch interaction with web content. Among other things, W3touch allows designers to relate misclicks to the web content they were likely intended to hit.

Before analysis, log data needs to be preprocessed. It needs to be cleaned and transformed it into a form that is easy to describe and analyze. Such data wrangling is complicated and the topic of much work [e.g., 672]. Among the topics are dealing with missing data, obviously wrong data, and noisy data. The data may then need to be segmented, or divided into some categories of interest. An example is given in the sidebox.

Paper Example 14.1.1 : Log study of real-world pointing

How well does Fitts' law apply to real-world pointing movements with a mouse? To find out, Chapuis et al. [145] logged the movement kinematics of 24 users in 36 computer configurations over several months. The data consisted of timestamped clicks and movement coordinates and velocities. To run a Fitts' law analysis on the data, they needed to first decide what "a movement" is. Alas, most hitherto studies had taken place in the laboratory, where the pointing task was a given priority for users, and often the only task they do. There was therefore no applicable definition available.

The authors needed to decide what constitutes the beginning and what is the ending of a pointing movement. Analyzing their data, they learned that real-world pointing trajectories contain numerous *pauses*, or time segments where velocity drops to zero. Based on a closer inspection, they decided that pauses that are longer than 300 ms are *stops*. Any pause that occurs after a stop and before a click would be ignored. This way they divided the data into 2 million pointing movements.



The results, shown in Paper Example 14.1.1 are telling. On the one hand, they show that real-world pointing movements adhere well to Fitts' law. On the other hand, there is remarkable variance in movement times unaccounted for by the law. This is visible in the "fat grey cloud" enveloping the trend line. What might cause this variance? Many things. For example, movements done in different regions of the display, and at different distances, cause variance. Moreover, users may have secondary tasks to take care of that compromise performance. Third, users may not care about "being fast and accurate" unlike in experiments where they are paid for that.

The first thing to do with the obtained dataset is to describe and visualize it. For instance, in user research on the web we may unobtrusively describe many aspects of the visits for a simple page, including: the number of visits, the average duration of visits, the number of revisits, the percentage of users that leave the page within 2s, the number of visitors that enter through this particular page, and many others. *Descriptive statistics* of such data are relevant in as far as they are indicators of use of the page or values to the users. We will later discuss this more fully under the heading of usability (see Chapter 19).

After description, the analyst may want to *identify critical incidents* in data, *cluster the data*, or *build a model of the data*. First, critical incidents are occurrences that are of interest for some reason. For example, a user getting lost on a web site is important for designers to know about. To recognize such incidents, one needs to develop a hypothesis on how those incidents map to data and, based on that, write either a rule-based script of a machine learning -based detector for those events. Second, data clustering refers to statistical methods that aim to redescribe a high-dimensional dataset using a few lower-dimensional clusters. For example, users can be clustered based on which features of a system they use [851]. Third, models of data can for example involve the sequences that users take through a service [335]. Data can also be used to train a supervised learning -based model. Supervised learning is a machine learning method where the task is to predict one label (a variable in data, e.g. a click) given a number of other labels (e.g., user's age, previous purchase etc).

Log analysis is often an inferential process. That is, the analyst needs to conclude, based on recorded events, something about the events that produced those events. Sometimes this inferential leap is impossible given the data. For example, click events on a web site say very little about the emotions of a user. Any attempt to infer emotion would be tenous. However, some usability-related events can be inferred from logs, such as *usability problems*. This has been much studied and many aspects of usability problems may be inferred from log files (see [370, 335]). Undo events for example, have been used to successfully infer usability problems [14].

14.2. Instrumenting People, Things and Places

Another approach to unobtrusively collecting data is to equip people, things, or place with digital recording devices, an act known as *instrumenting*. In short, people, things or places are turned into measuring instruments.

Instrumentation has a long history in mobile computing, where sensors in mobile phones – such as accelerometers, gyroscopes, and geopositioning – have been used to understand, for example, whether people are walking or standing still throughout the day [446]. Presentday mobile devices offer many sensors ranging from heart rate to the GPS.

Environments can be instrumented, too. Studies on ubiquitous computing have instrumented everyday environments using everything from pressure sensors to video cameras and microphones.

Finally, on-person measurements can be done. For example, everyday emotion sensing can be done by embedding skin galvanisation sensors to a driving wheel. Such instrumentations can complement log analysis, which focuses on how people use an interactive system. Real-world instrumentation is weaker in terms of direct access to the interface, but it allows richer understanding of interaction in its context.

14.3. Archival Data

Unobtrusive research may also be done from material that already exist. Rather than collecting traces of use through log files of ongoing use or by instrumenting people, places or things, such data need simply be gathered. We call such data *archival* because they are found in archives, be they online, in public repositories, or in individual's records. Archival data may be analyzed to teach us about users, their activities, their context, and the use of existing interactive systems.

Archival data of relevance for HCI take many forms. We already discussed the example of using YouTube videos to learn about the use of touch screens by people with impaired motion. Other examples include the following:

- Reviews of interactive systems, for instance from discussion forums, publications, or comments from people who have bought a system. Such reviews reflect users' attitudes towards interactive systems and may be a valuable way to learn about users.
- Reports on errors and bugs. Those may be actual error reports from failures in medical devices, errors dealt with in call centers or by support teams, or reports from formal investigation into errors for cars or planes.
- User-generated content from social media. Such content might include comments, discussions, pictures, or videos. For instance, [834] studied twitter usage during two natural disasters. The purpose was to learn from tweets how people share information and understand the bigger picture of emergencies. They did so by obtaining about five millions tweets from immediately before and after the disasters and manually coded the tweets on whether they were on-topic and on categories of the information being communicated.

The key benefit of using archival data for user research is that they are usually inexpensive to obtain. Archival data may be analysed at large scales. Archival data also have the benefit that it can easily be worked on by multiple researchers for as long as they like.

One challenge specific to archival data is that, in contrast to the data discussed in the two previous sections, they are often created for another purpose than that of the user research. For instance, social media content may be uploaded with a variety of aims in mind, none of which is to create research content. The inferential leap is longer. Thus, archival data need to be analysed differently and more in depth to be useful for user research.

14.3.1. Content analysis

Content analysis is a frequently used technique for dealing with archival data [426]. Content analysis is a technique that helps the researcher make a subjective interpretation of the data using a systematic processes of coding, aggregating, and describing archival data. Such descriptions may focus primarily on the meaning of data (e.g., intention, deeper message) or on the frequency of particular data (e.g., asking how many, how frequently, in what way). Typically, content analysis results in a classification of content, descriptions of types of content, and frequencies of content. For instance, a content analysis of complaints to a call center about an interactive system may use attempt to describe the main types of written complaints and their causes. It may also enter into a deeper analysis of the dynamics of the conversations, breakdowns in talk, and expressions of frustration.

Content analysis may be done in many ways; next we describe one account of the steps involved in using archival data to answers research questions. More elaborate descriptions of the technique may be found in [866], [62], and [426].

1. Clarify research aim. As with all other user research, content analysis of archival data starts by having a clear research question in mind that is appropriate to answer with that research method. The researcher needs to determine what topic or topics are in focus and what is the boundaries of the research.

2. Identify appropriate archival data. The researcher needs to determine which data will help arrive at the aim of the investigation.

3. Select sample and unit of analysis. The appropriate archival data may be sampled in many ways, for instance randomly, delimited by data, size or content, or by search. Unit of analysis refers to the smallest unit of the archival data whose content is analyzed. For instance, in analysis of social-media content, the unit of analysis may be a full post, specific words, or a sentence within a post.

4. Decide on the depth of analysis. The analyst may deal with the archival data in one of two main ways. In manifest analysis, the research focuses on the surface properties of the archival data. In latent analysis, the researcher tries to interpret the archival data further to make inferences about intention, context, and background.

5. Develop a coding scheme. Codes represent the researchers' reading of the units, captured so that they may be compared, discussed, and further analyzed in different ways. Typically, coding is done by individual researchers but it may be done automatically, or in groups of researchers. Codes may taken from earlier work, developed by coding a subset of units, or developed through a preliminary coding of the entire sample.

6. Coding the units and check that the relevant aspects of them is captures.

7. Reporting findings relative to original aims.

Three good practices of content analysis are: 1) include several people representing different viewpoints as analysts to reduce biases; 2) try to contextualize the findings within assumptions made about the users and the culture. 3) document coding units and coding schemes; Two benefits of this are *transparency* and replicability [426]. The codes used can usually be inspected and it is clear to which units they are applied.

Automated and computerized tools have been developed to facilitate content analysis. This can involve, among others,

1. machine learning tools to transcribe spoken speech into text
2. rapid video annotation tools that allow watching videos in fast forward while marking events using a keymapping
3. collaborative annotation tools that allow a team of annotators to work on a multi-modal dataset and obtain quality indices (e.g., interrated reliability)
4. tools for natural language analysis, e.g. sentiment analysis methods can tell if reviews of events in social media have positive or negative valence

14.4. Why Are We Not Doing Unobtrusive Research All The Time?

Unobtrusive research has the appeal of removing the possibly confounding effect of the research itself. It mitigates reactivity. Why is it then that we do not always use it?

There are there common challenges in unobtrusive research. The first is the gap between what is logged and what needs to be known. Data is often created for other purposes than than what the user research aims to study. For instance, although YouTube videos may be used to study touch-enabled mobiles, the motivations for uploading the videos differ and the content of the videos may or may not concern the topic to be studied. Thus, even in the study of [26], the videos available for analysis is a limited sample of all the difficulties that people with motor impairments encounter. Similarly, weblogs may not contain all the actions that a researcher want to study and records of errors with devices may not concern sufficient detail.

Second, sometimes data collected by unobtrusive research is difficult to interpret. This is particularly the case when data only serve as an indicator of what we are interested in or when we aim to understand the mechanisms that has produced certain data. For instance, logs of activity on web pages may produce indicators such as time-on-page. However, it is difficult to know *why* users spend a lot of time on the page—is it because the page is very interesting or just something the user must do? The threat is that an indicator that is readily available in data becomes a variable that is optimized. For example, one could make users stay longer on a site by making it very difficult to navigate.

Another instances arises when we use sensor readings, for example those from mobile phones, as predictors for tasks or user characteristics. Such readings may predict the tasks that users engage in, which may be useful. However, it is often difficult to *understand* anything about the task from the sensor readings themselves. Thus, unobtrusive research often require difficult inferences to conclude anything about users and their activities.

Third, the ethics of unobtrusive research is challenging. Unobtrusive research is sometimes covert, meaning that to prevent reactivity, participants are required to not know about the data collection. In a case where it is difficult – or impossible – to obtain the consent of those on the record, can a study be carried out at all? What if the data was produced long ago, in a different place, and by unknown people? Because unobtrusive research may allow powerful predictions about users, it may be misused against them. Researchers using unobtrusive research need to be conscious about these ethical difficulties.

Summary

- Unobtrusive research minimizes the confounding effect of reactivity; in other words, the potentially biasing effect that the research itself can have on participants.
- Unobtrusive research methods look at traces – direct and indirect, archival records, and instrumented recordings.
- Interpreting unobtrusively collected datasets is hard, because they often lack triangulating (additional data sources) information.
- Content analysis is an approach that helps make sense of unobtrusively collected datasets.

Exercises

1. The value of unobtrusive research. In which situations would you use unobtrusive research vs. other user research methods? Why? (a) Learning how museum visitors use audio guides. (b) Learning how museum visitors experience audio guides. (c) Learning what museum visitors need from audio guides.
2. Inference gap. Log data is often just a 'shadow' of what we really want to learn about. List 10 things that user researchers may commonly want to infer about users (e.g,. needs, requirements, ...). Then come up with a source of log or archival data that you think might best help infer each.
3. Ethics. What are some ethical issues one must consider when carrying out unobtrusive research?

15. Representations of User Research

User research aims to inform design decisions. Its aim is to gather knowledge about the prospective users of a system, their activities, the contexts in which they work, and existing technologies they use. Previous chapters in this part have been on the *methods* of user research. In contrast, this chapter is about the *outcomes* of user research, in particular how we *represent* research data using diagrams, models, and text. Figure 15.1 shows a couple of examples of such representations.

The previous chapters detailed how to analyze interviews, how to do ethnography, and how to analyze surveys using statistics. Why fuss further about representations of data? The main reason is that how we represent user research matters, just as how we represent a logical puzzle or lay out a math problem impacts how easy it is to solve. Data in itself do not *do* anything. A dataset is simply a collection of observations. This chapter concerns how we can represent it to inform us, provide insight, and act. In particular, we care about the following applications of user research data:

- For summarization. Here, the representations consolidate extensive data from user research. For instance, Tychsen and Canossa [816] used metrics collected in a computer game to create representations of how players interact with the game. Those representations pull together many metrics on how people engage with the game, but also made clear higher-level goals of players (e.g., some main groups of play style).

- For spelling out requirements, so that they may be used in design briefs and software production contracts (see also item VII).

- For inspiration: Finding new opportunities and concepts to drive design.

- For communicating within a development team. For example, results of usability tests are often summarized as criticality-ranked issues, which can be shown to developers together with example data such as videos (see Chapter 43).

- For checking insights with stakeholders: For example, task analyses are essentially hypotheses about how users structure their activities, and should be verified with stakeholders [25].

We may use user data for all of this; the techniques described in this chapter help do that.

User research informs the design, the topic of Part VI. That Part will use insights from user research to create ideas about future interactive systems. Some of the techniques that we discuss here can be used also for ideas of future interactive systems. For instance, we may represent the sequence of steps to be taken in a new system in a sequence model similar to that of Figure 15.1. However, the aims of the parts are different: user research describes current affairs, design imagines a possible future. It is important to keep clear on whether a scenario is based on interviews (user research) or reflects an imagination of

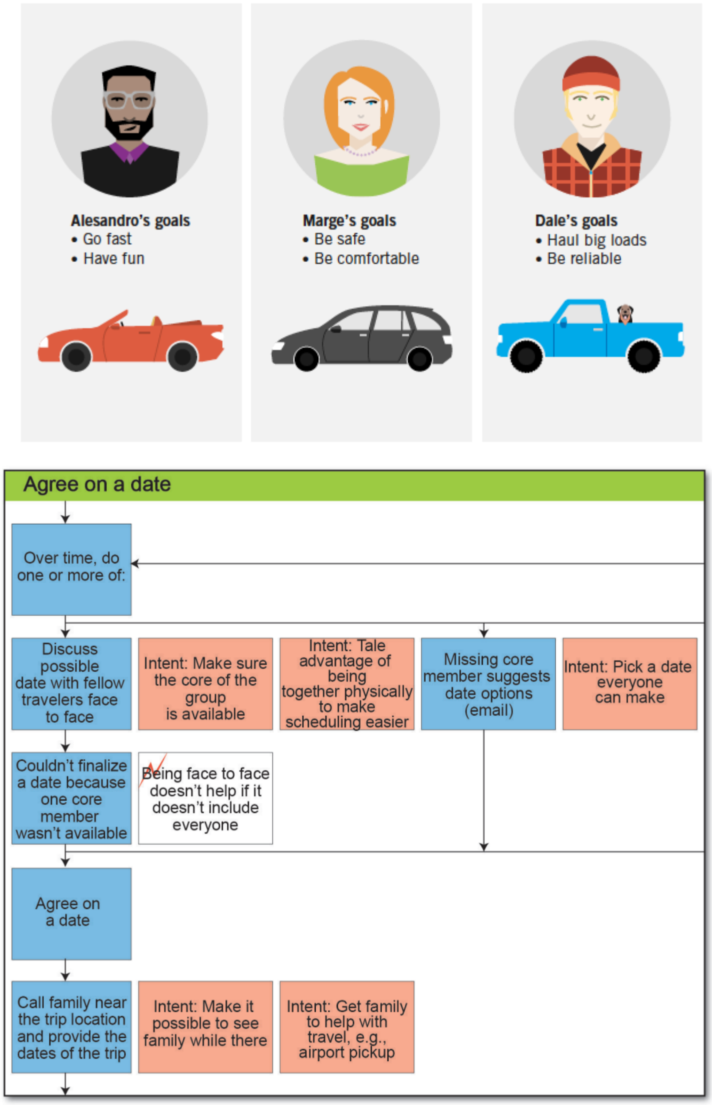


Figure 15.1.: For user research to inform design and decision-making, the acquired data must be summarized in some useful way. Panel (a) shows a few simple personas, representations of user groups as archetypes by Cooper et al. [169]; Panel (b) shows a summary of how people select a date when they go travelling, based on field work of Holtzblatt and Beyer [342]. Both representation types help understand, apply, and communicate findings of user research.

how people might work (design). One of the most critical challenges in design is to go beyond the familiar and propose a change, a new way of acting in the world. This requires understanding of the familiar, the starting point and the constraints and expectations it poses to the future.

User research also informs engineering, which we discuss in Part VII. User research obtains information on what users want, how they work, and what they value. In engineering processes, these are translated into requirements. Requirements, in turn, are used as objectives and criteria for deciding when a system is 'good enough'. Results of user research can also be translated into algorithms, for example as objectives and constraints in optimization and learning algorithms that drive intelligent user interfaces.

The results of user research are also part of collaborative efforts. As such, they have an important communicative purpose: they help form joint understanding among stakeholders. To this end, user research has the obligation to be well-formed and well-communicated.

In this chapter, we discuss focal points for representing user data. We will talk about, in turn, representing knowledge about users, including their needs, preferences and capabilities; about what users do and how these activities are organized; artefacts: the influence of physical layout of the work environment; and contexts: understanding of the situations and their use-defining factors. We also discuss user requirements for technical functionality and performance of systems.

15.1. Representations of People

A big part of user research is to understand who users are, what their needs are, what they try to achieve, and how technology might fit into their lives and work. At the beginning of this part (see Section 10.2), where we talked about establishing *who* the users are. For instance, we may create user segments based on market research, or we may spell out our initial ideas on who users are and what their goals are. In contrast, representations of people aim to assist drawing actionable conclusions about them. Design and development teams needs to be able to relate to users and make inferences about users on issues that might not have been covered by user research. The most popular tool for doing this is personas.

15.1.1. Personas

A *persona* is the description of an idealized (non-existing) person that represents a group or type of users. The term was popularized by Cooper et al. [169], and has since been widely used in HCI. The idea is to construct archetypes of users in the form of fictional (but representative) individuals with specific characteristics. Personas, however, are based on user research—thus, they are more a synthesis or aggregate than a fiction. Below is shown an example of a persona taken from Nielsen [577].

> Camilla and Jesper live on the outskirts of Copenhagen. They are 35 and 39 respectively, and they have enough on their plate with children and careers. They have lived together for the past five years. Two years ago, they had their

> son Storm. Jesper has two children from his previous marriage, Christian and Caroline, 11 and 8 years old. The children live with Jesper and Camilla every second week. Camilla and Jesper prefer to use self-service solutions, and they are curious about what information the public sector stores on them, and how it is stored.

Descriptions such as the above may be combined with pictures or sketches to make them more memorable, engaging, and concrete. Socio-demographic information can be added. Cooper suggested that the goals of the personas are also described. Nielsen [577] further suggested that personas be combined with scenarios. Scenarios may also be used without personas and is discussed further in subsection 15.2.1.

Personas can summarize general tendencies in the data, being like a 'mode or median' member in that group. They can also be stereotypifications of that group, embellished versions that help take perspectives and ensure sufficient diversity in design.

The purpose of representing user research as personas, such as the above, are several. First, creating specific individuals. Personas can help avoid too "flexible" (or "elastic") personas. A degree of inflexibility can keep design in check. It ensures that when we take design decisions, we do not distort user data to the benefit of the design. Second, personas help avoid self-centeredness. Technology is too often developed for developers by developers. Third, personas prioritize data. Personas require selecting which user segments are most important in design. Fourth, they synthesize data. Raw data is too complex to deal with in design. A few well-selected personas can efficiently distill and communicate the essence of a dataset. Fifth, personas can drive empathy. It is easier to relate to another person's (even if hypothetical) viewpoint than to quantitative summaries.

Creating Personas

There are numerous persona methods. Pruitt and Grudin [660], for example, offer a rich persona template for software engineering projects. The template contains slots for describing, among others, the user's basic properties – like name, age, and occupation, a day in the life of that person, work activities, goals and fears, market size, technology attitudes, and quotes. However, long templates strike the risk of being just 'catalogues' with no clear relevance for design.

We believe that personas that go beyond observable attributes ("Jane is 25, she attends a university") to drivers of such behavior ("Jane is fascinated by theoretical physics.") are valuable [169]. How to obtain such characteristics without conjuring them up? Cooper et al. [169] offers one such method, which is here adapted as follows:

1. Group participants in data by their role. For example: parents, students, managers, and administrators are examples of roles.

2. Identify potential drivers of behavior relevant in this dataset. Cooper et al. [169] calls for diving behind observable behavior and try to identify *behavioral variables*, such as those describing attitudes, skills, beliefs, motivations, and aptitudes. Those variables capture some aspects of why people did what they did. They may be

continuous (e.g., ”motivated by social connections [weak-strong]” or ”excited about a new product feature [weak-strong]”), others will be binary (e.g., ”uses Windows 11 [yes/no]”).

3. Map observational data about participants to behavioral variables. That is, take snippets of data (e.g., interview quotes) and associate them with behavioral variables, for example using post-it notes.

4. Identify significant clusters forming around behavioral variables.

5. Synthesize characteristics of roles: find clusters where a certain role has a particular behavioral variable strongly represented. These are candidates for key characteristics of that role.

6. Check for completeness; that is, ensure that a reasonable proportion of data belonging to a user role is covered in the clusters that form its characteristics.

7. Select a persona type and name it.

8. Verbally explain key attributes and behaviors rooted on the clusters.

The method requires iteration between observational data, roles, behavioral variables, and clusters. This method ensures that personas can be rooted back to observational data: the clusterings of data ensure that one can trace back what a characteristic is based on.

Drawbacks of personas

The persona method is often haphazardly applied. From a statistical perspective, a persona is a summary statistic analogous to a mean, median, or mode in some distribution. However, it is not always applied from this perspective; the questions of (1) how to cluster users and (2) select such persons to represent whole groups are dismissed. The consequence is that personas can be almost anything and the connection to original data can be broken.

Pruitt and Grudin [660] reported issues they encountered in the use of personas at Microsoft. The characters given in persona descriptions were not believable, they were ’designed by committee’ or lacked link to original data. Personas were also poorly communicated, for example via full-blown poster-sized lists, which made little effort to help understand the main points. There was also no understanding of how to use personas, because they were not properly linked to the different stages of product development.

Personas have a relationship to user groups. *User groups* are segments of user population. Personas are often thought to represent groups or segments. However, defining a user group is tricky. From a statistical perspective, the problem is a clustering problem: Given a population of m samples (users), each depicted by a feature vector x, The goal is to assign every sample (user) to one of n groups such that distance of two samples within a group is minimized and the distance to other groups is maximized. Unfortunately, although this is generally understood, personas and groups are often created based on intuition and not on systematic comparison of users not to mention statistical analysis.

This has the unwanted consequence that such representations can be unconvincing to stakeholders and - worse - misleading in the design process. Systematic methods like the one presented above have a better chance of avoiding these issues. New data-driven approaches are also emerging, as illustrated in the sidebox.

Paper Example 15.1.1 : Automated generation of personas from log data

Generating personas has traditionally been a labor-intensive process. To form a persona, data on users must be collected – for example via surveys or focus groups. It then needs to be labeled, clustered, and so on. Despite the high amount of effort, there is no guarantee that the personas represent well the targeted user population. Moreover, they can become outdated.

Automated persona generation refers to the use of online datasets to computationally generate up-to-date personas. These descriptions are automatically kept updated as changes occur in the user population

Jung et al. [392] present a case for automatically generating personas from data available from social media. They analyzed data from users who engage online with Al Jazeera, a popular news channel. The data came from 180,000 users in 181 countries and encompassed over 30 million interactions Al Jazeera's YouTube channel. Large datasets like this would normally be impactical for persona development.

The authors present a statistical method for identifying user segments from such datasets. They first identify user interaction patterns from data. These patterns tell which products users used and how. These clusters are then linked to demographic data, such as age distributions. Finally, persona descriptions are created around these attributes, and then enriched with descriptive features, such as names and photos by drawing from using generic databases (e.g., typical names of people of certain age). An example of a resulting persona in Al Jazeera's case:

> Samantha is a 25-year old female living in the US. She likes to read articles about society, environment, and racism on a computer terminal. She usually watches about 2 minutes of video.

Any number of such personas can be created, while ensuring that they are rooted in original data. They can be further augmented by representative comments (to videos) created by this group. The personas can also be updated on demand.

15.2. Representations of Activities

In previous chapters in this part, we have covered how difficult it is to obtain valid and reliable insights about the activities in which people participate. In contrast, this section will describe how we can represent these insights in a way that helps to clarify people's tasks.

Before we start, we need to be more clear about the terminology for activities. An *activity* is the total of what a user or a group of users wishes to achieve. It is purposeful interactions among actors that develop over time. A *task* is some piece of work to be done. As an analytic tool in thinking about user research, tasks have several characteristics:

1. A description of a task is often about things to be done, not about how they are done. When represented like this, tasks can be compared across different interactive systems. Part IV will explain why this is a slightly simplified picture by arguing that tasks and tools are always intertwined.

2. A description of a task can alternatively be a description of how things are done. Hierarchical task analysis, as we learn here, offers a decomposition of a task (when carried out with a particular tool).

3. A description of a task should be as specific and concrete as possible.

4. A task is a meaningful unit. That is, getting a task done is a meaningful achievement for the user. They care about being able to complete the task.

Next, let us look at some ways in which to work systematically with representing tasks.

15.2.1. Scenarios

Scenarios are narrative accounts of an activity or task. Scenarios use storytelling to communicate relationships between users and events. They are less structured than more analytical tools, such as task analysis or user requirements. Scenarios are almost always written from a user point-of-view. This helps emphatizing with users. They are almost never prescriptive in the sense that they would tell how things *should* occur. As a consequence of lacking a rigid predefined structure, scenarios can be underspecified as accounts of what happened. However, a well-written narrative may help us imagine the experience of a person from their perspective. The sidebox provides an example of how explainable AI was studied using scenarios.

Example scenario by Rosson and Carroll [694]:

> After three years at Virginia Tech, Sharon has learned to take advantage of her free time in-between classes. In her hour between her morning classes, she stops by the computer lab to visit the science fiction club. She has been meaning to do this for a few days because she knows she'll miss the next meeting later this week. As she opens a Web browser, she realizes that this computer will not have her bookmarks stored, so she starts at the homepage of the Blacksburg Electronic Village. She sees local news and links to categories of community resources (businesses, town government, civic organizations). She selects "Organizations", and sees an alphabetical list of community groups. She is attracted by a new one, the Orchid Society, so she quickly examines their Web page before going back to select the Science Fiction Club page. When she gets to the club page, she sees that there are two new comments in the discussion on Asimov's Robots and Empire, one from Bill and one from Sara. She browses each comment in turn, then submits a reply to Bill's

comment, arguing that he has the wrong date associated with discovery of the Zeroth Law.

Scenarios are written as stories. They consist of basic elements of a story: a setting, one or more actors with their goals and characteristics, and some tools and contexts that are available for that. The scenario then describes the sequences of actions and events that lead to an outcome. It is important to describe the actors' thinking and experiencing related to technology: how they set goals for, think about, react to, and experience it.

Paper Example 15.2.1 : Scenario-based development of explainable AI

It has become almost a truism that AI systems must be designed to be trustworthy and transparent. Alas, realizing this vision has turned out to be difficult. One reason is that it hard to explain to users reasons behind AI's actions. Simple approaches to explainable AI, such as presenting a simplified model of the AI's reasoning to users, have failed: users have very little interest in learning detailed accounts of how the AI draws its conclusions. There is a need to rethink what explanations are and how they are interacted with.

Wolf [889] turned to scenario-based design to address this problem. As the case, they studied aging-in-place monitoring; that is, AI systems that help aging people maintain autonomy by monitoring their daily activities. These are systems that for most of the time do not need to explain themselves; the question is when should they, and how.

The authors generated what they call explainability scenarios based on ethnographic studies of AI and machine learning professionals. In two field studies, they observed these professionals across three projects. To form scenarios, these primary data were augmented with results reported in studies of caretakers and other stakeholders.

The following example scenario imagines the experience of Jody and Catherine using an aging-in-place system [p.254]:

> Jody checks the app on her phone that displays her mom's daily activity report – the outcome of in-unit monitoring that captures and analyzes her mom's activities of daily living (ADLs). Her mom, Catherine, recently moved into an aging care facility and together with Jody decided to enroll in the facility's new activity monitoring program.
>
> [...]
>
> The system is designed to help keep Jody aware of her mom's daily activity levels (and alerting Jody of any abnormal changes or deviances in her mom's behavior) by providing daily reports in the app. To maintain Catherine's privacy, the reports do not document each and every activity of the day. Instead, it classifies her activity for the day into three levels (green, yellow, red) and provides a general explanation of why the system applied that label. [...]
>
> Several days go by and Catherine's reports all turn up green. Then one day, Jody sees that her mom's been given a yellow report. ”No kitchen activity recorded today.” Jody calls her mom to ask how her day was and if anything was wrong. ”I went for a long walk with my new neighbor, Vicki. It was nice, we have a lot in common. We ended up going out for lunch,” Catherine explained. Jody was relieved to hear that ”no kitchen activity recorded” wasn't anything alarming, she was actually excited to hear her mom was making friends.

The scenario continues to an event where these explanations do not suffice:

> As Jody opens the app again she is greeted with a new message: ”Your Monthly Activity Report is available.” The report provides a dashboard view of Catherine's activity levels for the past month. It also provides a prediction – based on the predictive model from pilot data, the app has high confidence that seniors with activity levels similar to Catherine's will need to transition to an assisted living level of care within the next three months. Okay, Jody thinks, why would that be? She clicks on the ”Tell me more” link next to the prediction, which says the top influences on the prediction are: age and frequency of ADLs. Her mom has mostly been getting ”green” reports so Jody is confused why she might need escalated care. Unsure of who she is supposed to ask for help, she wonders

15.2.2. Customer Journeys

Customer journey mapping is an account of events where a user encounters a service or product. This map is presented in the form of a journey, or a path that goes via *touchpoints*. Whereas scenarios regularly focus on events during use of technology, customer journey mapping also cover events that precede actual use, for example advertisements, creation of accounts, onboarding, etcetera. The term 'customer' depicts this view: Touchpoints are typically depicted from the perspective of evolving relationship between a customer and a business, where a might-be-customer transforms into a loyal customer via the touchpoints. Actual use of technology may be only in a minor role in this. At every touchpoint, there are goals, thoughts, and experiences reported.

Rosenbaum et al. [689] describe a method for customer journey maps. It calls for analysis of two dimensions:

1. Touchpoints: the sequence of events through which customers interact with a service
2. Relevant strategic actions associated with touchpoints.

Touchpoints are typically divided into three periods: pre-service, service, and post-service. The pre-service is about experience before registering or purchasing the service. For example, one may encounter advertisements of the service, or hear from friends about it. The post-service period takes place after the actual service, for example when posting to social media about the experience or complaining about service to customer service. Strategic actions refer to enabling actions that different participant groups should take to make the touchpoint successful. For example, strategic actions can concern what a manager, customer-facing personnel, or design is required to do.

15.2.3. Task analysis

Task analysis is a method for decomposing tasks and presenting them as hierarchically organized sequences of subtasks. It decomposes and exposes the *structure* of interaction [25]. Task analysis is to human-computer interaction what requirements engineering is to software engineering. Many analytical evaluation methods we cover later in this book (Chapter 41) build on task analysis. Task analysis can also inform design; an example is given in the sidebox. The reason for its popularity is its systematicity. Consequently, carrying out task analysis properly, especially when it involves validating the analysis itself with the stakeholders, is time-consuming.

Task analysis is *a functional description* of behavior. It begins by outlining goals before considering actions by which the task is accomplished. "Complex tasks are defined in terms of a hierarchy of goals and subgoals nested within higher order goals, each goal, and the means of achieving it being represented as an operation.The key features of an operation are the conditions under which the goal is activated and the conditions which satisfy the goal together with the various actions which may be deployed to attain the goal. These actions may themselves be defined in terms of subgoals." [25, p.68].

The output of task analysis is a *task model*. It describes the requirements of carrying out a task successfully. It can be expressed in multiple ways: as a sequence or as a hierarchical diagram.

A sequential task model describes a task as consisting of a linear progression of subtasks. It is often described verbally. For example, a task model of logging into a web service could consist of the following steps:

1. Enter login name
2. Enter password
3. Click 'Login'

But how does one 'enter' something? The two first rows can be expanded into two more subgoals, resulting in a higher-fidelity task model:

1. Click 'Username' field with mouse
2. Type login name with keyboard
3. Click 'Password' field with mouse
4. Type password with keyboard
5. Click 'Login' with mouse

A limitation of linear models is that the relationship that clicking and typing have is not captured well. Hierarchical task analysis can better account for this.

Hierarchical Task Analysis (HTA) is an established task analysis method that assumes two kinds of relationships between subtasks: (1) order: task A precedes task B and (2) part-whole relationship: task B is a subtask of task A. Every subtask is thought to require operations for completing it. Operations are any actions that the user must take. An example of a simple task is given in Figure 15.2.

The core concept of task analysis is a *task*. The given main task is recursively split into *subtasks*. A task (or sub-task) consist of a well-defined beginning state, goal state, and operations that transform the beginning state to the goal state. Some operations are conditional; that is, they can only be executed when some conditions apply. How many levels of subtasks one wants to do is a practical decision.

Task analysis can be done in two ways: empirically and analytically. The goal of *empirical task analysis*, which is used in user research, is to understand tasks in *unconstrained, real-world behavior*. Typical data collection methods include interviews, analysis of systems and their documentations like manuals, and observations in the field recorded with video cameras and field notes. In *analytical task analysis*, modeling is speculative; it is carried out as part of design or evaluation. It can also in itself expose potential problems, such as complicated task structures or need for training. Task analysis is modeling activity that depends on the human analysts, presently it cannot be adequately automated, although there are attempts based on for example data mining techniques from log data.

To do a hierarchical task analysis, one can start with a sequential model of the highest-level subtask. First, split the main task into a sequence of subtasks following each other. Consider a vending machine: you first need to decide which soda to drink, then find the corresponding button, then find wallet and decide the payment method, etcetera. Second, subtasks that are mutually dependent can then be clustered to form a hierarchical structure.

General steps in task analysis include [25]:

1. Decide the purpose(s) of the analysis
2. Get agreement between stakeholders on the definition of task goals and criterion measures
3. Identify sources of task information and select means of data acquisition
4. Acquire data and draft decomposition table/diagram
5. Re-check validity of decomposition with stakeholders
6. Identify significant operations in light of purpose of analysis
7. Generate and, if possible, test hypotheses concerning factors affecting learning and performance.

In addition to sequences and hierarchies, HTA offers additional notation to express *plans*, *conditions*, and *parallelisms*. Plans are routine procedures by people, "do this, then

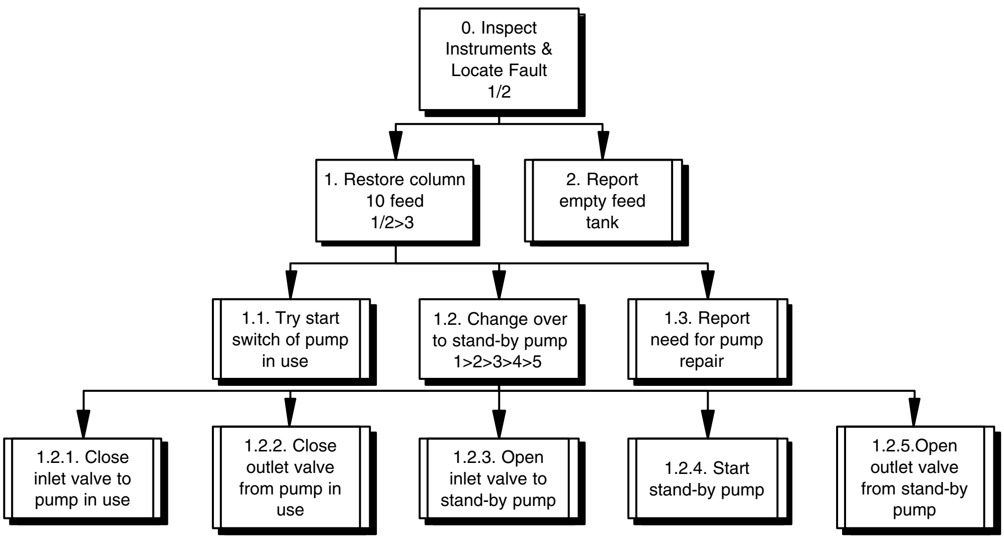


Figure 15.2.: Hierarchical task analysis for the task of inspecting faults. Adopted from Annett [25].

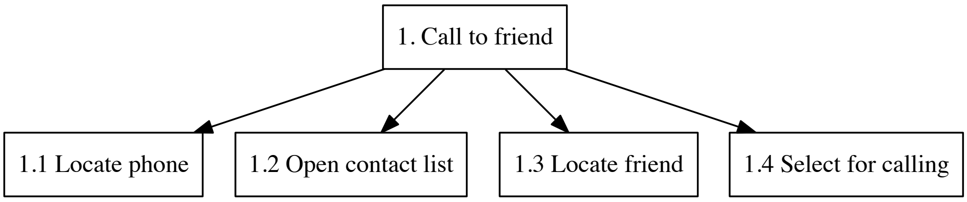


Figure 15.3.: Hierarchical task analysis diagram for the task of calling a friend.

this, then this..”. Conditions are if-then rules (”if condition X, do this”) that describe conditions for actions. Parallelisms are subtasks that can be pursued in parallel. Moreover, stop rules can be expressed: telling when a user would stop executing a subtask. These are annotated in the diagram.

Example

Let us consider the task of calling a friend using a smartphone. Figure 15.3 shows a diagram with four subtasks ordered sequentially. In order to complete a sub-task, more operations need to be carried out. For example, sub-task 1.3. "Locate friend" can be further split into Figure 15.4.

Like this, one can continue splitting down the subtasks, noting their relative orders etc. For example, sub-task 1.3.2 demands the following steps:

1. Point at vertical bar
2. Drag down finger
3. Confirm the right letter

When we continue this for all sub-tasks all the way to the third level, we obtain a full HTA diagram.

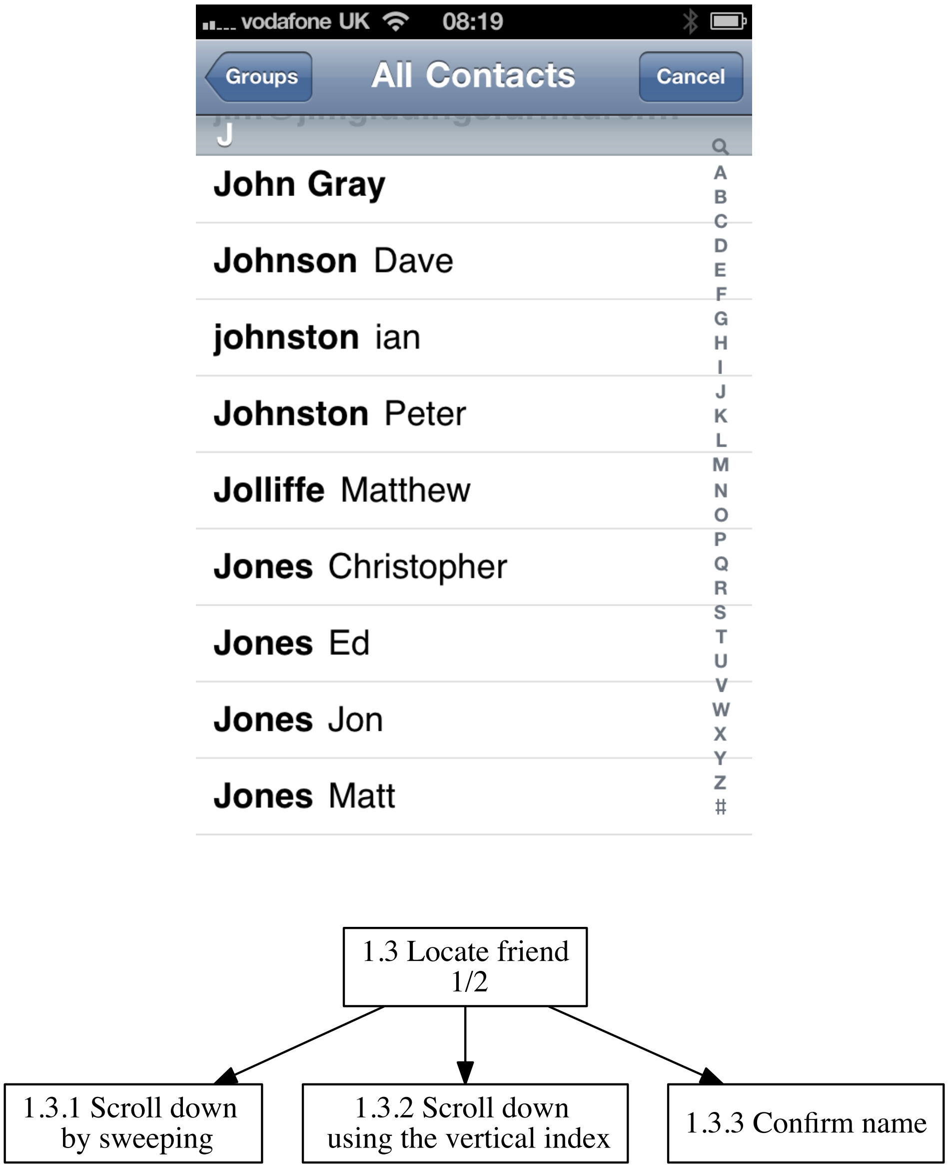


Figure 15.4.: Diagram for the sub-task of locating the friend in contact list. HTA introduces conventions to mark the order in which sub-tasks can be executed. Here, either "/" 1.3.1 or 1.3.2 is executed.

Paper Example 15.2.2 : Task analysis informs layout design

Task analysis is a systematic representation of users' activities. The method is among of the most common in the study of safety-critical systems, where it is used to identify potentials for errors and needs for training personnel. In user-centered design, task analysis forms the basis of several analytical evaluation methods, such as cognitive walkthrough and KLM (see Chapter 41). However, task analysis can inform lower-level decisions in design, too.

Sears [733] presented *Layout appropriateness* (LA), a metric to inform the design of widget layouts. A widget layout is a graphically laid out set of interactive widgets, such as text fields, scrollbar, toggles, buttons, etcetera. To compute LA, one needs two inputs: 1) tasks sequences that describe users' widget-level actions and 2) estimations of their frequencies. The former can be obtained via task analysis, the latter using methods like questionnaires, from logs, or by asking designers or stakeholders to rate the priority of tasks. This knowledge is expressed in a transition graph where nodes are widgets and links transition probabilities. For example, this graph shows an analysis of the task of opening a file in a word editor:

15.3. Representations of Context

Contextual inquiry (see Chapter 11) offers a rich set of models to describe contexts and situations. Note that these are not models in mathematical sense, but conceptual models. They describe dependencies and causal relationships verbally or diagrammatically. A key concept is that of an *actor*. An actor can be anything – human or non-human, digital or physical – that participates in some active sense in the activity. The context models describe flows, sequences, artefacts, as well as cultural and physical circumstances related to those actors.

A flow model enumerates the main actors in context and their relationships to the user. Actors can be humans, technology, or even physical objects like a kitchen table. Flow model is drawn as a graph where actors are nodes and actions are directional arrows connecting them. The direction of an arrow indicates who takes initiative. Arrows are annotated to describe what happens.

A sequence model is an ordered list of actions needed to carry out to complete a task (similar to what we described above). To get from an initial state to a goal state, a step of actions is carried out. On the way to the end goal there can be sub-goals. A sequence model is less rich representation of tasks than hierarchical task analysis (HTA). A sequence model provides an ordered list of subtasks at one level of depth in the HTA tree.

An artefact model describes interactions among artefacts in a work flow. An artefact can be a physical or digital object, such as a paper note, a social media message, presentation template, or anything that is interacted with during work. In contextual inquiry, such artefacts are documented, and the way they are manipulated and passed on is traced and modeled. An artefact can be something as mundane as a receipt or a post-it note, or a recording of some sound generated by a system. In contextual inquiry, artefact models are sometimes just collections of descriptions of artefacts. Sometimes they are organized into descriptions of flow.

A cultural model is an expression of different beliefs, values, and practices that the involved actors may have. The model is typically drawn as a Venn diagram, where each circle corresponds to a 'culture'.

A physical model is a rendition of the physical space and the actors' movement in the space.

15.3.1. Rich Picture

The rich picture is a way of representing the insights from user research in a diagram; it focuses on the relations among people and technologies, as well as the richness of the current use situation. The rich pictures was devised by as part of Soft Systems Methodology and later used in HCI by, for instance, Monk and Howard [542].

Figure 15.5 shows an example. An effective rich picture shows some of the structure of the situation, some of the key processes, and some of the concerns of the stakeholders. A rich picture can use the words and phrases of actors, but can otherwise use any means of communication. There is no set syntax or set of notations. The use of rich pictures here is to synthesize data from user research.

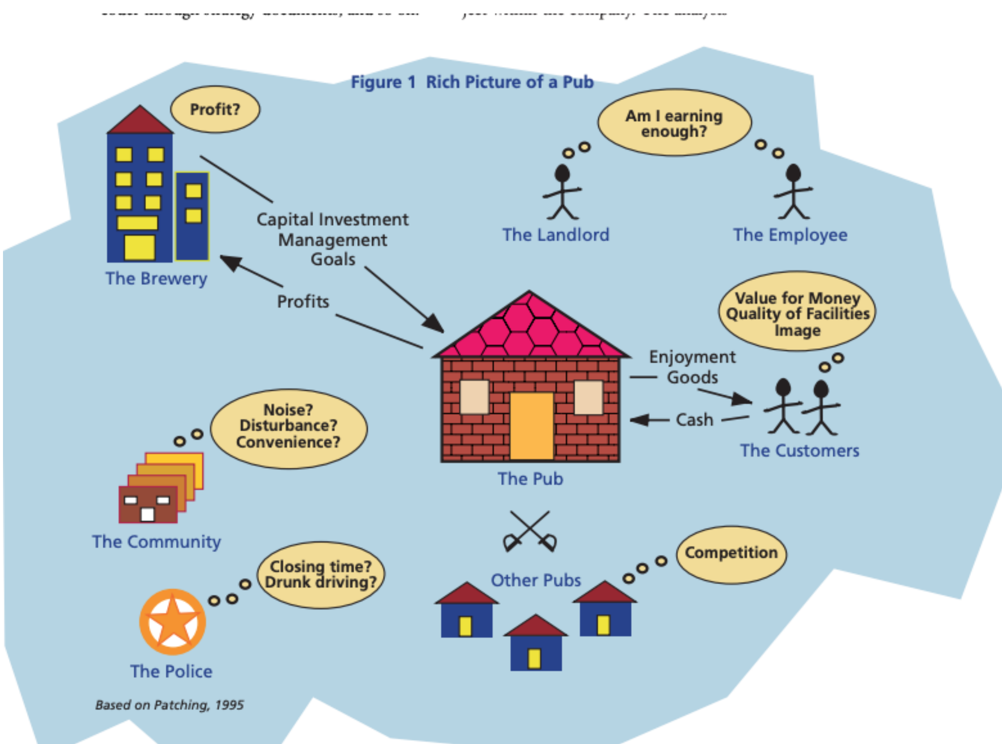


Figure 15.5.: A rich picture of a pub [542]. This picture shows main stakeholders, some of their considerations, and some links between them.

The process for creating rich pictures are as follows.

1. Gather people that knows about the topic of the user research.
2. Outline the key stakeholders, including people and organizations, and some of the structure (e.g., geographical, physical, organizational).
3. Detail processes and concerns among those stakeholders and structures

Remember, a rich picture is about capturing some of the complexity of the situation; it is not about a system or a solution to a problem.

15.4. Requirements

User requirements are used in software engineering to specify user's expectations on what the software is able to do (see also item VII). Unlike system requirements, which express features with no necessary reference to users, user requirements are user-centered. They express criteria for system's functioning from a user's perspective. Requirements for a medical device might state, for example, that the device's must be able report when its sensor or battery is not operational, or allow quick and hygienic cleaning. Requirements can be expressed in many ways, either non-technically for a client, or technically for a developer. They are normally prioritized, both from business and end-user perspectives.

User requirements must be based on data and traceable. Triangulation is commonly recommended for identifying requirements. Different data sources provide complementary information that should be cross-checked when forming requirements. For example, contextual inquiry may provide background for a requirement, which can be then elaborated in a focus group or interview. In larger software projects, requirements elicitation include interviews, observations, focus groups and workshops, brainstorming, and possibly prototyping. Requirements are at times gathered continuously, first during product launch, and then via user feedback. User requirements are in some cases reviewed by users themselves. Confirming requirements with stakeholders, in general, is recommendable especially in professional domains. Traceability means that we know the observations a requirement is based on.

A requirement should be associated with *verification criteria*: a way of identifying whether a product fulfils the requirement. A verification criterion may be empirically defined, for example *'95 % of representative users are able to complete task A within time T with no significant error'*. Vague criteria – like *'Feature F is easy to use'* – are problematic, because they leave fulfilment open to disagreement. Being as precise as possible is important, because user requirements can be contractual: a client and the developer agree on what a system must be able to do via requirements. After agreeing on the user requirements, the client is not expected to demand new features. On the other hand, the product is not considered ready if it does not meet the requirements. User requirements are also essential for planning, management, and communication of complex projects.

Use cases are a technique for eliciting and specifying requirements. The concept of use case was developed in mid-1990s to help develop interactive systems. In software engineering, use cases are also used to denote cases where the 'user' is non-human, typically an external system. *User stories* are informally expressed use cases. They are narrative-like descriptions of system use written from the perspective of an end user.

A use case enumerates actions that a user must take to achieve a goal. Several frameworks have been proposed for expressing them, but in the minimum, a case should state the initial state or preconditions for the case to occur, the user's goal, the flow of events, and post-conditions, anything that must be done after achieving the goal. So-called *abstract use cases* define user intents but no sequences (flow). Use cases are expressed diagrammatically, in tabular, in storyboards, or in a verbal format. Similarly as user requirements, there is often a need to associate verification criteria.

15.5. Can Representations be Both Valid and Informative?

Representations of user research data are supposed to inform practical decisions. How do personas, scenarios, task analyses etc. actually achieve this? They achieve this via two quite different routes: the realist and the instrumentalist. Balancing the two is a key challenge for user research.

According to the first view, the realist view, user research produces empirically grounded statements about users. A representation of user data is essentially a proposition, a claim about users. The claim that 'User Segment A has a high need for security' is a claim that that can be true or false. That the statements we make about users are valid is important for efforts at design and engineering. Decisions about which features to include or how to design can be thought as hypotheses about users. When those hypotheses are wrong, the design will fail, too.

According to the second view, the instrumentalist view, user research has instrumental value in inspiring design. The results of user research do not need to be valid in the realist sense – whatever helps design is valuable. In creative efforts where the goal is to envision novel possibilities, it may be detrimental to be tied too tightly to the present reality. Taken to extreme, from this viewpoint, even exaggerations and embellishments of data are fine, as long as they inspire new ideas. According to this view, user research does not need to seek a firm connection between a representation and the world it represents.

In practice, user research often serves both purposes. In projects where the goal is to inform decisions, observational data is routinely selected, redescribed, and augmented using other available knowledge. We saw this for example in how personas and scenarios are created. Data are also interpreted. An analyst may flag something as surprising, or write in points about users that are 'valuable'. Such 'inferential leaps', even if small, pose a challenge: On the one hand, leaps may be necessary to produce knowledge that informs practical decisions; On the other, such leaps dissociate the analyst from observational data, they make the conclusions less grounded. Practical decisions that are based on unsubstantiated or subjective inferences may be not only wrong but also possibly damaging, for example by disregarding or simplifying certain user groups.

How to balance between these views? We hold that verifiability and traceability are important in all user research projects, even those with instrumentalist goals. *Verifiability* refers to how well the claim can be cross-checked with observations that are independent of the original dataset. For example, in task analysis, it is generally recommended to have task analysis diagrams checked by those stakeholders whose work it describes. If verification is hard or impossible, it is important to be able to trace the reasoning that led to the interpretation. *Traceability* refers to the documentation of the reasonings that led from original data to the final claim. For example, when writing up personas, one should document the data and other knowledge they were based on. Good practices like these lend credibility to user research and ensures that stakeholders pay heed to its results. Traceability is also important for accountability and proper representation of the involved people.

Summary

- Representations of user research serve to capture and reflect upon findings and help prepare for design.
- They summarize basic aspects of user data: users, events, tasks, activities, and

contexts.

- Representations should be traceable and verifiable.

Exercises

1. Representation techniques. Take any recent product launch by an IT company. Write up a 1) persona, 2) scenario, 3) task analysis diagram, 4) customer journey map, and 5) a rich picture diagram for the product.

2. The value of representations. If you have original data, why bother with second-order representations like personas and scenarios? Are we not always bound to lose information when doing representations? Discuss the pros and cons of representations.

3. LLMs and user research. Large language models (e.g,. ChatGPT) can be used to generate personas and scenarios. For example, you can prompt ChatGPT to generate a persona with desired characteristics or even with real transcripts from interviews. Discuss how valid and informative such representations can be. Discuss both pros and cons.