Chapter 12

Observing users

12.1 Introduction

Observation involves watching and listening to users. Observing users interacting with software, even casual observing, can tell you an enormous amount about what they do, the context in which they do it, how well technology supports them, and what other support is needed. In Chapter 9 we discussed the role of observation and ethnography in informing design, particularly early in the process. In this chapter we describe how to observe and do ethnography and discuss their role in evaluation.

Users can be observed in controlled laboratory-like conditions, as in usability testing, or in the natural environments in which the products are used—i.e., the field. How the observation is done depends on why it is being done and the approach adopted. There is a variety of structured, less structured, and descriptive

observation techniques for evaluators to choose from. Which they select and how their findings are interpreted will depend upon the evaluation goals, the specific questions being addressed, and practical constraints. This chapter focuses on how to select appropriate observation techniques, how to do observation, and how to analyze the data and present findings from it. We also discuss the benefits and practicalities associated with each technique. An interview with interaction design consultant Sara Bly at the end of the chapter discusses how she uses observation in her work.

The main aims of this chapter are to:

- Discuss the benefits and challenges of different types of observation.
- Describe how to observe as an on-looker, a participant, and an ethnographer.
- Discuss how to collect, analyze and present data from observational evaluation.
- Examine key issues for doing think-aloud evaluation, diary studies and interaction logging.
- Give you experience in selecting and doing observational evaluation.

In general, observing and talking to users usually go together, but we leave the details of interview techniques until Chapter 13.

12.2 Goals, questions, and paradigms

Goals and questions provide a focus for observation, as the DECIDE framework points out. Even studies that use "quick and dirty" observations have a goal; for example, to identify or confirm usability and user experience goals in a prototype. *Goals and questions should guide all evaluation studies.* Just because some evaluators do not make their goals obvious does not mean that they don't have goals. Expert evaluators sometimes don't articulate their goals, but as you will read in Sara Bly's interview they do have them. Even in field studies and ethnography there is a careful balance between being guided by goals and being open to modifying, shaping, or refocusing the study as you learn about the situation. Being able to keep this balance is a skill that develops with experience.

ACTIVITY 12.1

(a) Find a small group of people who are using any kind of technology (e.g., computers, household or entertainment appliances, etc.) and try to answer the question, "What are these people doing?" Watch for three to five minutes and write down what you observe. When you have finished, note how you felt doing this.

(b) If you were to repeat the exercise what would you look for when you next observe the group? How would you refine your goals?

Comment

(a) What was the group doing? Were they talking, working, playing or something else? How were you able to decide? Did you feel awkward or embarrassed watching? Did you wonder whether you should tell them that you were observing them? What problems did you encounter doing this exercise? Was it hard to watch everything and re-

member what happened? What were the most important things? Did you wonder if you should be trying to identify and remember just those things? Was remembering the order of events tricky? Perhaps you naturally picked up a pen and paper and took notes. If so, was it difficult to record fast enough? How do you think the people being watched felt? Did they know they were being watched? Did knowing affect the way they behaved? Perhaps some of them objected and walked away. If you didn't tell them, do you think you should have?

(b) Your questions should be more focused. For example, you might ask, what are the people specifically trying to do and how is the technology being used? Is everyone in the group using the technology? Is it supporting or hindering the users' goals?

Having a goal, even a very general goal, helps to guide the observation because there is always so much going on.

12.2.1 What and when to observe

Observing is useful at any time during product development. Early in design, observation helps designers understand users' needs. Other types of observation are done later to examine whether the developing prototype meets users' needs.

Depending on the type of study, evaluators may be onlookers, participant observers, or ethnographers. Remember Christian Heath's and Paul Luff's ethnographic study of the London Underground discussed in Chapter 4 (Heath and Luff, 1992)? This study demonstrates the power of insightful observation to improve the redesign of a system. However, in order to understand how London Underground workers do their jobs the authors needed "insider" knowledge. The degree of immersion that evaluators adopt varies across a broad outsider-insider spectrum. Where a particular study falls along this spectrum depends on its goal and on the practical and ethical issues that constrain and shape it.

ACTIVITY 12.2 To understand this notion of an outsider-insider spectrum better, read the scenarios below and answer the questions that follow.

Scenario 1. A usability consultant joins a group who have been given WAP phones to test on a visit to Washington, DC. Not knowing the restaurants in the area, they use the WAP phone to find a list of restaurants within a five-mile radius of their hotel. Several are listed and while the group waits for a taxi, they find the telephone numbers of a couple, call them to ask about their menus, select one, make a booking, and head off to the restaurant. The usability consultant observes some problems keying instructions because the buttons seem small. She also notices that the screen seems rather small, but the person using it is able to get the information needed and call the restaurant, etc. Discussion with the group supports the evaluator's impression that there are problems with the interface, but on balance the device is useful and the group is pleased to get a table at a good restaurant nearby.

Scenario 2. A usability consultant observes how participants perform a pre-planned task using the WAP phone in a usability laboratory. The task requires the participants to find the telephone number of a restaurant called Matisse. It takes them several minutes to do this

and they appear to have problems. The video recording and interaction log suggest that the screen is too small for the amount of information they need to access and this is supported by participants' answers on a user satisfaction questionnaire.

(a) In which situation does the observer take the most control?

(b) What are the advantages and disadvantages of these two types of observation?

(c) When might each type of observation be useful?

Comment

(a) The observer takes most control in the second study. The task is predetermined, the participant is instructed what to do, and she is located in a controlled laboratory environment.

(b) The advantages of the field study are that the observer got to see how the device could be used in a real situation to solve a real problem. She experienced the delight expressed with the overall concept and the frustration with the interface. By watching how the group used the device "on the move," she gained an understanding of what they liked and needed. The disadvantage is that the observer was an "insider" in the group, so how objective could she be? The data is qualitative and while anecdotes can be very persuasive, how useful are they in evaluation? Maybe she was having such a good time that her judgment was clouded and she missed hearing negative comments and didn't notice some people's annoyance. Another study could be done to find out more, but it is not possible to replicate the exact situation, whereas the laboratory study is easier to replicate.

The advantages of the laboratory are that several users performed the same task, so different users' performance could be compared and averages calculated. The observer could also be more objective because she was more of an outsider. The disadvantage is that the study is artificial and says nothing about how the device would be used in the real environment.

(c) Both types of studies have merits. Which is better depends on the goals of the study. The laboratory study is useful for examining details of the interaction style to make sure that usability problems with the interface and button design are diagnosed and corrected. The field study reveals how the phone is used in a real world context and how it integrates with or changes users' behavior. Without this study, it is possible that developers **might** not have discovered the enthusiasm for the phone because the reward for doing laboratory tasks is not as compelling as a good meal!

Table 12.1 Type of observation

Observation	Controlled environment (i.e., lab-like)	Field environment (i.e., natural)
Outsider looking on	"Quick and dirty" In usability testing	"Quick and dirty" In field studies
Insider	(Not applicable)	Participant observation (e.g., in ethnography)

Table 12.1 summarizes this insider-outsider discussion, how it relates to different types of environments, and how much control evaluators take over the evaluation process.

12.2.2 Approaches to observation

Observers can be outsiders in the field and in the controlled environments, but they can't be insiders in a controlled environment. In the field it is possible to have varying degrees of "insider-outsiderness." In practice these distinctions are more difficult to describe than to experience!

"Quick and dirty" observation

"Quick and dirty" observations can occur anywhere, anytime. For example, evaluators often go into a school, home, or office to watch and talk to users in a casual way to get immediate feedback about a prototype or product. Evaluators can also join a group for a short time, which gives them a slightly more insider role. Quick and dirty observations are just that, ways of finding out what is happening quickly and with little formality.

Observation in usability testing

Video and interaction logs capture everything that the user does during a usability test including keystrokes, mouse clicks, and their conversations. In addition, observers can watch through a one-way mirror or via a remote TV screen. The observational data is used to see and analyze what users do and how long they spend on different aspects of the task. It also provides insights into users' affective reactions. For example, sighs, tense shoulders, frowns, and scowls speak of users' dissatisfaction and frustrations. The environment is controlled but users often forget that they are being observed. In addition, many evaluators also supplement findings from the laboratory with observations in the field.

Observation in field studies

In field studies, as we have said, observers may be anywhere along the outsider-insider spectrum. Looking on as an outsider, being a participant observer, or being an ethnographer brings a philosophy and practices that influence what data is collected, how data collection is done, and how the data is analyzed and reported. Colin Robson (1993) summarizes the possible levels of participation as: complete participants, more marginal participants, observers who also participate, and people who observe from the outside and do not participate.

Whether and in what ways observers influence those being observed depends on the type of observation and the observer's skills. The goal is to cause as little disruption as possible. An example of outsider observation is when an observer is interested only in the presence of certain types of behavior. For instance, in a study

of the time spent by boys and girls using technology in the classroom, an observer may go into the classroom to note when technology is used by boys and when by girls. She could do this by standing at the back of the room with a data sheet on which she notes the gender of the children who use the computer and how long they spend using it. In contrast, if the goal is to understand how the computer integrates with other artifacts and social interactions in the classroom, a more holistic approach would be better. In this situation the evaluator might take more of an insider perspective in which she talks to participants as well as observes. The observer mixes and integrates with participants more, but there is no illusion that she is anything other than an observer.

Inside observers may be participant observers or ethnographers. In participant observation evaluators participate with users in order to learn what they do and how and why they do it. A fully participant observer observes from the inside as a member of the group, which means she must not only be present to share experiences, but also learn the social conventions of the group, including beliefs and protocols, dress codes, communication conventions, use of language, and non-verbal communication. "Participant observation combines participation in the lives of the people under study with maintenance of a professional distance that allows adequate observation and recording of data" (Fetterman, 1998, p. 34–35).

Ethnographers can be thought of as participant observers or not, depending on your point of view. Ethnographers themselves debate this issue. Some see participant observation as virtually synonymous with ethnography (Atkinson and Hammersley, 1994). Others view participant observation as a technique that is used in ethnography along with informants from the community, interviews with community members, and the study of community artifacts (Fetterman, 1998). Ethnographic evaluation is derived from ethnography. Ethnographic studies typically take weeks, months, or even longer to gain an "inside" understanding of what is going on in a community. Much shorter studies are usual in interaction design because of the time constraints imposed by development schedules.

As in any evaluation study, goals and questions determine whether the observation will be "quick and dirty," in a controlled environment or in the field, and the extent to which the observers are outsiders or insiders. Determining goals, exploring questions, and choosing techniques are necessary steps in the DECIDE framework. Practical and ethical issues also have to be identified and decisions made about how to handle them.

12.3 How to observe

The same basic data-collection tools are used for laboratory and field studies (i.e., direct observation, taking notes, collecting video, etc.) but the way in which they are used is different. In the laboratory the emphasis is on the details of what individuals do, while in the field the context is important and the focus is on how people interact with each other, the technology, and their environment. Furthermore, the equipment in the laboratory is usually set up in advance and is relatively static, whereas in the field it usually must be moved around. In this section we discuss how to observe, and then examine the practicalities and compare data-collection tools.

12.3.1 In controlled environments

The role of the observer is to first collect and then make sense of the stream of data on video, audiotapes, or notes made while watching users in a controlled environment. Many practical issues have to be thought about in advance, including the following.

- It is necessary to decide where users will be located so that the equipment can be set up. Many usability laboratories, for example, have two or three wall-mounted, adjustable cameras to record users' activities while they work on test tasks. One camera might record facial expressions, another might focus on mouse and keyboard activity, and another might record a broad view of the participant and capture body language. The stream of data from the cameras is fed into a video editing and analysis suite where it is annotated and partially edited. Another form of data that can be collected is an interaction log. This records all the user's key presses. Mobile usability laboratories, as the name suggests, are intended to be moved around, but the equipment can be bulky. Usually it is taken to a customer's site where a temporary laboratory environment is created.
- The equipment needs testing to make sure that it is set up and works as expected, e.g., it is advisable that the audio is set at the right level to record the user's voice.
- An informed consent form should be available for users to read and sign at the beginning of the study. A script is also needed to guide how users are greeted, and to tell them the goals of the study, how long it will last, and to explain their rights. It is also important to make users feel comfortable and at ease.

Whether in a real or make-do laboratory one of the problems with this type of observation is that the observer doesn't know what users are thinking, and can only guess from what she sees.

Think-aloud technique Imagine observing someone who has been asked to evaluate the interface of the web search engine Northernlight. The user, who has used the web only once before, is told to find a list of the books written by the well-known biologist Stephen Jay Gould. He is told to type http://www.northernlight.com and then proceed however he thinks best. He types the URL and gets a screen similar to the one in Figure 12.1.

Next he goes to the search box but types Stephen Jay Gouild without realizing that he has made a typing error and added an 'i'. He presses return and gets a screen similar to the one in Figure 12.2.

He is silent. What is going on, you wonder? What is he thinking? One way around this problem is to collect a think-aloud protocol, using a technique developed by Erikson and Simon for examining people's problem-solving strategies (Erikson and Simon, 1985). The technique requires people to say out loud everything that they are thinking and trying to do, so that their thought processes are externalized.

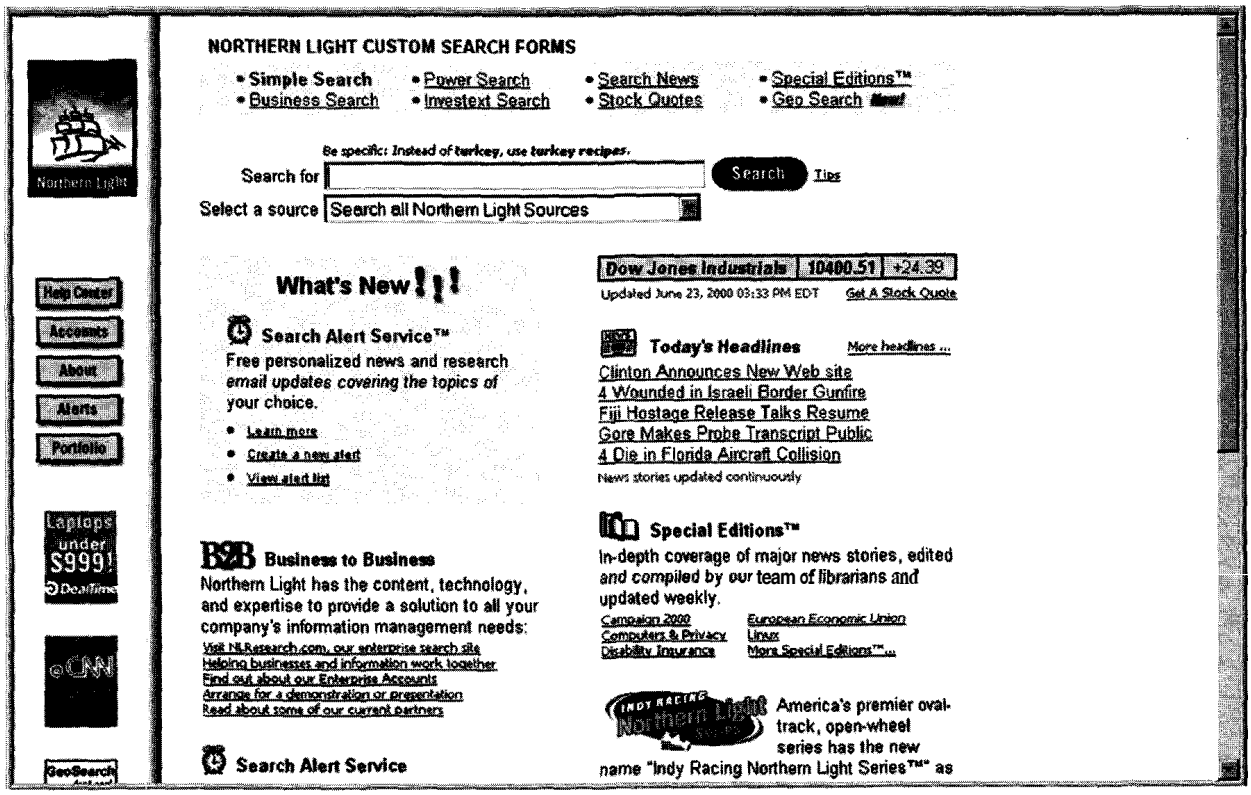


Figure 12.1 Home page of Northernlight search engine (www.northernlight.com).

So, let's imagine an action replay of the situation just described, but this time the user has been instructed to think aloud:

I'm typing in http://www.northernlight.comas you told me. (types)
Now I press the enter key, right? (presses enter key)
(pause and silence)
It's taking a few moments to respond.
Oh! Here it is. (Figure *12.1* appears)
Gosh, there's a lot of stuff on this screen, hmmm, I wonder what I do next. (pauses and looks at the screen) *Probably a simple search. What's apower search and there's all these others too?*
I just want to find Stephen Jay Gould, right, and then it's bound to have a list of his books? (pause) *Well, it looks like I should type his name in this box here.* (moves cursor towards the search box. Positions cursor. Types 'Stephen Jay Gouild'. Pauses, but does not notice that he has incorrectly included an "i" in Gould, then clicks the search button.) *Well, something seems to be happening...* (Watches) *something is happening. Ah! What's this...* (Looks at screen and Figure 12.2 appears)
Silence...

Now you know more about what the user is trying to achieve but he is silent again. You can see that he has spelled Gould incorrectly and that he doesn't realize that he has typed Gouild. What you don't know is what he is thinking now or what is he

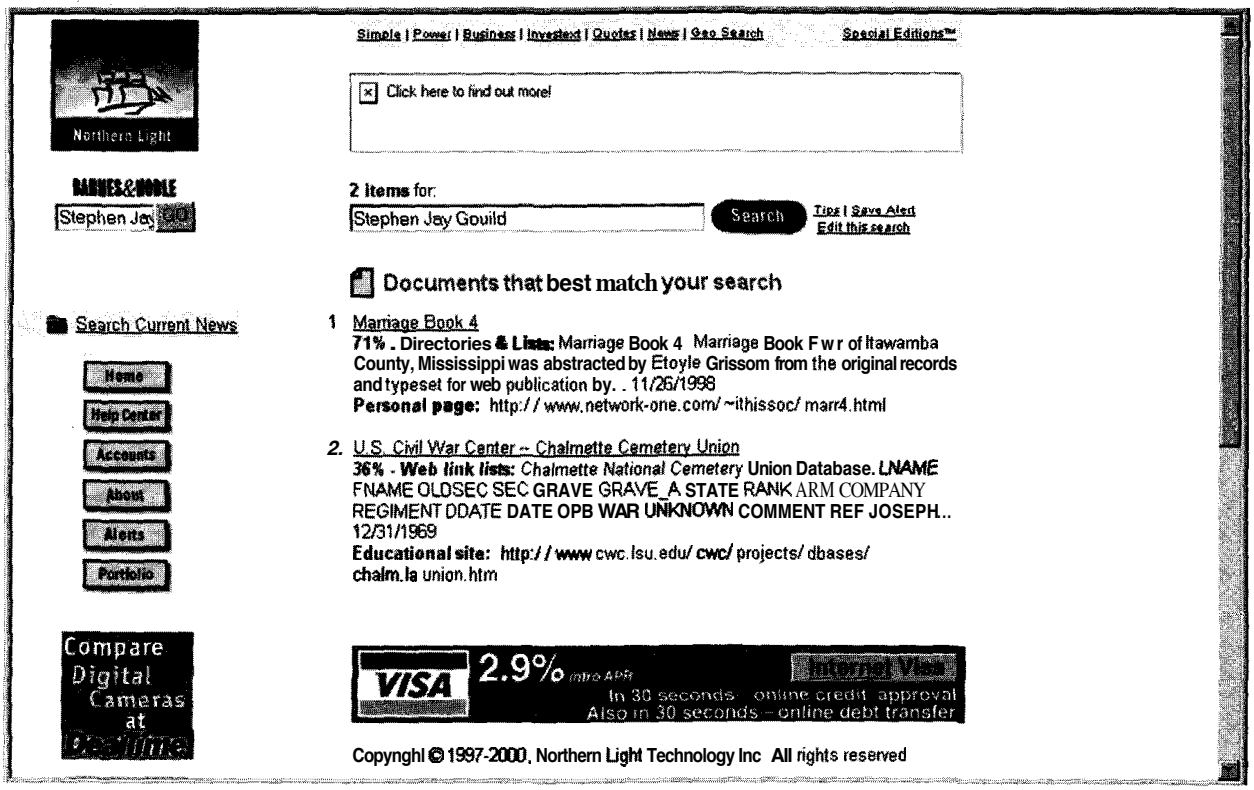


Figure 12.2 The screen that appears in response to searching for Stephen Jay Gould.

looking at. Has he noticed his typing error or the Barnes and Noble box at the top left that says "Stephen Jay"?

ACTIVITY 12.3 Try a think-aloud exercise yourself. Go to an e-commerce website, such as Amazon.com or BarnesandNoble.com, and look for something that you want to buy. Think aloud as you search and notice how you feel and behave. Did you find it difficult to keep speaking all the way through the task? Did you feel awkward? Did you stop when you got stuck?

Comment You probably felt self-conscious and awkward doing this. Some people say they feel really embarrassed. At times you may also have started to forget to speak out loud because it feels like talking to yourself, which most of us don't do. You may also have found it difficult to think aloud when the task got difficult. In fact, you probably stopped speaking when the task became demanding, and that is exactly the time when an evaluator is most eager to hear your comments.

The occurrence of these silences is one of the biggest problems with the think-aloud technique.

If a user is silent during a think-aloud protocol, the evaluator could interrupt and remind him to think out loud, but that would be intrusive. Another solution is

to have two people work together so that they talk to each other. Working with another person is often more natural and revealing because they talk in order to help each other along. This technique has been found particularly successful with children. It is also very effective when evaluating systems intended to be used synchronously by groups of users, **e.g.**, shared whiteboards.

12.3.2 In the field

Whether the observer sets out to be an outsider or an insider, events in the field can be complex and rapidly changing. There is a lot for evaluators to think about, so many experts have a framework to structure and focus their observation. The framework can be quite simple. For example, this is a practitioner's framework that focuses on just three easy-to-remember items to look for:

- ***The person.*** Who is using the technology at any particular time?
- ***The place.*** Where are they using it?
- ***The thing.*** What are they doing with it?

Frameworks like the one above help observers to keep their goals and questions in sight. Experienced observers may, however, prefer more detailed frameworks, such as the one suggested by Goetz and LeCompte (1984) below, which encourages observers to pay greater attention to the context of events, the people and the technology:

- ***Who*** is present? How would you characterize them? What is their role?
- ***What*** is happening? What are people doing and saying and how are they behaving? Does any of this behavior appear routine? What is their tone and body language?
- ***When*** does the activity occur? How is it related to other activities?
- ***Where*** is it happening? Do physical conditions play a role?
- ***Why*** is it happening? What precipitated the event or interaction? Do people have different perspectives?
- How is the activity organized? What rules or norms influence behavior?

Colin Robson (1993) suggests a slightly longer but similar set of items:

- ***Space.*** What is the physical space like and how is it laid out?
- ***Actors.*** What are the names and relevant details of the people involved?
- ***Activities.*** What are the actors doing and why?
- ***Objects.*** What physical objects are present, such as furniture?
- ***Acts.*** What are specific individuals doing?

  Events. Is what you observe part of a special event?

  Goals. What are the actors trying to accomplish?
- ***Feelings.*** What is the mood of the group and of individuals?

ACTIVITY 12.4

(a) Look at Goetz's and LeCompte's framework. Apart from there being more items than in the first framework, what is the other main difference?

(b) Now compare this framework with Robson's. What does Robson's attend to that is not obvious in Goetz's and LeCompte's framework?

(c) Which of the three frameworks do you think would be easiest to remember and why?

Comment

(a) The Goetz and LeCompte framework pays much more attention to the context of the observation.

(b) There is considerable overlap between the two frameworks despite differences in wording. The main difference is that Robson's framework pays attention to the mood of the group.

(c) The three-item framework is likely to be easy, but so is the Goetz and LeCompte framework because it adopts the much used organizing principle "who, what, when, where, why, how." Robson's framework has two extra items and no obvious way of remembering them. However, having said that, to me it is more explicit. Which is used for a particular study depends on the study goals and how much detail is needed, and to a degree, it is also a matter of personal preference.

These frameworks are useful not only for providing focus but also for organizing the observation and data-collection activity. Below is a checklist of things to plan before going into the field:

- State the initial study goal and questions clearly.
- Select a framework to guide your activity in the field.

  Decide how to record events—i.e., as notes, on audio, or on video, or using a combination of all three. Make sure you have the appropriate equipment and that it works. You need a suitable notebook and pens. A laptop computer might be useful but could be cumbersome. Although this is called observation, photographs, video, interview transcripts and the like will help to explain what you see and are useful for reporting the story to others.
- Be prepared to go through your notes and other records as soon as possible after each evaluation session to flesh out detail and check ambiguities with other observers or with the people being observed. This should be done routinely because human memory is unreliable. A basic rule is to do it within 24 hours, but sooner is better!
- As you make and review your notes, try to highlight and separate personal opinion from what happens. Also clearly note anything you want to go back to. Data collection and analysis go hand in hand to a large extent in fieldwork.
- Be prepared to refocus your study as you analyze and reflect upon what you see. Having observed for a while, you will start to identify interesting

phenomena that seem relevant. Gradually you will sharpen your ideas into questions that guide further observation, either with the same group or with a new but similar group.

- Think about how you will gain the acceptance and trust of those you observe. Adopting a similar style of dress and finding out what interests the group and showing enthusiasm for what they do will help. Allow time to develop relationships. Fixing regular times and venues to meet is also helpful, so everyone knows what to expect. Also, be aware that it will be easier to relate to some people than others, and it will be tempting to pay attention to those who receive you well, so make sure you attend to everyone in the group.
- Think about how to handle sensitive issues, such as negotiating where you can go. For example, imagine you are observing the usability of a portable home communication device. Observing in the living room, study, and kitchen is likely to be acceptable, but bedrooms and bathrooms are probably out of bounds. Take time to check what participants are comfortable with and be accommodating and flexible. Your choice of equipment for data collection will also influence how intrusive you are in people's lives.
- Consider working as a team. This can have several benefits; for instance, you can compare your observations. Alternatively, you can agree to focus on different people or different parts of the context. Working as a team is also likely to generate more reliable data because you can compare notes among different evaluators.

  Consider checking your notes with an informant or members of the group to ensure that you are understanding what is happening and that you are making good interpretations.
- Plan to look at the situation from different perspectives. For example, you may focus on particular activities or people. If the situation has a hierarchical structure, as in many companies, you will get different perspectives from different layers of management—e.g., end-users, marketing, product developers, product managers, etc.

12.3.3 Participant observation and ethnography

Being a participant observer or an ethnographer involves all the practical steps just mentioned, but especially that the evaluator must be accepted into the group. An interesting example of participant observation is provided by Nancy Baym's work (1997) in which she joined an online community interested in soap operas for over a year in order to understand how the community functioned. She told the community what she was doing and offered to share her findings with them. This honest approach gained her their trust, and they offered support and helpful comments. As Baym participated she learned about the community, who the key characters were, how people interacted, their values, and the types of discussions that were generated. She kept all the messages as data to be referred to later. She also

adapted interviewing and questionnaire techniques to collect additional information. She summarizes her data collection as follows (Baym, 1997, p. 104):

> ***The data for this study were obtained from three sources. In October 1991, I saved all the messages that appeared. . . . I collected more messages in 1993. Eighteen participants responded to a questionnaire Iposted. . . . Personal email correspondence with 10 other . . . participants provided further information. I posted two notices to the group explaining the project and offeringto exclude posts by those who preferred not to be involved. No one declined to participate.***

Using this data, Baym examined the group's technical and participatory structure, its emergent traditions, and its usage with the technology. As the work evolved, she shared its progress with the group members, who were supportive and helpful.

ACTIVITY 12.5 Drawing on your experience of using email, bulletin boards, UseNet News, or chat rooms, how might participant observation online differ from face-to-face participant observation?

Comment In online participant observation you don't have to look people in the eye, deal with their skepticism, or wonder what they think of you, as you do in face-to-face situations. What you wear, how you look, or the tone of your voice don't matter. However, what you say or don't say and how you say it are central to the way others will respond to you. Online you only see part of people's context. You usually can't see how they behave off line, how they present themselves, their body language, how they spend their day, their personalities, who is present but not participating, etc.

As we said the distinction between ethnography and participant observation is blurred. Some ethnographers believe that ethnography is an open interpretivist approach in which evaluators keep an open mind about what they will see. Others, such as David Fetterman from Stanford University, see a stronger role for a theoretical underpinning: "before asking the first question in the field the ethnographer begins with a problem, a theory or model, a research design, specific data collection techniques, tools for analysis, and a specific writing style" (Fetterman, 1998, p. 1). This may sound as if ethnographers have biases, but by making assumptions explicit and moving between different perspectives, biases are at least reduced. Ethnographic study allows ***multiple*** interpretations of reality; it is ***interpretivist.*** Data collection and analysis often occur simultaneously in ethnography, with analysis happening at many different levels throughout the study. The question being investigated is refined as more understanding about the situation is gained.

The checklist below (Fetterman, 1998) for doing ethnography is similar to the general list just mentioned:

- Identify a problem or goal and then ask good questions to be answered by the study, which may or may not invoke theory depending on your philosophy of ethnography. The observation framework such as those mentioned above can help to focus the study and stimulate questions.

- The most important part of fieldwork is just being there to observe, ask questions, and record what is seen and heard. You need to be aware of people's feelings and sensitive to where you should not go.
- Collect a variety of data, if possible, such as notes, still pictures, audio and video, and artifacts as appropriate. Interviews are one of the most important data-gathering techniques and can be structured, semi-structured, or open. So-called ***retrospective interviews*** are used after the fact to check that interpretations are correct.
- As you work in the field, be prepared to move backwards and forwards between the broad picture and specific questions. Look at the situation holistically and then from the perspectives of different stakeholder groups and participants. Early questions are likely to be broad, but as you get to know the situation ask more specific questions.
- Analyze the data using a ***holistic*** approach in which observations are understood within the broad context—i.e., they are *contextualized.* To do this, first synthesize your notes, which is best done at the end of each day, and then check with someone from the community that you have described the situation accurately. Analysis is usually iterative, building on ideas with each pass.

ACTIVITY 12.6 Look at the steps listed for doing ethnography and compare them with the earlier generic set for field observation (see Section 12.3.2). What is the main difference?

Comment

Both sets of steps involve structuring observations and refining goals and questions through knowledge gained during the study. Both use similar data collection techniques and rely on the trust and cooperation of those being observed. Ethnographers tend to be deeply immersed in the group, whereas not everyone doing field studies takes this approach. Some ethnographers, such as David Fetterman, are guided by theory; others are strongly against this and believe that ethnography should be approached open-mindedly.

During the last ten years ethnography has gained credibility in interaction design because if products are to be used in a wide variety of environments designers must know the context and ecology of those environments (Nardi and O'Day, 1999). However, for those unfamiliar with ethnography and general field observation there are two dilemmas. The first dilemma is, "When have I observed

DILEMMA When Should I Stop Observing?

Knowing when to stop doing *any* type of evaluation can be difficult for novice evaluators, but it is particularly tricky in observational studies and ethnography because there is no obvious ending. Schedules often dictate when your study ends. Otherwise, stop when you stop learning new things. Two indications of having done enough are when you start to see similar patterns of behavior being repeated, or when you have listened to all the main stakeholder groups and understand their perspectives.

DILEMMA How Can I Adapt Ethnography to Fit the Development Process?

Many developers are unsure how to integrate ethnographic evaluation into development cycles. In addition, most developers have a technical training that does not encourage them to value qualitative data. We discussed the use of ethnography to inform design in Chapter 9. Here is an example where it has been adapted for evaluation.

In a project for the Department of Juvenile Justice, Ann Rose and her colleagues developed a procedure to be used by technical design teams with limited ethnographic training (Rose et al., 1995). This applied form of ethnography acknowledges the comparatively small amounts of time available for any kind of user study. By making the process more structured the amount of time needed for the study can be reduced. It also emphasizes that taking time to become familiar with the intricacies of a system enhances the evaluator's credibility during the field study and promotes productive fieldwork. The procedures this group advocates are highly structured, and while they may seem contrary to ethnographic practice, this structure helps to make it possible for some development teams to benefit from an applied ethnographic approach. There are four stages, as follows:

1 Preparation
- Understand organization policies and work culture.
- Familiarize yourself with the system and its history.
- Set initial goals and prepare questions.
- Gain access and permission to observe and interview.

2 Field study
- Establish a rapport with managers and users.
- Observe and interview users in their workplace and collect data.
- Follow any leads that emerge from the visits.
- Record your visits.

3 Analysis
- Compile the collected data in numerical, textual, and multimedia databases.
- Quantify data and compile statistics.
- Reduce and interpret the data.
- Refine the goals and processes used.

4 Reporting
- Consider multiple audiences and goals.
- Prepare a report and present the findings.

enough?" The second dilemma is, "How can I adapt ethnography so that it better fits the short development cycles and the mindset of the developers?"

ACTIVITY 12.7 What are the main differences between the stages that Rose et al. (1995) describe and the steps suggested by Fetterman (1998)?

Comment The list in the "How Can I Adapt Ethnography" dilemma suggests that the evaluators are not as immersed in the study as Fetterman's process suggests. One aim of the Rose procedure is radically to reduce the time needed to do a study so that it is compatible with system development. Another aim is to reduce the data to a quantifiable form so that it is familiar and acceptable to the developers.

12.4 Data collection

Data collection techniques (i.e., taking notes, audio recording, and video recording) are used individually or in combination and are often supplemented with

photos from a still camera. When different kinds of data are collected, evaluators have to coordinate them; this requires additional effort but has the advantage of providing more information and different perspectives. Interaction logging and participant diary studies are also used, as we discuss later in Section 12.5. Which techniques are used will depend on the context, time available, and the sensitivity of what is being observed. In most settings, audio, photos, and notes will be sufficient. In others it is essential to collect video data so as to observe in detail the intricacies of what is going on.

12.4.1 Notes plus still camera

Taking notes is the least technical way of collecting data, but it can be difficult and tiring to write and observe at the same time. Observers also get bored and the speed at which they write is limited. Working with another person solves some of these problems and provides another perspective. Handwritten notes are flexible in the field but must be transcribed. However, this transcription can be the first step in data analysis, as the evaluator must go through the data and organize it. A laptop computer can be a useful alternative but it is more obtrusive and cumbersome, and its batteries need recharging every few hours. If a record of images is needed, photographs, digital images, or sketches are easily collected.

12.4.2 Audio recording plus still camera

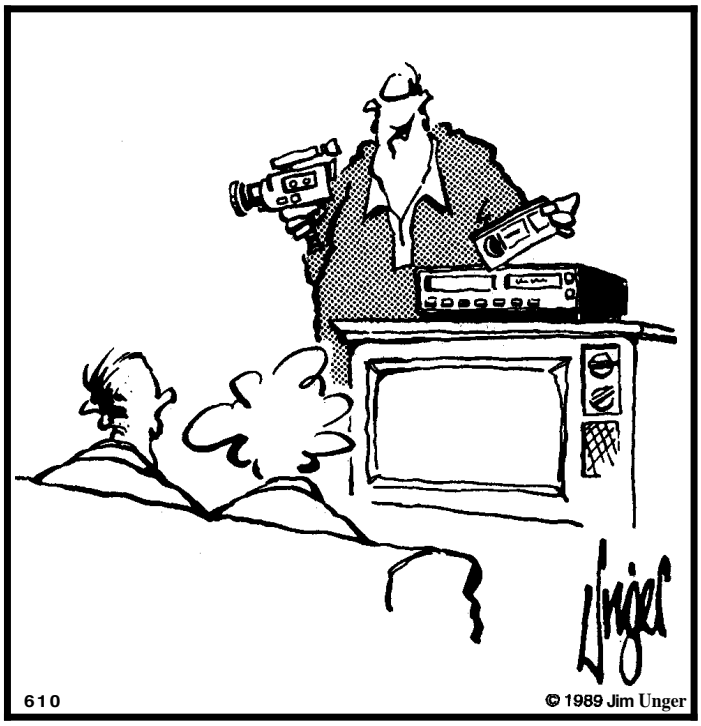
Audio can be a useful alternative to note taking and is less intrusive than video. It allows evaluators to be more mobile than with even the lightest, battery-driven video cameras, and so is very flexible. Tapes, batteries, and the recorder are now relatively inexpensive but there are two main problems with audio recording. One is the lack of a visual record, although this can be dealt with by carrying a small camera. The second drawback is transcribing the data, which can be onerous if the contents of many hours of recording have to be transcribed; often, however, only sections are needed. Using a headset with foot control makes transcribing less onerous. Many studies do not need this level of detail; instead, evaluators use the recording to remind them about important details and as a source of anecdotes for reports.

12.4.3 Video

Video has the advantage of capturing both visual and audio data but can be intrusive. However, the small, handheld, battery-driven digicams are fairly mobile, inexpensive and are commonly used.

A problem with using video is that attention becomes focused on what is seen through the lens. It is easy to miss other things going on outside of the camera view. When recording in noisy conditions, e.g., in rooms with many computers running or outside when it is windy, the sound may get muffled.

Analysis of video data can be very time-consuming as there is so much to take note of. Over 100 hours of analysis time for one hour of video recording is common for detailed analyses in which every gesture and utterance is analyzed. However, this



"This is a video of you two watching the video of our vacation."

level of detail is usually not needed because evaluators often focus on particular episodes and use the whole recording only for contextual information and reference.

In Table 12.2 we summarize the key features, advantages and drawbacks of these three combinations of data collection techniques.

ACTIVITY 12.8

Imagine you are a consultant who is employed to help develop a new computerized garden-planning tool to be used by amateur and professional garden designers. Your goal is to find out how garden designers use an early prototype as they walk around their clients' gardens sketching design ideas, taking notes, and asking the clients about what they like and how they and their families use the garden. What are the advantages and disadvantages of the three types of data-collection techniques in this environment?

Comment

Handwritten notes do not require specialist equipment. They are unobtrusive and very flexible but difficult to do while walking around a garden. If it starts to rain there is no equipment to get wet, but taking notes is tiring, people lose concentration, biases creep in, and handwriting can be difficult to decipher. Video captures more information (e.g., the landscape, where the designers are looking, sketches, comments, etc.) but it is more intrusive, you must also carry equipment and film and what happens if it starts to rain? You also need access to

Table 12.2 Comparison of the three main data-collection techniques used in observation

Criterion	Notes plus camera	Audio plus camera	Video
Equipment	Paper, pencil and camera are easily available.	Inexpensive, handheld recorder with a good microphone. Headset useful for easy transcription.	More expensive. Editing, mixing and analysis equipment needed.
Flexibility of use	Very flexible. Unobtrusive.	Flexible. Relatively unobtrusive.	Needs positioning and focusing camera lens. Even portable versions can be bulky.
Completeness of data	Only get what note-taker thinks is important and can record in the time available. Problem with inexperienced evaluators.	Can obtain complete audio recording but visual data is missing. Notes, photographs, sketches can augment recording but need coordinating with the recording.	Most complete method of data collecting, especially if more than one camera used, but coordination of video material is needed.
Disturbance to users	Very low.	Low but cassette must be changed and microphone positioned.	Can be very obtrusive. Care needed to avoid Hawthorne effect.
Reliability of data	May be low. Relies on humans making a good record and knowing what to record.	High but external noise, e.g. fans in computers can muffle what is said.	Can be high but depends on what camera is focused on.
Analysis	Relatively easy to transcribe. Rich descriptions can be produced. Transcribing data can be onerous or a useful first step in data analysis.	Critical discussions can be identified. Transcription needed for detailed analysis. Permanent original record that can be revisited.	Critical incidents can be identified and tagged. Automated support needed for detailed analysis. Permanent original record that can be revisited.
Feedback to design team	Relies strongly on the authority of the evaluator.	Material captured on tape is more convincing than notes but feedback relies on authority of evaluator.	Hard to dispute material captured on video. Video clips are very powerful for communicating ideas.

playback and editing facilities. Audio could be a good compromise, but integrating sketches and other artifacts later can be a burden and garden planning is a highly visual, aesthetic activity. You could also supplement notes and audio with a still camera.

12.5 Indirect observation: tracking users' activities

Sometimes direct observation is not possible because it is obtrusive or evaluators cannot be present over the duration of the study, and so users' activities are tracked indirectly. Diaries and interaction logs are two techniques for doing this. From the records collected evaluators reconstruct what happened and look for usability and user experience problems.

12.5.1 Diaries

Diaries provide a record of what users did, when they did it, and what they thought about their interactions with the technology. They are useful when users are scattered and unreachable in person, as in many Internet and web evaluations. Diaries are inexpensive, require no special equipment or expertise, and are suitable for long-term studies. Templates can also be created online to standardize entry format and enable the data to go straight into a database for analysis. These templates are like those used in open-ended online questionnaires. However, diary studies rely on participants being reliable and remembering to complete them, so incentives are needed and the process has to be straightforward and quick. Another problem is that participants often remember events as being better or worse than they really were, or taking more or less time than they actually did.

Robinson and Godbey (1997) asked participants in their study to record how much time Americans spent on various activities. These diaries were completed at the end of each day and the data was later analyzed to investigate the impact of television on people's lives. In another diary study, Barry Brown and his colleagues from Hewlett Packard collected diaries form 22 people to examine when, how, and why they capture different types of information, such as notes, marks on paper, scenes, sounds, moving images, etc. (Brown, et al., 2000). The participants were each given a small handheld camera and told to take a picture every time they captured information in any form. The study lasted for seven days and the pictures were used as memory joggers in a subsequent semi-structured interview used to get participants to elaborate on their activities. Three hundred and eighty-one activities were recorded. The pictures provided useful contextual information. From this data the evaluators constructed a framework to inform the design of new digital cameras and handheld scanners.

12.5.2 Interaction logging

Interaction logging in which key presses, mouse or other device movements are recorded has been used in usability testing for many years. Collecting this data is

usually synchronized with video and audio logs to help evaluators analyze users' behavior and understand how users worked on the tasks they set. Specialist software tools are used to collect and analyze the data. The log is also time-stamped so it can be used to calculate how long a user spends on a particular task or lingered in a certain part of a website or software application.

Explicit counters that record visits to a website were once a familiar sight. Recording the number of visitors to a site can be used to justify maintenance and upgrades to it. For example, if you want to find out whether adding a bulletin board to an e-commerce website increases the number of visits, being able to compare traffic before and after the addition of the bulletin board is useful. You can also track how long people stayed at the site, which areas they visited, where they came from, and where they went next by tracking their Internet Service Provider (I.S.P.) address. For example, in a study of an interactive art museum by researchers at the University of Southern California, server logs were analyzed by tracking visitors in this way (McLaughlin et al., 1999). Records of when people came to the site, what they requested, how long they looked at each page, what browser they were using, and what country they were from, etc., were collected over a seven-month period. The data was analyzed using Webtrends, a commercial analysis tool, and the evaluators discovered that the site was busiest on weekday evenings. In another study that investigated lurking behavior in listserver discussion groups, the number of messages posted was compared with list membership over a three-month period to see how lurking behavior differed among groups (Nonnecke and Preece, 2000).

An advantage of logging user activity is that it is unobtrusive, but this also raises ethical concerns that need careful consideration (see the dilemma about observing without being seen). Another advantage is that large volumes of data can be logged automatically. However, powerful tools are needed to explore and analyze this data quantitatively and qualitatively. An increasing number of visualization tools are being developed for this purpose; one example is WebLog, which dynamically shows visits to websites, as illustrated in Figure 12.3 (Hochheiser and Shneiderman, 2000).

DILEMMA **They Don't Know We Are Watching. Shall We Tell Them?**

If you have appropriate algorithms and sufficient computer storage, large quantities of data about Internet usage can be collected and users need never know. Furthermore, if we tell users that we are logging their behavior they may react or change their behavior. So, what should we do? It depends on the context, how much personal information is collected, and how the information will be used. Many companies now tell you that your computer activity and phone calls may be logged for quality assurance and other purposes. Most people do not object to this practice. However, should we be concerned about logging personal information (e.g., discussions about health or financial information)? Should users be worried? How can we exploit the ability to log user behavior when visiting websites without overstepping a person's civil rights? Where should we draw the line?

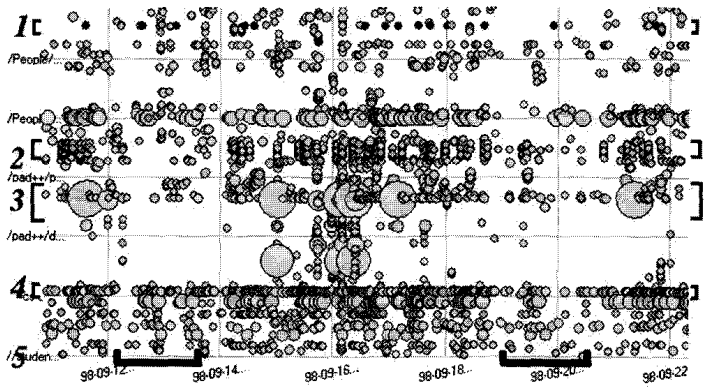


Figure 12.3 A display from WebLog, time vs. URL (Hochheiser and Shneiderman, 2001). The requested URL is on the y-axis, with the date and time on the x-axis. The dark lines on the x-axis correspond to weekends. Each circle represents a request for a single page, and the size of the circle indicates the number of bytes delivered for a given request. (Color, which is not shown here, indicates the Http status response.)

12.6 Analyzing, interpreting, and presenting the data

By now you should know that many, indeed most observational evaluations generate a lot of data in the form of notes, sketches, photographs, audio and video records of interviews and events, various artifacts, diaries, and logs. Most observational data is qualitative and analysis often involves interpreting what users were doing or saying by looking for patterns in the data. Sometimes qualitative data is categorized so that it can be quantified and in some studies events are counted.

Dealing with large volumes of data, such as several hours of video, is daunting, which is why it is particularly important to plan observation studies very carefully before starting them. The DECIDE framework suggests identifying goals and questions first before selecting techniques for the study, because the goals and questions help determine which data is collected and how it will be analyzed.

When analyzing any kind of data, the first thing to do is to "eyeball" the data to see what stands out. Are there patterns or significant events? Is there obvious evidence that appears to answer a question or support a theory? Then proceed to analyze it according to the goals and questions. The discussion that follows focuses on three types of data:

- *Qualitative data* that is *interpreted* and used to tell "the story" about what was observed.

  Qualitative data that is *categorized* using techniques such as content analysis.
- *Quantitative data* that is collected from interaction and video logs and presented as values, tables, charts and graphs and is treated statistically.

12.6.1 Qualitative analysis to tell a story

Much of the power of analyzing descriptive data lies in being able to tell a convincing story, illustrated with powerful examples that help to confirm the main points and will be credible to the development team. It is hard to argue with well-chosen video excerpts of users interacting with technology or anecdotes from transcripts.

In the interview with Sara Bly you will read about how she and her colleagues use data from several sources. At the end of each observation period they review their data, discuss what they observed, and construct a story from the data. This story evolves as more data is collected and more insights are generated. Teamwork plays an important role in this process because it provides different perspectives that can be compared. A large part of the analysis involves making "collections" of incidents or anecdotes that illustrate similar issues. For example, if several people comment at different times that it is hard to track down a manager in a particular work setting, these examples are powerful evidence of the need for better communication.

To summarize, the main activities involved in working with qualitative data to tell a story are:

- Review the data after each observation session to synthesize and identify key themes and make collections.
- Record the themes in a coherent yet flexible form, with examples. While post-its enable you to move ideas around and group similar ones, they can fall off and get lost and are not easily transported, so capture the main points in another form, either on paper or on a laptop, or make an audio recording.
- Record the date and time of each data analysis session. (The raw data should already be systematically logged with dates.)
- As themes emerge, you may want to check your understanding with the people you observe or your informants.
- Iterate this process until you are sure that your story faithfully represents what you observed and that you have illustrated it with appropriate examples from the data.
- Report your findings to the development team, preferably in an oral presentation as well as in a written report. Reports vary in form, but it is always helpful to have a clear, concise overview of the main findings presented at the beginning.

Analyzing and reporting ethnographic data Ethnographers work in a similar way but emphasize understanding events within the context in which they happen. Data is collected from participant observation, interviews, and artifacts, and analysis is continuous with great attention to detail. Ethnographers reconstruct knowledge to produce detailed descriptions known as ***rich*** or ***thick descriptions.*** In these descriptions, quotes, pictures, and anecdotes play a convincing role in communicating the findings to others. The main activities in analyzing ethno-

graphic data are similar to those just mentioned but notice the emphasis on detail (Fetterman, 1998):

- Look for key events within a group that speak about what drives the group's activity.
- Look for patterns of behavior in various situations and among different players. With experience, ethnographers build up sets of knowledge from various sources, asking questions, listening, probing, comparing and contrasting, synthesizing, and evaluating information.
- Compare sources of data against each other to provide consistent explanations.

Finally, report your findings in a convincing and honest way. Writing is part of the analysis since it helps to crystallize ideas.

Software tools, such as NUDIST and Ethnograph, allow ethnographers to code their notes and artifact descriptions so that they can be sorted, searched, and retrieved. For example, using NUDIST, field notes can be searched for key words or phrases and a report printed listing every occasion the word or phrase is used. The information can also be printed out as a tree showing the relationship of occurrences. Similarly, NUDIST can be used to search a body of text to identify specific predetermined categories or words for content analysis. The more copious the notes, the more useful tools like NUDIST are. Furthermore, many exploratory searches can be done to test hypotheses among different categories of data.

Other computerized tools support basic statistical analysis. For example, some data can be analyzed using statistical tests (such as chi-square contingency table analysis or rank correlation) to determine whether particular trends are significant.

12.6.2 Qualitative analysis for categorization

Data from think-aloud protocols, video, or audio transcripts can be analyzed in different ways. These can be coarse-grained or detailed analyses of excerpts from a protocol in which each word, phrase, utterance, or gesture is analyzed. Sometimes examining the comment or action in the context of other behavior is sufficient. In this section we discuss a selection of techniques. Some are used more often in research while others are used more for product development.

Looking for incidents or patterns

Analyzing even a short half-hour videotape would be very time-consuming if evaluators studied every comment or action in detail. Furthermore, such fine-grained analyses are often not necessary. A common strategy is to look for critical incidents, such as times when users were obviously stuck. Such incidents are usually marked by a comment, silence, looks of puzzlement, etc. Evaluators focus on these incidents and review them in detail, using the rest of the video as context to inform their analysis. For example, Jurgen Koenemann-Belliveau et al. (1994) used this approach to compare the efficacy of two versions of a Smalltalk

programming manual for supporting novice programmers. They used a form of critical incident analysis to examine breakdowns or problems in achieving a programming task and also to identify possible threats of incidents. This enabled them to identify specific problems that might otherwise have been overlooked. Taking this approach, they were able to trace through a sequence of incidents and achieve a more holistic understanding of the problem. For example they found that they needed to emphasize how objects interact in teaching object-oriented programming.

Theory may also be used to guide the study. Wendy Mackay et al. (2000) took this approach in analyzing a four-minute excerpt from a video of users working with a new software tool. Using Activity Theory to guide their analysis, they identified 19 shifts in attention between different parts of the tool interface and the task at hand. (In fact, some users spent so much time engaged in these shifts that they lost track of their original task.) Using the theory helped the evaluators to focus on relevant incidents.

Whether your analysis is coarse-grained or finer, whether you are guided by theory or are just looking for incidents and patterns of behavior, you need a way of handling your data and recording your analysis. For example, in another part of their study, Wendy Mackay et al. (2000) collected and analyzed video excerpts of users interacting with their tool and constructed a form of paper storyboards. The series of images taken from the video illustrated the changes made through the task, while the accompanying text descriptions provided details about the precise operations performed and the difficulties encountered.

A variety of tools are available to record, manipulate and search the data. NUDIST was mentioned above and Box 12.1 briefly describes the Observer Video-Pro tool. Typically reports from these analyses are fed back to the development team, often accompanied by video clips.

BOX 12.1 The Observer Video-Pro: An Automated Data Analysis Tool

The Observer Video-Pro provides the following features (Noldus, 2000):

- During preparation of a video tape recording, a *time code generator* adds an invisible time code to each video frame.
- During a data-collection session, a *time code reader* retrieves the time code from the tape, allowing frame-accurate event timing independent of the playback speed of the video cassette recorder (VCR).
- Each keyboard entry is firmly anchored to the video frame displayed at the instant the evaluator presses the first key of a behavior code or free-format note. The evaluator can also use a mouse to score events.
- Observational data can be reviewed and edited, with synchronized display of the corresponding video images.
- For optimal visual feedback during coding, the evaluator can display the video image in a window on the computer screen.
- The VCR can be controlled by the computer, allowing software-controlled "jog", "shuttle", and "search" functions.
- Video images can be captured and saved as disk files for use as illustrations in documents, slides for presentations, etc.
- Marked video episodes can be copied to an Edit Decision List for easy creation of highlight tapes.

ACTIVITY 12.9 What does the Observer Video-Pro tool allow you to search for in the data collected?

Comment Depending on how the logs have been annotated, using the Observer Video-Pro product, you can search the data for various things including the following:

Video time—A specific time, e.g., 02:24:36.04 (hh:mm: ss.dd).

Marker—A previously entered free-format annotation.

Event—A combination of actor, behavior, and modifiers, with optional wildcards (e.g., the first occurrence of "glazed look" or "Sarah approaches Janice").

Text—Any word or alphanumeric text string occurring in the coded event records or free-format notes.

Analyzing data into categories

Content analysis provides another fine grain way of analyzing video data. It is a systematic, reliable way of coding content into a meaningful set of mutually exclusive categories (Williams et al., 1988). The content categories are determined by the evaluation questions and one of its most challenging aspects is determining meaningful categories that are orthogonal—i.e., do not overlap each other in any way.

Deciding on the appropriate granularity is another issue to be addressed. The content categories must also be reliable so that the analysis can be replicated. This can be demonstrated by training a second person to use the categories. When training is complete, both researchers analyze the same data sample. If there is a large discrepancy between the two analyses, either training was inadequate or the categorization is not working and needs to be refined. By talking to the researchers you can determine the source of the problem, which is usually with the categorization. If so, then a better categorization scheme needs to be devised and re-tested by doing more inter-researcher reliability tests. However, if the researchers do not seem to know how to carry out the process then they probably need more training.

When a high level of reliability is reached, it can be quantified by calculating an ***inter-research reliability rating.*** This is the percentage of agreement between the two researchers, defined as the number of items that both categorized in the same way expressed as a percentage of the total number of items examined. It provides a measure of the efficacy of the technique and the categories.

Content analysis ***per se*** is not used very often in evaluations because it is very labor-intensive and time-consuming but a study by Maria Ebling and Bonnie John (2000) showed how useful it can be. They developed a hierarchical content classification for analyzing data when evaluating a graphical interface for a distributed file system.

Analyzing discourse

Another approach to video, and audio analysis is to focus on the dialog, i.e., the meaning of what is said, rather than the content. Discourse analysis is strongly interpretive, pays great attention to context, and views language not only as reflecting psychological and social aspects but also as constructing it (Coyle, 1995). An

underlying assumption of discourse analysis is that there is no objective scientific truth. Language is a form of social reality that is open to interpretation from different perspectives. In this sense, the underlying philosophy of discourse analysis is similar to that of ethnography. Language is viewed as a constructive tool and discourse analysis provides a way of focusing upon how people use language to construct versions of their worlds (Fiske, **1994).**

Small changes in wording can change meaning, as the following excerpts indicate (Coyle, **1995):**

> ***Discourse analysis is what you do when you are saying that you are doing discourse analysis....***
>
> ***According to Coyle, discourse analysis is what you do when you are saying that you are doing discourse analysis....***

By adding just three words "According to Coyle," the sense of authority changes, depending on what the reader knows about Coyle's work and reputation. Some analysts also suggest that a useful approach is to look for variability either within or between individuals.

Analyzing discourse on the Internet (e.g., in chatrooms, bulletin boards, and virtual worlds) has started to influence designers' understanding about users' needs in these environments. Conversation analysis is a very fine-grained form of discourse analysis that can be used for this purpose. In conversational analysis the semantics of the discourse are examined in fine detail. The focus is on how conversations are conducted. This technique is used in sociological studies and examines how conversations start, how turntaking is structured, and other rules of conversation. It can also be very useful when comparing conversations that take place during video-mediated sessions or in computer-mediated communication such as chatrooms as discussed in Chapter **4.**

12.6.3 Quantitative data analysis

Video data collected in usability laboratories is usually annotated as it is observed. Small teams of evaluators watch monitors showing what is being recorded in a control room out of the users' sight. As they see errors or unusual behavior, one of the evaluators marks the video and records a brief remark. When the test is finished evaluators can use the annotated recording to calculate performance times so they can compared users' performance on different prototypes. The data stream from the interaction log is used in a similar way to calculate performance times. Typically this data is further analyzed using simple statistics such as means, standard deviations, T-tests, etc. Categorized data may also be quantified and analyzed statistically, as we have said.

12.6.4 Feeding the findings back into design

The results from an evaluation can be reported to the design team in several ways, as we have indicated. Clearly written reports with an overview at the beginning and detailed content list make for easy reading and a good reference document. Includ-

ing anecdotes, quotations, pictures, and video clips helps to bring the study to life, stimulate interest, and make the written description more meaningful. Some teams like quantitative data, but its value depends on the type of study and its goals. Verbal presentations that include video clips can also be very powerful. Often both qualitative and quantitative data analysis are useful becuase they provide alternative perspectives.

Assignment

The aim of this assignment is for you to learn to do field obsewation. To do the assignment you will need to find a group of people or a single individual engaged in using one of the following: a mobile phone, a VCR, a photocopying machine, computer software, or some other type of technology that interests you. Assume that you have been employed to improve the product, either by doing a redesign or by creating a completely new product. You can observe people in your family, your friends, or people in your class or local community group.

For this assignment you should:

(a) Consider what the basic goal of "improving the product" means. What initial questions might you ask?

(b) Watch the group (or person) casually to get an understanding of issues that might create challenges for you doing this assignment and information that might enable you to refine your questions.

(c) Then plan your study:

(i) Think again about what questions will help direct your observation. What are you evaluating?

(ii) Decide where on the outsider-insider spectrum of observers you wish to be.

(iii) Prepare an informed consent form and any scripts that you need to introduce yourself and your study.

(iv) Decide how you will collect data and prepare any data-collection sheets needed; acquire and test any equipment needed.

(v) Decide how you will analyze the data that you collect.

(vi) Think through the DECIDE framework. Is everything covered?

(vii) If so, do a pilot study to check your preparation.

(d) Carry out your study but limit its scope. For example, plan two half-hour observation periods.

(e) Now analyze your data using the method chosen above.

(f) Write a report about what you did and why; describe your data, how you analyzed it, and your findings.

(g) Suggest some ways in which the product might be improved.

Summary

Observing users in the field enables designers to see how technology is used in context. It is valuable for confirming designers' understanding of users' needs and for exploring new design ideas. Various amounts of control, intervention, and involvement with users are possible.

At one end of the spectrum, laboratory studies offer a strongly controlled environment with little evaluator involvement; at the other, participant observation and ethnography require deeper involvement with users and understanding of context. Diaries and data-logging techniques provide a way of tracking user activity without intruding.

Key points

- Observation in usability testing tends to be objective, from the outside. The observer watches and analyzes what happens.
- In contrast, in participant observation the evaluator works with users to understand their activities, beliefs and feelings within the context in which the technology is used.
- Ethnography uses a set of techniques that include participant observation and interviews. Ethnographers immerse themselves in the culture that they study.
- The way that observational data is collected and analyzed depends on the paradigm in which it is used: quick and dirty, user testing, or field studies.
- Combinations of video, audio and paper records, data logging, and diaries can be used to collect observation data.
- In participant observation, collections of comments, incidents, and artifacts are made during the observation period. Evaluators are advised to discuss and summarize their findings as soon after the observation session as possible.

  Analyzing video and data logs can be difficult because of the sheer volume of data. It is important to have clearly specified questions to guide the process and also access to appropriate tools.
- Evaluators often flag events in real time and return to examine them in more detail later. Identifying key events is an effective approach. Fine-grained analyses can be very time-consuming.

Fuither reading

BLY, S. (1997) Field work: Is it product work? ***Interactions,*** January and February, 25–30. This article provides additional information to supplement the interview with Sara Bly. It gives a broad perspective on the role of participant observation in product development.

BOGDEWIC, S. P. (1992) Participant observation. In B. F. Crabtree and W. L. Miller (eds.), ***Doing Qualitative Research.*** Newbury Park, CA: Sage, 45-69. This chapter provides an introduction to participant observation.

BROWN, B. A., SELLEN, A. J., AND O'HARA, K. P. (2000). ***A diary study of information capture in working life.*** In the Proceedings of CHI2000, The Hague, Holland, 438–445. This paper discusses how cameras were used in a diary study, followed by semi-structured interviews, to inform the design of handheld storage devices.

FETTERMAN, D. M. (1998). ***Ethnography: Step by Step*** (2nd ed.). (Vol. 17). Thousand Oaks, CA: SAGE. This book provides an introduction to the theory and practice of ethnography and is an excellent guide for beginners. In addition, it has a useful section on computerized tools for ethnography.

ROBSON, C. (1993). ***Real World Research.*** Oxford, ***UK:*** Blackwell. Chapter 8 discusses a range of observation methods. There is a section on doing participant observation and also on observing from the outside using coding schemes.

INTERVIEW with Sara Bly

Sara Bly is a user-centered design consultant who specializes in the design and evaluation of distributed group technologies and practices. As well as having a Ph.D. in computer science, Sara pioneers the development of rich, qualitative observational techniques for analyzing group interactions and activities that inform technology design. Prior to becoming a consultant, Sara managed the Collaborative Systems Group at Xerox Palo Alto Research Center (PARC). While at PARC, Sara also contributed to ground-breaking work on shared drawing, awareness systems, and systems that used non-speech audio to represent information, and to the PARC Media Space project, in which video, audio, and computing technologies are uniquely combined to create a trans-geographical laboratory.

JP: Sara, tell us about your work and what especially interests you.

SB: I'm interested in the ways that qualitative studies, particularly based on ethnographic methods, can inform design and development of technologies. My work spans the full gamut of user-centered design, from early conceptual design through iterative prototypes to final product deployment. I've worked on a wide range of projects from complex collaborative systems to straightforward desktop applications, and a variety of new technologies. My recent projects include a cell phone enhancement, a web-based video application, and the integration of text-based virtual environments with documents.

JP: Why do you think qualitative methods are so important for evaluating usability?

SB: I strongly believe that technical systems are closely bound with the social setting in which they are used. An important part of evaluation is to look "beyond the task." Too often we think of computer systems in isolation from the rest of the activities in which the people are involved. It's important to be able to see the interface in the context of ongoing practice. Usually the complexities and "messiness" of everyday life do not lend themselves to constraining the evaluation to only a few variables for testing. Qualitative methods are particularly helpful for evaluating complex systems that involve several tasks, embedded in other activities that include multiple users.

JP: Can you give me an example?

SB: Recently I was asked to design and evaluate an application for setting up personal preferences and purchasing services on the web. I was told it would be hard to test the interface "in the field" because it was difficult to get a 45–60 minute test period when the user wasn't being interrupted. When I pointed out that interruptions were normal in the environment in which the product would be used and therefore should occur in the evaluation too, the client looked aghast. There was a moment of silence as he realized, for the first time, that this hadn't been taken into account in the design and that the interface timed out after 60 seconds. It was unusable because the user would have to start all over again after each timeout. This should have been noticed at the requirements stage. So why wasn't it? It sounds like such an obvious thing, but the team was so busy with the intricacies of the design that they failed to realize what the real world would be like in which the system would be used. This might sound extreme, but you'd be surprised how often such things happen.

JP: Collaborative applications seem particularly difficult to evaluate out of context.

SB: Yes, you have to evaluate collaborative systems integrated within an organizational culture in which working relationships are taken into account. We know that work practice impacts system design and that the introduction of a new system impacts work practice. Consequently, the system and the practice have to evolve together. Understanding the task or the interface is impossible without understanding the environment in which the system will be used.

JP: Much of what you've described involves various forms of observation. How do you collect and analyze this data?

SB: It's important that qualitative methods are not seen as just *watching.* Any method we use has at least three critical phases. First, there is the initial assess-

ment of the domain and/or technology and the determination of the questions to address in the evaluation. Second is the data collection, analysis, and representation, and third, the communication of the findings with the development team. I try to start with a clear understanding of what I need to focus on in the field. However, I also try hard not to start with assumptions about what will be true. So, 1 start with a ***well-defined*** focus but ***not*** a hypothesis. In the field (or even in the lab), I primarily use interviews and observations with some self-reporting that often takes the form of diaries, etc. The data typically consists of my notes, the audio and/or videotapes from interviews and observation time, still pictures, and as many artifacts as I can appropriately gather (e.g., a work document covered with post-its, a page from an old calendar). I also prefer to work with at least one other colleague so that there is a minimum of two perspectives on the events and data.

JP: It sounds like keeping track of all this data could be a problem. How do you organize and analyze it?

SB: Obviously it's critical not to end with the data collection. Whenever possible, I do immediate debriefs after each session in the field with my colleague, noting individually and collectively whatever jumped out at us. Subsequently, I use the interview notes (from everyone involved) and the tapes and artifacts to construct as much of a picture of what happened as possible, without putting any judgment on it. For example, in a recent study six of us were involved in interviews and observations. We worked in pairs and tried to vary the pairings as often as possible. Thus. we had lots of conversations about the data and the situations before we ever came together. First, we wrote up the notes from each session (something I try to do as soon as possible). Next we got together and began looking across the data. That is, we created representations of important events (tables, maps, charts) together. Because we collectively had observed all the events and because we could draw upon our notes, we could feed the data from each observation into each finding. Oftentimes, we create collections, looking for common behaviors or events across multiple sessions. A collection will highlight activities that are crucial to the design of the system being evaluated. Whatever techniques we use, we always come back to the data as a reality and validity check.

JP: Is it difficult to get development teams and managers to listen to you? How do you feed your findings back?

SB: As often as possible, development teams are involved in the process along the way. They participate in setting the initial goals of the evaluation, occasionally in observation sessions, and as recipients of a final report. My goal with any project is to ensure that the final report is not a handoff but rather an interaction that offers a chance to work together on what we've found.

JP: What are the main challenges you face?

SB: It's always difficult to conduct a field study with as much time and participation as would be ideal. Most product cycles are short and the evaluation is just one of many necessary steps. So it's always a challenge to do an evaluation that is timely, useful, and yet based on solid methodology.

A gnawing question for me is how to evaluate a system in the context of the customer's own environment and experience when the system is not fully developed and ready to deploy? If we can't bring a product to the field, can we bring the field to the product? For example, a client recently had a prototype interface for a system that was intended to provide a new approach to person-to-person calls. But using the interface made sense only in the context of actual real-world interactions. So, while we certainly could do a standard usability study of the interface, this approach wouldn't get at the questions of how well the product would fit into an actual work situation.

JP: Finally, what about the future? Any comments?

SB: I think the explosion of computing technology is both exciting and overwhelming. We now have so much new information constantly available and so many new devices to master that it's hard to keep up. Evaluation is going to become ever more critical and complex and we should use all the techniques at our disposal as appropriate. I think an increasingly important aspect of new interfaces will be not only how well they support performance, satisfaction, and experience, but the way in which a user is able to grasp a conceptual model that is compatible with, but does not overwhelm their ongoing practice.