Chapter 11

An evaluation framework

11.1 Introduction

Designing useful and attractive products requires skill and creativity. As products evolve from initial ideas through conceptual design and prototypes, iterative cycles of design and evaluation help to ensure that they meet users' needs. But how do evaluators decide *what* and *when* to evaluate? The HutchWorld case study in the previous chapter described how one team did this, but the circumstances surrounding every product's development are different. Certain techniques work better for some than for others.

Identifying usability and user experience goals is essential for making every product successful, and this requires understanding users' needs. The role of evaluation is to make sure that this understanding occurs during all the stages of the product's development. The skillful and sometimes tricky part of doing this is knowing what to focus on at different stages. Initial requirements get the design process started, but, as you have seen, understanding requirements tends to happen by a process of negotiation between designers and users. As designers understand users' needs better, their designs reflect .this understanding. Similarly, as users see and experience design ideas, they are able to give better feedback that enables the designers to improve their designs further. The process is cyclical, with evaluation playing a key role in facilitating understanding between designers and users.

Evaluation is driven by questions about how well the design or particular aspects of it satisfy users' needs. Some of these questions provide high-level goals to guide the evaluation. Others are much more specific. For example, can users find a particular menu item? Is a graphic useful and attractive? Is the product engaging? Practical constraints also play a big role in shaping evaluation plans: tight schedules, low budgets, or little access to users constrain what evaluators can do. You read in chapter 10 how the HutchWorld team had to plan its evaluation around hospital routines and patients' health.

Experienced designers get to know what works and what doesn't, but those with little experience can find doing their first evaluation daunting. However, with careful advance planning, problems can be spotted and ways of dealing with them can be found. Planning evaluation studies involves thinking about key issues and asking questions about the process. In this chapter we propose the DECIDE framework to help you do this.

The main aims of this chapter are to:

- Continue to explain the key concepts and terms used to discuss evaluation.
- Describe the evaluation paradigms and techniques used in interaction design.
- Discuss the conceptual, practical, and ethical issues to be considered when planning evaluation.
- Introduce the DECIDE framework to help you plan your own evaluation studies.

11.2 Evaluation paradigms and techniques

Before we describe the techniques used in evaluation studies, we shall start by proposing some key terms. Terminology in this field tends to be loose and often confusing so it is a good idea to be clear from the start what you mean. We start with the much-used term ***user*** studies, defined by Abigail Sellen in her interview at the end of Chapter 4 as follows: "user studies essentially involve looking at how people behave either in their natural [environments], or in the laboratory, both with old technologies and with new ones." Any kind of evaluation, whether it is a user study or not, is guided either explicitly or implicitly by a set of beliefs that may also be underpinned by theory. These beliefs and the practices (i.e., the methods or techniques) associated with them are known as an evaluation paradigm, which you should not confuse with the "interaction paradigms" discussed in Chapter 2. Often evaluation paradigms are related to a particular discipline in that they strongly influence how people from the discipline think about evaluation. Each paradigm has particular methods and techniques associated with it. So that you are not confused, we want to state explicitly that we will not be distinguishing between methods and techniques. We tend to talk about techniques, but you may find that other books call them methods. An example of the relationship between a paradigm and the techniques used by evaluators following that paradigm can be seen for usability testing, which is an applied science and engineering paradigm. The techniques associated with usability testing are: user testing in a controlled environment; observation of user activity in the controlled environment and the field; and questionnaires and interviews.

11.2.1 Evaluation paradigms

In this book we identify four core evaluation paradigms: (1) "quick and dirty" evaluations; (2) usability testing; (3) field studies; and (4) predictive evaluation. Other texts may use slightly different terms to refer to similar paradigms.

"Quick and dirty" evaluation

A "quick and dirty" evaluation is a common practice in which designers informally get feedback from users or consultants to confirm that their ideas are in line with users' needs and are liked. "Quick and dirty" evaluations can be done at any stage and the emphasis is on fast input rather than carefully documented findings. For example, early in design developers may meet informally with users to get feedback on ideas for a new product (Hughes et al., 1994). At later stages similar meetings may occur to try out an idea for an icon, check whether a graphic is liked, or confirm that information has been appropriately categorized on a webpage. This approach is often called "quick and dirty" because it is meant to be done in a short space of time. Getting this kind of feedback is an essential ingredient of successful design.

As discussed in Chapter 9, any involvement with users will be highly informative and you can learn a lot early in design by observing what people do and talking to them informally. The data collected is usually descriptive and informal and it is fed back into the design process as verbal or written notes, sketches and anecdotes, etc. Another source comes from consultants, who use their knowledge of user behavior, the market place and technical know-how, to review software quickly and provide suggestions for improvement. It is an approach that has become particularly popular in web design where the emphasis is usually on short timescales.

Usability testing

Usability testing was the dominant approach in the 1980s (Whiteside et al., 1998), and remains important, although, as you will see, field studies and heuristic evaluations have grown in prominence. Usability testing involves measuring typical users' performance on carefully prepared tasks that are typical of those for which the system was designed. Users' performance is generally measured in terms of number of errors and time to complete the task. As the users perform these tasks, they are watched and recorded on video and by logging their interactions with software. This observational data is used to calculate performance times, identify errors, and help explain why the users did what they did. User satisfaction questionnaires and interviews are also used to elicit users' opinions.

The defining characteristic of usability testing is that it is ***strongly controlled*** by the evaluator (Mayhew, 1999). There is no mistaking that the evaluator is in charge! Typically tests take place in laboratory-like conditions that are controlled. Casual visitors are not allowed and telephone calls are stopped, and there is no possibility of talking to colleagues, checking email, or doing any of the other tasks that most of us rapidly switch among in our normal lives. Everything that

the participant does is recorded—every keypress, comment, pause, expression, etc., so that it can be used as data.

Quantifying users' performance is a dominant theme in usability testing. However, unlike research experiments, variables are not manipulated and the typical number of participants is too small for much statistical analysis. User satisfaction data from questionnaires tends to be categorized and average ratings are presented. Sometimes video or anecdotal evidence is also included to illustrate problems that users encounter. Some evaluators then summarize this data in a usability specification so that developers can use it to test future prototypes or versions of the product against it. Optimal performance levels and minimal levels of acceptance are often specified and current levels noted. Changes in the design can then be agreed and engineered—hence the term "usability engineering." User testing is explained further in Chapter 14, how to observe users is described in Chapter 12, and issues concerned with interviews and questionnaires are explored in Chapter 13.

Field studies

The distinguishing feature of field studies is that they are done in natural settings with the aim of increasing understanding about what users do naturally and how technology impacts them. In product design, field studies can be used to (1) help identify opportunities for new technology; (2) determine requirements for design; (3) facilitate the introduction of technology; and (4) evaluate technology (Bly, 1997).

Chapter 9 introduced qualitative techniques such as interviews, observation, participant observation, and ethnography that are used in field studies. The exact choice of techniques is often influenced by the theory used to analyze the data. The data takes the form of events and conversations that are recorded as notes, or by audio or video recording, and later analyzed using a variety of analysis techniques such as content, discourse, and conversational analysis. These techniques vary considerably. In content analysis, for example, the data is analyzed into content categories, whereas in discourse analysis the use of words and phrases is examined. Artifacts are also collected. In fact, anything that helps to show what people do in their natural contexts can be regarded as data.

In this text we distinguish between two overall approaches to field studies. The first involves observing explicitly and recording what is happening, as an ***outsider*** looking on. Qualitative techniques are used to collect the data, which may then be analyzed qualitatively or quantitatively. For example, the number of times a particular event is observed may be presented in a bar graph with means and standard deviations.

In some field studies the evaluator may be an ***insider*** or even a participant. Ethnography is a particular type of insider evaluation in which the aim is to explore the details of what happens in a particular social setting. "In the context of human-computer interaction, ethnography is a means of studying work (or other activities) in order to inform the design of information systems and understand aspects of their use" (Shapiro, 1995, p. 8).

Predictive evaluation

In predictive evaluations experts apply their knowledge of typical users, often guided by heuristics, to predict usability problems. Another approach involves theoretically-based models. The key feature of predictive evaluation is that users need not be present, which makes the process quick, relatively inexpensive, and thus attractive to companies; but it has limitations.

In recent years heuristic evaluation in which experts review the software product guided by tried and tested heuristics has become popular (Nielsen and Mack, 1994). As mentioned in Chapter 1, usability guidelines (e.g., always provide clearly marked exits) were designed primarily for evaluating screen-based products (e.g. form fill-ins, library catalogs, etc.). With the advent of a range of new interactive products (e.g., the web, mobiles, collaborative technologies), this original set of heuristics has been found insufficient. While some are still applicable (e.g., speak the users' language), others are inappropriate. New sets of heuristics are also needed that are aimed at evaluating different classes of interactive products. In particular, specific heuristics are needed that are tailored to evaluating web-based products, mobile devices, collaborative technologies, computerized toys, etc. These should be based on a combination of usability and user experience goals, new research findings and market research. Care is needed in using sets of heuristics. As you will see in Chapter 13, designers are sometimes led astray by findings from heuristic evaluations that turn out not to be as accurate as they at first seemed.

Table 11.1 summarizes the key aspects of each evaluation paradigm for the following issues:

the role of users

- who controls the process and the relationship between evaluators and users during the evaluation
- the location of the evaluation
- when the evaluation is most useful
- the type of data collected and how it is analyzed

how the evaluation findings are fed back into the design process

- the philosophy and theory that underlies the evaluation paradigms

Some other terms that you may encounter in your reading are shown in Box 11.1.

ACTIVITY 11.1 Think back to the HutchWorld case study.

(a) Which evaluation paradigms were used in the study and which were not?

(b) How could the missing evaluation paradigms have been used to inform the design and why might they not have been used?

Comment

(a) The team did some "quick and dirty" evaluation during early development but this is not stressed in their report. Usability testing played a strong role, with some tests being carried out at the Fred Hutchinson Center and later tests in Microsoft's usability laboratories. Field studies are not strongly featured, but the team does mention

Table 11.1 Characteristics of different evaluation paradigms

Evaluation paradigms	"Quick and dirty"	Usability testing	Field studies	Predictive
Role of users	Natural behavior.	To carry out set tasks.	Natural behavior.	Users generally not involved.
Who controls	Evaluators take minimum control.	Evaluators strongly in control.	Evaluators try to develop relationships with users.	Expert evaluators.
Location	Natural environment or laboratory.	Laboratory.	Natural environment.	Laboratory-oriented but often happens on customer's premises.
When used	Any time you want to get feedback about a design quickly. Techniques from other evaluation paradigms can be used–e.g., experts review software.	With a prototype or product.	Most often used early in design to check that users' needs are being met or to assess problems or design opportunities.	Expert reviews (often done by consultants) with a prototype, but can occur at any time. Models are used to assess specific aspects of a potential design.
Type of data	Usually qualitative, informal descriptions.	Quantitative. Sometimes statistically validated. Users' opinions collected by questionnaire or interview.	Qualitative descriptions often accompanied with sketches, scenarios, quotes, other artifacts.	List of problems from expert reviews. Quantitative figures from model, e.g., how long it takes to perform a task using two designs.
Fed back into design by . . .	Sketches, quotes, descriptive report.	Report of performance measures, errors etc. Findings provide a benchmark for future versions.	Descriptions that include quotes, sketches, anecdotes, and sometimes time logs.	Reviewers provide a list of problems, often with suggested solutions. Times calculated from models are given to designers.
Philosophy	User-centered, highly practical approach.	Applied approach based on experimentation, i.e., usability engineering.	May be objective observation or ethnographic.	Practical heuristics and practitioner expertise underpin expert reviews. Theory underpins models.

BOX 11.1 Some Definitions to Help You

Objective and subjective Objective evaluations are based on techniques that use quantitative measurement rather than users' or experts' opinions. Subjective evaluations are based on opinions and anecdotes.

Quantitative and qualitative Quantitative evaluations involve measurements, whereas qualitative evaluations involve descriptions and anecdotes. Quantitative evaluations tend to be seen as objective and impartial, whereas many qualitative evaluations tend to be seen as subjective but this isn't necessarily true.

Laboratory and field or naturalistic studies Laboratory studies occur in controlled environments. They may be done in a specially built laboratory or in a space that is specially adapted for the purpose. Field or naturalistic studies are situated in the real-world context in which the system is or will be used.

observing how patients used HutchWorld in the Center. Field studies were planned in which patients, who have access to HutchWorld and the web, could be systematically compared with another group who does not have these facilities. However, distinguishing between evaluation paradigms isn't always clear-cut. In practice elements typically found in one may be transferred to another (e.g., the controlled approach the HutchWorld team planned to use in the field). The only evaluation paradigm that is not mentioned in the study is predictive evaluation.

(b) Expert reviews could have been done any time during its development but the team may have thought they were not needed, or there wasn't time, or perhaps they were performed but not reported.

11.2.2 Techniques

There are many evaluation techniques and they can be categorized in various ways, but in this text we will examine techniques for:

- observing users
- asking users their opinions
- asking experts their opinions
- testing users' performance
- modeling users' task performance to predict the efficacy of a user interface

The brief descriptions below offer an *overview* of each category, which we discuss in detail in the next three chapters. Be aware that some techniques are used in different ways in different evaluation paradigms.

Observing users

Observation techniques help to identify needs leading to new types of products and help to evaluate prototypes. Notes, audio, video, and interaction logs are well-known ways of recording observations and each has benefits and drawbacks. Obvious challenges for evaluators are how to observe without disturbing the people being observed and how to analyze the data, particularly when large quantities of

video data are collected or when several different types must be integrated to tell the story (e.g., notes, pictures, sketches from observers). You met several observation techniques in Chapter 7 in the context of the requirements activity; in Chapter 12 we will focus on how they are used in evaluation.

Asking users

Asking users what they think of a product—whether it does what they want; whether they like it; whether the aesthetic design appeals; whether they had problems using it; whether they want to use it again—is an obvious way of getting feedback. Interviews and questionnaires are the main techniques for doing this. The questions asked can be unstructured or tightly structured. They can be asked of a few people or of hundreds. Interview and questionnaire techniques are also being developed for use with email and the web. We discuss these techniques in Chapter 13.

Asking experts

Software inspections and reviews are long established techniques for evaluating software code and structure. During the 1980s versions of similar techniques were developed for evaluating usability. Guided by heuristics, experts step through tasks role-playing typical users and identify problems. Developers like this approach because it is usually relatively inexpensive and quick to perform compared with laboratory and field evaluations that involve users. In addition, experts frequently suggest solutions to problems. In Chapter 13 you will learn a few inspection techniques for evaluating usability.

User testing

Measuring user performance to compare two or more designs has been the bedrock of usability testing. As we said earlier when discussing usability testing, these tests are usually conducted in controlled settings and involve typical users performing typical, well-defined tasks. Data is collected so that performance can be analyzed. Generally the time taken to complete a task, the number of errors made, and the navigation path through the product are recorded. Descriptive statistical measures such as means and standard deviations are commonly used to report the results. In Chapter 14 you will learn the basics of user testing and how it differs from scientific experiments.

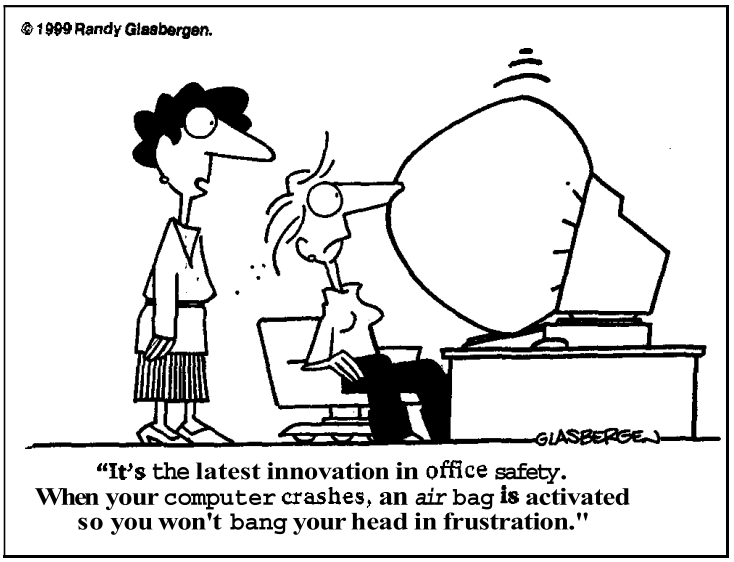
Modeling users' task performance

There have been various attempts to model human-computer interaction so as to predict the efficiency and problems associated with different designs at an early stage without building elaborate prototypes. These techniques are successful for systems with limited functionality such as telephone systems. GOMS and the keystroke model are the best known techniques. They have already been mentioned in Chapter 3 and in Chapter 14 we examine their role in evaluation.

Table 11.2 summarizes the categories of techniques and indicates how they are commonly used in the four evaluation paradigms.

Table 11.2 The relationship between evaluation paradigms and techniques.

	Evaluation paradigms			
Techniques	**"Quick and dirty"**	**Usability testing**	**Field studies**	**Predictive**
Observing users	Important for seeing how users behave in their natural environments.	Video and interaction logging, which can be analyzed to identify errors, investigate routes through the software, or calculate performance time.	Observation is the central part of any field study. In ethnographic studies evaluators immerse themselves in the environment. In other types of studies the evaluator looks on objectively.	N/A
Asking users	Discussions with users and potential users individually, in groups or focus groups.	User satisfaction questionnaires are administered to collect users' opinions. Interviews may also be used to get more details.	The evaluator may interview or discuss what she sees with participants. Ethnographic interviews are used in ethnographicstudies.	N/A
Asking experts	To provide critiques (called "crit reports") of the usability of a prototype.	N/A	N/A	Experts use heuristics early in design to predict the efficacy of an interface.
User testing	N/A	Testing typical users on typical tasks in a controlled laboratory-like setting is the cornerstone of usability testing.	N/A	N/A
Modeling users' task performance	N/A	N/A	N/A	Models are used to predict the efficacy of an interface or compare performance times between versions.



"It's the latest innovation in office safety. When your computer crashes, an air bag is activated so you won't bang your head in frustration."

11.3 DECIDE: A framework to guide evaluation

Well-planned evaluations are driven by clear goals and appropriate questions (Basili et al., 1994). To guide our evaluations we use the DECIDE framework, which provides the following checklist to help novice evaluators:

1. Determine the overall *goals* that the evaluation addresses.
2. Explore the specific *questions* to be answered.
3. Choose the *evaluation paradigm* and *techniques* to answer the questions.
4. Identify the *practical issues* that must be addressed, such as selecting participants.
5. Decide how to deal with the *ethical issues.*
6. Evaluate, interpret, and present the *data.*

11.3.1 Determine the goals

What are the high-level goals of the evaluation? Who wants it and why? An evaluation to help clarify user needs has different goals from an evaluation to determine the best metaphor for a conceptual design, or to fine-tune an interface, or to examine how technology changes working practices, or to inform how the next version of a product should be changed.

Goals should guide an evaluation, so determining what these goals are is the first step in planning an evaluation. For example, we can restate the general goal statements just mentioned more clearly as:

- Check that the evaluators have understood the users' needs.
- Identify the metaphor on which to base the design.

- Check to ensure that the final interface is consistent.
- Investigate the degree to which technology influences working practices.
- Identify how the interface of an existing product could be engineered to improve its usability.

These goals influence the evaluation approach, that is, which evaluation paradigm guides the study. For example, engineering a user interface involves a quantitative engineering style of working in which measurements are used to judge the quality of the interface. Hence usability testing would be appropriate. Exploring how children talk together in order to see if an innovative new groupware product would help them to be more engaged would probably be better informed by a field study.

11.3.2 Explore the questions

In order to make goals operational, questions that must be answered to satisfy them have to be identified. For example, the goal of finding out why many customers prefer to purchase paper airline tickets over the counter rather than e-tickets can be broken down into a number of relevant questions for investigation. What are customers' attitudes to these new tickets? Perhaps they don't trust the system and are not sure that they will actually get on the flight without a ticket in their hand. Do customers have adequate access to computers to make bookings? Are they concerned about security? Does this electronic system have a bad reputation? Is the user interface to the ticketing system so poor that they can't use it? Maybe very few people managed to complete the transaction.

Questions can be broken down into very specific sub-questions to make the evaluation even more specific. For example, what does it mean to ask, "Is the user interface poor?": Is the system difficult to navigate? Is the terminology confusing because it is inconsistent? Is response time too slow? Is the feedback confusing or maybe insufficient? Sub-questions can, in turn, be further decomposed into even finer-grained questions, and so on.

11.3.3 Choose the evaluation paradigm and techniques

Having identified the goals and main questions, the next step is to choose the evaluation paradigm and techniques. As discussed in the previous section, the evaluation paradigm determines the kinds of techniques that are used. Practical and ethical issues (discussed next) must also be considered and trade-offs made. For example, what seems to be the most appropriate set of techniques may be too expensive, or may take too long, or may require equipment or expertise that is not available, so compromises are needed.

As you saw in the HutchWorld case study, combinations of techniques can be used to obtain different perspectives. Each type of data tells the story from a different point of view. Using this triangulation reveals a broad picture.

11.3.4 Identify the practical issues

There are many practical issues to consider when doing any kind of evaluation and it is important to identify them ***before*** starting. Some issues that should be considered include users, facilities and equipment, schedules and budgets, and evaluators' expertise. Depending on the availability of resources, compromises may involve adapting or substituting techniques.

Users

It goes without saying that a key aspect of an evaluation is involving ***appropriate*** users. For laboratory studies, users must be found and screened to ensure that they represent the user population to which the product is targeted. For example, usability tests often need to involve users with a particular level of experience e.g., novices or experts, or users with a range of expertise. The number of men and women within a particular age range, cultural diversity, educational experience, and personality differences may also need to be taken into account, depending on **the** kind of product being evaluated. In usability tests participants are typically screened to ensure that they meet some predetermined characteristic. For example, they might be tested to ensure that they have attained a certain skill level or fall within a particular demographic range. Questionnaire surveys require large numbers of participants so ways of identifying and reaching a representative sample of participants are needed. For field studies to be successful, an appropriate and accessible site must be found where the evaluator can work with the users in their natural setting.

Another issue to consider is how the users will be involved. The tasks used in a laboratory study should be representative of those for which the product is designed. However, there are no written rules about the length of time that a user should be expected to spend on an evaluation task. Ten minutes is too short for most tasks and two hours is a long time, but what is reasonable? Task times will vary according to the type of evaluation, but when tasks go on for more than 20 minutes, consider offering breaks. It is accepted that people using computers should stop, move around and change their position regularly after every 20 minutes spent at the keyboard to avoid repetitive strain injury. Evaluators also need to put users at ease so they are not anxious and will perform normally. Even when users are paid to participate, it is important to treat them courteously. At no time should users be treated condescendingly or made to feel uncomfortable when they make mistakes. Greeting users, explaining that it is the system that is being tested and not them, and planning an activity to familiarize them with the system before starting the task all help to put users at ease.

Facilities and equipment

There are many practical issues concerned with using equipment in an evaluation. For example, when using video you need to think about how you will do the recording: how many cameras and where do you put them? Some people are dis-

turbed by having a camera pointed at them and will not perform normally, so how can you avoid making them feel uncomfortable? Spare film and batteries may also be needed.

Schedule and budget constraints

Time and budget constraints are important considerations to keep in mind. It might seem ideal to have 20 users test your interface, but if you need to pay them, then it could get costly. Planning evaluations that can be completed on schedule is also important, particularly in commercial settings. However, as you will see in the interview with Sara Bly in the next chapter, there is never enough time to do evaluations as you would ideally like, so you have to compromise and plan to do a good job with the resources and time available.

Expertise

Does the evaluation team have the expertise needed to do the evaluation? For example, if no one has used models to evaluate systems before, then basing an evaluation on this approach is not sensible. It is no use planning to use experts to review an interface if none are available. Similarly, running usability tests requires expertise. Analyzing video can take many hours, so someone with appropriate expertise and equipment must be available to do it. If statistics are to be used, then a statistician should be consulted before starting the evaluation and then again later for analysis, if appropriate.

ACTIVITY 11.2 Informal observation, user performance testing, and questionnaires were used in the HutchWorld case study. What practical issues are mentioned in the case study? What other issues do you think the developers had to take into account?

Comment No particular practical issues are mentioned for the informal observation, but there probably were restrictions on where and what the team could observe. For example, it is likely that access would be denied to very sick patients and during treatment times. Not surprisingly, user testing posed more problems, such as finding participants, putting equipment in place, managing the tests, and underestimation of the time needed to work in a hospital setting compared with the fast production times at Microsoft.

11.3.5 Decide how to deal with the ethical issues

The Association for Computing Machinery (ACM) and many other professional organizations provide ethical codes (Box 11.2) that they expect their members to uphold, particularly if their activities involve other human beings. For example, people's privacy should be protected, which means that their name should not be associated with data collected about them or disclosed in written reports (unless they give permission). Personal records containing details about health, employment, education, financial status, and where participants live should be confidential. Similarly,

BOX 11.2 ACM Code of Ethics

The ACM code outlines many ethical issues that professionals are likely to face. Section 1 outlines fundamental ethical considerations, while section 2 addresses additional, more specific considerations of professional conduct. Statements in section 3 pertain more specifically to individuals who have a leadership role. Principles involving compliance with the code are given in section 4. Two principles of particular relevance to this discussion are:

- Ensure that users and those who will be affected by a system have their needs clearly articulated during the assessment of requirements; later the system must be validated to meet requirements.
- Articulate and support policies that protect the dignity of users and others affected by a computing system.

it should not be possible to identify individuals from comments written in reports. For example, if a focus group involves nine men and one woman, the pronoun "she" should not be used in the report because it will be obvious to whom it refers.

Most professional societies, universities, government and other research offices require researchers to provide information about activities in which human participants will be involved. This documentation is reviewed by a panel and the researchers are notified whether their plan of work, particularly the details about how human participants will be treated, is acceptable.

People give their time and their trust when they agree to participate in an evaluation study and both should be respected. But what does it mean to be respectful to users? What should participants be told about the evaluation? What are participants' rights? Many institutions and project managers require participants to read and sign an informed consent form similar to the one in Box 11.3. This form explains the aim of the tests or research and promises participants that their personal details and performance will not be made public and will be used only for the purpose stated. It is an

BOX 11.3 Informed Consent Form

I state that I am over 18 years of age and wish to participate in a program of research being conducted by Dr. Hoo Hah and his colleagues at the College of Extraordinary Research, University of Highland, College Estate.

The purpose of the research is to assess the usability of HighFly, a website developed at the National Library to provide information to the general public.

The procedures involve the monitored use of HighFly. I will be asked to perform specific tasks using HighFly. I will also be asked open-ended questions about HighFly and my experience using it.

All information collected in the study is confidential, and my name will not be identified at any time.

I understand that I am free to ask questions or to withdraw from participation at any time without penalty.

_____ _____

Signature of Participant Date

(Adapted from Cogdill, 1999.)

agreement between the evaluator and the evaluation participants that helps to confirm the professional relationship that exists between them. If your university or organization does not provide such a form it is advisable to develop one, partly to protect yourself in the unhappy event of litigation and partly because the act of constructing it will remind you what you should consider.

The following guidelines will help ensure that evaluations are done ethically and that adequate steps to protect users' rights have been taken.

- Tell participants the goals of the study and exactly what they should expect if they participate. The information given to them should include outlining the process, the approximate amount of time the study will take, the kind of data that will be collected, and how that data will be analyzed. The form of the final report should be described and, if possible, a copy offered to them. Any payment offered should also be clearly stated.
- Be sure to explain that demographic, financial, health, or other sensitive information that users disclose or is discovered from the tests is confidential. A coding system should be used to record each user and, if a user must be identified for a follow-up interview, the code and the person's demographic details should be stored separately from the data. Anonymity should also be promised if audio and video are used.
- Make sure users know that they are free to stop the evaluation at any time if they feel uncomfortable with the procedure.
- Pay users when possible because this creates a formal relationship in which mutual commitment and responsibility are expected.
- Avoid including quotes or descriptions that inadvertently reveal a person's identity, as in the example mentioned above, of avoiding use of the pronoun "she" in the focus group. If quotes need to be reported, e.g., to justify conclusions, then it is convention to replace words that would reveal the source with representative words, in square brackets. We used this convention in Boxes 9.2 and 9.3.
- Ask users' permission in advance to quote them, promise them anonymity, and offer to show them a copy of the report before it is distributed.

The general rule to remember when doing evaluations is ***do unto others only what you would not mind being done to you.***

ACTIVITY 11.3 Think back to the HutchWorld case study. What ethical issues did the developers have to consider?

Comment The developers of HutchWorld considered all the issues listed above. In addition, because the study involved patients, they had to be particularly careful that medical and other personal information was kept confidential. They were also sensitive to the fact that cancer patients may become too tired or sick to participate so they reassured them that they could stop at any time if the task became onerous.

ACTIVITY 11.4 Usability laboratories often have a one-way mirror that allows evaluators to watch users doing their tasks in the laboratory without the users seeing the evaluators. Should users be told that they are being watched?

Comment Yes, users should be told that they will be observed through a one-way mirror. It is unethical not to. This honest approach will not compromise the study because users forget about the mirror as they get more absorbed in their tasks. Telling users what is happening helps to build trust.

The recent explosion in Internet and web usage has resulted in more research on how people use these technologies and their effects on everyday life. Consequently, there are many projects in which developers and researchers are logging users' interactions, analyzing web traffic, or examining conversations in chatrooms, bulletin boards, or on email. Unlike most previous evaluations in human-computer interaction, these studies can be done without users knowing that they are being studied. This raises ethical concerns, chief among which are issues of privacy, confidentiality, informed consent, and appropriation of others' personal stories (Sharf, 1999). People often say things online that they would not say face to face. Furthermore, many people are unaware that personal information **they share** online can be read by someone with technical know-how years later, even after they have deleted it from their personal mailbox (Erickson et al., 1999).

ACTIVITY 11.5 Studies of user behavior on the Internet may involve logging users' interactions and keeping a copy of their conversations with others. Should users be told that this is happening?

DILEMMA What Would You Do?

There is a famous and controversial story about a 1961–62 experiment by Yale social psychologist Stanley Milgram to investigate how people respond to orders given by people in authority. Much has been written about this experiment and details have been changed and embellished over the years, but the basic ethical issues it raises are still worth considering, even if the details of the actual study have been distorted.

The subjects were ordinary residents of New Haven who were asked to administer increasingly high levels of electric shocks to victims when they made errors in the tasks they were given. As the electric shocks got more and more severe, so did the apparent pain of the victims receiving them, to the extent that some appeared to be on the verge of dying. Not surprisingly, those administering the shocks became more and more disturbed by what they were being asked to do, but several continued, believing that they should do as their superiors told them. What they did not realize was that the so-called victims were, in fact, very convincing actors who were not being injured at all. Instead, the shock administrators were themselves the real subjects. It was their responses to authority that were being studied in this deceptive experiment.

This story raises several important ethical issues. First, this experiment reveals how power relationships can be used to control others. Second and equally important, this experiment relied on deception. The experimenters were, in fact, the subjects and the fake subjects colluded with the real scientists to deceive them. Without this deception the experiment would not have worked.

Is it acceptable to deceive subjects to this extent? What do you think?

Comment Yes, it is better to tell users in advance that they are being logged. As in the previous example, the users' knowledge that they are being logged often ceases to be an issue as they become involved in what they are doing.

11.3.6 Evaluate, interpret, and present the data

Choosing the evaluation paradigm and techniques to answer the questions that satisfy the evaluation goal is an important step. So is identifying the practical and ethical issues to be resolved. However, decisions are also needed about what data to collect, how to analyze it, and how to present the findings to the development team. To a great extent the technique used determines the type of data collected, but there are still some choices. For example, should the data be treated statistically? If qualitative data is collected, how should it be analyzed and represented? Some general questions also need to be asked (Preece et al., 1994): Is the technique reliable? Will the approach measure what is intended, i.e., what is its validity? Are biases creeping in that will distort the results? Are the results generalizable, i.e., what is their scope? Is the evaluation ecologically valid or is the fundamental nature of the process being changed by studying it?

Reliability

The reliability or consistency of a technique is how well it produces the ***same*** results on separate occasions under the ***same*** circumstances. Different evaluation processes have different degrees of reliability. For example, a carefully controlled experiment will have high reliability. Another evaluator or researcher who follows exactly the same procedure should get similar results. In contrast, an informal, unstructured interview will have low reliability: it would be difficult if not impossible to repeat exactly the same discussion.

Validity

Validity is concerned with whether the evaluation technique measures what it is supposed to measure. This encompasses both the technique itself and the way it is performed. If for example, the goal of an evaluation is to find out how users use a new product in their homes, then it is not appropriate to plan a laboratory experiment. An ethnographic study in users' homes would be more appropriate. If the goal is to find average performance times for completing a task, then counting only the number of user errors would be invalid.

Biases

Bias occurs when the results are distorted. For example, expert evaluators performing a heuristic evaluation may be much more sensitive to certain kinds of design flaws than others. Evaluators collecting observational data may consistently fail to notice certain types of behavior because they do not deem them important.

Put another way, they may selectively gather data that they think is important. Interviewers may unconsciously influence responses from interviewees by their tone of voice, their facial expressions, or the way questions are phrased, so it is important to be sensitive to the possibility of biases.

Scope

The scope of an evaluation study refers to how much its findings can be generalized. For example, some modeling techniques, like the keystroke model, have a narrow, precise scope. The model predicts expert, error-free behavior so, for example, the results cannot be used to describe novices learning to use the system.

Ecological validity

Ecological validity concerns how the environment in which an evaluation is conducted influences or even distorts the results. For example, laboratory experiments are strongly controlled and are quite different from workplace, home, or leisure environments. ***Laboratory*** experiments therefore have low ecological validity because the results are unlikely to represent what happens in the real world. In contrast, ethnographic studies do not impact the environment, so they have high ecological validity.

Ecological validity is also affected when participants are aware of being studied. This is sometimes called the ***Hawthorne effect*** after a series of experiments at the Western Electric Company's Hawthorne factory in the US in the 1920s and 1930s. The studies investigated changes in length of working day, heating, lighting, etc., but eventually it was discovered that the workers were reacting positively to being given special treatment rather than just to the experimental conditions.

11.4 Pilot studies

It is always worth testing plans for an evaluation by doing a pilot study before launching into the main study. **A** pilot study is a small trial run of the main study. The aim is to make sure that the plan is viable before embarking on the real study. For example, the equipment and instructions for its use can be checked. It is also an opportunity to practice interviewing skills, or to check that the questions in a questionnaire are clear or that an experimental procedure works properly. **A** pilot study will identify potential problems in advance so that they can be corrected. Sending out 500 questionnaires and then being told that two of the questions were very confusing wastes time, annoys participants, and is expensive.

Many evaluators run several pilot studies. As in iterative design, they get feedback, amend the procedure, and test it again until they know they have a good study. If it is difficult to find people to participate or if access to participants is limited, colleagues or peers can be asked to comment. Getting comments from peers is quick and inexpensive and can save a lot of trouble later. In theory, at least, there is no limit to the number of pilot studies that can be run, although there will be practical constraints.

Assignment

Find a journal or conference publication that describes an interesting evaluation study or select one using www.hcibib.org. Then use the DECIDE framework to determine which paradigms and techniques were used. Also consider how well it fared on ethical and practical issues.

(a) Which evaluation paradigms and techniques are used?

(b) Is triangulation used? How?

(c) Comment on the reliability, validity, ecological validity, biases and scope of the techniques described.

(d) Is there evidence of one or more pilot studies?

(e) What are the strengths and weakness of the study report? Write a 50–100 word critique that would help the author(s) improve their report.

Summary

This chapter has introduced four core evaluation paradigms and five categories of techniques and has shown how they relate to each other. The DECIDE framework identifies the main issues that need to be considered when planning an evaluation. It also introduces many of the basic concepts that will be revisited and built upon in the next three chapters: Chapter 12, which discusses observation techniques; Chapter 13, which examines techniques for gathering users' and experts' opinions; and Chapter 14, which discusses user testing and techniques for modeling users' task performance.

Key points

- An evaluation paradigm is an approach in which the methods used are influenced by particular theories and philosophies. Four evaluation paradigms were identified:
  1. "quick and dirty"
  2. usability testing
  3. field studies
  4. predictive evaluation
- Methods are combinations of techniques used to answer a question but in this book we often use the terms "methods" and "techniques" interchangeably. Five categories were identified:
  1. observing users
  2. asking users
  3. asking experts
  4. user testing
  5. modeling users' task performance
- The DECIDE framework has six parts:
  1. Determine the overall goals of the evaluation.
  2. Explore the questions that need to be answered to satisfy the goals.
  3. Choose the evaluation paradigm and techniques to answer the questions.
  4. Identify the practical issues that need to be considered.
  5. Decide on the ethical issues and how to ensure high ethical standards.
  6. Evaluate, interpret, and present the data.
- Drawing up a schedule for your evaluation study and doing one or several pilot studies will help to ensure that the study is well designed and likely to be successful.

Further reading

DENZIN, N. K. AND LINCOLN, Y. S. (1994) ***Handbook of Qualitative Research.*** London: Sage. This book is a collection of chapters by experts in qualitative research. It is an excellent reference source.

DIX, A., FINLAY, J., ABOWD, G. AND BEALE, R. (1998) ***Human-Computer Interaction*** (2d ed.). London: Prentice Hall Europe. This book provides a useful introduction to evaluation.

SHNEIDERMAN, B. (1998) ***Designing the User Interface: Strategies for Effective Human-Computer Interaction*** (3rd ed.). Reading, MA: Addison-Wesley. This text provides an alternative way of categorizing evaluation techniques and offers a good overview.

ROBSON, C. (1993) ***Real World Research.*** Oxford, ***UK:*** Blackwell. This book offers a practical introduction to applied research and evaluation. It is very readable.

WHITESIDE, J., BENNETT, J., AND HOLTZBLATT, K. (1998) Usability engineering: our experience and evolution. In M. Helander (ed.), ***Handbook of Human-Computer Interaction.*** Amsterdam: North Holland. This chapter reviews the strengths and weakness of usability engineering and explains why ethnographic techniques can provide a useful alternative in some circumstances,791–817.