Chapter 10

Introducing evaluation

10.1 Introduction

Recently I met two web designers who, proud of their newest site, looked at me in astonishment when I asked if they had tested it with users. "No," they said "but we know it's OK." So, I probed further and discovered that they had asked the "web whiz-kids" in their company to look at it. These guys, I was told, knew all the tricks of web design.

The web's presence has heightened awareness about usability, but unfortunately this reaction is all too common. Designers assume that if they and their colleagues can use the software and find it attractive, others will too. Furthermore, they prefer to avoid doing evaluation because it adds development time and costs money. So why is evaluation important? Because without evaluation, designers cannot be sure that their software is usable and is what users want. But what do we mean by evaluation? There are many definitions and many different evaluation techniques, some of which involve users directly, while others call indirectly on an understanding of users' needs and psychology. In this book we define evaluation as the process of systematically collecting data that informs us about what it is like for a particular user or group of users to use a product for a particular task in a certain type of environment.

As you read in Chapter 9, the basic premise of user-centered design is that users' needs are taken into account throughout design and development. This is achieved by evaluating the design at various stages as it develops and by amending

it to suit users' needs (Gould and Lewis, 1985). The design, therefore, progresses in iterative cycles of design-evaluate redesign. Being an effective interaction designer requires knowing how to evaluate different kinds of systems at different stages of development. Furthermore, developing systems in this way usually turns out to be less expensive than fixing problems that are discovered after the systems have been shipped to customers (Karat, 1993). Studies also suggest that the business case for using systems with good usability is compelling (Dumas and Redish, 1999; Mayhew, 1999): thousands of dollars can be saved.

Many techniques are available for supporting design and evaluation. Chapter 9 discussed techniques for involving users in design and part of this involvement comes through evaluation. In this and the next four chapters you will learn how different techniques are used at different stages of design to examine different aspects of the design. You will also meet some of the same techniques that are used for gathering user requirements, but this time used to collect data to evaluate the design. Another aim is to show you how to do evaluation.

This chapter begins by discussing ***what*** evaluation is, ***why*** evaluation is important, and ***when*** to use different evaluation techniques and approaches. Then a case study is presented about the evaluation techniques used by Microsoft researchers and the Fred Hutchinson Cancer Research Center in developing HutchWorld (Cheng et al., 2000), a virtual world to support cancer patients, their families, and friends. This case study is chosen because it illustrates how a range of techniques is used during the development of a new product. It introduces some of the practical problems that evaluators encounter and shows how iterative product development is informed by a series of evaluation studies. The HutchWorld study also lays the foundation for the evaluation framework that is discussed in Chapter 11.

The main aims of this chapter are to:

- Explain the key concepts and terms used to discuss evaluation.
- Discuss and critique the HutchWorld case study.
- Examine how different techniques are used at different stages in the development of HutchWorld.
- Show how developers cope with real-world constraints in the development of HutchWorld.

10.2 What, why, and when to evaluate

Users want systems that are easy to learn and to use as well as effective, efficient, safe, and satisfying. Being entertaining, attractive, and challenging, etc. is also essential for some products. ***So,*** knowing what to evaluate, why it is important, and when to evaluate are key skills for interaction designers.

10.2.1 What to evaluate

There is a huge variety of interactive products with a vast array of features that need to be evaluated. Some features, such as the sequence of links to be followed to find an item on a website, are often best evaluated in a laboratory, since such a

setting allows the evaluators to control what they want to investigate. Other aspects, such as whether a collaborative toy is robust and whether children enjoy interacting with it, are better evaluated in natural settings, so that evaluators can see what children do when left to their own devices.

You may remember from Chapters 2, 6 and 9 that John Gould and his colleagues (Gould et al., 1990; Gould and Lewis, 1985) recommended three similar principles for developing the 1984 Olympic Message System:

- focus on users and their tasks
- observe, measure, and analyze their performance with the system
- design iteratively

Box 10.1 takes up the evaluation part of the 1984 Olympic Messaging System story and lists the many evaluation techniques used to examine different parts of the OMS during its development. Each technique supported Gould et al.'s three principles.

Since the OMS study, a number of new evaluation techniques have been developed. There has also been a growing trend towards observing how people interact with the system in their work, home, and other settings, the goal being to obtain a better understanding of how the product is (or will be) used in its intended setting. **For example, at work people are frequently being interrupted by phone calls, others** knocking at their door, email arriving, and so on—to the extent that many tasks are interrupt-driven. Only rarely does someone carry a task out from beginning to end without stopping to do something else. Hence the way people carry out an activity (e.g., preparing a report) in the real world is very different from how it may be observed in a laboratory. Furthermore, this observation has implications for the way products should be designed.

10.2.2 Why you need to evaluate

Just as designers shouldn't assume that everyone is like them, they also shouldn't presume that following design guidelines guarantees good usability, Evaluation is needed to check that users can use the product and like it. Furthermore, nowadays users look for much more than just a usable system, as the Nielsen Norman Group, a usability consultancy company, point out (www.nngroup.com):

> *"User experience" encompasses all aspects of the end-user's interaction . . . the first requirement for an exemplary user experience is to meet the exact needs of the customer, without fuss or bother. Next comes simplicity and elegance that produce products that are a joy to own, a joy to use."*

Bruce Tognazzini, another successful usability consultant, comments (www.asktog.com) that:

> *"Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results. If you don't have user-testing as an integral part of your design process you are going to throw buckets of money down the drain."*

BOX 10.1 The Story of the 1984 Olympic Messaging System

The 1984 Olympic Message System (OMS), a voice mail system, was developed by IBM so that Olympic Games contestants and their families and friends could send and receive messages (Gould et al., 1990). They could hear the message and the actual voice of the sender exactly as it was spoken. This system could be used from almost any push-button phone system around the world: this may not sound amazing when compared with today's technology, but in 1983 it was highly innovative.

Non-Olympians called their own country's National Olympic Committee using either push-button or dial telephones and spoke in their own language. They were helped to connect to OMS so that they could leave their messages. The voice message was immediately transferred by a central telephone operator to the message boxes of the Olympian for whom it was intended. The OMS worked in 12 languages. The kiosks looked like the one in Figure 10.1 and the dialog is shown in Figure 10.2

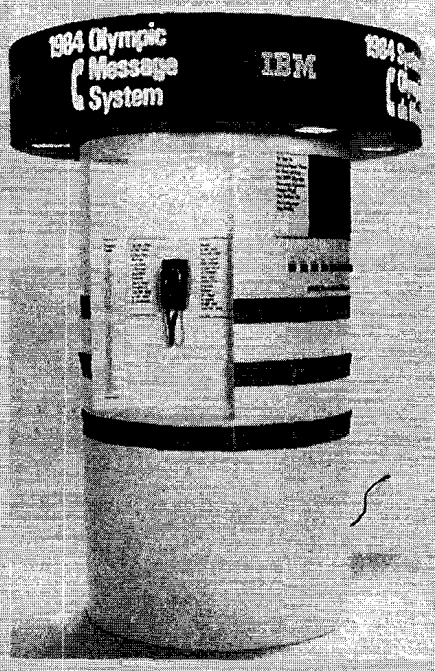

Figure 10.1 The Olympic message system kiosk.

During development, the evaluation activities included:

- Use of printed scenarios of the screens to get feedback from the Olympic committee and the Olympians themselves.
- Iteratively testing the user guides for the OMS with the Olympians, their families, and friends.
- Developing early simulations of a telephone keypad with a person speaking the commands back. These simulations really tested how much a user needed to know about the system, what feedback was needed, and any incorrect assumptions about user behavior made by the designers.
- Developing early demonstrations to test the reactions of people outside the US who did not know much about computers.
- An Olympian joining the design team to discuss ideas and provide feedback.
- Interviews with Olympians to make sure that the system being developed was what the users wanted.
- Overseas tests of the interface with friends and family.
- Free coffee and donut tests: 65 people were enticed to test the system in return for these treats.
- More traditional usability tests (discussed in Chapter 14) of the prototype involving about 100 participants.
- A 'try-to-destroy-it' test in which 24 computer science students were challenged to bring down the system. One of these tests involved all the students calling into the OMS at the same time. The students enjoyed the challenge and didn't need any other motivation!
- A pre-Olympic field test of the interface at an international event with competitors from 65 countries. The outcome of this test was surprising because, despite all the other testing, 57 different usability problems were recorded by the end of the five-day test period. The lesson for the design team was that the results of

field tests could be surprising. In this case they discovered that strong cultural differences affected how users from different countries used the OMS. Testers from Oman, Colombia, Pakistan, Japan, and Korea were unable to use the system. Gould and his colleagues comment that "watching this helplessness and hopelessness had a far greater impact than reading about it. It was embarrassing . . ." (Gould et al., 1990, p. 274).

- Two other tests examined the reliability of the system with heavy traffic generated by 2800 and 1000 people respectively.

This extensive evaluation was needed because the Olympics was such a high-profile event and IBM's reputation was at stake. Less intensive evaluation is more normal. However, the take-away message from this study is that the more evaluation with users, the better the final product.

Caller:	(Dials 213-888-8888.)
Operator:	Irish National Olympic Committee. Can I help you?
Caller:	I want to leave a message for my son, Michael.
Operator:	Is he from Ireland?
Caller:	Yes.
Operator:	How do you spell his name?
Caller:	*K-E-L-L-Y.*
Operator:	Thank you. Please hold for about 30 seconds while I connect you to the Olympic Message System.
Operator:	Are you ready?
Caller:	Yes.
OMS:	When you have completed your message, hang up and it will be automatically sent to Michael Kelly. Begin talking when you are ready.
Caller:	'Michael, your Mother and I will be hoping you win. Good luck.' (Caller hangs up.)

Figure 10.2 Parent leaving a voice message for an Olympian.

Tognazzini points out that there are five good reasons for investing in user testing:

1. Problems are fixed before the product is shipped, not after.
2. The team can concentrate on real problems, not imaginary ones.
3. Engineers code instead of debating.
4. Time to market is sharply reduced.
5. Finally, upon first release, your sales department has a rock-solid design it can sell without having to pepper their pitches with how it will all actually work in release 1.1 or 2.0.

Now that there is a diversity of interactive products, it is not surprising that the range of features to be evaluated is very broad. For example, developers of a new web browser may want to know if users find items faster with their product. Government authorities may ask if a computerized system for controlling traffic lights

results in fewer accidents. Makers of a toy may ask if six-year-olds can manipulate the controls and whether they are engaged by its furry case and pixie face. A company that develops the casing for cell phones may ask if the shape, size, and color of the case is appealing to teenagers. A new dotcom company may want to assess market reaction to its new home page design.

This diversity of interactive products, coupled with new user expectations, poses interesting challenges for evaluators, who, armed with many well tried and tested techniques, must now adapt them and develop new ones. As well as usability, user experience goals can be extremely important for a product's success, as discussed in Chapter 1.

ACTIVITY 10.1 Think of examples of the following systems and write down the usability and user experience features that are important for the success of each:

(a) a word processor

(b) a cell phone

(c) a website that sells clothes

(d) an online patient support community

Comment

(a) It must be as easy as possible for the intended users to learn and to use and it must be satisfying. Note, that wrapped into this are characteristics such as consistency, reliability, predictability, etc., that are necessary for ease of use.

(b) A cell phone must also have all the above characteristics; in addition, the physical design (e.g., color, shape, size, position of keys, etc.) must be usable and attractive (e.g., pleasing feel, shape, and color).

(c) A website that sells clothes needs to have the basic usability features too. In particular, navigation through the system needs to be straightforward and well supported. You may have noticed, for example, that some sites always show a site map to indicate where you are. This is an important part of being easy to use. So at a deeper level you can see that the meaning of "easy to use and to learn" is different for different systems. In addition, the website must be attractive, with good graphics of the clothes—who would want to buy clothes they can't see or that look unattractive? Trust is also a big issue in online shopping, so a well-designed procedure for taking customer credit card details is essential: it must not only be clear but must take into account the need to provide feedback that engenders trust.

(d) An online patient support group must support the exchange of factual and emotional information. So as well as the standard usability features, it needs to enable patients to express emotions either publicly or privately, using emoticons. Some 3D environments enable users to show themselves on the screen as avatars that can jump, wave, look happy or sad, move close to another person, or move away. Designers have to identify the types of social interactions that users want to express (i.e., sociability) and then find ways to support them (Preece, 2000).

From this selection of examples, you can see that success of some interactive products depends on much more than just usability. Aesthetic, emotional, engaging, and motivating qualities are important too.

Usability testing involves measuring the performance of typical users on typical tasks. In addition, satisfaction can be evaluated through questionnaires and interviews. As mentioned in Chapter 1, there has been a growing trend towards developing ways of evaluating the more subjective user-experience goals, like emotionally satisfying, motivating, fun to use, etc.

10.2.3 When to evaluate

The product being developed may be a brand-new product or an upgrade of an existing product. If the product is new, then considerable time is usually invested in market research. Designers often support this process by developing mockups of the potential product that are used to elicit reactions from potential users. As well as helping to assess market need, this activity contributes to understanding users' needs and early requirements. As we said in Chapter 8, sketches, screen mockups, and other low-fidelity prototyping techniques are used to represent design ideas. Many of these same techniques are used to elicit users' opinions in evaluation (e.g., questionnaires and interviews), but the purpose and focus of evaluation is different. The goal of evaluation is to assess how well a design fulfills users' needs and whether users like it.

In the case of an upgrade, there is limited scope for change and attention is focused on improving the overall product. This type of design is well suited to usability engineering in which evaluations compare user performance and attitudes with those for previous versions. Some products, such as office systems, go through many versions, and successful products may reach double-digit version numbers. In contrast, new products do not have previous versions and there may be nothing comparable on the market, so more radical changes are possible if evaluation results indicate a problem.

Evaluations done during design to check that the product continues to meet users' needs are know as *formative evaluations*. Evaluations that are done to assess the success of a finished product, such as those to satisfy a sponsoring agency or to check that a standard is being upheld, are know as *summative evaluation*. Agencies such as National Institute of Standards and Technology (NIST) in the USA, the International Standards Organization (ISO) and the British Standards Institute (BSI) set standards by which products produced by others are evaluated.

ACTIVITY 10.2 Re-read the discussion of the 1984 Olympic Messaging System (OMS) in Box 10.1 and briefly describe some of the things that were evaluated, why it was necessary to do the evaluations, and when the evaluations were done.

Comment Because the Olympic Games is such a high-profile event and IBM's reputation was at stake, the OMS was intensively evaluated throughout its development. We're told that early evaluations included obtaining feedback from Olympic officials with scenarios that used printed screens and tests of the user guides with Olympians, their friends, and family. Early evaluations of simulations were done to test the usability of the human-computer dialog. These were done first in the US and then with people outside of the US. Later on, more formal tests investigated how well 100 participants could interact with the system. The system's robustness was also

tested when used by many users simultaneously. Finally, tests were done with users from minority cultural groups to check that they could understand how to use the OMS.

So how do designers decide ***which*** evaluation techniques to use, ***when*** to use them, and ***how*** to use the findings? To address these concerns, we provide a case study showing how a range of evaluation techniques were used during the development of a new system. Based on this, we then discuss issues surrounding the "which, when, and how" questions relating to evaluation.

10.3 HutchWorld case study

HutchWorld is a distributed virtual community developed through collaboration between Microsoft's Virtual Worlds Research Group and librarians and clinicians at the Fred Hutchinson Cancer Research Center in Seattle, Washington. The system enables cancer patients, their caregivers, family, and friends to chat with one another, tell their stories, discuss their experiences and coping strategies, and gain emotional and practical support from one another (Cheng et. al., 2000). The design team decided to focus on this particular population because caregivers and cancer patients are socially isolated: cancer patients must often avoid physical contact with others because their treatments suppress their immune systems. Similarly, their caregivers have to be careful not to transmit infections to patients.

The big question for the team was how to make HutchWorld a useful, engaging, easy-to-use and emotionally satisfying environment for its users. It also had to provide privacy when needed and foster trust among participants. **A** common approach to evaluation in a large project like Hutchworld is to begin by carrying out a number of informal studies. Typically, this involves asking a small number of users to comment on early prototypes. These findings are then fed back into the iterative development of the prototypes. This process is then followed by more formal usability testing and field study techniques. Both aspects are illustrated in this case study. In addition, you will read about how the development team managed their work while dealing with the constraints of working with sick people in a hospital environment.

10.3.1 How the design team got started: early design ideas

Before developing this product, the team needed to learn about the patient experience at the Fred Hutchinson Center. For instance, what is the typical treatment process, what resources are available to the patient community, and what are the needs of the different user groups within this community? They had to be particularly careful about doing this because many patients were very sick. Cancer patients also typically go through bouts of low emotional and physical energy. Caregivers also may have difficult emotional times, including depression, exhaustion, and stress. Furthermore, users vary along other dimensions, such as education and experience with computers, age and gender and they come from different cultural backgrounds with different expectations.

It was clear from the onset that developing a virtual community for this population would be challenging, and there were many questions that needed to be an-

swered. For example, what kind of world should it be and what should it provide? What exactly do users want to do there? How will people interact? What should it look like? To get answers, the team interviewed potential users from all the stakeholder groups—patients, caregivers, family, friends, clinicians, and social support staff—and observed their daily activity in the clinic and hospital. They also read the latest research literature, talked to experts and former patients, toured the Fred Hutchinson (Hutch) research facilities, read the Hutch web pages, and visited the Hutch school for pediatric patients and juvenile patient family members. No stone was left unturned.

The development team decided that HutchWorld should be available for patients any time of day or night, regardless of their geographical location. The team knew from reading the research literature that participants in virtual communities are often more open and uninhibited about themselves and will talk about problems and feelings in a way that would be difficultin face-to-face situations. On the downside, the team also knew that the potential for misunderstanding is higher in virtual communities when there is inadequate non-verbal feedback (e.g., facial expressions and other body language, tone of voice, etc.). On balance, however, research indicates that social support helps cancer patients both in the psychological adjustments needed to cope and in their physical wellbeing. For example, research showed that women with breast cancer who received group therapy lived on average twice as long as those who did not (Spiegel, et **al.,** 1989). The team's motivation to create HutchWorld was therefore high. The combination of information from research literature and from observations and interviews with users convinced them that this was a worthwhile project. But what did they do then?

The team's informal visits to the Fred Hutchinson Center led to the development of an early prototype. They followed a user-centered development methodology. Having got a good feel for the users' needs, the team brainstormed different ideas for an organizing theme to shape the conceptual design—a conceptual model possibly based on a metaphor. After much discussion, they decided to make the design resemble the outpatient clinic lobby of the Fred Hutchinson Cancer Research Center. By using this real-world metaphor, they hoped that the users would easily infer what functionality was available in HutchWorld from their knowledge of the real clinic. The next step was to decide upon the kind of communication environment to use. Should it be synchronous or asynchronous? Which would support social and affective communications best? **A** synchronous chat environment was selected because the team thought that this would be more realistic and personal than an asynchronous environment. They also decided to include 3D photographic avatars so that users could enjoy having an identifiable online presence and could easily recognize each other.

Figure 10.3 shows the preliminary stages of this design with examples of the avatars. You can also see the outpatient clinic lobby, the auditorium, the virtual garden, and the school. Outside the world, at the top right-hand side of the screen, is a list of commands in a palette and a list of participants. On the right-hand side at the bottom is a picture of participants' avatars, and underneath the window is the textual chat window. Participants can move their avatars and make them gesture to tour the virtual environment. They can also click on objects such as pictures and interact with them.

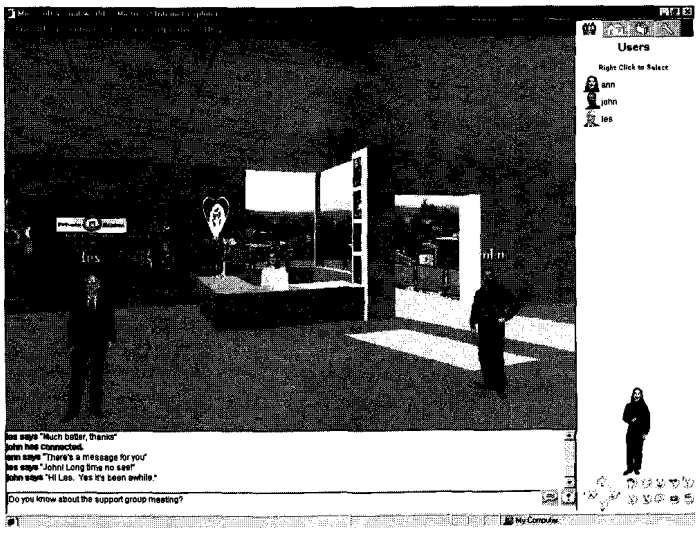

Figure 10.3 Preliminary design showing a view of the entrance into Hutch-World.

The prototype was reviewed with users throughout early development and was later tested more rigorously in the real environment of the Hutch Center using a variety of techniques. A Microsoft product called V-Chat was used to develop a second interactive prototype with the subset of the features in the preliminary design shown in Figure 10.3; however, only the lobby was fully developed, not the auditorium or school, as you can see in the new prototype in Figure 10.4.

Before testing could begin, the team had to solve some logistical issues. There were two key questions. Who would provide training for the testers and help for the patients? And how many systems were needed for testing and where should they be placed? As in many high-tech companies, the *Microsoft* team was used to short, market-driven production schedules, but this time they were in for a shock. Organizing the testing took much longer than they anticipated, but they soon

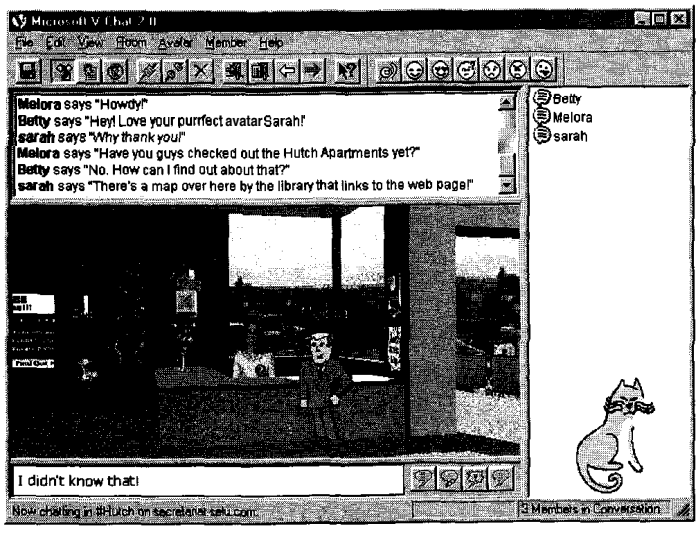

Figure 10.4 The Hutch V-Chat prototype.

learned to set realistic expectations that were in synch with hospital activity and the unexpected delays that occur when working with people who are unwell.

10.3.2 How was the testing done?

The team ran two main sets of user tests. The first set of tests was informally run onsite at the Fred Hutchinson Center in the hospital setting. After observing the system in use on computers located in the hospital setting, the team redesigned the software and then ran formal usability tests in the usability labs at Microsoft.

Test 1: Early observations onsite

In the informal test at the hospital, six computers were set up and maintained by Hutch staff members. A simple, scaled-back prototype of HutchWorld was built using the existing product, Microsoft V-Chat and was installed on the computers, which patients and their families from various hospital locations used. Over the course of several months, the team trained Hutch volunteers and hosted events in the V-Chat prototype. The team observed the usage of the space during unscheduled times, and they also observed the general usage of the prototype.

Test 1: What was learned?

This V-Chat test brought up major usability issues. First, the user community was relatively small, and there were never enough participants in the chat room for successful communication—a concept known as *critical mass.* In addition, many of the patients were not interested in or simultaneously available for chatting. Instead, they preferred asynchronous communication, which does not require an immediate response. Patients and their families used the computers for email, journals, discussion lists, and the bulletin boards largely because they could be used at any time and did not require others to be present at the same time. The team learned that a strong asynchronous base was essential for communication.

The team also observed that the users used the computers to play games and to search the web for cancer sites approved by Hutch clinicians. This information was not included in the virtual environment, and so users were forced to use many different applications. A more "unified" place to find all of the Hutch content was desired that let users rapidly swap among a variety of communication, information, and entertainment tasks.

Test 1: The redesign

Based on this trial, the team redesigned the software to support more asynchronous communication and to include a variety of communication, information, and entertainment areas. They did this by making HutchWorld function as a portal that provides access to information-retrieval tools, communication tools, games, and other types of entertainment. Other features were incorporated too, including email, a bulletin board, a text-chat, a web page creation tool, and a way of checking to see if anyone is around to chat with in the 3D world. The new portal version is show in Figure 10.5.

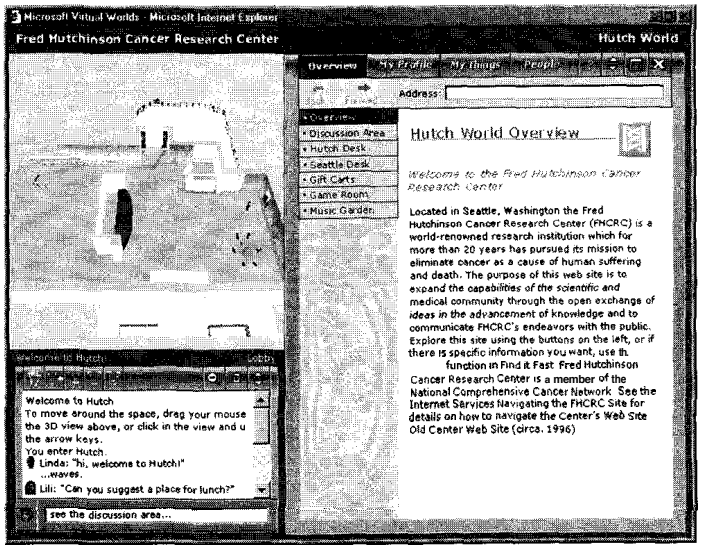

Figure 10.5 HutchWorld portal version.

Test 2: Usability tests

After redesigning the software, the team then ran usability tests in the Microsoft usability labs. Seven participants (four male and three female) were **tested. Four** of these participants had used chat rooms before and three were regular users. All had browsed the web and some used other communications software. The participants were told that they would use a program called HutchWorld that was designed to provide support for patients and their families. They were then given five minutes to explore HutchWorld. They worked independently and while they explored they provided a running commentary on what they were looking at, what they were thinking, and what they found confusing. This commentary was recorded on video and so were the screens that they visited, so that the Microsoft evaluator, who watched through a one-way mirror, had a record of what happened for later analysis. Participants and the evaluator interacted via a microphone and speakers. When the five-minute exploration period ended, the participants were asked to complete a series of *structured tasks* that were designed to test particular features of the HutchWorld interface.

These tasks focused on how participants:

- dealt with their virtual identity; that is, how they represented themselves and were perceived by others
- communicated with others
- got the information they wanted
- found entertainment

Figure 10.6 shows some of the structured tasks. Notice that the instructions are short, clearly written, and specific.

Welcome to the HutchWorld Usability Study

For this study we are interested in gaining a better understanding of the problems people have when using the program HutchWorld. HutchWorld is an all-purpose program created to offer information and social support to patients and their families at the Fred Hutchinson Cancer Research Center.

The following pages have tasks for you to complete that will help us achieve that better understanding.

While you are completing these tasks, it is important for us know what is going on inside your mind. Therefore, as you complete each task please tell us what you are looking at, what you are thinking about, what is confusing to you, and so forth.

Task #1: Explore HutchWorld

Your first task is to spend five minutes exploring HutchWorld.

A. First, open HutchWorld.

B. Now, explore!

Remember, tell us what you are looking at and what you are thinking about as you are exploring HutchWorld

Task #2: All about Your Identity in HutchWorld

A. Point to the 3 dimensional (3D) view of HutchWorld.

B. Point at yourself in the 3D view of HutchWorld.

C. Get a map view in the 3D view of HutchWorld.

D. Walk around in the 3D view: go forward, turn left and turn right.

E. Change the color of your shirt.

F. Change some information about yourself, such as where you are from.

Task #3: All about Communicating with Others

A. Send someone an email.

B. Read a message on the HutchWorld Bulletin Board.

C. Post a message on the HutchWorld Bulletin Board.

D. Check to see who is currently in HutchWorld.

E. Find out where the other person in HutchWorld is from.

F. Make the other person in HutchWorld a friend.

G. Chat with the other person in HutchWorld

H. Wave to the other person in HutchWorld.

I. Whisper to the other person in HutchWorld.

Task #4: All about Getting Information

A. Imagine you have never been to Seattle before. Your task is to find something to do.

B. Find out how to get to the Fred Hutchinson Cancer Research Center.

C. Go to your favorite website. [Or go to Yahoo: www.yahoo.com]

D. Once you have found a website, resize the screen so you can see the whole web page.

Figure 10.6 A sample of the structured tasks used in the HutchWorld evaluation.

Task #5: All about Entertainment

A. Find a game to play.
B. Get a gift from a Gift Cart and send yourself a gift.
C. Go and open your gift.

Figure 10.6 *(continued).*

During the study, a member of the development team role-played being a participant so that the real participants would be sure to have someone with whom to interact. The evaluator also asked the participants to fill out a short questionnaire after completing the tasks, with the aim of collecting their opinions about their experiences with HutchWorld. The questionnaire asked:

- What did you *Like* about HutchWorld?
- What did you not like about HutchWorld?
- What did you find confusing or difficult to use in HutchWorld?
- How would you suggest improving HutchWorld?

Test 2: What was learned from the usability tests?

When running the usability tests, the team collected masses of data that they had to make sense of by systematical analysis. The following discussion offers a snapshot of their findings. Some participants' problems started right at the beginning of the five-minute exploration. The login page referred to "virtual worlds" rather than the expected HutchWorld and, even though this might seem trivial, it was enough to confuse some users. This isn't unusual; developers tend to overlook small things like this, which is why evaluation is so important. Even careful, highly skilled developers like this team tend to forget that users do not speak their language. Fortunately, finding the "go" button was fairly straightforward. Furthermore, most participants read the welcome message and used the navigation list, and over half used the chat buttons, managed to move around the 3D world, and read the overview. But only one-third chatted and used the navigation buttons. The five-minute free-exploration data was also analyzed to determine what people thought of HutchWorld and how they commented upon the 3D view, the chat area, and the browse area.

Users' performance on the structured tasks was analyzed in detail and participant ratings were tabulated. Participants rated the tasks on a scale of 1–3 where 1 = easy, 2 = OK, 3 = difficult, and bold = needed help. Any activity that received an average rating above 1.5 across participants was deemed to need detailed review by the team. Figure 10.7 shows a fragment of the summary of the analysis.

In addition, the team analyzed all the problems that they observed during the tests. They then looked at all their data and drew up a table of issues, noting whether they were a priority to fix and listing recommendations for changes.

Structured Tasks

Participant number:	1	2	3	4	5	6	7	Average
Background Information								
Sex	F	F	M	M	F	M	M	3F, 4M
Age	37	41	43	54	46	44	21	40.9
years of college	4	2	4	4	4	1	2	3.0
hours of chat use in past year	0	3	0	0	365	200	170	105.4
hours of web use in past year	9	11	36	208	391	571	771	285.3
Structured Tasks								
Identify 3D view	1	1	1	1	1	1	1	1.0
Identity self in 3D view	1	2	1	1	1	1	1	1.1
Get a map view of 3D view	1	2	2	1	2	**3**	1	1.7
Walk in 3D view	1	3	2	1	**3**	2	1	1.9
Change color of shirt	1	1	**3**	**3**	2	3	2	2.1
Change where self is from	1	1	**3**	1	1	3	1	1.6
Find place to send email	1	**3**	3	1	3	2	2	2.1
Read a bulletin board message	2	1	**3**	1	1	1	–	1.5
Post a bulletin board message	1	**3**	3	3	2	2	–	2.3
Check to see who is currently on	1	3	1	**3**	2	3	2	2.1
Find out where the other person is from	1	1	2	1	1	3	2	1.6
Make the other person a friend	1	1	3	1	1	2	1	1.4
Chat with the other person	**3**	1	**3**	1	1	**3**	1	1.9
Wave to the other person	1	1	1	1	1	1	1	1.0
Whisper to the other person	1	**3**	2	2	1	2	1	1.7
Find something to do in Seattle	2	1	2	1	1	1	2	1.4
Find out how to get to FHCRC	1	3	3	2	1	1	2	1.9
Go to a website	1	**3**	2	**3**	**3**	1	1	2.0
Resize web screen	1	**3**	2	2	2	**3**	1	2.0
Find a game to play	1	1	2	1	1	1	2	1.3
Send self a gift	1	**3**	**3**	**3**	**3**	**3**	3	2.7
Open gift	**3**	1	2	*3*	**3**	**3**	**3**	2.6
Participant Average:	1.3	1.9	2.2	1.7	**1.7**	2.0	1.6	

The following descriptions provide examples of some of the problems participants experience.

Get map **view.** People generally did not immediately know how to find the map view. However, they knew to look in the chat buttons, and by going through the buttons they found the map view.

Walk in *3D* view. People found the use of the mouse to move the avatar awkward, especially when they were trying to turn around. However, once they were used to using the mouse they had no difficulty. For a couple of people, it was not clear to them that they should click on the avatar and drag it in the desired direction. A couple of people tried to move by clicking the place they wanted to move to.

Figure 10.7 Participant information and ratings of difficulty in completing the structured tasks. 1 = easy, 2 = okay, 3 = difficult and bold = needed help.

Issue#	Issue Priority	Issue	Recommendation
1	high	Back button sometimes not working.	Fix back button.
2	high	People are not paying attention to navigation buttons.	Make navigation buttons more prominent.
3	low	Fonts too small, hard to read for some people.	Make it possible to change fonts. Make the font colors more distinct from the background color.
4	low	When navigating, people were not aware overview button would take them back to the main page.	Change the overview button to a home button, change the wording of the overview page accordingly.
5	medium	"Virtual worlds" wording in login screen confusing.	Change wording to "HutchWorld".
6	high	People frequently clicking on objects in 3D view expecting something to happen.	Make the 3D view have links to web pages. For example, when people click on the help desk the browser area should show the help desk information.
7	low	People do not readily find map view button.	Make the icon on the map view button more map-like.
8	medium	Moving avatar with mouse took some getting used to.	Encourage the use of the keyboard. Mention clicking and dragging the avatar in the welcome.
9	low	People wanted to turn around in 3D view, but it was awkward to do so.	Make one of the chat buttons a button that lets you turn around.
10	medium	Confusion about the real worldlvirtual world distinction.	Change wording of overview description, to make clear Hutch-World is a "virtual" place made to "resemble" the FHCRC, and is a place where anybody can go.
11	high	People do not initially recognize that other real people could be in HutchWorld, that they can talk to them and see them.	Change wording of overview description, to make clear Hutch-World is a place to "chat" with others who are "currently in" the virtual HutchWorld.
12	high	People not seeing/finding the chat window. Trying to chat to people from the people list where other chat-like features are (whisper, etc.)	Make chat window more prominent. Somehow link chat-like features of navigation list to chat window. Change wording of chat window. Instead of type to speak here. type to chat here.

Figure 10.8 A fragment of the table showing problem rankings.

13	low	Who is here list and who has been here list confused.	Spread them apart more in the people list.
14	medium	Difficulty in finding who is here.	Change People button to "Who is On" button.
15	low	Went to own profile to make someone a friend.	Let people add friends at My profile
16	low	Not clear how to append/reply to a discussion in the bulletin board.	Make an append button pop up when double clicking on a topic. Change wording from "post a message" to "write a message" or "add a message".
17	low	Bulletin board language is inconsistent.	Change so it is either a bulletin board, or a discussion area.

Figure 10.8 *(continued).*

Figure 10.8 shows part of this table. Notice that issues were ranked in priority: low, medium, **and** high. **There were** just five high-ranking problems that absolutely had to be fixed:

- The back button did not always work.
- People were not paying attention to navigation buttons, so they needed to be more prominent.
- People frequently clicked on objects in the 3D view and expected something to happen. A suggestion for fixing this was to provide links to a web page.
- People did not realize that there could be other real people in the 3D world with whom they could chat, so the wording in the overview description had to be changed.
- People were not noticing the chat window and instead were trying to chat to people in the participant list. The team needed to clarify the instructions about where to chat.

In general, most users found the redesigned software easy to use with little instruction. By running a variety of tests, the informal onsite test, and the formal usability test, key problems were identified at an early stage and various usability issues could be fixed before the actual deployment of the software.

10.3.3 Was it tested again?

Following the usability testing, there were more rounds of observation and testing with six new participants, two males and four females. These tests followed the same general format as those just described but this time they tested multiple users at once, to ensure that the virtual world supported multiuser interactions. The tests were also more detailed and focused. This time the results were more positive, but

DILEMMA When Is It Time to Stop Testing?

Was HutchWorld good enough after these evaluations? When has enough testing been done? This frequently asked question is difficult to answer. Few developers have the luxury of testing as thoroughly as John Gould and his colleagues when developing the 1984 Olympic Messaging System (Gould and Lewis, 1990), or even as much as Microsoft's HutchWorld team. Since every test you do will reveal some area where improvement can be made, you cannot assume that there will be a time when the system is perfect: no system is ever perfect. Normally schedule and budget constraints determine when to stop. Joseph Dumas and Ginny Redish, established usability consultants, point out that for iterative design and testing to be successful, each test should take as little time as possible while still yielding useful information and without burdening the team (Dumas and Redish, 1999).

of course there were still usability problems to be fixed. Then the question arose: what to do next? In particular, had they done enough testing (see Dilemma)?

After making a few more fixes, the team stopped usability testing with specific tasks. But the story didn't end here. The next step was to show HutchWorld to cancer patients and caregivers in a focus-group setting at the Fred Hutchinson Cancer Research Center to get their feedback on the final version. Once the **team made** adjustments to HutchWorld in response to the focus-group feedback, the final step was to see how well HutchWorld worked in a real clinical environment. It was therefore taken to a residential building used for long-term patient and family stays that was fully wired for Internet access. Here, the team observed what happened when it was used in this natural setting. In particular, they wanted to find out how HutchWorld would integrate with other aspects of patients' lives, particularly with their medical care routines and their access to social support. This informal observation allowed them to examine patterns of use and to see who used which parts of the system, when, and why.

10.3.4 Looking *to* the future

Future studies were planned to evaluate the effects of the computers and the software in the Fred Hutchinson Center. The focus of these studies will be the social support and wellbeing of patients and their caregivers in two different conditions. There will be a control condition in which users (i.e., patients) live in the residential building without computers and an experimental condition in which users live in similar conditions but with computers, Internet access, and HutchWorld. The team will evaluate the user data (performance and observation) and surveys collected in the study to investigate key questions, including:

- How does the computer and software impact the social wellbeing of patients and their caregivers?
- What type of computer-based communication best supports this patient community?
- What are the general usage patterns? i.e., which features were used and at what time of day were they used, etc.?

- How might any medical facility use computers and software like HutchWorld to provide social support for its patients and caregivers?

There is always more to learn about the efficacy of a design and how much users enjoy using a product, especially when designing innovative products like HutchWorld for new environments. This study will provide a longer-term view of how HutchWorld is used in its natural environment that is not provided by the other evaluations. It's an ambitious plan because it involves a comparison between two different environmental settings, one that has computers and HutchWorld and one that doesn't (see Chapter 13 for more on experimental design).

ACTIVITY 10.3

(a) The case study does not say much about early evaluation to test the conceptual design shown in Figure 10.5. What do you think happened?

(b) The evaluators recorded the gender of participants and noted their previous experience with similar systems. Why is this important?

(c) Why do you think it was important to give participants a five-minute exploration period?

(d) ***Triangulation*** is a term that describes how different perspectives are used to understand a problem or situation. Often different techniques are used in triangulation. Which techniques were triangulated in the evaluations of the HutchWorld prototype?

(e) The evaluators collected participants' opinions. What kinds of concerns do you think participants might have about using HutchWorld? Hints: personal information, medical information, communicating feelings, etc.

Comment

(a) There was probably much informal discussion with representative users: patients, medical staff, relatives, friends, and caregivers. The team also visited the clinic and hospital and observed what happened there. They may also have discussed this with the physicians and administrators.

(b) It is possible that our culture causes men and women to react differently in certain circumstances. Experience is an even more important influence than gender, so knowing how much previous experience users have had with various types of computer systems enables evaluators to make informed judgments about their performance. Experts and novices, for example, tend to behave very differently.

(c) The evaluators wanted to see how participants reacted to the system and whether or not they could log on and get started. The exploration period also gave the participants time to get used to the system before doing the set tasks.

(d) Data was collected from the five-minute exploration, from performance on the structured tasks, and from the user satisfaction questionnaire.

(e) Comments and medical details are personal and people want privacy. Patients might be concerned about whether the medical information they get via the computer and from one another is accurate. Participants might be concerned about how clearly and accurately they are communicating because non-verbal communication is reduced online.

10.4 Discussion

In both HutchWorld and the 1984 Olympic Messaging System, a variety of evaluation techniques were used at different stages of design to answer different questions. "Quick and dirty" observation, in which the evaluators informally examine how a prototype is used in the natural environment, was very useful in early design. Following this with rounds of usability testing and redesign revealed important usability problems. However, usability testing alone is not sufficient. Field studies were needed to see how users used the system in their natural environments, and sometimes the results were surprising. For example, in the OMS system users from different cultures behaved differently. A key issue in the HutchWorld study was how use of the system would fit with patients' medical routines and changes in their physical and emotional states. Users' opinions also offered valuable insights. After all, if users don't like a system, it doesn't matter how successful the usability testing is: they probably won't use it. Questionnaires and interviews were used to collect user's opinions.

An interesting point concerns not only how the different techniques can be used to address different issues at different stages of design, but also how these techniques complement each other. Together they provide a broad picture of the system's usability and reveal different perspectives. In addition, some techniques are better than others for getting around practical problems. This is a large part of being a successful evaluator. In the HutchWorld study, for example, there were not many users, so the evaluators needed to involve them sparingly. For example, a technique requiring 20 users to be available at the same time was not feasible in the HutchWorld study, whereas there was no problem with such an approach in the OMS study. Furthermore, the OMS study illustrated how many different techniques, some of which were highly opportunistic, can be brought into play depending on circumstances. Some practical issues that evaluators routinely have to address include:

- what to do when there are not many users
- how to observe users in their natural location (i.e., field studies) without disturbing them

  having appropriate equipment available
- dealing with short schedules and low budgets
- not disturbing users or causing them duress or doing anything unethical
- collecting "useful" data and being able to analyze it
- selecting techniques that match the evaluators' expertise

There are many evaluation techniques from which to choose and these practical issues play a large role in determining which are selected. Furthermore, selection depends strongly on the stage in the design and the particular questions to be answered. In addition, each of the disciplines that contributes to interaction design has preferred bodies of theory and techniques that can influence this choice. These issues are discussed further in the next chapter.

Assignment

1. Reconsider the HutchWorld design and evaluation case study and note *what* was evaluated, *why* and *when,* and *what* was learned at each stage?
2. How was the design advanced after each round of evaluation?
3. What were the main constraints that influenced the evaluation?
4. How did the stages and choice of techniques build on and complement each other (i.e., triangulate)?
5. Which parts of the evaluation were directed at usability goals and which at user experience goals? Which additional goals not mentioned in the study could the evaluations have focused upon?

Summary

The aim of this chapter was to introduce basic evaluation concepts that will be revisited and built on in the next four chapters. We selected the HutchWorld case study because it illustrates how a team of designers evaluated a novel system and coped with a variety of practical constraints. It also shows how different techniques are needed for different purposes and how techniques are used together to gain different perspectives on a product's usability. This study highlights how the development team paid careful attention to usability and user experience goals as they designed and evaluated their system.

Key points

- Evaluation and design are very closely integrated in user-centered design.
- Some of the same techniques are used in evaluation as in the activity of establishing requirements and identifying users' needs, but they are used differently (e.g., interviews and questionnaires, etc.).
- Triangulation involves using combinations of techniques in concert to get different perspectives or to examine data in different ways.

  Dealing with constraints, such as gaining access to users or accommodating users' routines, is an important skill for evaluators to develop.

Further reading

CHENG, L., STONE, L., FARNHAM, S., CLARK, A. M., AND ZANER-GODSEY, M. (2000) ***Hutchworld: Lessons Learned. A Collaborative Project: Fred Hutchinson Cancer Research Center & Microsofi Research.*** In the Proceedings of the Virtual Worlds Conference 2000, Paris, France. This paper describes the HutchWorld study and, as the title suggests, it discusses the design lessons that were learned. It also describes the evaluation studies in more detail.

GOULD, J. D., BOIES, S. J., LEVY, S., RICHARDS, J. T., AND SCHOONARD, J. (1990). The 1984 Olympic Message System: A test of behavioral principles of system design. In J. Preece and L. Keller (eds.), ***Human-Computer Interaction (Readings).*** Prentice Hall International Ltd., Hemel Hempstead, UK: 260–283. This edited paper tells the story of the design and evaluation of the OMS.

GOULD, J. D., BOIES, S. J., LEVY, S., RICHARDS, J. T., AND SCHOONARD, J. (1987). The 1984 Olympic Message System: a test of behavioral principles of systems design. ***Communications of the ACM,*** 30(9), 758–769. This is the original, full version of the OMS paper.