CHAPTER VII

7. Evaluation Techniques

7. Evaluation Techniques

- Evaluation tests the usability, functionality and acceptability of an interactive system.
- Evaluation may take place:
  – in the laboratory
  – in the field.
- Some approaches are based on expert evaluation:
  – analytic methods
  – review methods
  – model-based methods.
- Some approaches involve users:
  – experimental methods
  – observational methods
  – query methods.
- An evaluation method must be chosen carefully and must be suitable for the job.

7.1 What Is Evaluation?

Even if such a process is used, we still need to assess our designs and test our systems to ensure that they actually behave as we expect and meet user requirements. This is the role of evaluation.

7.2 Goals of Evaluation

Evaluation has three main goals:

- to assess the extent and accessibility of the system's functionality,
- to assess users' experience of the interaction,
- and to identify any specific problems with the system.

Evaluation at this level may also include measuring the user's performance with the system, to assess the effectiveness of the system in supporting the task. In addition to evaluating the

system design in terms of its functional capabilities, it is important to assess the user's experience of the interaction and its impact upon him.

The final goal of evaluation is to identify specific problems with the design. These may be aspects of the design which, when used in their intended context, cause unexpected results, or confusion amongst users.

7.3 Evaluation Through Expert Analysis

If the design itself can be evaluated, expensive mistakes can be avoided, since the design can be altered prior to any major resource commitments. Typically, the later in the design process that an error is discovered, the more costly it is to put right and, therefore, the less likely it is to be rectified.

These depend upon the designer, or a human factors expert, taking the design and assessing the impact that it will have upon a typical user. The basic intention is to identify any areas that are likely to cause difficulties because they violate known cognitive principles, or ignore accepted empirical results.

We will consider four approaches to expert analysis:

- cognitive walkthrough,
- heuristic evaluation,
- the use of models and
- use of previous work.

7.3.1 Cognitive walkthrough

The origin of the cognitive walkthrough approach to evaluation is the code walk through familiar in software engineering. Walkthroughs require a detailed review of a sequence of actions. In the code walkthrough, the sequence represents a segment of the program code that is stepped through by the reviewers to check certain characteristics(for example, that coding style is adhered to, conventions for spelling variables versus procedure calls, and to check that system-wide invariants are not violated).

In the cognitive walkthrough, the sequence of actions refers to the steps that an interface will require a user to perform in order to accomplish some known task.

To do a walkthrough (the term walkthrough from now on refers to the cognitive walkthrough, and not to any other kind of walkthrough), you need four things:

1. A specification or prototype of the system. It doesn't have to be complete, but it should be fairly detailed. Details such as the location and wording for a menu can make a big difference.
2. A description of the task the user is to perform on the system. This should be are presentative task that most users will want to do.
3. A complete, written list of the actions needed to complete the task with the proposed system.

4. An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.

7.3.2 Heuristic evaluation

A heuristic is a guideline or general principle or rule of thumb that can guide a design decision or be used to critique a decision that has already been made. *Heuristic evaluation*, developed by Jakob Nielsen and Rolf Molich, is a method for structuring the critique of a system using a set of relatively simple and general heuristics.

The general idea behind heuristic evaluation is that several evaluators independently critique a system to come up with potential usability problems. It is important that there be several of these evaluators and that the evaluations be done independently.

Nielsen's ten heuristics are:
1. **Visibility of system status**
2. **Match between system and the real world**
3. **User control and freedom**
4. **Consistency and standards**
5. **Error prevention**
6. **Recognition rather than recall**
7. **Flexibility and efficiency of use**
8. **Aesthetic and minimalist design**
9. **Help users recognize, diagnose and recover from errors**
10. **Help and documentation**

Once each evaluator has completed their separate assessment, all of the problems are collected and the mean severity ratings calculated. The design team will then determine the ones that are the most important and will receive attention first.

7.3.3 Model-based evaluation

A third expert-based approach is the use of models. Certain cognitive and design models provide a means of combining design specification and evaluation into the same framework.

Dialog models can also be used to evaluate dialog sequences for problems, such as unreachable states, circular dialogs and complexity. Models such as state transition networks are useful for evaluating dialog designs prior to implementation.

7.3.4 Using previous studies in evaluation

Experimental psychology and human–computer interaction between them possess a wealth of experimental results and empirical evidence. Some of this is specific to a particular domain, but much deals with more generic issues and applies in a variety of situations.

A final approach to expert evaluation exploits this inheritance, using previous results as evidence to support (or refute) aspects of the design. It is expensive to repeat experiments continually and an expert review of relevant literature can avoid the need to do so. It should be noted that experimental results cannot be expected to hold arbitrarily across contexts.

7.4 Evaluation through user Participation

The techniques we have considered so far concentrate on evaluating a design or system through analysis by the designer, or an expert evaluator, rather than testing with actual users. However, useful as these techniques are for filtering and refining the design, they are not a replacement for actual usability testing with the people for whom the system is intended: the users.

These include:

- empirical or experimental methods,
- observational methods,
- query techniques, and
- methods that use physiological monitoring, such as eye tracking and measures of heart rate and skin conductance.

7.4.1 Styles of evaluation

Before we consider some of the techniques that are available for evaluation with users, we will distinguish between two distinct evaluation styles: those performed under laboratory conditions and those conducted in the work environment or 'in the field'.

Laboratory studies

In the first type of evaluation studies, users are taken out of their normal work environment to take part in controlled tests, often in a specialist usability laboratory (although the 'lab' may simply be a quiet room). This approach has a number of benefits and disadvantages.

A well-equipped usability laboratory may contain sophisticated audio/visual recording and analysis facilities, two-way mirrors, instrumented computers and the like, which cannot be replicated in the work environment.

There are, however, some situations where laboratory observation is the only option, for example, if the system is to be located in a dangerous or remote location, such as a space station. Also some very constrained single-user tasks may bead equate performed in a laboratory.

Field studies

The second type of evaluation takes the designer or evaluator out into the user's work environment in order to observe the system in action. Again this approach has its pros and cons.

High levels of ambient noise, greater levels of movement and constant interruptions, such as phone calls, all make field observation difficult. However, the very 'open' nature of the situation means that you will observe interactions between systems and between individuals that would have been missed in a laboratory study.

7.4.2 Empirical methods: experimental evaluation

One of the most powerful methods of evaluating a design or an aspect of a design is to use a controlled experiment. This provides empirical evidence to support a particular claim or hypothesis. It can be used to study a wide range of different issue sat different levels of detail.

Any experiment has the same basic form. The evaluator chooses a hypothesis to test, which can be determined by measuring some attribute of participant behavior.

Participants

The choice of participants is vital to the success of any experiment. In evaluation experiments, participants should be chosen to match the expected user population as closely as possible. If participants are not actual users, they should be chosen to be of a similar age and level of education as the intended user group.

A second issue relating to the participant set is the sample size chosen. Often this is something that is determined by pragmatic considerations: the availability of participants is limited or resources are scarce.

Variables

Experiments manipulate and measure variables under controlled conditions, in order to test the hypothesis. There are two main types of variable: those that are 'manipulated' or changed (known as the independent variables) and those that are measured (the dependent variables).

Independent variables are those elements of the experiment that are manipulated to produce different conditions for comparison. Examples of independent variables in evaluation experiments are interface style, level of help, number of menu items and icon design.

Dependent variables, on the other hand, are the variables that can be measured in the experiment, their value is 'dependent' on the changes made to the independent variable.

The dependent variable must be measurable in some way, it must be affected by the independent variable, and, as far as possible, unaffected by other factors. Common choices of dependent variable in evaluation experiments are the time taken to complete a task, the number of errors made, user preference and the quality of the user's performance.

Hypotheses

A hypothesis is a prediction of the outcome of an experiment. It is framed in terms of the independent and dependent variables, stating that a variation in the independent Variable will cause a difference in the dependent variable. The aim of the experiment is to show that this prediction is correct.

Experimental design

In order to produce reliable and generalizable results, an experiment must be carefully designed. We have already looked at a number of the factors that the experimenter must consider in the design, namely the participants, the independent and dependent variables, and the hypothesis.

The first phase in experimental design is to choose the hypothesis: to decide exactly what it is you are trying to demonstrate.

The next step is to decide on the *experimental method* that you will use. There are two main methods: *between-subjects* and *within-subjects*.

In a between-subjects(or *randomized*) design, each participant is assigned to a different condition. There are at least two conditions: the experimental condition (in which the variable has been manipulated) and the control, which is identical to the experimental condition except for this manipulation.

There may, of course, be more than two groups, depending on the number of independent variables and the number of levels that each variable can take.

The advantage of a between-subjects design is that any learning effect resulting from the user performing in one condition and then the other is controlled: each user performs under only one condition.

The disadvantages are that a greater number of participants are required, and that significant variation between the groups can negate any results. Also, individual differences between users can bias the results.

The second experimental design is within-subjects (or *repeated measures*). Here each user performs under each different condition. This design can suffer from transfer of learning effects, but this can be lessened if the order in which the conditions are tackled is varied between users. There is also less chance of effects from variation between participants.

Statistical Measures

The first two rules of statistical analysis are to *look* at the data and to *save* the data. It is easy to carry out statistical tests blindly when a glance at a graph, histogram or table of results

would be more instructive. In particular, looking at the data can expose *outliers*, single data items that are very different from the rest.

Variables can be classified as either *discrete variables* or *continuous variables*. A discrete variable can only take a finite number of values or *levels*, for example, a screen color that can be red, green or blue.

A third sort of test is the contingency table, where we classify data by several discrete attributes and then count the number of data items with each attribute combination.

Examples of questions one might ask about the data are as follows:
Is there a difference?
How big is the difference?
How accurate is the estimate?

Identify your hypothesis, participant group, dependent and independent variables, experimental design, task and analysis approach.

Answer The following is only an example of the type of experiment that might be devised.
Participants Taken from user population.
Hypothesis Color coding will make selection more accurate.
IV (Independent Variable) Color coding.
DV (Dependent Variable) Accuracy measured as number of errors.
Design Between-groups to ensure no transfer of learning (or within-groups with appropriate safeguards if participants are scarce).
Task The interfaces are identical in each of the conditions, except that, in the second, color is added to indicate related menu items.
Analysis *t* test.

Studies of groups of users

So far we have considered the experimental evaluation of single-user systems. Experiments to evaluate elements of group systems bring additional problems. Given the complexities of human–human communication and group working, it is hardly surprising that experimental studies of groups and of groupware are more difficult than the corresponding single-user experiments already considered.

The participant groups To organize, say, 10 experiments of a single-user system requires 10 participants.

The experimental task Choosing a suitable task is also difficult. We may want to test a variety of different task types: creative, structured, information passing, and so on. Also, the tasks must encourage active cooperation, either because the task requires consensus, or because information and control is distributed among the participants.

Data gathering Even in a single-user experiment we may well use several video cameras as well as direct logging of the application. In a group setting this is replicated for each participant. So for a three-person group, we are trying to synchronize the recording of six or more video sources and three keystroke logs.

Field studies with groups There are, of course, problems with taking groups of users and putting them in an experimental situation. If the groups are randomly mixed, then we are effectively examining the process of group formation, rather than that of a normal working group.

7.4.3 Observational Techniques

A popular way to gather information about actual use of a system is to observe users interacting with it.

Think aloud and cooperative evaluation

Think aloud process has a number of advantages:

- the process is less constrained and therefore easier to learn to use by the evaluator
- the user is encouraged to criticize the system
- the evaluator can clarify points of confusion at the time they occur and so maximize the effectiveness of the approach for identifying problem areas.

The usefulness of think aloud, cooperative evaluation and observation in general is largely dependent on the effectiveness of the recording method and subsequent analysis. The record of an evaluation session of this type is known as a *protocol*, and there are a number of methods from which to choose.

Protocol Analysis

Methods for recording user actions include the following:

Paper and pencil This is primitive, but cheap, and allows the analyst to note interpretations and extraneous events as they occur. However, it is hard to get detailed information, as it is limited by the analyst's writing speed.

Audio recording This is useful if the user is actively 'thinking aloud'. However, it may be difficult to record sufficient information to identify exact actions in later analysis, and it can be difficult to match an audio recording to some other form of protocol (such as a handwritten script).

Video recording This has the advantage that we can see *what* the participant is doing (*as long as* the participant stays within the range of the camera).

Computer logging It is relatively easy to get a system automatically to record user actions at a keystroke level, particularly if this facility has been considered early in the design.

User notebooks The participants themselves can be asked to keep logs of activity / problems. This will obviously be at a very coarse level – at most, records every few minutes and, more likely, hourly or less.

Automatic Protocol Analysis Tools

Analyzing protocols, whether video, audio or system logs, is time consuming and tedious by hand. It is made harder if there is more than one stream of data to synchronize. One solution to this problem is to provide automatic analysis tools to support the task.

7.4.4 Query Techniques

Another set of evaluation techniques relies on asking the user about the interface directly. Query techniques can be useful in eliciting detail of the user's view of a system. They embody the philosophy that states that the best way to find out how a system meets user requirements is to 'ask the user'.

There are a number of styles of question that can be included in the questionnaire. These include the following:

General These are questions that help to establish the background of the user and his place within the user population. They include questions about age, sex, occupation, place of residence, and so on.

Open-ended These ask the user to provide his own unprompted opinion on a question, for example 'Can you suggest any improvements to the interface?'.

Scalar These ask the user to judge a specific statement on a numeric scale, usually corresponding to a measure of agreement or disagreement with the statement.

Multi-choice Here the respondent is offered a choice of explicit responses, and may be asked to select only one of these, or as many as apply.

Ranked These place an ordering on items in a list and are useful to indicate a user's preferences. Answer Assume that all users have used both systems.

Questionnaire
Consider the following questions in designing the questionnaire:

- what information is required?
- how is the questionnaire to be analyzed?

You are particularly interested in user preferences so questions should focus on different aspects of the systems and try to measure levels of satisfaction. The use of scales will make responses for each system easier to compare.

7.4.5 Evaluation through monitoring physiological responses

One of the problems with most evaluation techniques is that we are reliant on observationand the users telling us what they are doing and how they are feeling. What if we were able to measure these things directly? Interest has grown recently in the use of what is sometimes called objective usability testing, ways of monitoring physiological aspects of computer use.

Eye tracking for usability evaluation

There are many possible measurements related to usability evaluation including:
Number of fixations
Fixation duration
Scan path

Physiological Measurements

Physiological measurement involves attaching various probes and sensors to the user

These measure a number of factors:
Heart activity,
Activity of the sweat glands
Electrical activity in muscle
Electrical activity in the brain

7.5 Choosing an Evaluation Method

Factors Distinguishing Evaluation Techniques

We can identify at least eight factors that distinguish different evaluation techniques and therefore help us to make an appropriate choice. These are:

- the stage in the cycle at which the evaluation is carried out
- the style of evaluation
- the level of subjectivity or objectivity of the technique
- the type of measures provided
- the information provided
- the immediacy of the response
- the level of interference implied
- the resources required.

1. **Design vs. implementation**
2. **Laboratory vs. field studies**
3. **Subjective vs. objective**
4. **Qualitative vs. quantitative measures**
5. **Information provided**

6. **Immediacy of response**
7. **Intrusiveness**
8. **Resources**

7.6 Universal Design

- Universal design is about designing systems so that theycan be used by anyone in any circumstance.
- Multi-modal systems are those that use more than onehuman input channel in the interaction.
- These systems may, for example, use:
  – speech
  – non-speech sound
  – touch
  – handwriting
  – gestures.
- Universal design means designing for diversity, including:
  – people with sensory, physical or cognitive impairment
  – people of different ages
  – people from different cultures and backgrounds.

Universal design is the process of designing products so that they can be used by as many people as possible in as many situations as possible. In our case, this means particularly designing interactive systems that are usable by anyone, with any range of abilities, using any technology platform. This can be achieved by designing systems either to have built in redundancy or to be compatible with assistive technologies.

7.7 Universal Design Principles

In the late 1990s a group at North Carolina State University in the USA proposed seven general principles of universal design. These were intended to cover all areas of design and are equally applicable to the design of interactive systems. These principles give us a framework in which to develop universal designs.

- *equitable use*: the design is useful to people with a range of abilities and appealing to all. No user is excluded or stigmatized. Where appropriate, security, privacy and safety provision should be available to all.
- *flexibility in use*: the design allows for a range of ability and preference, through choice of methods of use and adaptivity to the user's pace, precision and custom.
- *simple and intuitive to use*, regardless of the knowledge, experience, language or level of concentration of the user.
- *perceptible information*: the design should provide effective communication of information regardless of the environmental conditions or the user's abilities. Presentation should support the range of devices and techniques used to access information by people with different sensory abilities.

- *tolerance for error*: minimizing the impact and damage caused by mistakes or unintended behavior. Potentially dangerous situations should be removed or made hard to reach. Potential hazards should be shielded by warnings.
- *low physical effort*: systems should be designed to be comfortable to use, minimizing physical effort and fatigue. The physical design of the system should allow the user to maintain a natural posture with reasonable operating effort.
- *size and space for approach and use*: the placement of the system should be such that it can be reached and used by any user regardless of body size, posture or mobility.

7.8 Multi-Modal Interaction

In addition, such multi-sensory or multi-modal systems support the principle of redundancy required for universal design, enabling users to access the system using the mode of interaction that is most appropriate to their abilities.

The majority of interactive computer systems are predominantly visual in their interactive properties; often WIMP based, they usually make use of only rudimentary sounds while adding more and more visual information to the screen.

By utilizing the other sensory channels, the visual channel can be relieved of the pressure of providing all the information required and so interaction should improve.

The use of multiple sensory channels increases the *bandwidth* of the interaction between the human and the computer, and it also makes human–computer interaction more like the interaction between humans and their everyday environment, perhaps making the use of such systems more natural.

Usable sensory inputs

In computing, the visual channel is used as the predominant channel for communication, but if we are to use the other senses we have to consider their suitability and the nature of the information that they can convey.

Sound is already used, to a limited degree, in many interfaces: beeps are used as warnings and notification, recorded or synthesized speech and music are also used. Tactile feedback, as we have already seen, is also important in improving interactivity and so this represents another sense that we can utilize more effectively.

Sound in the interface

Sound is an important contributor to usability. There is experimental evidence to suggest that the addition of audio confirmation of modes, in the form of changes in key clicks, reduces errors. Video games offer further evidence, since experts tend to score less well when the sound is turned off than when it is on; they pick up vital clues and information from the sound while concentrating their visual attention on different things.

Speech in the interface

Language is rich and complex. We learn speech naturally as children 'by example' –by listening to and mimicking the speech of those around us. This complexity makes speech recognition and synthesis by computer very difficult.

Structure of speech If we are fully to appreciate the problems involved with the computer-based recognition and generation of speech, we need first to understand the basic structure of speech.

The English language is made up of 40 *phonemes*, which are the atomic elements of speech. Each phoneme represents a distinct sound, there being 24 consonants and 16 vowel sounds.

Speech recognition There have been many attempts at developing speech recognition systems, but, although commercial systems are now commonly and cheaply available, their success is still limited to single-user systems that require considerable training.

Speech synthesis Complementary to speech recognition is speech synthesis. The notion of being able to converse naturally with a computer is an appealing one for many users, especially those who do not regard themselves as computer literate, since it reflects their natural, daily medium of expression and communication.

Un interpreted speech Speech does not have to be recognized by a computer to be useful in the interface.