Heuristic Evaluation and Usability Testing

Ms. Sana Ashraf

Evaluation tests the usability, functionality and acceptability of an interactive system.

- Evaluation may take place: in the laboratory and in the field.
- Some approaches are based on expert evaluation: analytic methods, review methods and model-based methods.
- Some approaches involve users: experimental methods, observational methods and query methods.

An evaluation method must be chosen carefully and must be suitable for the job.

Goals of evaluation

Evaluation is an integral part of the design process, which should **take place throughout the design life cycle**. It should not be thought of as a single phase in the design process. Its aim is to **test the functionality and usability of the design and to identify and rectify any problems**. It can also try to determine the user's attitude and response to the system.

Evaluation has three main goals: 1) to assess the extent and accessibility of the system's functionality, 2) to assess users' experience of the interaction, and 3) to identify any specific problems with the system.

Evaluation through expert analysis

The basic intention of expert analysis is to identify any areas that are likely to cause difficulties because they violate known cognitive principles, or ignore accepted empirical results. Four approaches are considered here: cognitive walk- through, heuristic evaluation, the use of models and use of previous work.

1) Cognitive walkthrough

CW is a detailed **review of a sequence of actions,** in this case, the steps that an interface will require the user to perform in order to accomplish some known task. The **evaluators go through each step and provide a story about why that step is not good for new users.**

To do a CW, you need four things:

1) a specification or prototype of the system
2) a description of the task the user is to perform on the system
3) a complete written list of the actions needed to complete the task with the system
4) An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.

For each step, the evaluators try to answer the following questions: Is the effect of the action the same as then user's goal at that point? Will the users see that the action is available? Once the users have found the correct action, will they know it is the one they need? After the action is taken, will users understand the feedback they get?

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. HE is a method for structuring the critique of a system using a set of relatively simple and general heuristics. Several evaluators independently critique a system to come up with potential usability problems. Each evaluator assesses the system and notes violations of any of the following heuristics and the severity of each of these violations based on four factors: how common is the problem, how easy is it for users to overcome, will it be a one-off problem or a persistent one, and, how seriously will the problem be perceived. The overall result is a severity rating on a scale of 0-4.

The 10 heuristics of Nielsen are: Visibility of the system status, match between system & real world, user control & freedom, consistency & standards, error prevention, recognition rather than recall, flexibility & efficiency of use, aesthetic & minimalist design, help users recognize, diagnosis & recover from errors, and help & documentation.

3) Model-based evaluation

Certain cognitive and design models provide a means of combining design specification and evaluation into the same framework.

4) Using previous studies in evaluation

A similar experiment conducted earlier can cut some of the costs of a new design evaluation by reusing the data gained from it.

Evaluation through user participation

1) Styles of evaluation

Laboratory studies: In LS, users take part in controlled tests, often in specialist usability laboratories. The advantages are advanced laboratory equipment and the interruption-free environment. The disadvantage is the lack of context, which may result in unnatural situations.

Field studies: in FS, the user is observed using the system in its own work environment. The advantage is the 'natural' use of the system that can hardly be achieved in the lab. However, the interruptions that come with this natural situation may make the observations more difficult.

2) Empirical methods:

Experimental evaluation: Any experiment has the same basic forms: the evaluator chooses a hypothesis to test, which can be determined by measuring some attribute of participant behavior. A number of experimental conditions are considered which differ only in the values of certain controlled variables. Any changes in the behavioral measures are attributed to the different conditions. Some factors in the experiment must be considered carefully: the participants chosen, the variables tested and manipulated and the hypothesis tested.

Statistical measures: the data should first of all be save to enable performing multiple analysis on the same data. The choice of statistical analysis depends on the type of data and the questions we want to answer. Variables can be classified as discrete (which can take a finite number of values and levels) and continuous variables (which can take any value between a lower and upper limit).

3) Observational techniques

Think aloud and cooperative evaluation

Think aloud is a form for observation where the user is asked to talk through what he is doing as he is being observed. It has the advantage of simplicity, but the information provided is often subjective and may be selective. A variation is cooperative evaluation, in which the user and evaluator work together to evaluate the system.

Protocol analysis

Methods for recording user actions include paper and pencil, audio recording, video recording, computer logging and user notebooks. In practice, a mixture of the different methods is used. With recordings, the problem is transcription.

Automatic protocol analysis tools

Using Experimental Video Annotator, an evaluator can use predefined tags to write an audio to video transcription in real time. Using Workplace Project, this can be done while supporting the analysis and synchronization of information from different data streams. DRUM supports the same facilities.

Post-task walkthrough

A walkthrough after the observation reflects the participants 'actions' back to them after the event. The participant is asked to comment it and to answer questions by the evaluator in order to collect missing information.

4) Query techniques

There are two main types of query technique: interviews and questionnaires.

5) Evaluation through monitoring physiological responses

Eye tracking for usability evaluation

Eye tracking has been possible for many years, but recent improvements in hardware and software have made it more viable as an approach to measuring usability. The original eye trackers required highly invasive procedures where eye caps were attached

to the cornea under anesthetic. Clearly inappropriate for usability testing! Modern systems vary: some use a head-mounted camera to monitor the eye, but the most sophisticated do not involve any contact between the equipment and the participant, with the camera and light sources mounted in desk units.

Physiological measurements

Emotional response is closely tied to physiological changes. These include changes in heart rate, breathing and skin secretions. Measuring these physiological responses may therefore be useful in determining a user's emotional response to an interface [288, 363]. Could we determine which interaction events really cause a user stress or which promote relaxation?

Choosing an evaluation method

Range of techniques is available for evaluating an interactive system at all stages in the design process. So **how do we decide which methods are most appropriate for our needs**? There are no hard and fast rules in this – each method has its particular strengths and weaknesses and each is useful if applied appropriately. However, there are a **number of factors that should be taken into account when selecting evaluation techniques**. These also provide a way of categorizing the different methods so that we can compare and choose between them. In this final section we will consider these factors.

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:

1) **Design vs. implementation** - the stage in the cycle at which the evaluation is carried out

   Early evaluation, whether of a design or an early prototype or mock-up, will bring the **greatest pay-off** since problems can be easily resolved at this stage. As more commitment is made to a particular design in the implementation, it becomes increasingly difficult for changes to be made, no matter what the evaluation suggests.

Roughly speaking, **evaluation at the design stage needs to be quick and cheap** so might involve design experts only and be analytic, whereas **evaluation of the implementation needs to be more comprehensive, so brings in users as participants.**

2) **Laboratory vs. field studies** - the style of evaluation

   Laboratory studies **allow controlled experimentation and observation while losing something of the naturalness** of the user's environment. **Field studies retain the latter but do not allow control over user activity**. Ideally, the **design process should include both styles of evaluation**, probably with laboratory studies dominating the early stages and field studies conducted with the new implementation.

3) **Subjective vs. objective** - the level of subjectivity or objectivity of the technique

   Evaluation techniques also vary according to their objectivity – some techniques rely heavily on the interpretation of the evaluator, others would provide similar information for anyone correctly carrying out the procedure.

   The more subjective techniques, such as cognitive walkthrough or think aloud, rely to a large extent on the knowledge and expertise of the evaluator, who must recognize problems and understand what the user is doing. They can be powerful if used correctly and will provide information that may not be available from more objective methods. However, the problem of evaluator bias should be recognized and avoided. One way to decrease the possibility of bias is to use more than one evaluator.

   Objective techniques, on the other hand, should produce repeatable results, which are not dependent on the persuasion of the particular evaluator. Controlled experiments are an example of an objective measure.

4) **Qualitative vs. quantitative measures** - the type of measures provided

   The quantitative measurements is usually numeric and can be easily analyzed using statistical techniques. The other is non-numeric and is therefore more difficult to analyze, but can provide important detail that cannot be determined from numbers.

5) **Information provided**

6) **Immediacy of the response**

Another factor distinguishing evaluation techniques is the immediacy of the response they provide. Some methods, such as think aloud, record the user's behavior at the time of the interaction itself. Others, such as post-task walkthrough, rely on the user's recollection of events. Such recollection is liable to suffer from bias in recall and reconstruction, with users interpreting events according to their preconceptions. Recall may also be incomplete. However, immediate techniques can also be problematic, since the process of measurement can actually alter the way the user works.

7) **Intrusiveness - The level of interference implied**

Related to the immediacy of the response is the intrusiveness of the technique itself. Certain techniques, particularly those that produce immediate measurements, are obvious to the user during the interaction and therefore run the risk of influencing the way the user behaves. Sensitive activity on the part of the evaluator can help to reduce this but cannot remove it altogether. Most immediate evaluation techniques are intrusive, with the exception of automatic system logging. Unfortunately, this is limited in the information that it can provide and is difficult to interpret.

8) **Resources required**

The final consideration when selecting an evaluation technique is the availability of resources. Resources to consider include equipment, time, money, participants, expertise of evaluator & context. Some decisions are forced by resource limitations.

For example, time and money may be limited, forcing a choice between two possible evaluations. In these circumstances, the evaluator must decide which evaluation tactic will produce the most effective and useful information for the system under consideration.