Lecture 8: Usability evaluation methods

Summative Evaluation techniques

Summative evaluation, often performed under the umbrella of ***usability testing*** is carried out at the end of a project after the system has been built, to assess whether it meets its specification, or whether a project was successful. This is in contrast to ***formative*** evaluation, where the main objective is to contribute to the design of the product, by assessing specifications or prototypes before the system has been built. Formative evaluation is often ***analytic*** (it proceeds by reasoning about the design), while summative evaluation is often ***empirical*** (it proceeds by making observations or measurements).

Controlled experiments

The most common empirical method used in HCI research, derived from its origins in human factors and experimental psychology, is the controlled experiment. An experiment is based on a number of ***observations*** (measurements made while someone is using an experimental interface). A typical measurement might be "How long did Fred take to finish task A?" or "How many errors did he make"?

A single observation of speed is not very interesting, We therefore collect sets of measurements, and compare averages. The sets might be multiple observations of one person performing a task over many ***trials***, or of a range of people (experimental ***participants***) performing the same task under controlled conditions. As with most human performance, the measured results will usually be found to have a ***normal distribution***.

A typical HCI experiment involves one or more experimental ***treatments*** that modify the user interface. A very simple example might test the question: "How long does Fred take to finish task A when using a good UI, compared to a bad UI?" The result will often be that the good UI is *usually* faster to use than the bad, but not in *every* trial. We need to know whether the difference between the averages is the result of ordinary random variation, or the effect of the changes we made to the user interface.

In HCI research, we usually insist that the probability of the result being due to random variation (***p***) is less than 0.05, or 5%.

Good quality research results are normally based on experiments with significance values $p < 0.01$, which can be expressed as 'we reject the null hypothesis, with 99% confidence'.

Variation in controlled experiments

Some computer scientists find it surprising that one can draw scientific conclusions from measurements that are different every time we make them, and even offer the opinion that the basis of HCI in probability and statistics instead of mathematical proof is a fundamental flaw in HCI research. This is rather fatalistic. Everyone knows that people are different. If there were no way to measure the value of a user interface for a wide range of different people, there would be no chance of progress in user interface development. It is important, however, that we are aware of the sources of variation in the measurements.

These include:

- Variations in the task participant (changing with day of the week and time of day);
- The effect of the treatment (i.e. the user interface improvements that we made);
- Individual differences between experimental subjects (e.g. IQ);
- Different stimuli for each task;
- Distractions during the trial (sneezing, dropping things);
- Motivation of the subject;
- Accidental hints or intervention by the experimenter;
- Or other random factors.

Think aloud studies

Although not really experiments (they are often conducted without a hypothesis, and the data is qualitative rather than quantitative), controlled studies in HCI often use the think aloud technique to gain insight into the way the user has interpreted a prototype.

When used as a rigorous scientific technique, a great deal of care is taken to ensure that the users vocalizes every thought they are aware of, and the recording is transcribed and analysed in detail for evidence of particular mental processes.

However in a commercial context, the think-aloud protocol can seem much more like realtime evaluation feedback, in which users are simply asked to make as many comments as possible on the user interface. This may not provide very much scientific insight, but at least it avoids the problem of users who spend an hour using a new system, then say almost nothing in the way of feedback.

Other empirical techniques

Hypothesis testing is a very useful technique for making quantifiable statements about improvements in a user interface. It also hides a lot of useful information, however.

Experimental subjects usually have a lot of useful feedback about the interface that they are trying, but there is no easy way to incorporate this into statistical analyses. Instead, we use a range of other techniques to capture and aggregate interpretative reports from system users.

Surveys

Surveys include a range of techniques for collecting report data from a population. The most familiar types of survey are public opinion polls and market research surveys, but there are a much greater range of survey applications. Surveys are usually composed of a combination of ***closed*** and ***open*** questions. Closed questions require a yes/no answer, or a choice on a ***Likert*** scale - this is the familiar 1 to 5 scale asking respondents to rank the degree to which they agree with a statement. Closed questions are useful for statistical comparisons of different groups of respondents. In open questions the respondent is asked to compose a free response to the question. The latter requires a methodical ***coding*** technique to structure the content of the responses across the population, and is particularly useful for discovering information that the investigator was not expecting.

Questionnaires

Questionnaires are a particular type of survey. (Interview studies of a sample population are also a form of survey). Questionnaires are generally used to gather responses from a larger sample, and can be administered by email as well as on paper.

Field tests

Some very successful software companies have carried out ***field testing*** of their products in

addition to field studies at the specification phase. A well-documented example is the "follow-me-home" programme carried out by Intuit Inc. after the release of their Quicken product. Company researchers selected customers at random, when they were buying a shrink-wrapped copy of Quicken in a store. The researcher then went home with the customer in order to observe them as they read the manuals, installed the product, and used it for their home financial management. Intuit directly attribute the impressive success of the product to this type of exercise, and to the observational studies they carried out during initial product planning.

Bad evaluation techniques

Some user interface developers use evaluation techniques that are practically useless. Unfortunately these techniques can even be found in some published research in computer science. This section is included as a warning to interpret such results with great care.

Simple ***subjective reports*** seldom give useful information about interface usability. When users are shown a shiny new interface next to a tatty old one, they will often say that they like the new one better, regardless of its usability. There are many circumstances in which a person's introspective feelings about their mental performance is not a good predictor of actual performance, so this type of report is unreliable as well as open to bias.

Some research proposes a usability hypothesis, then does not test it at all. "It was proposed that more colours should be used in order to increase usability". This type of statement is speculation rather than science; designing novel user interfaces without any kind of experimental testing is rather pointless.

There is a great deal of variation between different people in their ability to use different interfaces. This may result from different mental models, different cognitive skills, different social contexts and many other factors. Any conclusions drawn from an observation of only one person must therefore be very suspect. Unfortunately, many user interfaces are developed based on observations of a single person - the programmer. The ***introspection*** of the user interface developer about his or her performance is seldom relevant to users.

The word "intuitive" is often used in discussion of user interfaces to summarise theories based on all the above, so should be considered a danger sign, if it is used to describe the

advantages of a particular user interface design, without any further specific detail or empirical evidence.

Formative evaluation techniques

In system testing development costs can be minimized by finding bugs as early as possible in the software development cycle. ***Formative evaluation*** describes studies that are carried out as part of the design process.

To some extent, formative evaluation can be carried out simply by inviting usability experts, or representative users, to review product plans and specifications, and offer their opinion. A more formalized approach to soliciting user opinions is ***participatory design*** methods, where representative users take part in design activities, perhaps structured in a way that means they do not have to learn too much technical jargon, but can concentrate on the way they are likely to interact with the user interface. A more formalized approach to engaging with usability experts is via techniques such as ***heuristic evaluation***, where a panel of experts review a proposed user interface one screen at a time, assessing whether it meets some predefined set of 'heuristic' criteria for good usability.

Formative evaluation using Cognitive Dimensions of Notations

There are also more theoretically motivated techniques for formative evaluation. The Cognitive Dimensions of Notations can be applied in 'checklist' style, as with Heuristic Evaluation. However, it is more useful to apply the dimensions more broadly, to consider both user needs and potential general approaches to the design. At this level, all design work can be considered 'evaluative', in the sense that designers are always having to evaluate which are the best options or trade-offs in the final product.

Formative evaluation using Cognitive Walkthrough

The Cognitive Walkthrough method is a structured analytic approach to assessing usability early in the project.

Behaviour model

The model of a user carrying out a task through exploratory learning involves four basic phases:

1) The model describes how a notional user sets a ***goal*** to be accomplished with the system. A typical goal will be expressed in terms of the expected capabilities of the system, such as "check spelling of this document".

2) The model describes how the notional user searches the interface for currently available ***actions***. The availability of actions may be observable as the presence of menu items, of buttons, of available command-line inputs, etc.

3) The model describes how the notional user ***selects*** the action that seems likely to make progress toward the goal.

4) The model describes how the notional user ***performs*** the selected action and ***evaluates*** the system's feedback for evidence that progress is being made toward the current goal.

Evaluation procedure

The evaluation procedure is based on a manual ***simulation*** of a notional user iteratively carrying out the stages of the behavioural model. Before evaluation can start, the evaluators need to have access to the following information:

1) A general description of the ***type of users*** who would be expected to use the system, and the ***relevant knowledge*** that these users would be expected to have.

2) A description of one or more ***representative tasks*** to be used in the evaluation.

3) For each of the tasks, a list of the ***correct actions*** that should be performed in order to complete the task.

The evaluation is conducted by the interface designer, and by a ***group of peers***. This group includes a nominated ***scribe*** who records the results of the evaluation and a ***facilitator*** who is responsible for the smooth running of the evaluation process. The scribe and the facilitator are also active members of the evaluation group.

The group of evaluators move through each of the tasks, considering the user interface at each step. At each step, they examine the interface and tell a ***story*** about why the notional user would choose that action. These stories are then evaluated according to an information-processing model derived from the exploratory learning behavioural model:

1) consider what the notional user's current ***goal*** would be;

2) evaluate the ***accessibility*** of the correct control;

3) evaluate the quality of the ***match*** between the control's label and the goal; and
4) evaluate the ***feedback*** that would be provided to the notional user after the action.

Evaluation of Part II projects

A substantial proportion of the marks for a Part II project are awarded for proper evaluation. In most projects, this tends to be summative – formative evaluation work could be reported in the 'preparation' or even 'implementation' sections of the dissertation.

Non-HCI projects

In all projects, whether or not they include a user interface component, empirical measures are considered to offer stronger evidence for the quality of your work, and a higher degree of scientific rigour. Empirical evaluation involves taking measurements (perhaps of compile times, or network traffic estimates). Most empirical measurements are not exact, so it will be necessary to make a number of measurements, and report the degree of variance as well as the mean. Empirical results are particularly convincing if they offer a comparison – either comparing performance of your system to an existing one, or comparing earlier and later versions of your work. Where a comparison is being made, and there is some variance in measurements, it is necessary to give some statistical evidence to support the claim that the observed difference was not the result of random variation.

Projects with user interfaces

Evaluation of user interfaces need not be quantitative, of course. In cases where speed and accuracy measurements are not meaningful ways to assess your project, any of the qualitative methods described above could provide useful evaluations: think-aloud studies, interviews, questionnaires, observational studies, or even field trials. Where there is a lot of variation between users, it can be a good idea to interview them and ask why.

Ethical issues with human participants!

There are a number of simple precautions that you should take, when conducting research involving human participants. Fortunately, these are fairly straightforward for routine technology evaluation studies.

Last resort evaluation

As a last resort (for example, if your project is incomplete), you could carry out a formative

evaluation and report this in the evaluation section of your dissertation. Either Cognitive Walkthrough, or Cognitive Dimensions of Notations, could be used for this purpose. Otherwise, your results are likely to be obviously biased and subjective, and will not impress examiners. It is also worth remembering that choosing an analytic technique rather than an empirical one (i.e. not basing your conclusions on measurements or observations of the system in use), will always plant suspicion in the minds of the examiners that you have chosen to do this because the system doesn't work, meaning that proper summative evaluation wasn't an option.