CHAPTER 7

HOW TO EVALUATE THE DESIGN?

OVERVIEW

- Evaluation is one of the main process in the interactive systems design.
- By evaluation we mean reviewing, trying out or testing a design idea, a piece of software, a product or a service to discover whether it meets some criteria.
- These criteria will often be summed up by the guidelines for good design, namely that the system is learnable, effective and accommodating.
- But at other times the designer might be more interested in some other characteristic of the design.
- The designer is concerned not just with surface features such as the meaningfulness of icons, but also with whether the system is fit for its purpose, enjoyable, engaging and so on.

AIMS OF THE EVALUATION

Deciding the aim(s) for evaluation helps to determine the type of data required. It is useful to write down the main questions you need to answer. Early concept evaluation, in DISCOVER for instance, entailed questions such as:

- Do the trainers understand and welcome the basic idea of the virtual training environment?
- Would they use it to extend or replace existing training courses?

INTRODUCTION

- The techniques in this chapter will allow you to evaluate many types of product, system or service.
- Evaluation of different types of system, or evaluation in different contexts may offer particular challenges. For example evaluating mobile devices, or services delivered by a mobile device or evaluating in mixed reality, ubiquitous computing environments
- In particular many of the techniques for understanding are applicable to evaluation.
- Evaluation is also critically dependent on the form of envisionment used to represent the system.
- You will only be able to evaluate features that are represented in a form appropriate for the type of evaluation.
- There are also issues concerning who is involved in the evaluation.

PARTICIPANTS

- Where possible these people should be representative of the people who the system is aimed at (sometimes called ‘end-users’).
- This is the preferred choice for ‘in house’ systems where the designer has access to the target population.
- Alternatively the participants may be other people (perhaps other designers, students or whoever happens to be around), who are invited to play the part of the people who will use the system.
- The characteristics of the target population can be captured through personas.

EVALUATION THROUGH DESIGN

- At different stages, different methods will be more or less effective.
- The form of envisionment of the future systems is also critical to what can be evaluated.
- You may need to evaluate initial concepts, especially if the application is novel.
- Here quick paper prototypes can help, or even software if this can be produced rapidly.
- Evaluations of competitor products or previous versions of technology can also feed into the design process at this stage.
- Deciding between different design options
- During development, designers have to decide between options, for example between voice input or touch screen interaction for a shared electronic household wall-planner or between different sequences for order processing functions.

CHECKING FOR USABILITY PROBLEMS

- Testing will identify potential problems once a stable version of the technology is available.
- This needs to respond when a participant activates a function, but it does not require the whole system be fully operational (a horizontal prototype).
- Alternatively, the system may be completely functional, but only in some parts (a vertical prototype).
- What is important is that there is still time to fix problems.

FORMATIVE EVALUATION

- What happens all too frequently is that you are asked to check that interaction is ‘user friendly’ just before development is completed.
- Very often all that can be changed are minor issues such as the position, color or labeling of on-screen buttons.
- It is best to be helpful in these circumstances.
- If you make a note of problems that could have been solved easily if identified sooner, you can exploit these examples (tactfully) to justify evaluation work at an earlier stage in the next project.

SUMMATIVE EVALUATION

- Is about assessing the usability of a finished product
- This may be to test against in-house guidelines, or formal usability standards such as ISO 9241, or to provide evidence of usability required by a customer, for example the time to complete a particular set of operations.
- Government departments and other public bodies often require suppliers to conform to accessibility standards and health and safety legislation.

PARTICIPATORY DESIGN

- Evaluation as a means of involving people in the design process
- In participatory design approach, stakeholders help designers set the goals for the evaluation work.
- Involving stakeholders has great benefits in terms of eventual uptake and use of the technology. (Of course, this applies only to technology that is tailor-made for defined communities, rather than off-the-shelf products.)
- A number of methods of involving people in the design process were discussed in understanding and envisionment (prototyping).
- Susanne Bodker's Design Collabatorium (Bodker, 2002) is another example.

EXPERT EVALUATION

- A simple, relatively quick and effective method of evaluation is to get an interaction design, or usability, expert to look at the system and try using it.
- This is no substitute for getting real people to use your design, but expert evaluation is effective particularly early in the design process. Experts will pick up common problems based on this experience, and will identify factors that might otherwise interfere with an evaluation by non-experts.
- However, to help the experts structure their evaluation, it is useful to adopt a particular approach.
- This will help to focus the expert's critique on the most relevant aspects for the purpose.

HEURISTIC EVALUATION (CONT.)

- Heuristic evaluation refers to a number of methods in which a person trained in HCI and interaction design examines a proposed design it measures up against a list of principles, guidelines or ‘heuristics’ for good design.
- This review may be a quick discussion over the shoulder of a colleague, or may be a formal, carefully documented process.
- There are many sets of heuristics to choose from, both general purpose and those relating to particular application domains, for example heuristics for Web design.

1. Visibility 2. Consistency 3. Familiarity 4. Affordance 5. Navigation 6. Control

7. Feedback 8. Recovery 9. Constraints 10. Flexibility 11. Style 12. Conviviality

HEURISTIC EVALUATION (MORE..)

- Ideally, several people with expertise in interactive systems design should review the interface.
- Each expert notes the problems and the relevant heuristic, and suggests a solution where possible.
- It is also helpful if a severity rating, say on a scale of 1–3, is added, according to the likely impact of the problem as recommended
- However, Dumas and Fox (2008) note the disappointing level of correlation amongst experts in rating severity of problems.

- Evaluators work independently and then combine results.
- They may need to work through any training materials and be briefed by the design team about the functionality.
- The scenarios used in the design process are valuable here.
- It is now used for any ‘quick and dirty’ approach to evaluation where the aim is to get useful, informed feedback as soon as possible.
- Once again a number of usability experts ‘walk through’ concrete scenarios, preferably accompanied by personas, and inspect the design for difficulties.

HEURISTIC EVALUATION

If time is very short, a quick review of the design against this triad can produce reasonably useful results. This approach to evaluation was pioneered by Jakob Nielsen (1993) and enthusiastically followed by many time-pressured evaluation practitioners.

INDEPENDENCE

- Unless there is no alternative, you should not evaluate your own designs.
- It is extremely difficult to ignore your knowledge of how the system works, the meaning of icons or menu names and so on, and you are likely to give the design the ‘benefit of the doubt’ or to find obscure flaws which few users will ever happen upon.
- Woolrych and Cockton (2000) conducted a large-scale trial of heuristic evaluation.
- Evaluators were trained to use the technique, then evaluated the interface to a drawing editor.
- The editor was then trialed by customers.
- Comparison of findings showed that many of the issues identified by the experts were not experienced by people (false positives), while some severe difficulties were missed by the inspection against heuristics.

EXPERT REVIEW

- Many false positives stemmed from a tendency by the experts to assume that people had no intelligence or even commonsense.
- As for ‘missing’ problems, these tended to result from a series of mistakes and misconceptions, often relating to a set of linked items, rather than isolated misunderstandings.
- Sometimes heuristics were misapplied, or apparently added as an afterthought. Woolrych and Cockton (2000) conclude that the heuristics add little advantage to an expert evaluation and the results of applying them may be counter-productive.
- They (and other authors) suggest that more theoretically informed techniques such as the cognitive walkthrough offer more robust support for problem identification.
- It is very evident that heuristic evaluation is not a complete solution.
- At the very least, the technique must be used together with careful consideration of people and their real-life skills and that participant evaluation is required to get a realistic picture of the success of a system.

HEURISTIC EVALUATION (AGAIN)

- Heuristic evaluation therefore is valuable as formative evaluation, to help the designer improve the interaction at an early stage.
- It should not be used as a summative assessment, to make claims about the usability and other characteristics of a finished product.
- If that is what we need to do, then we must carry out properly designed and controlled experiments with a much greater number of participants.
- However, the more controlled the testing situation becomes, the less it is likely to resemble the real world, which leads us to the question of ‘ecological validity’.

ECOLOGICAL VALIDITY

- In real life, people multitask, use several applications in parallel or in quick succession, are interrupted, improvise, ask other people for help, use applications intermittently and adapt technologies for purposes the designers never imagined.
- We have unpredictable, complex but generally effective coping strategies for everyday life and the technologies supporting it.
- The small tasks which are the focus of most evaluations are usually part of lengthy sequences directed towards aims which change according to circumstances.
- All of this is extremely difficult to reproduce in testing, and often deliberately excluded.
- So the results of most user testing can only ever be indicative of issues in real-life usage.
- Practitioners and researchers are not unaware of this problem and a number of solutions have been proposed.

COGNITIVE WALKTHROUGH

- Cognitive walkthrough is a rigorous paper-based technique for checking through the detailed design and logic of steps in an interaction.
- It is derived from the human information processor view of cognition and closely related to task analysis
- In essence, the cognitive walkthrough entails a usability analyst stepping through the cognitive tasks that must be carried out in interacting with technology.
- Originally developed by Lewis *et al. (1990*) for applications where people browse and explore information, it has been extended to interactive systems in general (Wharton et al.,1994).
- Aside from its systematic approach, the great strength of the cognitive walkthrough is that it is based on well-established theory rather than the trial and error or a heuristically based approach.

COGNITIVE WALKTHROUGH (CONT..)

- It defined as an understanding of the people who are expected to use the system.
- A set of concrete scenarios representing both (a) very common and (b) uncommon but critical sequences of activities
- A complete description of the interface to the system – this should comprise both a representation of how the interface is presented, e.g. screen designs, and the correct sequence of actions for achieving the scenario tasks, usually as a hierarchical task analysis (HTA).
- Having gathered these materials together, the analyst asks the following four questions for each individual step in the interaction:
  - ✓ Will the people using the system try to achieve the right effect?
  - ✓ Will they notice that the correct action is available?
  - ✓ Will they associate the correct action with the effect that they are trying to achieve?
  - ✓ If the correct action is performed, will people see that progress is being made towards the goal of their activity?

USABILITY ISSUES

- If any of the questions is answered in the negative, then a usability problem has been identified and is recorded, but redesign suggestions are not made at this point.
- If the walkthrough is being used as originally devised, this process is carried out as a group exercise by analysts and designers together.
- The analysts step through usage scenarios and the design team are required to explain how the user would identify, carry out and monitor the correct sequence of actions.
- Software designers in organizations with structured quality procedures in place will find some similarities to program code walkthroughs.

CUT-DOWN VERSION

- The ‘cognitive jogthrough’ (Rowley and Rhoades, 1992) – video records (rather than conventional minutes) are made of walkthrough meetings, annotated to indicate significant items of interest, design suggestions are permitted, and low-level actions are aggregated wherever possible.
- The ‘streamlined cognitive walkthrough’ (Spencer, 2000) – designer defensiveness is defused by engendering a problem-solving ethos, and the process is streamlined by not documenting problem-free steps and by combining the four original questions into two (*ibid., p. 355):*
  - ✓ Will people know what to do at each step?
  - ✓ If people do the right thing, will they know that they did the right thing, and are making progress towards their goal?
- Other checklist approaches have been suggested such as the Activity Checklist (Kaptelinin, Nardi and Macaulay, 2004), but have not been widely taken up by other practitioners.

EXPERT BASED EVALUATION SUMMARY

- While expert-based evaluation is a reasonable first step, it will not find all problems, particularly those that result from a chain of 'wrong' actions or are linked to fundamental misconceptions.
- Experts even find problems that do not really exist – people overcome many minor difficulties using a mixture of common sense and experience.
- So it is really important to complete the picture with some real people trying out the interaction design.
- The findings will always be interesting, quite often surprising and occasionally disconcerting. From a political point of view, it is easier to convince designers of the need for changes if the evidence is not simply one 'expert' view, particularly if the expert is relatively junior.
- The aim is to trial the design with people who represent the intended target group in as near realistic conditions as possible.

- Whereas expert, heuristic evaluations can be carried out by designers on their own, there can be no substitute for involving some real people in the evaluation. Participant evaluation aims to do exactly that.
- There are many ways to involve people that involve various degrees of cooperation.
- The methods range from designers sitting with participants as they work through a system to leaving people alone with the technology and observing what they do through a 2-way mirror.

PARTICIPATION BASED EVALUATION

COOPERATIVE EVALUATION

- A means of maximizing the data gathered from a simple testing session.
- The technique is ‘cooperative’ because participants are not passive subjects but work as co-evaluators.
- It has proved a reliable but economical technique in diverse applications.

PARTICIPANT AND HEURISTIC

- The developers of participatory heuristic evaluation (Muller et al., 1998*)* claim that it extends the power of heuristic evaluation without adding greatly to the effort required.
- An expanded list of heuristics is provided, based on those of Nielsen and Mack (1994) – but of course you could use any heuristics.
- The procedure for the use of participatory heuristic evaluation is just as for the expert version, but the participants are involved as 'work-domain experts' alongside usability experts and must be briefed about what is required.

CO-DISCOVERY

- Co-discovery is a naturalistic, informal technique that is particularly good for capturing first impressions. It is best used in the later stages of design.
- The standard approach of watching individual people interacting with the technology, and possibly 'thinking aloud' as they do so, can be varied by having participants explore new technology in pairs.
- For example, a series of pairs of people could be given a prototype of a new digital camera and asked to experiment with its features by taking pictures of each other and objects in the room.
- This tends to elicit a more naturalistic flow of comment, and people will often encourage each other to try interactions that they may not have thought of in isolation.
- It is a good idea to use people who know each other quite well. As with most other techniques, it also helps to set participants some realistic tasks to try out.
- Depending on the data to be collected, the evaluator can take an active part in the session by asking questions or suggesting activities, or simply monitor the interaction either live or using a video-recording.
- Inevitably, asking specific questions skews the output towards the evaluator's interests, but does help to ensure that all important angles are covered.

LIVING LABS

- Living Labs is a European approach to evaluation that aims to engage as many people as possible in exploring new technologies. There are a number of different structures for Living Labs.
- For example Nokia has teamed up with academics and other manufacturers of mobile devices to hand out hundreds of early prototype systems to students to see how they use them.
- Other labs work with elderly people in their homes to explore new types of home technologies.
- Others work with travelers and migrant workers to uncover what new technologies can do for them.
- The key idea behind Living Labs is that people are both willing and able to contribute to designing new technologies and new services and it makes sense for companies to work with them.
- The fact that the discussions and evaluation take place in the life-context of people, and often with large numbers of people, gives the data a strong ecological validity.

CONTROLLED EXPERIMENT

- Another way of undertaking participant evaluation is to set up a controlled experiment.
- Controlled experiments are appropriate where the designer is interested in particular features of a design, perhaps comparing one design to another to see which is better.
- In order to do this with any certainty the experiment needs to be carefully designed and run.

VARIABLES

- The first thing to do when considering an controlled experiment approach to evaluation is to establish what it is that you are looking at. This is the independent variable.
- For example you might want to compare two different designs of a web site, or two different ways of selecting a function on a mobile phone application.
- An experiment that we carried out examined two different ways of presenting an audio interface to select locations of objects.
- The independent variable was the type of audio interface.
- One you have established what it is you are looking at you need to decide how you are going to measure the difference. These are the *dependent variables.*
- You might want to judge which web design is better based on the number of clicks need to achieve some task, speed of access could be the dependent variable for selecting a function. In the case of the audio interface, accuracy of location was the dependent variable.

EXPERIMENT DESIGN

- Once the independent and dependent variables have been agreed, the experiment needs to be designed to avoid anything getting in the way of the relationship between independent and dependent variables.
- Things that might get in the way are learning effects, the effects of different tasks, the effects of different background knowledge, etc.
- These are the *confounding variables.* You want to ensure a balanced and clear relationship between independent and dependent variables so that you can be sure you are looking at the relationship between them and nothing else.

CONFOUNDING VARIABLES

- One possible confounding variable is that the participants in any experiment are not balanced across the conditions.
- To avoid this participants are usually divided up across the conditions so that there are roughly the same number of people in each condition and there are roughly the same number of males and females, young and old, experienced and not.
- The next stage is to decide whether each participant will participate in all conditions (so-called within-subject design) or whether each participant will only perform one condition (so-called between-subject design).
- In deciding this you have to be wary of introducing compounding variables. For example, the learning effects that happen if people perform a similar task on more than one system.
- They start off slowly but soon get good at things so if time to complete a task is a measure they inevitably get quicker the more they do it.
- This effect can be controlled by randomizing the sequence in which people perform in the different conditions.

EXPERIMENTS

- Having got some participants to agree to participate in a controlled experiment it is tempting to try and find out as much as possible.
- There is nothing wrong with an experiment being set up to look at more than one independent variable, perhaps one being looked at between subjects and another being looked at within subjects.
- You just have to be careful how the design works and, of course, there is nothing wrong with interviewing them afterwards, or using focus groups afterwards to find out other things about the design.
- People can be videoed and perhaps talk aloud during the experiments (so long as this does not count as a confounding variable) and this data can also prove useful for the evaluation.

EVALUATION IN PRACTICE

- A controlled experiment will often result in some quantitative data; the measures of the dependent values. This data can then be analyzed using statistics, for example comparing the average time to do something across two conditions, or the average number of clicks.
- So, to undertake controlled experiments you will need some basic understanding of probability theory, of experimental theory, and of course of statistics.
- A survey of 103 experienced practitioners of human center design conducted (Vredenburg et al., 2002) indicates that around 40 percent of those surveyed conducted 'usability evaluation', around 30 percent used 'informal expert review' and around 15 percent used 'formal heuristic
- These figures do not indicate where people used more than one technique. As the authors note, some kind of cost–benefit trade-off seems to be in operation
- For busy practitioners, the relative economy of review methods often compensates for the better information obtained from user testing.
- Clearly the community remains in need of methods that are both light on resources and productive of useful results.

THE MAIN STEPS

1. Establish the aims of the evaluation, the intended participants in thee valuation, the context of use and the state of the technology; obtain or construct scenarios illustrating how the application will be used.
2. Select evaluation methods. These should be a combination of expert-based methods review and participant methods.
3. Carry out expert review.
4. Plan participant testing; use the results of the expert review to help focus this.
5. Recruit people and organize testing venue and equipment.
6. Carry out the evaluation.
7. Analyze results, document and report back to designers.

KEEP IN MIND

- There are three things to keep in mind when deciding metrics:
- Just because something can be measured, it doesn't mean it should be.
- Always refer back to the overall purpose and context of use of the technology.
- Consider the usefulness of the data you are likely to obtain against the resources it will take to test against the metrics.

PEOPLE

- The most important people in evaluation are the people who will use the system.
- Analysis work should have identified the characteristics of these people, and represented these in the form of personas.
- Relevant data can include knowledge of the activities the technology is intended to support, skills relating to input and output devices, experience, education, training and physical and cognitive capabilities.
- You need to recruit at least three and preferably five people to participate in tests.
- Nielsen's recommended sample of 3–5 participants has been accepted wisdom in usability practice for over a decade.
- However, some practitioners and researchers advise that this is too few.
- We consider that in many real-world situations obtaining even 3–5 people is difficult, so we continue to recommend small test numbers as part of a pragmatic evaluation strategy.

GROUPS

- However, testing such a small number only makes sense if you have a relatively homogeneous group to design for – for example, experienced managers who use a customer database system, or computer games players aged between 16 and 25.
- If you have a heterogeneous set of customers that your design is aimed at, then you will need to run 3–5 people *from each group through your tests.*
- For example, in DISCOVER we tested the virtual training environment with both tutors and trainees.
- If your product is to be demonstrated by sales and marketing personnel, it is useful to involve.
- Finding representative participants should be straightforward if you are developing an in-house application.
- Otherwise participants can be found through focus groups established for marketing purposes or, if necessary, through advertising.
- Students are often readily available, but remember that they are only representative of a particular segment of the population. If you have the resources, payment can help recruitment.
- Inevitably, your sample will be biased towards cooperative people with some sort of interest in technology, so bear this in mind when interpreting your results.

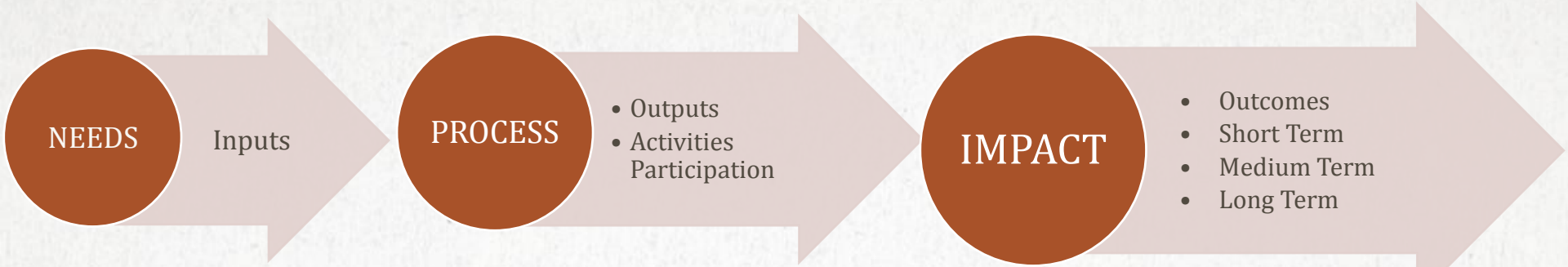
OTHER OPTION

- If you cannot recruit any genuine participants – people who are really representative of the target customers, and you are the designer of the software, at least have someone else try to use it.
- This could be one of your colleagues, a friend, your mother or anyone you trust to give you a brutally honest reaction. Almost certainly, they will find some design flaws.
- The data you obtain will be limited, but better than nothing.
- You will, however, have to be extremely careful as to how far you generalize from your findings.

YOUR ROLE

- Finally, consider your own role and that of others in the evaluation team if you have one.
- You will need to set up the tests and collect data, but how far will you become involved?
- Our recommended method for basic testing requires an evaluator to sit with each user and engage with them as they carry out the test tasks.
- We also suggest that for ethical reasons and in order to keep the tests running, you should provide help if the participant is becoming uncomfortable, or completely stuck.
- The amount of help that is appropriate will depend on the type of application (e.g. for an information kiosk for public use you might only provide very minimal help), the degree of completeness of the test application and in particular whether any help facilities have been implemented.

LOGIC MODEL



Needs/Asset Assessment
- What the characteristic of target population?
- What are the potential barrier/facilitator?
- What is most appropriate to do?

Process Evaluation
- How is program implemented?
- Are activities delivered as intended?
- What are user reactions?

Outcome Evaluation
- To what extent are desired changes occurring? Goals met?
- Who is benefiting? Not?
- What are unintended outcomes?

Impact Evaluation
- To what extent can changes be attributed to program?
- What are final consequences?
- Is program worth its cost?

TEST PLAN AND SPECIFICATION

A plan should be drawn up to guide the evaluation. The plan specifies:

- Aims of the test session.
- Practical details including where and when it will be conducted, how long each session will last, the specification of equipment and materials for testing and data collection and any technical support that may be necessary.
- Numbers and types of participant.
- Tasks to be performed with a definition of successful completion. This section also specifies what data should be collected and how it will be analyzed.
- You should now conduct a pilot session and fix any unforeseen difficulties. For example, task completion time is often much longer than expected, and instructions may need clarification.

REPORTING RESULT

- However competent and complete the evaluation, it is only worthwhile if the results are acted upon.
- Even if you are both designer and evaluator, you need an organized list of findings so that you can prioritize redesign work.
- If you are reporting back to a design/development team, it is crucial that they can see immediately what the problem is, how significant its consequences are, and ideally what needs to be done to fix it.

REPORTING WITHOUT BEING THERE

- With the arrival of Internet connectivity, people can participate in evaluations without being physically present.
- If the application itself is Web-based, or can be installed remotely, instructions can be supplied so that users can run test tasks and fill in and return questionnaires in soft or hard copy.
- On-line questionnaires and crowd sourcing methods described in are appropriate here.

EYE TRACKING

- Eye-movement tracking (or ‘eye tracking’) can show participants’ changing focus on different areas of the screen.
- This can indicate which features of a user interface have attracted attention, and in which order, or capture larger-scale gaze patterns indicating.
- Eye tracking is very popular with web site designers as it can be used to highlight which parts of the page are most looked at, so-called ‘hot spots’ and which are missed altogether.
- Eye-tracking equipment is head-mounted or attached to computer monitors.
- Eye tracking software is readily available to provide maps of the screen.

MEASUREMENT

- Some of it can also measure pupil dilation, which is taken as an indication of arousal. Your pupil dilates if you like what you see.
- Physiological techniques in evaluation rely on the fact that all our emotions – anxiety, pleasure, apprehension, delight, surprise and so on – generate physiological changes.
- The most common measures are of changes in heart rate, the rate of respiration, skin temperature, blood volume pulse and galvanic skin response (an indicator of the amount of perspiration).
- All are indicators of changes in the overall level of arousal, which in turn may be evidence of an emotional reaction.
- Sensors can be attached to the participants' body (commonly the fingertips) and linked to software which converts the results to numerical and graphical formats for analysis.
- But there are many unobtrusive methods too such as pressure sensors in the steering wheel of a games interface, or sensors that measure if the participant is on the edge of their seat.
- Which particular emotion is being evoked cannot be deduced from the level of arousal alone, but must be inferred from other data such as facial expression, posture or direct questioning.

PRESENCE

- Another current application is in the assessment of the degree of presence – the sense of 'being there' evoked by virtual environments
- Typically, startling events or threatening features are produced in the environment and arousal levels measured as people encounter them. Researchers have conducted a series of experiments when measuring arousal as participants approach a 'virtual precipice'.
- In these circumstances changes in heart rate correlated most closely with self-reports of stress.

EVALUATING PRESENCE

- Designers of virtual reality – and some multimedia – applications are often concerned with the sense of presence, of being ‘there’ in the virtual environment rather then ‘here’ in the room where the technology is being used.

- A strong sense of presence is thought to be crucial for such applications as games, those designed to treat phobias, or to allow people to ‘visit’ real places they may never see otherwise, or indeed for some workplace applications such as training to operate effectively under stress.

The difficulties include:

✓ The sense of presence is strongly entangled with individual dispositions, experiences and expectations. Of course this is also the case with reactions to any interactive systems, but presence is an extreme example of this problem.

✓ The concept of presence itself is ill-defined and the subject of much debate among researchers. Variants include the sense that the virtual environment is realistic, the extent to which the user is impervious to the outside world, the retrospective sense of having visited rather than viewed a location and a number of others.

✓ Asking people about presence while they are experiencing the virtual environment tends to interfere with the experience itself. On the other hand, asking questions retrospectively inevitably fails to capture the experience as it is lived.

✓ The measures used in evaluating presence adapt various strategies to avoid these problems, but none are wholly satisfactory.

PRESENCE QUESTIONNAIRE

- The various questionnaire measures developed by NASA scientists Witmer and Singer (1998) or the range of instruments developed at University College and Goldsmiths' College, London (Slater, 1999; Lessiter et al., 2001).
- These can be cross-referenced to measures which attempt to quantify how far a person is generally susceptible to being 'wrapped up' in experiences mediated by books, films, games and so on as well as through virtual reality.
- The Witmer and Singer Immersive Tendencies Questionnaire (Witmer and Singer, 1998) is the best known of such instruments. However, presence as measured by presence questionnaires is a slippery and ill-defined concept.
- In one experiment, questionnaire results showed that while many people did not feel wholly present in the virtual environment (a re-creation of an office), some of them did not feel wholly present in the real-world office either (Usoh et al., 2000).

OTHER MEASUREMENT

- Less structured attempts to capture verbal accounts of presence include having people write accounts of their experience, or inviting them to provide free-form comments in an interview.
- The results are then analyzed for indications of a sense of presence.
- The difficulty here lies in defining what should be treated as such an indicator, and in the layers of indirection introduced by the relative verbal dexterity of the participant and the interpretation imposed by the analyst.
- Other approaches to measuring presence attempt to avoid such layers of indirection by observing behavior in the virtual environment or by direct physiological measures.

EVALUATION AT HOME

- People at home are much less of a ‘captive audience’ for the evaluator than those at work.
- They are also likely to be more concerned about protecting their privacy and generally unwilling to spend their valuable leisure time in helping you with your usability evaluation.
- So it is important that data gathering techniques are interesting and stimulating for users, and make as little demand on time and effort as possible.
- This is very much a developing field and researchers continue to adapt existing approaches and develop new ones. Petersen, Madsen and Kjaer (2002 used conventional interviews at the time the technology (a new television) was first installed, but followed this by having families act out scenarios using it.
- Diaries were also distributed as a data collection tool, but in this instance the non-completion rate was high, possibly because of the complexity of the diary pro-forma and the incompatibility between a private diary and the social activity of television viewing.

TECHNOLOGY TOUR

- An effective example of this in early evaluation is reported in Baillie (2002, 2003).
- Here the investigator supplied users with Post-its to capture their thoughts about design concepts.
- An illustration of each different concept was left in the home in a location where it might be used, and users were encouraged to think about how they would use the device and any issues that might arise.
- These were noted on the Post-its, which were then stuck to the illustration and collected later.
- Where the family is the focus of interest, techniques should be engaging for children as well as adults – not only does this help to ensure that all viewpoints are covered, but working with children is a good way of drawing parents into evaluation activities.

CUSTOMER SATISFACTION

- Customer satisfaction is a marketing term that measures how products or services supplied by a company meet or surpass a customer's expectation.
- Within organizations, the collection, analysis and dissemination of these data send a message about the importance of tending to customers and ensuring that they have a positive experience with the company's goods and services (Farris, 2010)
- He found that 71 percent responded that a customer satisfaction metric very useful in managing and monitoring their businesses because several reasons:
  - ✓ It's a leading indicator of consumer repurchase intentions and loyalty
  - ✓ It's a point of differentiation
  - ✓ It reduces customer churn
  - ✓ It increases customer lifetime value
  - ✓ It reduces negative word of mouth
  - ✓ It's cheaper to retain customers than acquire new ones

Form Configuration

Clear Form
Back to Previous Form List

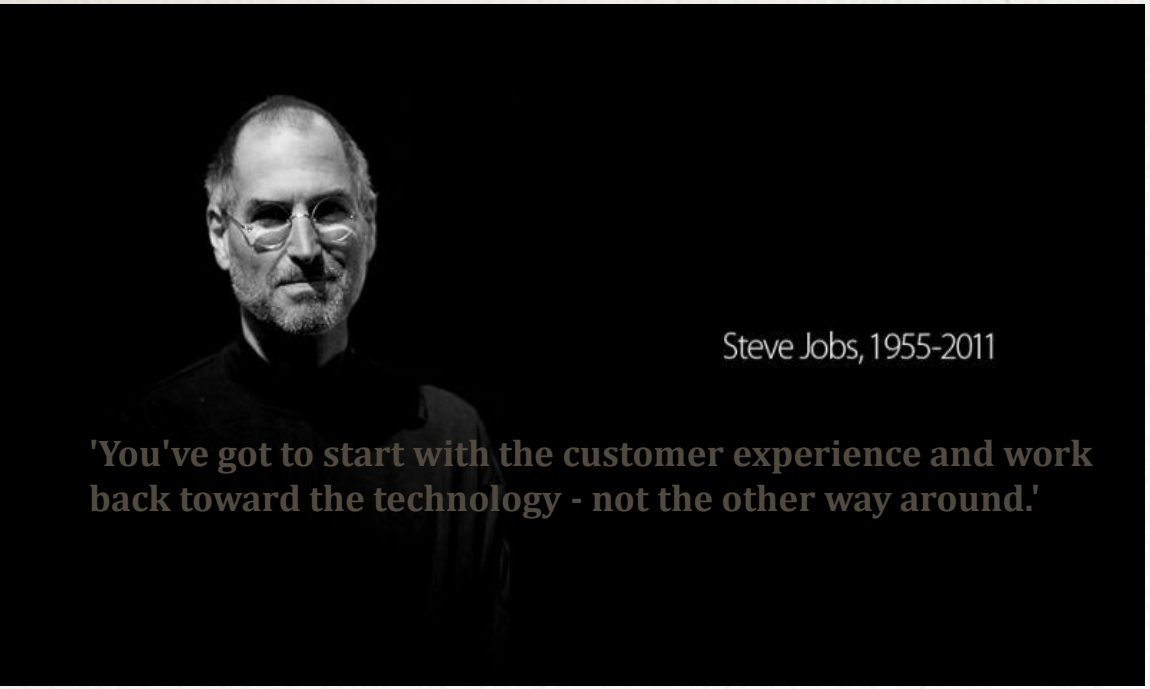
Form Data | Form Designer | Form Details | Call Details

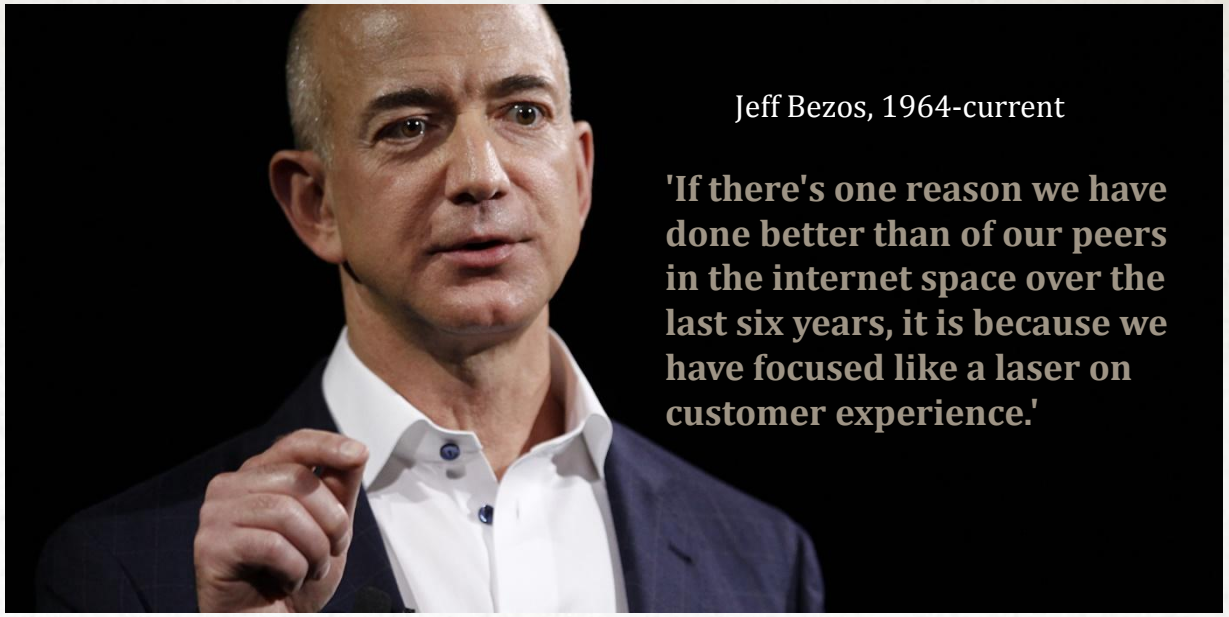
Total score					**170**
Greeting					**25 (15%)**
Did the agent say thank you for calling or apply a local greeting?	Yes	No	N/A		5 (3%)
Did the agent mention the company name?	Yes	No	N/A		5 (3%)
Did the agent mention his/ her name?	Yes	No	N/A		5 (3%)
Did the agent offer assistance to the caller?	Yes	No	N/A		5 (3%)
If the call was transferred did the agent adapt the greeting accordingly?	Yes	No	N/A		5 (3%)
Handle contact					**20 (12%)**
Did the agent ask for / confirm the caller's name?	Yes	No	N/A		5 (3%)
Did the agent ask for/ confirm the caller's company Name?	Yes	No	N/A		5 (3%)
Did the agent ask for / confirm the caller's telephone number?	Yes	No	N/A		5 (3%)
Did the agent ask for/ confirm the customer's account number?	Yes	No	N/A		5 (3%)
Transaction information					**30 (18%)**
Maintained a courteous, pleasant, and respectful tone throughout the call?	Yes	No	N/A		5 (3%)
Conveyed information clearly and confidently and in a manner that was ea	Yes	No	N/A		5 (3%)
Demonstrated effective listening skills?	Yes	No	N/A		5 (3%)
Expressed empathy and concern as appropriate?	Yes	No	N/A		5 (3%)
Efficiently managed time and call flow (call management)?	Yes	No	N/A		5 (3%)
Demonstrated professionalism (call etiquette)?	Yes	No	N/A		5 (3%)

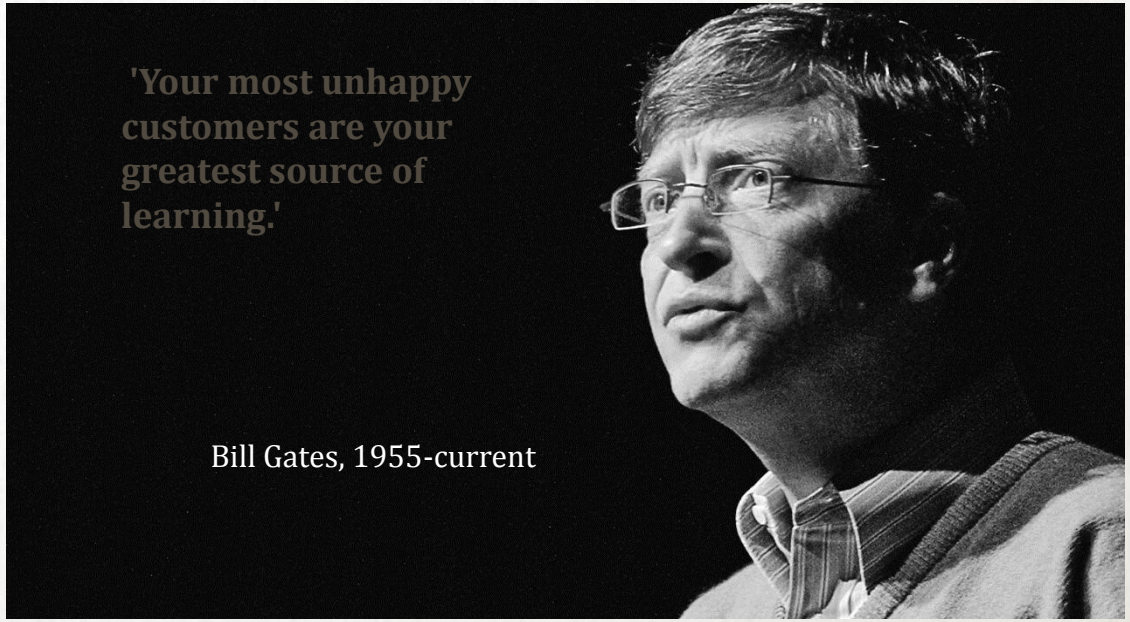
CUSTOMER SERVICE

- Customer service is the process of ensuring customer satisfaction with a product or service. Often, customer service takes place while performing a transaction for the customer, such as making a sale or returning an item.
- The perception of success of such interactions is dependent on employees "who can adjust themselves to the personality of the guest“ (Leigh, 2011).
- Most popular metric used by company involved:
  - first response time
  - average response time
  - total handle time
  - customer satisfaction score

QUOTE







'The customer experience is the next competitive battleground.' - Jerry Gregoire (DELL CIO)

- This presentation has presented a diverse selection of evaluation techniques to suit different circumstances, some of which are relatively well established, others the subject of continuing research.
- It is likely that none will fit your own evaluation circumstances exactly, so do as other practitioners do – consider what you really need to evaluate, review what is already available, and extend or adapt where necessary.
- Of course, this is most effective when you have a good working knowledge of both theoretical and practical aspects of your application context.
- Expert review and end-user testing are both effective, but should be used together as complementary methods.
- Almost any degree of participant testing can reveal useful insights, but care must be taken in generalizing from a small number of users.

SUMMARY

What all have in common, however, is that they ground evaluation in real contexts of use. Differing aims for evaluation require different questions to be answered.