

Chapter 4

Evaluating Interface Designs

Agenda

- To understand the usability evaluation process.
- To understand how to create an evaluation strategy.
- To understand how to create an evaluation plan.
- Experts reviews
- Usability Testing Laboratories
- Survey Instruments
- Acceptance test
- Evaluation During Active Use

Introduction

- Why often designers fail to evaluate their own designs?

Introduction

- Designers may fail to evaluate their designs adequately.
- Experienced designers know that extensive testing is a necessity.
- The determinants of the evaluation plan include:
 - stage of design (early, middle, late)
 - number of expected users
 - criticality of the interface (life-critical medical system vs. museum exhibit support)
 - costs of product and finances allocated for testing
 - time available
 - experience of the design and evaluation team

Introduction

- The range of evaluation plans might be from an ambitious two-year test to a few days test.
- The range of costs might be from 20% of a project down to 5%.
- Few years ago, evaluation was considered as “just a good idea”
- Failure to perform and document testing could lead to failed projects.

▶ Introduction

- One troubling aspect is the uncertainty that remains even after exhaustive testing.
- The following points should be in the designers mind:
 - Perfection is not possible in complex systems, so planning must include continuing methods to asses and repair problems during the lifecycle of an interface
 - At some point a decision has to be made about completing prototype testing and delivering the product
 - Most testing methods are appropriate for normal usage, but performance in unpredictable situations with high levels of input is extremely difficult to test

The five Es of usability

■ Effective

- The completeness and accuracy with which users achieve their goals.

■ Efficient

- The speed (with accuracy) with which users can complete their tasks.

■ Engaging

- The degree to which the tone and style of the interface makes the product pleasant or satisfying to use.

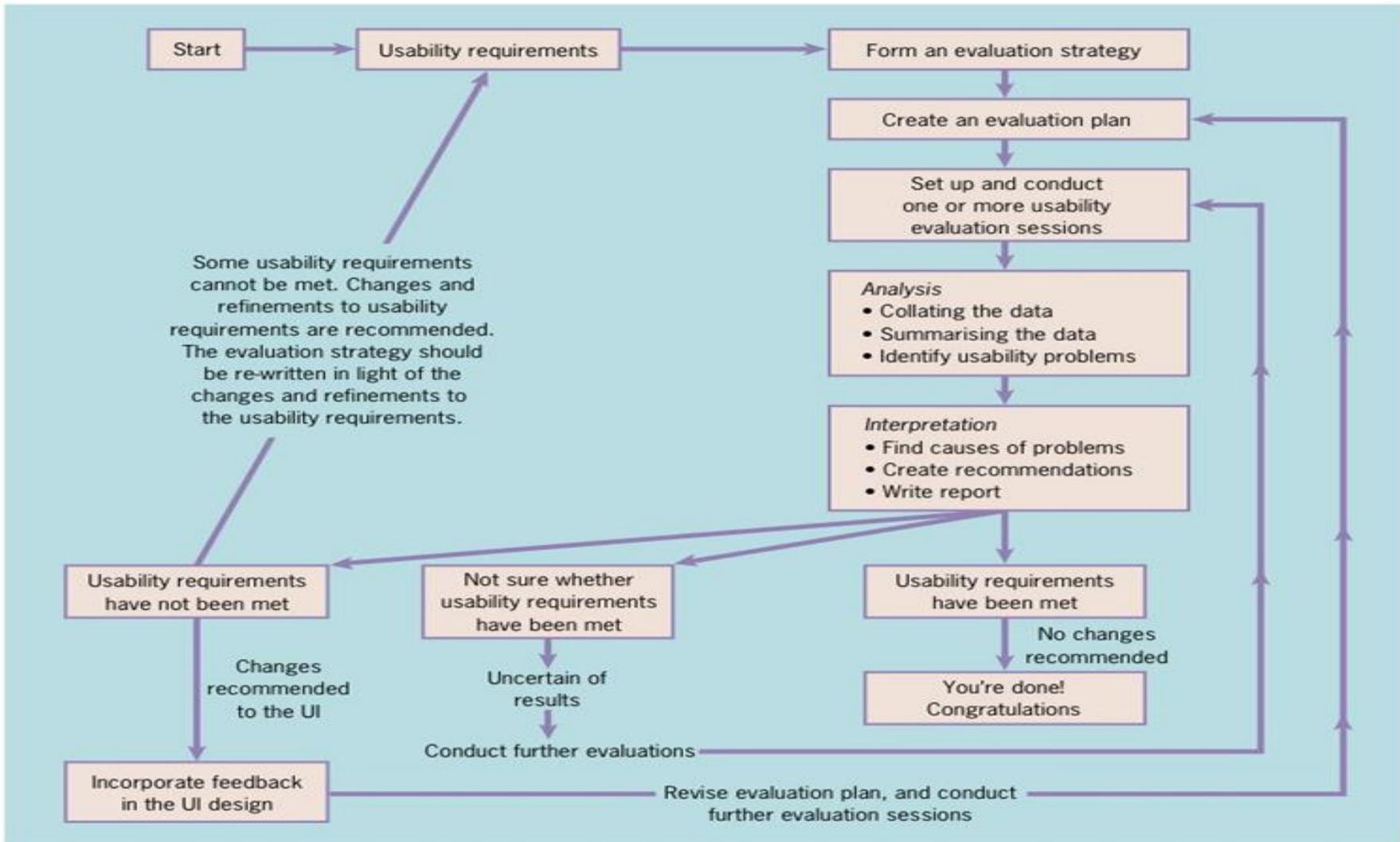
■ Error tolerant

- How well the design prevents errors or helps with recovery from those that do occur.

■ Easy to learn

- How well the product supports both initial orientation and deepening understanding of its capabilities.

Process of usability evaluation



Evaluation strategy

- **What** is the purpose of the evaluation? Are there any specific concerns or questions that you want to ask the participant about? Are there any usability requirements to explore? Define your evaluation criteria here
- **What** data do you need to collect?
- **What** product, system, or prototype are you testing?
- **What** constraints do you have?

What is the purpose of this evaluation?

- The key purpose of evaluation is to determine **whether a system meets its usability requirements** (define evaluation criteria)
 - 5Es, structure, graphics, readability, speed of performance, rate of errors by users, time to learn etc
- Qualitative usability requirements
 - They can be **subjective** and are not always easy to measure or quantify.
 - Here are two examples:
 - Railway clerks work in extremely noisy environments, so any warning messages to them should be **visually distinct** and highlighted on the screens.
 - The users on an e-shopping site should be able to order an item **easily** and without assistance.

What is the purpose of this evaluation?

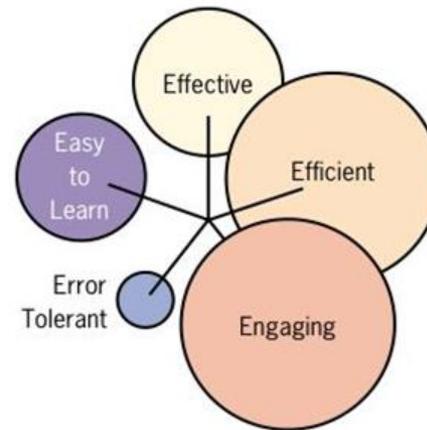
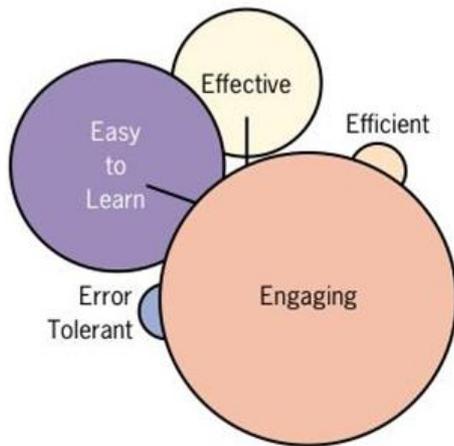
■ Quantitative usability requirements

- Usability requirements are quantitative when **explicit measures**, such as percentages, timings, or numbers **are specified**.
- These are referred to as **usability metrics**. Here are three examples:
 - It should be possible for the users to load any page of a web site **in 10 seconds** using a 56K modem.
 - It should take **no more than two minutes** for an experienced user (one who has domain knowledge and has undergone the prescribed level of training when the new system is introduced) to enter a customer's details in the hotel's database.
 - At least **four out of five** novices using the product must rate it as “easy to use” or “very easy to use” on a five-point scale where the points are “very easy to use,” “easy to use,” “neither easy nor difficult to use,” “difficult to use,” and “very difficult to use.”

Prioritizing usability requirements

- Usability requirements can be prioritized for design and evaluation.
- Knowledge about the domain, the users, their tasks, the environment, and any constraints regarding costs, budgets, timescales, and technology will help you to determine which usability requirements are most important to the success of the system.
- One way of helping stakeholders to think about and prioritize requirements is to get them to assign values to the five dimensions of usability, the Five Es.

Prioritizing usability requirements: Examples



What type of data should be collected?

■ Quantitative data

- **Numeric** data derived from taking measurements.
- For instance, if during evaluation, you are recording measurements such as the time taken by the participant to complete a task.

■ Qualitative data

- Data **without a numeric content**.
- For example, **subjective descriptions** of the difficulties that participants faced while interacting with the UI or users' stated likes or dislikes of UI features are qualitative data.

Evaluation data for the dimensions of usability

Dimension	Possible quantitative data to collect	Possible qualitative data to collect
Effective	Whether the task was completed accurately or not	User's views of whether the task was finished correctly or not
Efficient	Counting clicks/keystrokes or elapsed time on realistic tasks Analysis of navigational paths to see how often users made good choices	User's views of whether the task was easy or difficult
Engaging	Numeric measures of satisfaction	User satisfaction surveys or qualitative interviews to gauge user acceptance and attitudes toward the user interface

Evaluation data for the dimensions of usability *(Cont'd)*

Dimension	Possible quantitative data to collect	Possible qualitative data to collect
Easy to learn	Number of “false starts” — use of incorrect functions or routes Time spent in incorrect routes Time spent by a novice to complete a task compared to time spent by an experienced user to complete a task	Novice users’ reports about their level of confidence in using the interface
Error tolerant	Level of accuracy achieved in the task compared to time spent in false starts	Users reports of a feeling of confidence in the interface even if they make mistakes

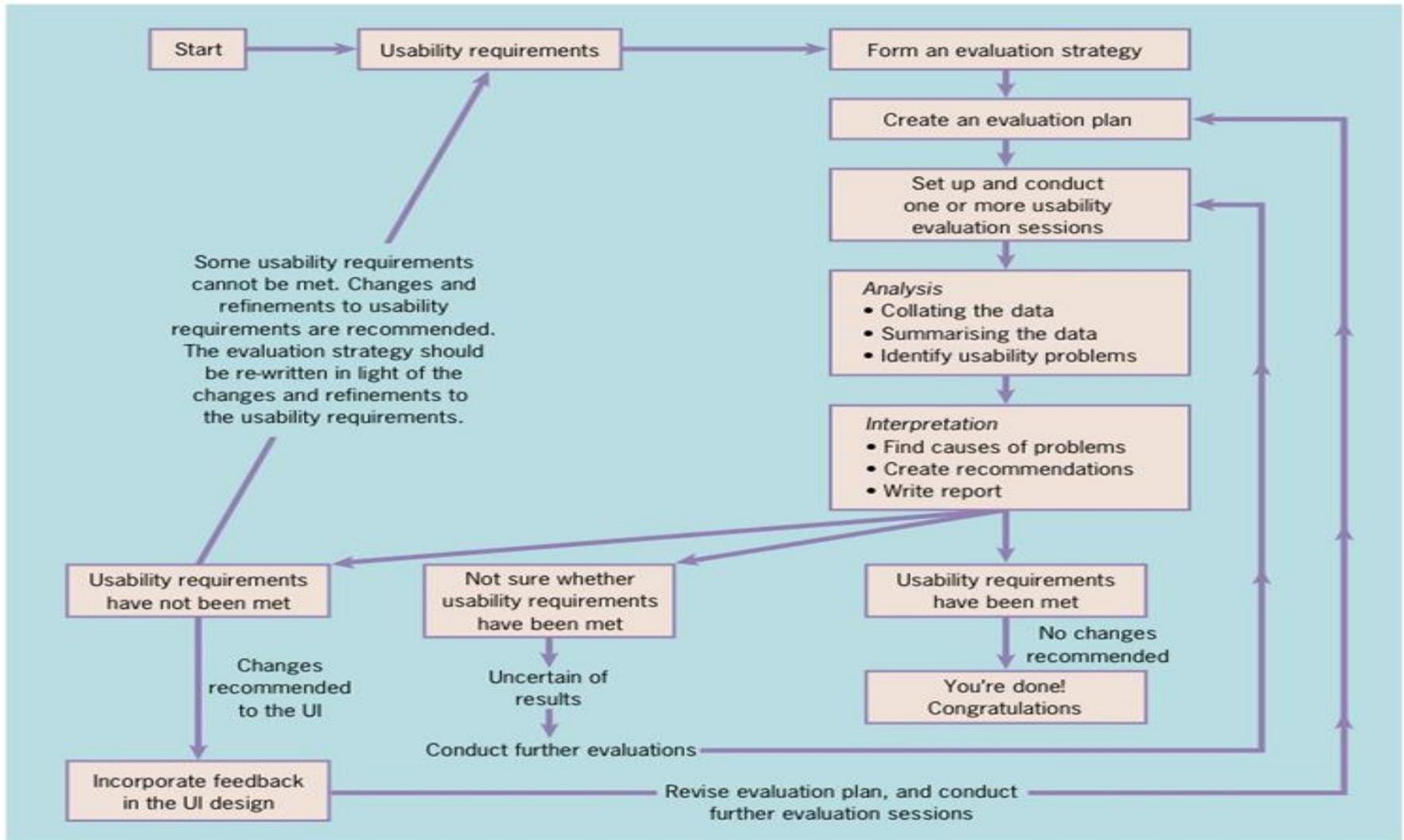
What is evaluated?

- Is it a low-fidelity prototype such as a storyboard (A sequence of drawings) or a content diagram or is it a high-fidelity interactive software prototype?
- For low-fidelity prototype
 - Useful to validate qualitative requirements and other usability concerns, but they are less useful for collecting quantitative data or for validating usability metrics.
- For high-fidelity interactive prototype
 - It is possible to take measurements to obtain quantitative data to validate the usability metrics.

What constraints do you have?

- While formulating an evaluation strategy, you should consider your constraints.
 - Money
 - Time
 - Availability of usability equipment
 - Availability of participants and the costs of recruiting them
 - Availability of evaluators

Process of usability evaluation



Evaluation plan

- Choosing your users (Who?)
- Creating a timetable (When?)
- Preparing task descriptions (What?)
- Deciding where to do the evaluation (Where?)

Choosing your users

- Who is a real user, and when is it acceptable to have someone else do your testing?
 - Find users who reflect the different skills, domain knowledge, and system experience
- Should you have one participant at a time, or would it be better for them to work in pairs?
- How many participants do you need?
- Why might it be advantageous to involve a usability expert in evaluation?
 - Usability experts are trained to understand usability issues and how to solve them, so they may be able to **identify common mistakes** more quickly than real users.
 - However, the overall aim is to ensure that real users can use the system, not that usability experts approve of it.

Creating a timetable

- There are two components to consider when drawing up a timetable for the evaluation:
 - How long do you need for each evaluation session?
 - **Between 30 and 90 minutes**, allowing time for greeting the participant and explanations before the tasks and for finishing up with your final questions.
 - How much time will the whole evaluation process take?

Preparing task descriptions

- Task descriptions represent the tasks the participant will perform while interacting with the prototype during the evaluation.
- Different kinds of tasks that you might consider:
 - Core tasks that are frequently performed by the users
 - Tasks that are very important to the users or to the business
 - Tasks that have some new design features or functionality added
 - Critical tasks, even though they may not be frequently used
 - A task that you feel has to be validated with the users for greater clarity and understanding of the design team
- Choose tasks that help you in validating the usability requirements or that focus on any particular design features or usability concerns you want to assess.
- A “task card” is simply a card with the task description on it.
 - Useful if you want to vary the order of tasks for each participant

Preparing task descriptions

■ **Example:**

■ **Structure**

- Is the site layout easy to understand and use?
- Can you navigate readily from page to page?
- Is it easy to get back to Home page or the top of a page?
- Is the loading time excessive?

■ **Graphics**

- Are graphics clear and attractive?
- Do graphics contribute to the purpose of the page?
- Are graphics excessive or distracting?
- Do graphics contribute to understanding?
- Do graphics contribute to excessive loading time?
- Do graphics aid visitor with navigation?

Deciding where to do the evaluation

- Evaluations that are undertaken in the user's own environment are called **field studies**.
- Evaluations conducted at a place somewhere else are known as **controlled studies**.
- Other techniques, e.g. experts reviews, Usability Testing and Laboratories, surveys etc

Field studies

- Useful in gathering data about the environment within which the users work as well as about the system.
 - For example, you might observe that the participant is being constantly interrupted by other colleagues' queries.
 - Based on this knowledge, in the next phase of design you might plan to incorporate reminders and status messages.
- Might be difficult to arrange and set up.

Controlled studies

- To make the controlled study more realistic, the evaluation sessions should closely simulate the user's actual work environment.

Expert Reviews

- While informal demos to colleagues or customers can provide some useful feedback, more formal expert reviews have proven to be effective
- The outcome can be a formal report with problems identified or recommendations for changes.
 - Alternatively, the review may result in a discussion with or presentation to designers or managers
- Expert reviews entail one-half day to one week effort
 - although a lengthy training period may sometimes be required to explain the task domain or operational procedures
- Expert reviews can be scheduled at several points in the development process

► Expert Reviews

- There are a variety of expert review methods to chose from:
 - Heuristic evaluation
 - Guidelines review
 - Consistency inspection
 - Cognitive walkthrough
 - Formal usability inspection



Expert Reviews: Heuristic Evaluation

- The expert reviewers critique an interface to determine conformance with a short list of design heuristics (principles), such as the eight golden rules.
- The experts should be familiar with the rules and able to interpret and apply them.
- Example heuristics (Nielsen's heuristics):
 - *"Recognition rather than recall"*
 - Are objects, actions and options always visible?
 - *"Flexibility and efficiency of use"*
 - Have accelerators (shortcuts) been provided that allow more experienced users to carry out tasks more quickly?

Expert Reviews: Guidelines Review

- The interface is checked for conformance with the organizational or other guidelines document.
- Because guidelines documents may contain hundreds of items, it may take a long time to master the guidelines and to review the interface.

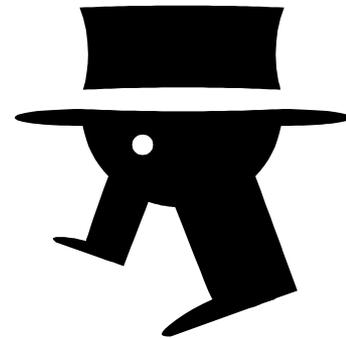


Expert Reviews: Consistency Inspection

- The experts verify consistency across a family of interfaces and help documents
- Checking for terminology, fonts, color schemes, layout, input/output formats, and so on.
- A ***bird's-eye view*** (printed screens laid out on the floor or pinned to walls) has proved to be fruitful in detecting inconsistencies and unusual patterns

Expert Reviews: Cognitive Walkthrough

- The experts simulate users walking through the interface to carry out typical tasks.
- High-frequency tasks are a starting point, but rare critical tasks should also be walked through.
- During a walkthrough, the expert should try to check:
 - will the users know what to do,
 - see how to do it, and
 - understand from feedback whether the action was correct or not?



Expert Reviews: Formal Usability Inspection

- The experts hold a courtroom-style meeting, with a moderator or judge, to present the interface and to discuss its merits and weaknesses. Design-team members comment on the design.
- Rarely used compared to other expert review methods

Expert Reviews (cont.)

- Expert reviews can be scheduled at several points in the development process when experts are available and when the design team is ready for feedback.
- Different experts tend to find different problems in an interface, so 3-5 expert reviewers can be highly productive, as can complementary usability testing.
- The dangers with expert reviews are that the experts may not have an adequate understanding of the task domain or user communities.
- Even experienced expert reviewers have great difficulty knowing how typical users, especially first-time users will really behave.

Usability Testing and Laboratories

- The emergence of usability testing and laboratories since the early 1980s
- The movement towards usability testing stimulated the construction of usability laboratories.
- A typical modest usability lab would have two 10 by 10 foot areas, one for the participants to do their work and another, separated by a half-silvered mirror, for the testers and observers.
- The Lab staff has experience in testing and user interface design.
- They may serve many projects in a year throughout an organization.
- They help the designers to make a test plan and to carry out a pilot test one week ahead of the actual test

► Usability Testing and Labs

- Participants should be chosen to represent the intended user communities,
 - with attention to background in computing, experience with the task, education, and ability with the natural language used in the interface.
- Participants should be treated with respect and should be informed that it is not *they* who are being tested; rather, it is the *interface* that is being tested
- They should be told about what they will be doing and how long they will be expected to stay.
- Participation should always be voluntary, and *informed consent* should be obtained.

► Usability Testing and Labs

- **Thinking-aloud** often leads to many spontaneous suggestions for improvements
- **Videotaping** participants performing tasks is often valuable for later review and for showing designers or managers the problems that users encounter.
- Many variant forms of usability testing have been tried:
 - Paper mockups
 - Competitive usability testing
 - Universal usability testing
 - Field test and portable labs
 - Remote usability testing
 - Can-you-break-this tests

► Usability Testing and Labs

■ Paper mockups

- It is conducted using paper mockups of screen displays to assess user reactions to wording, layout, and sequencing.
- A test administrator plays the role of the computer by flipping the pages while asking a participant user to carry out typical tasks.
- This informal testing is inexpensive, rapid, and usually productive.
- Good in early stages of design.

► Usability Testing and Labs

■ Competitive usability testing

- It compares a new interface to previous versions or to similar products from competitors.
- Needs care to construct parallel sets of tasks and to counterbalance the order of presentation of the interfaces
- Fewer participants are needed, although each is needed for a longer time period.

► Usability Testing and Labs

■ Universal usability testing

- It tests interfaces with highly diverse users, hardware, software platforms, and networks
 - consumer electronics products
 - web-based information services
 - e-government services
- Trials with the followings will raise the rate of customer success:
 - small and large displays
 - slow and fast networks
 - different operating systems and browsers

► Usability Testing and Labs

■ Field test and portable labs

- It puts new interfaces to work in realistic environments for a fixed trial period.
- Portable usability labs with videotaping and logging facilities have been developed
- A different kind of field testing is to supply users with test versions of new software or consumer products; tens or even thousands of users might receive beta versions and be asked to comment

► Usability Testing and Labs

■ Remote usability testing

□ Online usability tests

- no need to bring participants to a lab.

□ Larger numbers of participants with more diverse backgrounds

□ May add to the realism

- participants do their tests in their own environments, using their own equipment

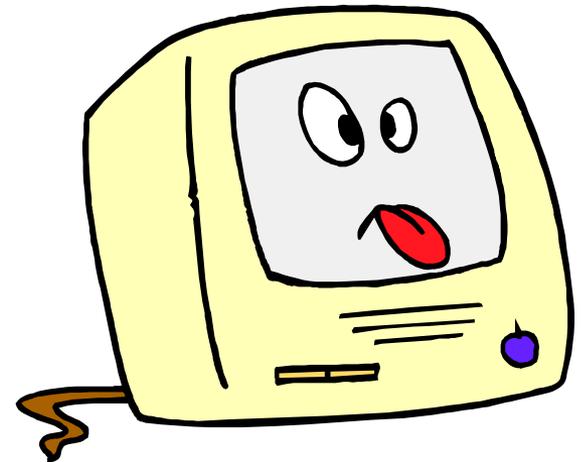
□ Less control over user behavior and less chance to observe their reactions

- Usage logs are useful supplements.

► Usability Testing and Labs

■ Can-you-break-this tests

- A destructive testing approach to usability testing by providing energetic teenagers with the challenge of trying to beat new games.
- The users try to find fatal flaws in the system or otherwise destroy it
- Pioneered by game designers; challenge of trying to beat new games



► Usability Testing and Labs

- Limitations of usability testing:
 - It emphasizes first-time usage
 - We cannot estimate how the performance will be after one week or one month of use?
 - It has limited coverage of interface features
 - Participants may get to use only a small fraction of the system's feature
- A good strategy might be:
 - Usability testing + expert reviews

Survey Instruments

- Written user surveys are a familiar, inexpensive and generally acceptable companion for usability tests and expert reviews.
- Large number of respondents offer a sense of authority compare to the potentially biased and variable results from small numbers of usability participants or expert reviewers
- Keys to successful surveys
 - Clear goals in advance
 - Development of focused items that help attain the goals.

► Survey Instruments

- Survey goals can be to ascertain the users'
 - background (age, gender, origins, education, income)
 - experience with computers (specific applications or software packages, length of time, depth of knowledge)
 - job responsibilities (decision-making influence, managerial roles)
 - reasons for not using an interface (inadequate services, too complex, too slow)
 - familiarity with features (printing, macros, shortcuts, tutorials)
 - feelings after using an interface (confused vs. clear, frustrated vs. in-control, bored vs. excited).

► Survey Instruments

- Online surveys avoid the cost of printing and the extra effort needed for distribution and collection of paper forms.
- Many people prefer to answer a brief survey displayed on a screen, instead of filling in and returning a printed form.
- Example
 - QUIS: Questionnaire for User Interaction Satisfaction
 - www.lap.umd.edu/quis/
 - WAMMI: Website Analysis and MeasureMent Inventory
 - www.wammi.com

Acceptance Test

- For large implementation projects, the customer or manager usually sets objective and measurable goals for hardware and software performance.
- If the completed product fails to meet these acceptance criteria, the system must be reworked until success is demonstrated.
- Rather than the vague and misleading criterion of "user friendly," measurable criteria for the user interface can be established for the following:
 - Time to learn specific functions
 - Speed of task performance
 - Rate of errors by users
 - Human retention of commands over time
 - Subjective user satisfaction



► Acceptance Test

- An acceptance test for a food-shopping web site might specify:
 - *The participants will be 35 adults (25-45 years old), native speakers with no disabilities, hired from an employment agency. They have moderate web-use experience: 1-5 hours/week for at least a year. They will be given a 5-minute demonstration on the basic features. At least 30 of the 35 adults should be able to complete the benchmark tasks, within 30 minutes.*
- Another testable requirement for the same interface might be this:
 - *Special participants in three categories will also be tested: (a) 10 older adults aged 55-65; (b) 10 adults users with varying motor, visual, and auditory disabilities; and (c) 10 adults users who are recent immigrants and use English as a second language.*
- A third item in the acceptance test plan might focus on retention:
 - *10 participants will be recalled after one week, and asked to carry out a new set of benchmark tasks. In 20 minutes, at least 8 of the participants should be able to complete the tasks correctly.*

► Acceptance Test

- In a large system, there may be 8 or 10 such tests to carry out on different components of the interface and with different user communities.
 - Other criteria may include: subjective satisfaction, system response time, installation procedures, printed documentation, graphical appeal, etc.
- The central goal is to verify adherence to requirements
- Once acceptance testing has been successful, there may be a period of field testing before national or international distribution.

Evaluation During Active Use

- Successful active use requires constant attention from dedicated managers, user-services personnel, and maintenance staff.

- Perfection is not attainable, but percentage improvements are possible.

- Interviews and focus group discussions
 - Interviews with individual users can be productive because the interviewer can pursue specific issues of concern.
 - Group discussions are valuable to ascertain the universality of comments.

Evaluation During Active Use (cont.)

- Continuous user-performance data logging
 - The software architecture should make it easy for system managers to collect data about
 - The patterns of system usage
 - Speed of user performance
 - Rate of errors
 - Frequency of request for online assistance
 - A major benefit is guidance to system maintainers in optimizing performance and reducing costs for all participants.

Evaluation During Active Use (cont.)

- Online suggestion box or e-mail trouble reporting
 - Electronic mail to the maintainers or designers.
 - For some users, writing a letter may be seen as requiring too much effort.

- Discussion group and newsgroup
 - Permit postings of open messages and questions
 - Some are independent, e.g. America Online and Yahoo!
 - Topic list
 - Sometimes moderators
 - Social systems
 - Comments and suggestions should be encouraged.

Summary

- To understand the usability evaluation process.
- To understand how to create an evaluation strategy.
- To understand how to create an evaluation plan.
- Experts reviews
- Usability Testing Laboratories
- Survey Instruments
- Acceptance test
- Evaluation During Active Use