Chapter 4

Evaluating Interface Designs
Introduction

- Designers may fail to evaluate their designs adequately.
- Experienced designers know that extensive testing is a necessity.
  - Few years ago, evaluation was considered as “just a good idea”
- Many determinants of the evaluation plan:
  - Stage of design, criticality, cost, time, experience, …
- 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%.
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 assess 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.
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 choose from:

- Heuristic evaluation
- Guidelines review
- Consistency inspection
- Cognitive walkthrough
- Formal usability inspection
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?
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 may rebut the evidence about problems in an adversarial format.

- Can be educational experiences for novice designers and managers, but they may take longer to prepare.

- Rarely used compared to other expert review methods
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.
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
- Discount usability testing
- 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

Discount usability testing
- Quick-and-dirty approach; widely influential
- Three to six test participants
  - Advocates say most serious problems are found with a few participants
  - Critics say that more participants are required to thoroughly test more complex systems.
- Should be used as a **formative** evaluation not as a **summative** evaluation
  - Formative evaluation: Throughout the design process; it identifies problems that guide redesign
  - Summative evaluation: Near the end of the design process; it provides evidence for product announcements
    - “94% of our 120 testers completed their shopping tasks without assistance”
    - “with 4 minutes of instruction, every participant successfully programmed the video recorder”
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.
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
Field test and portable labs

- It puts new interfaces to work in realistic environments for a fixed trial period.
- They can be made more fruitful if logging software is used to capture error, command, and help frequencies.
- 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.
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.
Can-you-break-this tests

- A destructive testing approach, in which 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

**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 compared 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 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).
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.

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.
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.

Because of the possible adversarial atmosphere, outside testing organizations are often appropriate to ensure neutrality.

The central goal is not to detect flaws, but rather to verify adherence to requirements.

Once acceptance testing has been successful, there may be a period of field testing before national or international distribution.
Skipped Sections

The following section have been skipped:

- 4.6 Evaluation During Active Use
- 4.7 Controlled Psychologically Oriented Experiments