Evaluating User Interfaces

Testing: What does testing not do?

- Guarantee perfection
- Difficult to test unusual situations
  - Military attack
  - Heavy load (e.g. voting)

Testing: Expert Review

- Colleagues or Customers
  - Ask for opinions
- Considerations:
  - What is an expert? User or designer?
- Half day to week

Heuristic Evaluation

- Give Expert heuristic, ask them to evaluate
  - Eight Golden Rules
- Specific to application area
    - Allow users to customize video and audio setting, difficulty, and game speed

Guidelines Review

- Interface is checked against organizational guidelines.
  - Military
  - Government
  - Security
  - Education

Consistency Inspection

- Verify consistency across family of interfaces
- Check terminology, fonts, color, layout, i/o formats
- Look at documentation and online help

Usability Testing and Labs

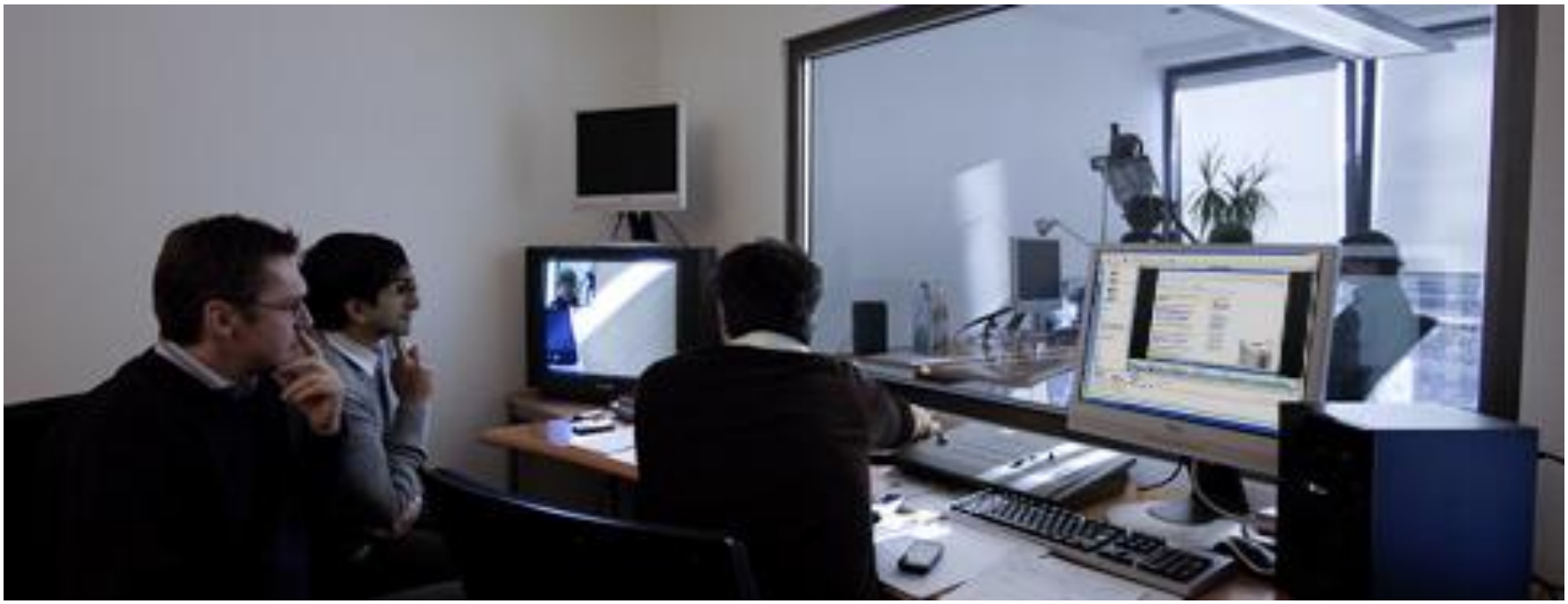
- 1980s, testing was luxury (but deadlines crept up)
- Sped up projects
- Cost savings
- Labs are different than academia
  - Less general theory
  - More practical studies

The Usability Lab

- Similar to TV studio: microphones, audio, video, one-way mirror

Usability Labs

- IBM early leader
- Microsoft next (>25 labs)
- Now hundreds of companies



Staff

- Expertise in testing (psych, hci, comp sci)
- 10 to 15 projects per year
- Meet with UI architect to plan testing
- 2-6 weeks, design and test plan
  - E.g. Who are participants? Beta testers, current customers, in company staff, advertising
- 1 week, pilot test (1-3 participants)

Participants

- Labs categorize users based on:
  - Computing background
  - Experience with task
  - Education
  - Ability with the language used in the interface
- Controls for
  - Physical concerns (e.g. eyesight, handedness, age)
  - Experimental conditions (e.g. time of day, physical surroundings, noise, temperature, distractions)

Recording Participants

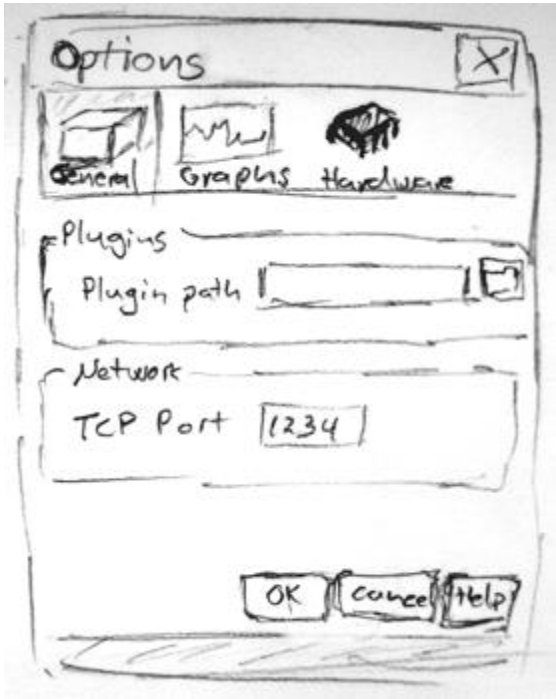
- Logging is important
  - Software to help (Live Logger, Morae, Spectator)
  - New approaches: eye tracking
  - Focus users on interface
  - Tell them the task, duration

Thinking Aloud

- think aloud
  - Invite users to *think aloud*
  - Nothing they say is wrong
  - Don’t interrupt, let the user talk
  - Spontaneous, encourages positive suggestions
  - Can be done in teams of participants
- Retrospective think aloud
  - Asks people afterwards what they were thinking
  - Does not interrupt users

Types of Usability Testing

- Paper mockups and prototyping
  - Inexpensive, rapid, very productive
  - Low fidelity is sometimes better (Synder, 2003)



Example Usability Test with a Paper Prototype

http://www.youtube.com/watch?v=9wQkLthhHKA

Types of Usability Testing

- Discount usability testing
  - Test early and often (with 3 to 6 testers)
  - Pros: Most serious problems can be found with 6 testers. Good for formative evaluation (early)
  - Cons: Complex systems can't be tested this way.
- Competitive usability testing
  - Compare against prior versions

Types of Usability Testing

- Universal usability testing
  - Test with highly diverse
    - Users (experience levels, ability, etc.)
    - Platforms (mac, pc, linux)
    - Hardware (old (how old is old?) -> latest)
    - Networks (dial-up -> broadband)
- Field tests and labs
  - Tests UI in realistic environments

Types of Usability Testing

- Remote usability testing (via web)
  - Recruited via online communities, email
  - Large n
  - Difficulty in logging, validating data
  - Software can help (NetMeeting, WebEx)

Limitations of Testing

- Focuses on first-time users
- Limited coverage of interface features
  - Emergency (military, medical, mission-critical)
- Difficult to simulate realistic conditions
  - Testing mobile devices
    - Signal strength
    - Batteries
    - User focus

Survey Instruments

- Questionnaires
  - Paper or online (e.g. surveymonkey.com)
  - Easy to grasp for many people
  - The power of many can be shown
    - 80% of the 500 users who tried the system liked Option A
    - 3 out of the 4 experts like Option B

Designing survey questions

- Ideally
  - Based on existing questions
  - Reviewed by colleagues
  - Pilot tested
- Direct activities are better than gathering statistics

Scenario 1 - Finding information

Imagine you are an alum that wants to know whether the organisation (which website you are testing) will have an interesting event while you are in town. Try to accomplish this task and come back to answer the following questions.

Could you complete the task? *

- ◯ Yes
- ◯ No

If you were successful, please rate how simple/confusing was to complete it. *

	1	2	3	4	5	
Very Simple	◯	◯	◯	◯	◯	Very Confusing

Independently of your success, could you describe what steps you followed to complete the task?

Do you have any suggestions on how we can improve the completion of this task?

Likert Scales

- Most common methodology
  - Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
- 5, 7, 9-point scales
- Examples
  - Improves my performance in book searching and buying
  - Enables me to search and by books faster
  - Makes it easier to search for an purchase books
- What does 1.5 mean?

Most Used Likert-scales

- E.g. questions
  - How long have you worked on this system?
  - Learning to operate
    - Difficult 1 2 3 4 5 6 7 8 9 Easy
- Computer System Usability Questionniare
- Software usability Measurement Inventory
- Website Analysis and Measurement Inventory
- Mobile Phone Usability Questionnaire

Acceptance Tests

- Set goals for performance
  - Objective
  - Measurable
- Examples
  - Mean time between failures
  - Test cases
    - Response time requirements
    - Readability (including documentation and help)
    - Satisfaction

Examples

- Test A
  - The participants will be
    - 35 adults (25-45 years old)
    - Native speakers with no disabilities
    - Hired from an employment agency
    - Moderate web-use experience (1-5 hours/week) for at least one year
  - >30 of the 35 should complete the benchmark tests within 30 minutes
- Test B
  - The participants will be
    - 10 older adults 55-65
    - 10 adult users with varying motor, visual, and auditory disabilities
    - 10 adult users who are recent immigrants and use English as a second language
- Test C
  - Ten participants will be recalled after one week
  - Carry out new set of benchmark tests
  - In 20 minutes, at least 8 should be able to complete tasks

Acceptance Tests

- Different than usability tests
  - More neutral
  - Users should complete the whole test

Evaluation during use

- Evaluation methods after a product has been released
  - Interviews with individual users
    - Get very detailed on specific concerns
    - Costly and time-consuming
  - Focus group discussions
    - Patterns of usage

Continuous Logging

- The system itself logs user usage
- examples
  - Track frequency of errors
  - Speed of performance
  - Track which features are used and which are not
  - Web Analytics
- Privacy? What gets logged?
- What about companies?

Online and Telephone Help

- Users enjoy having people ready to help (real-time chat online or via telephone)
- E.g. Netflix has 8.4 million customers, how many telephone customer service reps?
  - 375
  - Expensive, but higher customer satisfaction

Further Reading

- https://www.usabilityhome.com/