Usability Engineering

Nielson’s book on Usability Engineering, Chapter 6
Usability Testing

• Reliability: can the test results be replicated
  – Individual differences
  – Sampling issues
  – Statistics: which we will cover next week
  – How many users?

“elaborate usability tests are a waste of resources. The best results come from testing no more than 5 users and running as many small tests as you can afford.” - Nielsen.
Types of Evaluation

• Formative
  – Help iteration testing
  – Help improvement of the design

• Summative
  – Establish the quality of the design
  – Helps compare between two designs
Test Plan

- Goal
- Where and when?
- How long?
- What computer support will be needed?
- What software is needed?
- What is the state of the system?
- Who will serve as experimenters?
- Who are the test users? How are you recruiting them?
- How many test users are needed?
- How do you know when user are done?
- What user aids is needed?
- To what extend can the experimenter help?
- What data to collect?
- How would you measure success for the design?
Pay-off Number of Users

![Graph showing the ratio of benefits to cost against the number of test users. The graph has a peak at around 4 test users, after which the ratio decreases.]
Pilot Testing

• Trial of the method to see if they work
• Also helps answer all the questions needed for the test plan
Recruitment

• Representative sample
• Novice vs. expert users
Experimental designs

2 types: **Between-groups** versus **within-subjects** designs

**Between-groups (subjects) (independent measures)**
Each subject participates in **only one condition** of the study.
e.g. gender differences in memory.

**Within-subjects (repeated measures)**
Each subject does **all of the conditions** in a study.
e.g. effects of alcohol on memory (test w./w.o. alcohol)
Between subjects design - HCl

– **Users** divided into two groups

– One sees interface X, the other Y

– Results are compared between groups:
  • Is mean(Xi) >= mean (Yi)?
  • Test using statistics, e.g. T-test

– Eliminates e.g. ordering effects – users cannot learn from interface X to do better on Y
Within subjects design - HCI

• Each user sees both interface X and Y, in random order

• Results are compared within each user
  – For user I, compute the difference Xi-Yi
  – Is mean(Xi-Yi) > 0?

• Eliminates variation due to user differences as user is only compared with self
Ethical considerations

Before the test:
Have everything ready before the user shows up.
Emphasize that it is the system that is being tested, not the user.
Acknowledge that the software is new and untested, and may have problems.
Let users know that they can stop at any time.
Explain any recording, keystroke logging, or other monitoring that is used.
Tell the user that the test results will be kept completely confidential.
Make sure that you have answered all the user’s questions before proceeding.

During the test:
Try to give the user an early success experience.
Hand out the test tasks one at a time.
Keep a relaxed atmosphere in the test room, serve coffee and/or have breaks.
Avoid disruptions: Close the door and post a sign on it. Disable telephone.
Never indicate in any way that the user is making mistakes or is too slow.
Minimize the number of observers at the test.
Do not allow the user’s management to observe the test.
If necessary, have the experimenter stop the test if it becomes too unpleasant.
Test Tasks

• Representative
• Provide coverage
• Short
• Clear
• Go from simple to hard
• Identify:
  – task analysis
  – product identity listing
  – field observations
  – logs of use
What to do in the test?

• Introduction
  – Consent
  – Privacy and confidentiality
  – Stop at any point
  – Explain the study
  – Ask if the user has any question

• Test

• debrief
Usability Testing: Testing

- Performance Measurement
- Retrospective Testing (user review the recording of interaction recounting what he/she has done)
- Think Aloud Protocol
- Constructive interaction (like think aloud but with two users)
- Observation
Typical quantified variables to measure

- Time users take to complete a task
- The number of tasks that can be completed within a time limit
- The ratio between successful interactions and errors
- Time spent recovering from errors
- Number of user errors
- Number of immediately subsequent errors
- Number of commands or features that were utilized by the user
Typical quantified variables to measure

- The number of commands never used
- Number of system features the user can remember during debriefing
- Frequency of use of manuals and help and time spent on these
- Frequency of help solving user problems
- Ratio of positive User statements
- Number of times users express frustration
- No of users who prefer this system
- No of times a user worked around unsolvable problems
Typical quantified variables to measure

- Amount of dead time
- Number of times user sidetracked from focusing on task
Usability Testing: Inspection

• Cognitive Walkthroughs
• Heuristic Evaluation
Usability Testing: Inquiry

- Field Observation
- Focus Group
- Interviews
# Choosing a method

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Lifecycle Stage</th>
<th>Users Needed</th>
<th>Main Advantage</th>
<th>Main Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic evaluation</td>
<td>Early design, &quot;inner cycle&quot; of iterative design</td>
<td>None</td>
<td>Finds individual usability problems. Can address expert user issues.</td>
<td>Does not involve real users, so does not find &quot;surprises&quot; relating to their needs.</td>
</tr>
<tr>
<td>Performance measures</td>
<td>Competitive analysis, final testing</td>
<td>At least 10</td>
<td>Hard numbers. Results easy to compare.</td>
<td>Does not find individual usability problems.</td>
</tr>
<tr>
<td>Thinking aloud</td>
<td>Iterative design, formative evaluation</td>
<td>3–5</td>
<td>Pinpoints user misconceptions. Cheap test.</td>
<td>Unnatural for users. Hard for expert users to verbalize.</td>
</tr>
<tr>
<td>Observation</td>
<td>Task analysis, follow-up studies</td>
<td>3 or more</td>
<td>Ecological validity; reveals users’ real tasks. Suggests functions and features.</td>
<td>Appointments hard to set up. No experimenter control.</td>
</tr>
<tr>
<td>Questionnaires</td>
<td>Task analysis, follow-up studies</td>
<td>At least 30</td>
<td>Finds subjective user preferences. Easy to repeat.</td>
<td>Pilot work needed (to prevent misunderstandings).</td>
</tr>
<tr>
<td>Interviews</td>
<td>Task analysis</td>
<td>5</td>
<td>Flexible, in-depth attitude and experience probing.</td>
<td>Time consuming. Hard to analyze and compare.</td>
</tr>
<tr>
<td>Focus groups</td>
<td>Task analysis, user involvement</td>
<td>6–9 per group</td>
<td>Spontaneous reactions and group dynamics.</td>
<td>Hard to analyze. Low validity</td>
</tr>
<tr>
<td>Logging actual use</td>
<td>Final testing, follow-up studies</td>
<td>At least 20</td>
<td>Finds highly used (or unused) features. Can run continuously.</td>
<td>Analysis programs needed for huge mass of data. Violation of users’ privacy.</td>
</tr>
<tr>
<td>User feedback</td>
<td>Follow-up studies</td>
<td>Hundreds</td>
<td>Tracks changes in user requirements and views.</td>
<td>Special organization needed to handle replies.</td>
</tr>
</tbody>
</table>
Applying this to games

READINGS
Readings

• Games are different
  – Not productivity but for leisure
  – Market is crowded
  – What else?
Readings

• TRUE Method and instrumentation
• Advantages
• Disadvantages
Readings

• RITE (Rapid Iterative Test Evaluation) Method
• Advantages
• Disadvantages
Class Assignment

• Candy Crush Experiment Design

Goal: I want to know if the game interface and mechanics are usable

– What method would you use?
– Why?
Class Assignment

• Developing a new game for children
  – What method would you use? Why?
  – How would you design it?