User Interface
Evaluation Methods
• Usability is, according to ISO 9241-11:

“the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”

• How to measure it??
Evaluation Methods

- **Analytical (without users)**
  - Heuristic Evaluation
  - Cognitive Walkthrough
  - Model based methods
  - Review methods

- **Empirical (involving users)**
  - Observation
  - usability tests
  - Query
  - Controlled Experiments

(✓ - have used in Lab classes)
Heuristic Evaluation (Nielsen and Molich 1990)

• A “discount usability engineering method” for quick, cheap, and easy evaluation of a UI design

• The most popular of the usability inspection methods

• It is a systematic inspection of a design for usability

• Meant to find the usability problems in the design so that they can be attended to as part of an iterative design process.

• Involves a small set of analysts judging the UI against a list of usability principles ("heuristics").
• Is difficult for a single individual to do; one person will never be able to find all the problems

• Involving multiple evaluators improves the effectiveness of the method significantly

• Nielsen generally recommends to use three to five evaluators

• not much gain that much additional information by using larger numbers
Example:

• Heuristic evaluation of a banking system:
  – 19 evaluators
  – 16 usability problems

black square - problem found
white square – not found

http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/#sthash.OmTrV7Og.6ZrkgzXB.dpuf

Conclusion : in general 3 to 5 evaluators is reasonable
How to perform Heuristic Evaluation

• Each evaluator:

• First make a general analysis to get to know the UI

• Then, make a systematic analysis having in mind the heuristics

• Take note of each potential usability problem, the heuristic and the severity grade

• Finally, compile all the potential problems

http://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation
sthash.OmTrV7Og.dpuf
Ten Nielsen’s heuristics

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

https://www.nngroup.com/articles/ten-usability-heuristics/
• **Main advantages of heuristic evaluation:**
  
  – May produce useful results with modest investment
  
  – Simple to apply even by not very experienced evaluators

• **Main limitations:**
  
  – Subjective (partially overcome with more evaluators)
  
  – Can’t find all usability problems
    
    -> evaluation involving users are needed!
Cognitive Walkthrough  (Wharton, et al., 1992)

• Usability inspection method not involving users (analytical)

• Based on the fact that users usually prefer to learn a system by using it (e.g., instead of studying a manual)

• Focused on assessing learnability (i.e., how easy it is for new users to accomplish tasks with the system)

• May produce results quickly at a low cost

• Applicable at early phases, before any coding
How to perform a cognitive walkthrough

1- Task analysis: sequence of steps or actions required by a user to accomplish a task, and the system responses

2- Designers and developers walkthrough as a group, asking themselves a set of questions at each step

3- Data gathering during the walkthrough: answering the questions for each subtask usability problems are detected

4- Report of potential issues

5- UI redesign to address the issues identified
CW Four questions:

• Will the user try to achieve the effect that the subtask has?  
  (Does the user understand this subtask is needed to reach the goal?)

• Will the user notice that the correct action is available?  
  (E.g. is the button visible?)

• Will the user understand that the wanted subtask can be achieved by the action?  
  (E.g. the button is visible but the user doesn’t understand the text and will not click on it)

• Does the user get feedback?  
  Will the user know that they have done the right thing?
Common issues

• The evaluator doesn't know how to perform the task; the method involves the **optimal** sequence of actions

• Involves an extensive analysis and documentation and often too many potential issues are detected, resulting very **time consuming**

Thus:

**Lighter variants of Cognitive Walkthrough** were proposed to make it more applicable in S/W development companies
Streamlined Cognitive Walkthrough (Spencer, 2000)

- Only two questions:
  
  - Will the user know what to do at this step?
  
  - If the user does the right thing, will they know that they did the right thing, and are making progress towards their goal?

- And a set of rules to streamlining the walkthrough and trade-off granularity for coverage

comprises the 3 first questions of CW
According to Spencer the method can be applied successfully if the usability specialist:

• takes care to prepare the team for the walkthrough,

• avoids design discussions during the walkthrough,

• explicitly neutralizes defensiveness among team members,

• streamlines the procedure by collapsing the first three questions into one question,

• and captures data selectively
Example: Evaluation of a desktop photocopier UI

- Machine UI:
  - numeric keypad,
  - "Copy" button,
  - push button on the back to turn on the power

The machine automatically turns itself off after 5 min inactivity

- Task: copy a single page

- User: any office worker

- Actions needed: turn on the power, put the original on the machine, press the "Copy" button

http://hcibib.org/tcuid/chap-4.html#4-1
• Story for action number one:
  “the user wants to make a copy and knows that the machine has to be turned on. So she pushes the power button. Then she goes on to the next action”

Not convincing!

• why shouldn't the user assume that the machine is already on? That is often the case

• Will the user figure out that the machine is off, and find the power switch?
  etc. etc.
Practice the Streamlined Cognitive Walkthrough:

• Look for a phone number at the University of Aveiro Web site
  user: any student from the University

• Create a pdf of a PowerPoint file using the Print option but not
  printing the hidden slides
  user: anyone familiar with a previous version
Look for a person’s phone number at the University of Aveiro Web site

user: any student from the University

Task analysis:

- look for the icon 📚 (directório);
- input part of the person’s name and search
- get the phone number

But the defined user profile (any student from the University) includes foreign students, thus a previous action is needed:

- select the English version

For each action we need to ask the two questions and put ourselves in the shoes of the user!
First action: find the icon

Q1 - Will the user know what to do at this step?

Even reading the tooltip (directório) possibly the correct icon is not recognizable!

Q2 - If the user does the right thing (selects the icon), will they know that they did the right thing, and are making progress towards their goal?

Probably yes; this looks a familiar search bar and it is adequately labeled (lista telefónica; pesquisar)
Second action: input part of the person’s name

Q1 - Will the user know what to do at this step?

Probably yes; the tooltip lets the user know he/she should input the person’s name and select “pesquisar”

Q2 - If the user does the right thing (selects the icon), will they know that they did the right thing, and are making progress towards their goal?

Probably yes; however, some users might not recognize 24117 as a phone number (it only has 5 digits, as it is internal, and not 9 as possibly expected)
Evaluation Methods

- **Analytical (without users)**
  - Heuristic Evaluation ✓
  - Cognitive Walkthrough
  - Model based methods
  - Review methods

- **Empirical (involving users)**
  - Observation
  - usability tests ✓
  - Query
  - Controlled Experiments

( ✓ - have used in Lab classes)
Another example: Smart TV

How to access the Internet?

(before reading the manual?)

(we see the symbol at the screen only after pressing it on the control!)
Practice the Streamlined Cognitive Walkthrough:

Analysing interactive systems/applications that should be very intuitive:

• Turn on and off the video projector in your Lab using the remote control or directly on the projector
  user: any student from the University

• Change the Channel using the box of your TV service (not the remote control)
  user: anyone having a TV box
Limitations of Analytical Methods

- Are subjective
- Involve several usability experts
- Cannot find all usability problems

Thus, empirical methods (involving users) are needed

observation
query
controlled experiments (scientific approach)

Usability test (engineering approach)
Ethics in applying empirical methods

Involving users implies specific cautions:

- Asking for explicit consent
- Confidentiality
- Security (avoid any risk)
- Freedom (users may give up at any time)
- Limit stress

It’s the system that is under evaluation not the user!
Observation

Has many variants from very simple to very complex and expensive:

• Direct: observer takes notes

• Undirect: through audio/ video – more complex and time consuming

• Think Aloud: users are asked to explain what they are doing

• Logging: users activity is logged by the system

• Combinations of the previous, etc
Query

• Two main variants:
  – Questionnaire  (reach more people; less flexible)
  – Interview

• Should be carefully prepared and tested

• Collected data should be carefully analyzed
Usability tests

- Involve observation and query

- Main aspects:
  - Participants
  - Tasks
  - Test facilities and systems
  - Protocol
  - Usability measures
  - Data analysis

- Have a complex logistics

- Standard: **Common Industry Format (CIF)** for usability test reports
Participants

• The total number of participants to be tested
  (a valid statistical analysis implies a sufficient number of subjects)

• Segmentation of user groups tested, if more than one

• Key characteristics and capabilities of user group
  (user profile: age, gender, computing experience, product experience, etc.)

• How to select participants

• Differences between the participant sample and the user population
  (e.g. actual users might have training whereas test subjects were untrained)
Tasks

• The task scenarios for testing

• Why these tasks were selected
  (e.g. the most frequent tasks, the most troublesome tasks)

• The source of these tasks
  (e.g. observation of users using similar products, product specifications)

• Any task data given to the participants

• Completion or performance criteria established for each task
  (e.g. n. of clicks < N, time limit)
Test Facilities and equipment

• The setting and type of space in which the evaluation will be done  
  (e.g. usability lab, cubicle office, meeting room, home office, home family 
  room, manufacturing floor, etc.)

• Any relevant features or circumstances that can affect the results  
  (e.g. video and audio recording equipment, one-way mirrors, or automatic 
  data collection equipment)

• Participant’s computing environment  
  (e.g. computer configuration, including model, OS version, required libraries 
  or settings, browser name and version; relevant plug-in, etc.)

• Display and input devices characteristics

• Any questionnaires to be used
Protocol

• Procedure: the logical design of the test

• Participant general instructions and task instructions

• The usability measures to be used:
  a) for effectiveness (completeness rate, errors, assists)
  b) for efficiency (times)
  c) for satisfaction
Common Industry Format (CIF) for usability test reports
ISO/IEC 25062:2006

• Specifies the format for reporting the results of a **summative** evaluation

• The most common type of usability evaluation is **formative**, (i.e. designed to identify problems that can be fixed)

• A summative evaluation produces usability metrics that describe how usable a product is when used in a particular context of use

• The CIF report format and metrics are consistent with the ISO 9241-11
http://www.usabilitynet.org/prue/cif.htm
http://zing.ncsl.nist.gov/iusr/

• Top ten things to know about the CIF (a list of do and don’t)
http://zing.ncsl.nist.gov/iusr/top_ten.html
Controlled experiments

• The “work horse” of science ...

• Important issues to consider:
  
  – Hypothesis
  – Variables (input or independent; output or dependent)
  – Secondary variables
  – Experimental design (within groups; between groups)
  – Participants (number, profile)
  – Statistics
Controlled experiment

• Define an hypothesis

• Define input (independent), output (dependent) and secondary variables

• Define experimental design (within-groups / between groups)

• Select the participants

• Prepare all the documentation:
  - list of tasks and perceived difficulty
  - final questionnaire
  - list of tasks for the observer to take notes

• Run a pilot test

• Take care of the logistics … and after the experiment analyze data
“Effect of Hand-Avatar in a Selection Task using a Tablet as Input Device in an Immersive Virtual Environment”

L. Afonso, P. Dias, C. Ferreira, B. Sousa Santos
IEEE 3D UI, Los Angeles, March 2017

• Research question: How does the virtual representation of the user's hands influence the performance on a button selection task performed in a tablet-based interaction within an immersive virtual environment?

• Method: Controlled experiment

• 55 participants used three conditions:
  - no-hand avatar,
  - realistic avatar,
  - translucent avatar.

• Participants were faster but made slightly more errors with no-avatar.

• Considered easier to perform the task with the translucent avatar.
Experimental Design

Null Hypothesis: usability is independent of the hands representation

Independent variable (with 3 levels): representation of the hands

Dependent variable: usability (performance + satisfaction)

Within-groups: all participants used all experimental conditions in different sequences (to avoid bias)

Task: selecting as fast as possible a highlighted button from a group of twenty buttons (repeated measures)
Experimental Conditions

1- No avatar: the user only sees the virtual tablet;

2- Realistic avatar: a realistic representation of the hands movement is shown

3- Translucent avatar: a translucent hand model is used (to alleviate occlusion)
Experimental Set-up

- Laptop running the main application (in Unity)
- HMD (Oculus Rift DK2) providing head tracking
- Tablet (Google Nexus 7) as input device running the controller application (in Unity)
- Leap Motion (mounted on the HMD) to track the user’s hands
- Tablet camera tracking the position and orientation of an AR marker on the HMD to map tablet position in the virtual world (using Vuforia)
Results

Selection time:
Participants completed the button selections in average faster with no-avatar

Selection errors:
Participants made slightly less errors with avatar (realistic or translucent)

Participants’ opinion:
The translucent avatar:
- was more often preferred
- was considered as better than the realistic avatar (statistically significant)
Controlled experiment: comparing two versions of Meo Go

Altisse Labs User - UX Group
DETI – HCI course
April/2016
Controlled experiment: comparing two versions of Meo Go: Current version vs Version to be deployed

- Null Hypothesis: both are equally usable
- Input (independent) variable: version
- Output (dependent) variables: performance, opinion and satisfaction
- Secondary variable: participant profile
- Participants: 66 volunteer HCI -2016 students (12 female)
- Experimental design: between groups (one version per participant)
- Exploratory Data Analysis and non-parametric tests (ordinal variables)
Experimental Protocol

• Experimenter salutes participant and explains the experiment

• Participant performs tasks thinking aloud while Experimenter observes participant

• Experimenter registers participant’s performance

• Participant answers questionnaire

This controlled experiment involved observation and questionnaire
Obtained data

– total task time
– task completion rate
– difficulty stated by participants
– difficulty as observed by observers
– participants comments and suggestions
– observer comments
– participants’ profile
– most important characteristics for participants
– participants’ opinion
Main results

• Characteristics considered as more important for an application of this type:
  – Find the schedule for a specific TV channel for the next days – 94%
  – Search a specific TV show providing channel and schedule - 0 76%
  – TV shows currently shown in several channels 76%

• Satisfaction measure in a Likert-like scale (1-not at all ... 5- very much satisfied)

No significant difference between versions
## Application characteristics  (1- completely disagree  5- agree)

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Median values:</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Easy to navigate in the application</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Easy to find what I need in the application</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Easy to understand functionality of elements</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Application works as expected</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Application has all expected functionality</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Application is pleasant to use</td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Application is attractive</td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>I need help to perform some tasks</td>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Experience with similar applications is needed</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>I would recommend this application to friends</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>Application complements my TV experience</td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Application characteristics (1- completely disagree  5- completely agree)

<table>
<thead>
<tr>
<th></th>
<th>1 – The navigation model “Now/Later” is understandable</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>[Significantly different]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2 – The “Now” button is useful</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>5</td>
<td>[Significantly different]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3 – Automatic alignment of TV shows is helpful</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>4 – I fell lost after navigation in a specific channel</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>5 – The <em>timeline</em> is useful for positioning in time</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6 – The date filter is redundant in view of the <em>timeline</em></th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>7 – It is important to have a channel filter</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>8 – Images of current TV shows are relevant</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Task Difficulty reported by the participants
(1- very difficult ... 5- very easy)

<table>
<thead>
<tr>
<th>Task</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task 1</strong> – Get the timetable of the current afternoon</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>Task 2</strong> – Get the synopsis of the TV show you want to watch</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>Task 3</strong> – Navigate to the current TV show</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td><strong>Task 4</strong> – Visualize the same TV show tomorrow at 20h00</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Task 5</strong> – Look up current TV shows in the first TV channels</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

No significant differences were found between the medians of the two versions.
Task success reported by the observers
(1- did not succeed ... 5- easy to success)

<table>
<thead>
<tr>
<th>Task</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1 – Get the timetable of the current afternoon</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Task 2 – Get the synopsis of the TV show you want to watch</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Task 3 – Navigate to the current TV show</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Task 4 – Visualize the same TV show tomorrow at 20h00</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Task 5 – Look up current TV shows in the first TV channels</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

No significant differences were found between the medians of the two versions
Total time (to perform all tasks)

No significant difference between the medians corresponding to the two versions was found
Summary of results

• Participants were satisfied with both versions
  no significant difference between versions was observed

• Both versions got a good classification concerning application usage
  no significant difference between versions was observed

• Both versions got a good classification concerning difficulty
  no significant difference between versions was observed

• Both versions got a good classification concerning specific aspects of the application
  two aspects improved significantly in version 2

• No significant difference between the two versions is observed concerning total task time
Study main limitations

• Participants’ profile - students are not representative of all target users

• Tasks – are simple and performed in a controlled context - limit ecological validity

• Think aloud - total task time must be considered with caution

However:

HCI students are aware of usability issues and provided valuable feedback
• This data analysis was complemented by the analysis of comments and suggestions

• The combined analysis provided more insight concerning usability issues

• Allowed **identify easy to implement improvements** in the UI with potential **positive impact on the UX** of the final version
Bibliography – Books and links

- Jackob Nielsen, Usability Engineering, Morgan Kaufmann, 1993
- Standard ISO 9241-11 - *Ergonomic requirements for office work with visual display terminals Part 11: Guidance on usability*
- What is CIF? - [http://www.usabilitynet.org/prue/cif.htm](http://www.usabilitynet.org/prue/cif.htm)
Bibliography - Papers


http://www.usabilitynet.org/prue/cif.htm
Mandatory Readings for the Exam

• Slides

• Alan Dix t al., *Human-Computer Interaction*, 3rd ed., Prentice Hall, 2004 (at the Library)
  - Chapters: 1 to 4 (topics addressed in the slides)
  - Chapter 9 (topics addressed in the slides)
  - Chapters 12, 14 and 16 (topics addressed in the slides)

• Randolph Bias, Deborah Meyhew, *Cost Justifying usability*, Morgan Kaufmann, 2005
  (Chapter 3, pag. 42-55 in Moodle)

  (Chapter 29, in Moodle or at)