Universitat de Lleida

EVALUATION TECHNIQUES AND USABILITY TESTING

Theme II – Usability Testing Study

(4rt Part – Questionnaires; Setting and collecting metrics)

Montserrat Sendín
DIEI - UdL

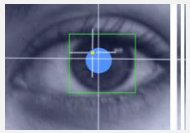
Contents

- Setting of metrics for non-standardized usability properties
- Questionnaires

Test Plan – *Documents to prepare*

- **Sections** to be included in the **Test Plan main document**:
  - Product definition
  - Purpose and test objectives
    - Test type
  - Product state and problems in which to incise
  - Target audience, user profile definition and target user sample
  - Participants recruitment process
  - Methodology
    - Procedure and protocol
    - Number of participants
    - Scenarios for the tasks to be evaluated
  - Test environment and equipment
  - **Metrics and data to be registered** (including qualitative data)
  - List of complementary documents

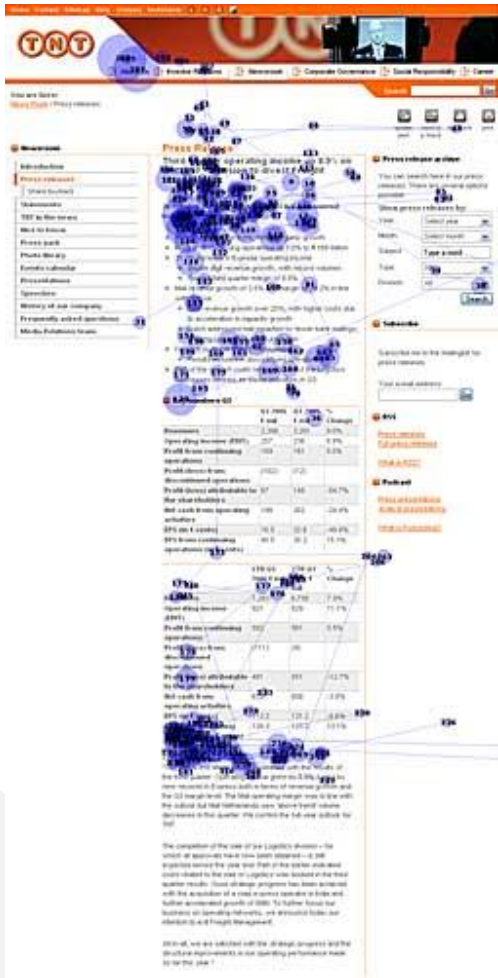
Setting of metrics - *Eyetracking*



Specific metrics for *eyetracking*

Remember *eyetracking* can help you knowing about **clarity** and **visibility** regarding some particular elements in the interface

Their specific metrics let you know about users' **visual attention**



Quantitative metrics

- **Duration of fixations** ➔ *Fixation duration mean*
- **Number of fixations** (*fixation count*): amount of visual fixations registered ➔ *Average fixation count*
- ***Regarding a specific AOI:***
  - Time to fist fixation, Fixations before, Fixation count, Fixation length, Observation length, Observation count

Qualitative metrics

- **Saccade path / gaze plot:** spatial graph that shows the sequence of fixations and saccade movements among them

With which usability objectives can they be associated ?

- With **all the 5Es except for Engaging** (*Be creative !*)

Look up: *eye_metrics_in_TobiiStudio.pdf* (*Eyetracking* folder in CV)

Setting of metrics

Test objectives

Select any combination of objectives from

- the **three standardized ISO usability properties** *or*
- The complete **typified usability goals set** (the *5Es*):

5 Es

ISO

- **Effectiveness**
- **Efficiency**
- **Engaging**

- **Error tolerance**
- **Easy to learn**

REMEMBER:

- *No standardized metrics !* So, *you need to be creative !*
- They can be measured both **qualitative** and **quantitatively**
- They also can be measured using ***eyetracking***

What information is going to be obtained ?

How test results are going to be measured ?

Setting of metrics – *Error tolerance*

Test objectives

Select any combination of objectives from the ISO or **the <u>typified usability goals</u> (the *5Es*)**:

5 Es

- **The three ISO dimensions**
- **Error tolerance:** How well <u>the system **reacts** to errors</u> (user's and/or system's errors), and **helps** the user recover from mistakes

  <u>Prevention</u> + <u>Reaction</u>

  <u>Ex</u>:

  Qual.) **Subjective appreciations of users about Error Tolerance**

  - ***Specific questions*** *in post-task and post-test Questionnaires*
    - *How easily have you been able to **recover** upon an **error/mistake** ?*
    - *How **clear** and **understandable** are the **error messages** for you ?*
    - *Are **corrective actions** available to you when a problem occurs ?*
    - *Have you **understood** the next actions to do when an error occurred ?*

  Quant.) **Quantitative measures**

- **Easy to learn**

Setting of metrics – *Error tolerance*

Test objectives

Select any combination of objectives from the ISO or **the <u>typified usability goals</u> (the *5Es*)**:

- **The three ISO dimensions**

- **Error tolerance:** How well <u>the system **reacts** to errors</u> (user's and/or system's errors), and **helps** the user recover from mistakes

  <u>Prevention</u> + <u>Reaction</u>

  <u>Ex</u>:

  5

  Qual.) **Subjective appreciations of users about Error Tolerance**

  - ***Specific questions*** *in post-task and post-test Questionnaires*

  Es

  Quant.) **Quantitative measures**

  - *<u>Time to recover from errors/mistakes</u>*
  - *<u>Number of assistances</u> after an error*
  - *<u>Ratio</u>: errors corrected / errors committed*

  ET- Visual attention regarding app <u>prevention mechanisms</u> (defining AOI)

  Are they visible ? Are they clear ?

  ET- Visual attention regarding <u>error messages</u>

  Are they visible ? Are they clear ?

- **Easy to learn**

Setting of metrics – *Easy to learn*

Test objectives

Select objectives from the ISO or **the <u>typified usability goals</u> (the *5Es*)**:

5 Es

- **The three ISO dimensions**
- **Error tolerance**
- **Easy to learn:** How well <u>the product **supports** both</u> the **initial orientation** and **continued learning** throughout the complete lifetime of use

<u>Ex</u>:

Qual.) **Subjective appreciations of users about Error Tolerance**

- ***Specific questions*** *in post-task and post-test Questionnaires*
  - *Have you been able to use the app properly **on the 1rst time**, without reading a help manual or asking for assistance ?*
  - *Have your abilities **gradually improved** as you have continued using the app ?*
  - *Has it been easy for you to **remember** the path used to complete a task ?*
  - *Have you detected **some similarities and patterns** while doing the different tasks ?*

Quant.) **Quantitative measures**

Setting of metrics – *Easy to learn*

Test objectives

Select objectives from the ISO or **the typified usability goals (the *5Es*)**:

- **The three ISO dimensions**
- **Error tolerance**
- **Easy to learn:** How well the product **supports** both the **initial orientation** and **continued learning** throughout the complete lifetime of use

Ex:

Qual.) **Subjective appreciations of users about Error Tolerance**

- ***Specific questions*** *in post-task and post-test Questionnaires*

Quant.) **Quantitative measures**

- *Comparing time* (*Be strategic at selecting tasks !*)
- *Reduction in the number of mistakes*

ET- Comparing visual attention –qualitative & quantitatively, overlapping gaze plots- in a certain screen at different task scenarios (*be strategic at selecting tasks !*)

ET- Visual attention regarding particular elements in the UI & app metaphors (defining AOI)

Are they visible ? Are they clear ?

5

Es

Test Plan: Defining objectives - *Satisfaction*

Satisfaction: Absence of discomfort and positive attitude by the user towards the use of the product

Extent to which the user's physical, cognitive and emotional responses that result from the use of an interactive system meet the user's needs and expectations

[9241-11, 08]

User satisfaction has an **outstanding importance**

Composed by **a)** subjective information (**Qual**itative data), **opinions regarding satisfaction**, but also **is quantified with b) questionnaire responses** (**Quant**itative)

a-*Ql)* Listen and **observe** for things that show if users were satisfied with the experience

- ***Qualitative data*** *and* ***Post-Tasks*** questionnaires

b-*Qt)* Ask questions after the test to register in which extend users liked the product

- *Standardized* ***Post-Test*** questionnaires

Satisfaction = ***Qual****itative data* **+** ***Quant****itative data from questionnaires*

Contents

- Setting of metrics for non-standardized usability properties
- Questionnaires

Test Plan – *Overview*

- **Documents** to prepare:
  - Test Plan main document
  - **Annexes**:
    - Screener
    - **Questionnaires and/or interviews**
    - Informed consent form about participation and recording session
    - Observation document
    - List of scenarios for tasks
    - Test script for the moderator
    - Test calendar
    - Additional documents to be used in the test: briefing guide and checklist

Pre-Test Questionnaire

Offers

an additional chance to validate the user profile, as well as his/her capacities

Allows

acquiring additional information of interest regarding the user (products he/she is used to use, in which conditions…)

➔ This information is going to be useful for **posterior data analysis** and **interpretation**

- Use **ranges** for **demographical questions** (gender, age, and other such as income, when necessary)
- As long as you can, *try to categorize answers !* So, prioritize **multiple choice answers** over open questions in this kind of inputs
- Destine **open questions** for letting the user to express some opinions or relate some concrete aspects [Bookify]

Always include **gratitude messages** and also **introduce** the questionnaire (and if necessary the type of information in each **section** too)

Same considerations** as for **screener form

Post-Task Questionnaire

Measuring satisfaction

Allows collecting

- **impressions** and **ratings** regarding each task carried out in the test, and especially **how difficulty has been undertaking each task**
- **relevant comments**

Why to use Post-Task Questionnaires ?

The aim is collecting data regarding **task satisfaction** both, **qualitative** and **quantitatively**

- Collecting **qualitative** data is always helpful
- They provide a **correlation** with *Effectiveness & Efficiency* data results ➔ ***False success***

They have a high **diagnostic power**, especially if accompanied by an **interview**

They help seeking **more detailed diagnosis of problem areas** in a UI than *Post-Test Q.*

- *Different styles:*
  - **Balanced questions** can also be used:

    *"**How difficult or easy** was it to complete the task ?"*

  - **Expectation ratings** Post-Task

    *"How difficult or easy **did you expect** this task to be ?"*
  - Sometimes they are broken down into **two profiles**:
    - Users who have completed the task; and Users who have not completed the task

"In your opinion, what has contributed to the fact that you were unable to complete the task ?"

Post-Task Questionnaire

Measuring satisfaction

- **Customized** or **standard** *Post-Task questionnaires* can be used:
  - **Standand** *Post-Task questionnaires, which provide quantitative results:*
    - ***ASQ** (**A**fter-**S**cenario **Q**uestionnaire)*

Strongly agree ..... Strongly disagree

			1	2	3	4	5	6	7		NA
1	Overall, I am satisfied with the ease of completing the tasks in this scenario.		O	O	O	O	O	O	O		O
2	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario.		O	O	O	O	O	O	O		O
3	Overall, I am satisfied with the support information (online help, messages, documentation) when completing the tasks.		O	O	O	O	O	O	O		O

    - ***SEQ** (**S**ingle **E**ase **Q**uestion)*

Overall, this task was:

Very difficult ○ ○ ○ ○ ○ ○ ○ Very easy

- **They contain** more specific questions than *Post-Test Questionnaires* and more standard than *Pre-test Questionnaire*
- **Consider to include** open questions. Invite users to justify answers [Reciclame] [Commed]

Post-Test Questionnaire

Measuring satisfaction quantitatively

They provide a great chance to inquiry the participant about his/her **general perception** regarding the **usability of the product**

The aim is to deepen in the **'*whys*'** (the reasons) **for participants' actions**

It assesses *Satisfaction* at a relatively **high level**

- **Question to decide:**
  - *Why not to combine (semi-formal tests)*:
    - **Written questionnaire** *(Quant.)*
      - **More agile** as it does not require the moderator presence
      - More **appropriate when** you are interested in
        - Obtaining **a set of standard information**
        - **Comparing** between two or more products (competence) or between different versions of the same product (*formal tests*)
    - **Oral interview** *(Qual.)*
      - More appropriate when it is intended to **deepen in complex and subjective questions**

Post-Test Questionnaire

Measuring satisfaction quantitatively

- **Consistence:**
  - **Graduated scales** (named ***Likert*** scales) are commonly used
    - They allow to **set the level of agreement or disagreement** with a statement

      Ex: *"I think that I would like to use this system frequently"*

      "I found the system unnecessarily complex"

      "I thought the system was easy to use"

      1. Strongly disagree
      2. Disagree
      3. Neither agree nor disagree
      4. Agree
      5. Strongly agree

      Strongly agree ..... Strongly disagree

1	2	3	4	5

    - They allow to realize a **quantitative study** about the level of satisfaction
  - Post-test questionnaires usually include **open questions** too, in order to tackle questions difficult to classify, such as likes, dislikes or general impressions
    - They always **can be added**

Post-Test Questionnaire

Measuring satisfaction quantitatively

Formulate a good questionnaire is not an easy task, and It is subject to misinterpretations and subjectivity

- **Recommendations (I) :**
  - Utilize **standardized opinion questionnaires** in order to avoid **subjectivity**, above all in **formal tests**
    - Standardization is essential for assuming a possible **generalization of the results**
    - They have a **score scheme** established and universally accepted (based on *Likert scales*)
  - In case of not using a **standardized questionnaire:** [Commed]
    - Focus questions on the **test purpose and objectives**
    - **Ensure** that each question is focused on **an only aspect** or problem
    - Invite participants to **justify their** answers (enable extra space to this kind of comments)
    - Realize a **pilot test** before using it, aiming to remove ambiguities

Post-Test Questionnaire

Measuring satisfaction quantitatively

- **Recommendations (II) :**
  - Realize **questionnaires in digital** (e.g. *Morae*, *Tobii Studio*, *Google Sheets*), as they streamline data treatment after the test
  - **Avoid** excessive extension and difficulty to understand questionnaires

Common questions in a post-test questionnaire:

- *Is the **design** (look and feel) of the user interface engaging ?*
- *Is the **amount of information** handled appropriate, excessive or scarce ?*
- *Is the **terminology** clear and easy to understand ?*
- *Are the **icons** easy to understand ?*
- *Are the **buttons and controls** easy to access and to operate ?*
- *Is **navigation** easy to understand ?*
- *Are **instructions** easy to follow ?*
- *Are **error messages** clear and easy to understand ?*

Post-Test Questionnaire

Measuring satisfaction quantitatively

- **Recommendations** for posterior analysis:
  - **Pay special attention** to possible reasons for **extreme positioning** in questionnaire answers (*high or low satisfaction*)
  - **The center point of the scale** serves when participants feel that they cannot respond to a particular item
  - **Analyze open questions answers** where users can justify their answers; incite them to fill in
  - **Check for a possible correlation** between the level of satisfaction expressed by participants on questionnaires and the number of successes and failures ?

    "Do you think you could have completed the tasks without any help ?"

➔ ***"False success"***

- **Websites of tools** to measure satisfaction:

  http://measuringuserexperience.com/Satisfaction/index.htm

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

Some of the most used questionnaires:

- **SUMI** (**S**oftware **U**sability **M**easurement **I**nventory) [University College Cork, 96]
  - **Standardized** by International Organization for Standardization (**ISO 9241**)
  - **Commercial** and currently available in 12 languages
  - 50 questions; *likert* scale of 3 (*Agree*, *Don't Know*, or *Disagree*)
  - It contains a mixture of positive and negative statements

Statements 1–10 of 50.	**Agree**	**Undecided**	**Disagree**
This software responds too slowly to inputs.	○	○	○

  - The SUMI considers five subscales:
    - » Efficiency
    - » Affect
    - » Helpfulness
    - » Control
    - » Learnability
  - A powerful feature of the SUMI:
    - » A **normative database** with which to compare results with similar products

http://sumi.uxp.ie/

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

Some of the most used questionnaires:

SUMI derivates:

- **MUMMS** (**M**easuring the **U**sability of **M**ulti-**M**edia **S**ystems)
  - Extension of SUMI focused on **multimedia** products
  - The MUMMS items produce five subscales

  https://www.sciencedirect.com/topics/computer-science/usability-questionnaire

- **WAMMI** (**W**ebsite **A**nalysis and **M**easurement **I**nventory)
  - Specialization of SUMI focused on **web** usability (web analytics service)
  - **Commercial** (visible online) and available in diverse languages
  - 20 questions; *likert* scale of 5
  - *It allows adding questions !*
  - It generates an easy-to-read hypertext report using qualitative and quantitative data, and also benchmarks websites

  http://www.wammi.com/

  http://www.allaboutux.org/wammi-website-analysis-and-measurement-inventory

SUMI and **WAMMI** have their own **scoring software** and **standard report**

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

Some of the most used questionnaires :

- **PSSUQ** (**P**ost **S**tudy **S**ystem **U**sability **Q**uestionnaire) [IBM, 1990]
  - ➔ **Free** and **available on the web** (*not required any license, only adding credits*)
  - Developed in IBM for **scenario-based usability assessment**
  - 19 questions; *likert* scale of 7
  - The PSSUQ items produce four scores (1 overall and 3 subscales):
    - » System quality
    - » Information quality
    - » Interface quality
  - **Flexibility** in its usage:
    - » practitioners can add items to the questionnaires if there is a need, or, to a limited extent, can remove items that do not make sense in a specific context

https://www.conetrees.com/ux-glossary/post-study-system-usability-questionnaire-pssuq/

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

Some of the most used questionnaires :

- **CSUQ** (**C**omputer **S**ystem **U**sability **Q**uestionnaire) [IBM, 1995]
  - **A variant** of the **PSSUQ**
  - ➔ **Free** and **available on the web** (*pdf* and *html*)
  - 19 questions; *likert* scale of 7, together with 2 open questions

  http://garyperlman.com/quest/quest.cgi?

- **QUIS** (**Q**uestionnaire for **U**ser **I**nterface **S**atisfaction) [University of Maryland, 88]
  - **Commercial** (requires a license) and available in 5 languages
    - » Available in https://garyperlman.com/quest/quest.cgi?form=QUIS
  - 27 questions; *likert* scale of 9, together with 2 open questions
  - Consists of 5 different sections

  http://www.cs.umd.edu/hcil/quis/

5.4 Messages which appear on screen:

	1	2	3	4	5	6	7	8	9		NA
Confusing	○	○	○	○	○	○	○	○	○	Clear	○

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

Some of the most popular and extended questionnaires

Recent research indicates that although it is quick, it is far from "dirty"

- **SUS** (**S**ystem **U**sability **S**cale) [A "*quick and dirty*" usability scale] [J. Brooke ,1986]
  - – **Low-cost** and **easy**, but also **valid**, **robust** and **reliable**
  - ➔ **Free** (*only add credits*) and considered an **industrial** and ***de facto* standard**
    - » with references in over 1300 publications, 5000 users and 500 different studies
  - – **10 questions**; *likert* scale of 5; scores a **range of 0 to 100**
    - » the odd-numbered items have a positive tone; the even's ones negative
      - » This alternation is intended **to avoid response biases**, especially as the questionnaire **invites to rapid responses**

– Provides a **general indication** of the **overall level of usability** (the perceived usability) in comparison to its competitors or predecessors

– **SUS scores** are not percentages. They must be interpreted as a **percentile rank**

https://measuringu.com/sus/

Different Post-Test Questionnaires

Measuring satisfaction quantitatively

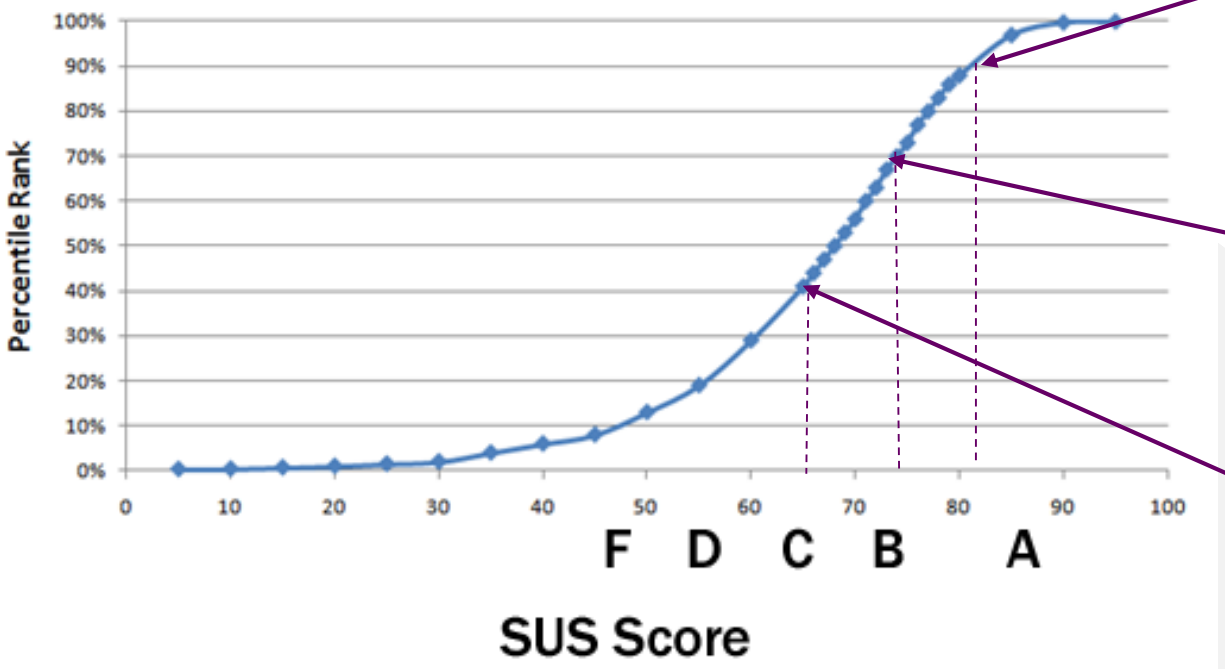
Some of the most popular and extended questionnaires

Recent research indicates that although it is quick, it is far from "dirty"

- **SUS** (**S**ystem **U**sability **S**cale) [A "*quick and dirty*" usability scale] [J. Brooke ,1986]

  To understand **how** a product **compares to others**, it is required to look at its **percentile ranking**

  SUS scores must be interpreted as **percentile ranks** (*normalizing* process)



Above an 80.3 gets an A *letter-grade* (the top 10% of scores)

A **SUS score of a 74** converts to a percentile rank of 70% (above the 70% of all products tested) -a *letter-grade* of a B

A **SUS score above a 68** converts to a percentile rank of 40% (a perceived usability below average) -a *letter-grade* of a C

What to do with all the data ?

Steps to cover just completed the test

- **Compilation** of registered data
  - Complete videos of users (RDG files from *Morae Recorder*)
  - Informed consent form
  - Questionnaires
  - Observation document (*notes from observers*)
- **Check** fulfillment of quantitative and qualitative established goals
- **Solve possible doubts** between members of group
- **Do a first judgement** about what has happen during the test

Data collecting

How to register some of these metrics ?

- The specific software for usability testing allows to define a set of **codes** ("*markers*") to collect the different users' behaviors

Example of Document for registration and observation:

The ***Observation document***

Test de Usabilidad de un programa de correo electrónico									
#participante:		M = error de menú						O = online help	
Fecha:		S = selección de la lista de errores						H = desk help	
Recorder:		E = otros errores						F = frustración	
Tarea	**Tiempo**	**M**	**S**	**E**	**O**	**H**	**F**	**Comentarios participantes**	**Notas**
Tarea 1	Inicio: Fin:								