Multimedia Retrieval
2018-2019
Evaluation, The user

Egon L. van den Broek
Accompanying literature

- The book’s chapters 10, 20, and 21
Evaluation, part II: The user, in three parts

*Introduction*

1. Performance metrics
2. Self-reported metrics (user perception)
3. Physiological metrics
INTRODUCTION
<table>
<thead>
<tr>
<th>Technique</th>
<th>Good for</th>
<th>Kind of data*</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviews</td>
<td>Exploring issues</td>
<td>Some quantitative but mostly qualitative</td>
<td>Interviewer can guide interviewee if necessary.</td>
<td>Time-consuming. Artificial environment may intimidate interviewee</td>
</tr>
<tr>
<td>Focus groups</td>
<td>Collecting multiple viewpoints</td>
<td>Some quantitative but mostly qualitative</td>
<td>Highlights areas of consensus and conflict.</td>
<td>Possibility of dominant characters</td>
</tr>
<tr>
<td>Questionnaires</td>
<td>Answering specific questions</td>
<td>Quantitative and qualitative</td>
<td>Can reach many people with low resource</td>
<td>The design is crucial. Response rate may be low. Responses may not be what you want</td>
</tr>
<tr>
<td>Direct observation in the field</td>
<td>Understanding context of user activity</td>
<td>Mostly qualitative</td>
<td>Observing actual work gives insights that other techniques can't give</td>
<td>Very time-consuming. Huge amounts of data</td>
</tr>
<tr>
<td>Direct observation in a controlled environment</td>
<td>Capturing the detail of what individuals do</td>
<td>Quantitative and qualitative</td>
<td>Can focus on the details of a task without interruption</td>
<td>Results may have limited use in the normal environment because the conditions were artificial</td>
</tr>
<tr>
<td>Indirect observation</td>
<td>Observing users without disturbing their activity; data captured automatically</td>
<td>Quantitative (logging) and qualitative (diary)</td>
<td>User doesn't get distracted by the data gathering; automatic recording means that it can extend over long periods of time</td>
<td>A large amount of quantitative data needs tool support to analyze (logging); participants’ memories may exaggerate (diary)</td>
</tr>
</tbody>
</table>

* For a discussion of qualitative and quantitative see Section 8.2

**Table 7.1 Overview of data gathering techniques and their use**
Usability criteria

Specific criteria to assess the usability of your system, via measurement of user performance. For example,

• Efficiency: Time needed to complete a task
• Learnability: Time needed to learn a task
• Memorability: Number of errors made (after a while)
Criteria for User eXperience (UX)

“User eXperience refers to all aspects of someone’s interaction with a product, application, or system” (Tullis & Albert, 2013) For example,

• How many errors do users make in trying to log onto a retrieval system?
• How many users get frustrated trying to read the advanced settings file trying to get the system working the way they want?
Usability Metrics

• Ways of measuring/evaluating the UX
• Reveal *something* about the UX; but,
  – What to measure?
  – When to measure?
  – How to gather, analyze and interpret data
• Three main types of usability metrics:
  – Performance metrics
  – Self-reported metrics (user perception)
  – Behavioural and physiological metrics
Measurements for Usability / UX

• Performance
e.g., criteria that explicitly measure effectiveness, efficiency, and learnability

• Perceived experience
e.g., satisfaction, expectation, perceived ease of use, perceived usefulness, awareness, and pleasure

• Physiology and behaviour
e.g. eye-movement, neural activity, facial expressions, and stress
Validity

Whether an instrument actually measures what it sets out to measure

- *Construct validity*: the degree to which a measure relates to other variables as expected within a system of theoretical relationships
- *Content validity*: the degree to which a measure corresponds to the content of the construct it was designed to cover
- *Criterion validity*: evidence that scores from an instrument correspond with concurrent external measures conceptually related to the measured construct
- *Ecological validity*: evidence that the results of a study can be applied to real-world conditions
Reliability

whether an instrument can be interpreted consistently across different situations
PERFORMANCE METRICS
Performance metrics

• **task-success** (measures effectiveness, efficiency, ...)
• **time-on-task** (measures efficiency, learnability...)
• **steps-to-completion** (measures efficiency, ...)
• **efficiency** (measures efficiency, ...)
• **lostness** (measures efficiency, ...)
• **errors**
Performance metrics: task-success

• “how effectively are users able to complete a given set of tasks?”

• clear end-state?
  – Find the best live performance of The Prodigy.
  versus
  – List all possible lists of courses to complete your MSc.
(Levels of) task sucesso

- Complete sucesso
  - With/Without assistance

- Partial sucesso
  - With/Without assistance

- Failure
  - Participant thought it was complete, but it wasn’t
  - Participant gave up
Table 2.3 Choosing the Right Statistics for Different Data Types and Usability Metrics

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Common Metrics</th>
<th>Statistical Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal (categories)</td>
<td>Task success (binary), errors (binary), top-2-box scores</td>
<td>Frequencies, crosstabs, Chi-square</td>
</tr>
<tr>
<td>Ordinal (ranks)</td>
<td>Severity ratings, rankings (designs)</td>
<td>Frequencies, crosstabs, chi-square, Wilcoxon rank sum tests, Spearman rank correlation</td>
</tr>
<tr>
<td>Interval</td>
<td>Likert scale data, SUS scores</td>
<td>All descriptive statistics, <em>t</em>-tests, ANOVAs, correlation, regression analysis</td>
</tr>
<tr>
<td>Ratio</td>
<td>Completion time, time (visual attention), average task success (aggregated)</td>
<td>All descriptive statistics (including geometric means), <em>t</em>-tests, ANOVAs, correlation, regression analysis</td>
</tr>
</tbody>
</table>
Measuring efficiency

• the amount of effort that a user expends to find what he searches for

• two types of effort
  – Cognitive effort – involves finding the right place to perform an action (e.g. finding a link on a web page), deciding what action is necessary (should I click on this link?), and interpreting the results of the action
  – Physical effort – involves the physical effort required to take an action

• simple and compound measures
Measuring efficiency

• Simple metrics
  – time-on-task
  – steps-to-completion
    (number of steps or actions to complete a task)

• Compound metrics
  – efficiency
  – lostness
Performance metrics: \textit{time-on-task}

- How much time is required to complete a task?
- How to measure?
- All tasks? Only successful tasks?

The faster a participant can complete a task, the better the UX?

- hotel /airplane ticket reservation
- searching nice games
Performance metrics: efficiency

- Compound efficiency metric:
  - combination of task success and time-on-task
  - typically measured per task; alternative = per participant

\[
\text{efficiency} = \frac{\text{task completion rate}}{\text{mean time per task}}
\]

or, alternatively:

\[
\text{efficiency} = \frac{\text{number of successfully completed tasks}}{\text{total time spent}}
\]
Performance metrics: \textit{lostness}

A compound efficiency metric (Smith, 1996)

\[
L = \sqrt{(N/S - 1)^2 + (R/N - 1)^2}
\]

\textit{N}: The number of \textit{different} web pages visited while performing the task.

\textit{S}: The \textit{total} number of pages visited while performing the task, counting revisits to the same page.

\textit{R}: The \textit{minimum} (optimum) number of pages that must be visited to accomplish the task.
• example
Measuring learnability

• measure how performance changes over time
• (how any efficiency metric changes over time)
• how much time and effort is required to become proficient using the product or application
• collecting data multiple times (trials)
• within-subjects design
Performance metrics: **time-on-task**

multiple trials for single subject (same task); gives ‘learning curve’

**FIGURE 4.13**
An example of how to present learnability data based on time-on-task.
Severity ratings of usability issues

A combination of:
• frequency
• impact
• persistence

0 = I don't agree that this is a usability problem at all.
1 = Cosmetic problem only: need not be fixed unless extra time is available on project.
2 = Minor usability problem: fixing this should be given low priority.
3 = Major usability problem: important to fix, so should be given high priority.
4 = Usability catastrophe: imperative to fix this before product can be released.
SELF-REPORTED METRICS
Self-reported metrics

• What to measure?
• How to measure?
  – Single-item formats
  – Multiple-item formats: indexes and scales: *general and usability-specific*
• Gathering self-reported data
  – Pre/Post-task
  – Pre/Post-test
• Analyzing self-reported data
What to measure?

• Characteristics (e.g., age, level of education, gender, and occupation)
• Attitudes: What people say they want
• Beliefs: What people think is true.
• Behaviors: What people say they do or what you observe they do.
Single-item formats

I think that I would like to use this system frequently:

___ Strongly Disagree
___ Disagree
___ Neither agree not disagree
___ Agree
___ Strongly Agree

e.g., well-known ‘Likert scale’*

* in fact a misnomer: it’s not a scale but a well-known question format
Guidelines single-items formats

• **Avoid "acquiescence bias":** people are more likely to agree with a statement than to disagree with it (Cronbach, 1946)
  – You need to balance positively-phrased statements (such as "I found this interface easy to use") with negative ones (such as "I found this interface difficult to navigate").

• **Use 5-9 levels in a rating**
  – You gain no additional information by having more than 10 levels

• **Include a neutral point in the middle of the scale**
  – Otherwise you lose information by forcing some participants to take sides

• **Don’t use numbers, but if so: use positive integers**
  – 1-7 instead of -3 to +3
    (Participants are less likely to go below 0 than they are to use 1-3)

• **Use word labels for at least the end points.**
  – Hard to create labels for every point beyond 5 levels
  – Having labels on the end points only also makes the data more “interval-like”
Models with validated usability scales: SUS

- System Usability Scale
- “A quick and dirty usability scale”

System Usability Scale


1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system
**System Usability Scale (SUS)**

1. I think that I would like to use this system frequently

2. I found the system unnecessarily complex

3. I thought the system was easy to use

4. I think that I would need the support of a technical person to be able to use this system

5. I found the various functions in this system were well integrated

6. I thought there was too much inconsistency in this system

7. I would imagine that most people would learn to use this system very quickly

8. I found the system very cumbersome to use

9. I felt very confident using the system

10. I needed to learn a lot of things before I could get going with this system

**Total score = 22**

SUS Score = 22 * 2.5 = 55
Measuring expectations:
Pre- and Post-Task Ratings

• Before the task:

How easy or difficult do you expect this task to be?

<table>
<thead>
<tr>
<th>Very easy</th>
<th>Very difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

• After the task:

How easy or difficult was this task to do?

<table>
<thead>
<tr>
<th>Very easy</th>
<th>Very difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Pre/Post- task ratings versus Pre/Post-session ratings

• task-level data: help identify areas that need improvement
  (quick ratings immediately after each task help pinpoint tasks and interface parts that are particularly problematic)

• session-level data: help to get a sense of overall usability
  (effective overall evaluation after each participant has had a chance to interact with the product more fully)
Post-session ratings: Examples

- Software Usability Scale (SUS) – 10 ratings
- Usefulness, Satisfaction, and Ease of use (USE)
- Questionnaire for User-Interface Satisfaction – QUIS * - 71 (long form), 26 (short form) ratings
- Software Usability Measurement Inventory (SUMI) * – 50 ratings
- After Scenario Questionnaire (ASQ) – three ratings
- Post Study System Usability Questionnaire (PSSOQ) - 19 ratings. Electronic version called the Computer System Usability Questionnaire (CSUQ)
- Website Analysis and MeasureMent Inventory (WAMMI) * – 20 ratings of website usability
- Computer System Usability Questionnaire (CSUQ)

* requires a license
PHYSIOLOGICAL METRICS
Physiological and behavioural metrics

- Verbal behaviours
  - Comments
  - Questions
  - Utterance of confusion / frustration

- Nonverbal behaviours
  - Facial expressions
  - Eye behaviour
  - Skin conductance
  - Heart rate
  - Blood flow
  - Temperature
  - Sleep / wake
Usability Test Observation Coding Form

Date: ___________________  Participant ID: ___________________  Task #: ___________________

Start Time: _______________  End Time: ___________________

**Verbal Behaviors**

- **Strongly positive comment**
- **Other positive comment**
- **Strongly negative comment**
- **Other negative comment**
- **Suggestion for improvement**
- **Question**
- **Variation from expectation**
- **Stated confusion**
- **Stated frustration**

**Non-verbal Behaviors**

- **Frowning/Grimacing/Unhappy**
- **Smiling/Laughing/Happy**
- **Surprised/Unexpected**
- **Furrowed brow/Concentration**
- **Evidence of Impatience**
- **Leaning in close to screen**
- **Variation from expectation**
- **Fidgeting in chair**
- **Random mouse movement**
- **Groaning/Deep sigh**
- **Rubbing head/eyes/neck**

**Task Completion Status:**

- **Incomplete:**
  - Participant gave up
  - Task “called” by moderator
  - Thought complete, but not

- **Complete:**
  - Fully complete
  - Complete with assistance
  - Partial completion

Notes:

---

Verbal Behaviors Notes

---

Non-verbal Behaviors Notes

---

Task Completion Status Notes:
Measuring physiological signals: equipment
Facial expressions

- Video-based systems
- Electromyogram sensors
pupils
Eye tracking (measuring attention)

Faces draw attention to them on webpages

Study 1: users are clearly drawn to faces when asked to look at pages and report what they remember

Are People Drawn to Faces on Webpages? – T. Tullis, M. Siegel & M. Sun In: CHI 2009, Boston, MA, USA.
Eye tracking (task-performance)

Study 2:
- a Portfolio Summary page was modified to contain either a photo of a woman’s face or no image
- tasks that had answers that could be found by reading information on the page
Eye tracking and task-performance

Study 2:
Contrary to expectation, a picture of a face in this context actually caused users to do worse on a task involving information adjacent to the face.
Thermal Imaging (measuring stress)

- Thermal imaging of the face
- Stresscam: a small thermal imaging camera
Thermal Imaging (measuring stress)

User stress is correlated with increased blood flow in the frontal vessel of the forehead. This increased blood flow dissipates convective heat.