Evaluation of Interactive Systems

Questionnaires

Caroline Appert - 2018/2019
Presentations

Make sure your name is in a group for presenting a research article
Questionnaires

Oral interviews

Questionnaires and surveys
Oral Interviews

Types of questions

If you ask too broad questions out of context, users will get confused

Conduct users’ interviews in their work environments

Capture users’ needs

Make them describe their typical tasks (do not hesitate to ask for details, _they_ are the experts)

Capture usability issues

Use the Critical Incident Technique
Oral Interviews
Critical Incident Technique

1/ Set the scope of your interview by asking the interviewee to report on a critical incident

2/ Collect details

Advantages
- Discount evaluation
- High-level of details

Disadvantages
- Not representative: only memorizable incidents

Tell me about the last big problem you had with Word
Oral Interviews
Critical Incident Technique

Ideally, critical incidents are real incidents reported by actual users doing real activities in their normal environments.

Define the activity you intend to study. Get access to the users as soon as possible after the activity has finished, and if possible in the same environment.

Oral Interviews

Focus groups

Group of 5 and 10 users who work with a moderator.

The moderator leads the group to make the discussion focused on a beforehand-prepared list of topics/issues (but must not intervene too much in order to keep the discussion fluid, and make users comfortable).

The collective view becomes established which is greater than the individual parts.
Oral Interviews

Log collections

Note taking

+ recorded information is filtered out (analysis is facilitated)
- recorded information is limited (low bandwidth, interviewer’s attention reduced if there is no additional note taking person)

Audio and video recording

+ No loss of information
- Difficulty for analyzing information (too much data!)
Questionnaires and Surveys

Preparation “expensive”, but administration “low-cost”

Can reach a wide group of users

  e.g., mail, email, web

Does not require presence of evaluator

Results can be quantified

**Warning:** Only as good as the questions asked
Questionnaires and Surveys

Do not waste participants’ time

- establish the purpose of the questionnaire to make it as short as possible. Do not ask questions if you don’t actually plan to analyze provided answers!
- **pre-test the questionnaire** before it is used in a full-scale survey

Make the collected information useful

- how will you analyze the results (statistics)?
- what will you do with your analysis (conclusions)?

Include demographic questions

- e.g., age, gender, education level, profession
Determine the audience you want to reach and the way you will deliver the questionnaire

web site with forms (targets computer users)

useful support: Google forms, SurveyMonkey

surface mail (pre-addressed reply envelope gives far better response)

telephone
Open-ended questions

Ask for unprompted opinions

Good for general subjective information

Difficult to analyze rigorously

“Can you suggest any improvement to the user interface?”

“Sketch the best symbol to depict command X?”

Keep in mind: Do not ask questions if you don’t actually plan to analyze provided answers!
Closed questions

Restrict participants’ responses by supplying alternative answers

Make questionnaire an easy task for participants to fill in

Can be easily analyzed

e.g., “Which internet browser do you usually use? (Apple Safari, Google Chrome, Firefox, Internet Explorer, ...)”
Closed questions

Watch out for hard to interpret responses
Alternative answers should be very specific

“Do you use computers at work:”

- [ ] often
- [ ] sometimes
- [ ] rarely

vs

“In your typical work day, do you use computers:“

- [ ] over 4 hrs a day
- [ ] between 2 and 4 hrs daily
- [ ] between 1 and 2 hrs daily
- [ ] less than 1 hr a day
Closed questions: scalar

Likert-type scale

participants specify their level of agreement or disagreement on a symmetric agree-disagree scale for a statement

e.g., “Characters on the computer screen are:”

hard to read easy to read

1 2 3 4 5

Note: in order to use statistics like median or mode, a Likert scale must be interval (equal distance between two consecutive items). A distribution is usually a more appropriate analysis tool.
Closed questions: multi-choice

Offer a choice of explicit responses

e.g. “How do you most often get help with the system? (tick one)”
- on-line manual
- paper manual
- ask a colleague

e.g. “Which types of software have you used? (tick all that apply)”
- word processor
- data base
- spreadsheet
- compiler
Closed questions: ranked

Place an ordering on items in a list

e.g., “Rank the usefulness of these methods of issuing a command (1 most useful, 2 next most useful..., 0 if not used)”

__2__ command line
__1__ menu selection
__3__ control key accelerator
Combining opened and closed questions

Get specific response, but allow room for users’ opinion

e.g. “It is easy to recover from mistakes:"

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>agree</td>
<td></td>
<td></td>
<td></td>
<td>❋</td>
<td></td>
</tr>
</tbody>
</table>

comment: the undo facility is really helpful
NASA TLX QUESTIONNAIRE

For assessing subjective workload

By using a multi-dimensional rating procedure that derives an **overall workload score** based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration
Interviewees are provided with the _written_ definition of scales that will be used in the questionnaire.

<table>
<thead>
<tr>
<th>Title</th>
<th>Endpoints</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MENTAL DEMAND</td>
<td>Low/High</td>
<td>How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?</td>
</tr>
<tr>
<td>PHYSICAL DEMAND</td>
<td>Low/High</td>
<td>How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?</td>
</tr>
<tr>
<td>TEMPORAL DEMAND</td>
<td>Low/High</td>
<td>How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?</td>
</tr>
<tr>
<td>PERFORMANCE</td>
<td>Good/Poor</td>
<td>How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?</td>
</tr>
<tr>
<td>EFFORT</td>
<td>Low/High</td>
<td>How hard did you have to work (mentally and physically) to accomplish your level of performance?</td>
</tr>
<tr>
<td>FRUSTRATION LEVEL</td>
<td>Low/High</td>
<td>How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?</td>
</tr>
</tbody>
</table>
## NASA TLX Questionnaire

The questionnaire is then a two-part evaluation procedure.

**Part 1** - The rater indicates the magnitude for each scale after having read the definition.

<table>
<thead>
<tr>
<th>Name</th>
<th>Task</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand</td>
<td>How mentally demanding was the task?</td>
<td></td>
</tr>
<tr>
<td>Physical Demand</td>
<td>How physically demanding was the task?</td>
<td></td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>How hurried or rushed was the pace of the task?</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>How successful were you in accomplishing what you were asked to do?</td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>How hard did you have to work to accomplish your level of performance?</td>
<td></td>
</tr>
<tr>
<td>Frustration</td>
<td>How insecure, discouraged, irritated, stressed, and annoyed were you?</td>
<td></td>
</tr>
</tbody>
</table>

The NASA Task Load Index (TLX) method assesses workload on five 7-point scales. Increments of high, medium, and low estimates for each point result in 21 gradations on the scales.

- **Very Low**
- **Very High**
- **Failure**
- **Perfect**

The questionnaire is a two-part evaluation procedure. Part 1 involves the rater indicating the magnitude for each scale after reading the definitions.
Part 2 - Assess the right definition of workload for the raters (what scales are more important than others to them).

Evaluator: “Circle the scale title that represents the more important contributor to workload for the specific task you performed in this experiment”

<table>
<thead>
<tr>
<th>Effort or Performance</th>
<th>Temporal Demand or Frustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Demand or Effort</td>
<td>Physical Demand or Frustration</td>
</tr>
</tbody>
</table>

15 pairwise comparisons
weight of a scale = number of times it has been circled
The rater circled “Mental Demand” 3 times.

The rater marked 12 on the Mental Demand scale.
The workload score can be graphically represented.

A bar width represents the weight of a scale (its relative importance in the overall workload), and a bar height represents how much this scale is “bad.”
Questionnaires

Take-away messages

Relatively low-cost and easy to analyze

Be careful about users’ interpretation

Refine wording and choose closed questions when possible

Constantly refer to the original research/usability questions