Qualitative Evaluation Techniques

Qualitative Approach

– Allows you to quickly evaluate certain things by observing people’s use of them and/or asking them questions.
– Tends to produce a description and/or set of anecdotes and usually in non-numeric terms.
– Can have issues with being subjective.
– Different methods might be selected based on the scenario/goals to help you discover what participants are thinking about.
Qualitative Methods for UI evaluation

Categories of Methods

– Query
  • interviews (structured and retrospective)
  • surveys and questionnaires
– Introspection
  • by designer
  • by users
– Direct observation
  • simple observation
  • think-aloud
  • constructive interaction

Qualitative Evaluation: Querying
Querying Users: Interviews

Can be very good for pursuing specific issues.
– vary questions to suit the context
– probe more deeply on interesting issues as they arise
– good for exploratory studies via open-ended questioning
– can lead to specific constructive suggestions

Potential problems with this include the fact that accounts are subjective, the gathering and coding of the data is going to be time consuming, an evaluator can easily bias the interview, and answers are prone to rationalization of events/thoughts by user (the user’s recollection or reconstruction could be wrong as a result of this).

Querying Users: Structured Interviews

Design a set of central questions
– could be based on results of user observations
– gets things started
– focuses the interview
– ensures a base of consistency

Try not to ask leading questions!
– “Now that was easy, wasn’t it?”
– “How hard would you say this task was?”

Might start with individual discussions to discover different perspectives, and then continue with group discussions. With groups you might learn more, especially if it encourages discussion between users.
Querying Users: Retrospective

With user studies in particular, after a participant has interacted with a system the researcher can interview them to clarify events that occurred during system use.

– perform an observational test and video-record it
  • there are differences between assigning users tasks and allowing them to define their own.

– have the participant then view the video with you and comment on what they did and why
  • watching the video helps ground a post-test interview and helps avoid erroneous reconstruction
  • the users will often offer concrete suggestions

Querying Users: Surveys/Questionnaires
Surveys/Questionnaires Overview

Preparing them can be “expensive” due to human-time costs, but the administration might be fairly “cheap” since you don’t need an evaluator present (or one evaluator can oversee a large group).

Need to consider how to reach your desired subject group; e-mail lists, postal mail, social media, etc. and be prepared for a low return rate (what’s in it for them?) or a biased sample (who will see it? who will take the time to answer?)

Even with qualitative questions, the results can be quantified. [http://www.jstor.org/stable/1177100](http://www.jstor.org/stable/1177100)

Surveys/Questionnaires Design Issues

As is the case with many things, a survey is only as good as the questions asked!!!

Surveys often have a low return rate (what’s in it for them?) or a biased sample (who will take the time to answer?)

Writing a full set of well-formed questions can be challenging so resources exist to help, such as QUIS (Questionnaire for User Interface Satisfaction).
- About the approach [http://www.lap.umd.edu/quis/](http://www.lap.umd.edu/quis/)
Surveys and Questionnaires Details (I)

Establish the *purpose* of the questionnaire.
- what information is sought?
- how would you analyze the results?
- what would you do with your analysis?

With usability (for example) you typically do not ask questions whose answers you will not use.
- this is unlike many other types of surveys and introduces a variety of issues

Surveys and Questionnaires Details (II)

Determine the *audience* and *demographic* you want to reach.
- typical survey might desire a random sample of anywhere between 50 and 1000 users of the product
- level of experience, age, income, etc.

Determine how would you will deliver and collect the questionnaire.
- dedicated website? page within survey-specific site? social media site?
- pay-to-answer model? Amazon Turk? Swag Bucks?
- surface mail (with pre-addressed reply envelope for better response rate)
Styles of Questions: Open Ended

*Open-ended* questions which can ask for “unprompted” opinions can be good for gathering general subjective information, but can be difficult to analyze rigorously.

“Can you suggest any improvements to the interfaces?”

Styles of Questions: Closed

*Closed* questions has the effect of restricting the respondent’s responses by supplying answer options which makes it easier to analyze, but can also make questionnaires a chore for respondent to fill in. When supplying the possible answer, you don’t want responses that end up being difficult to interpret.

Do you use computers at work:

- O often
- O sometimes
- O rarely

- vs -

In your typical work day, do you use computers:

- O over 4 hours
- O between 2 and 4 hours
- O between 1 and 2 hours
- O less than 1 hour
Bipolar Scaling

_Bipolar scaling_ is a closed approach that asks the participant to judge a specific statement on a numeric scale which usually assumes a linear scaling of feelings and corresponds with agreement or disagreement with a statement.

Characters on the computer screen are:

hard to read  1 2 3 4 5 easy to read

Using a scale of 1 to 7 or 1 to 9 might provide better results than 1 to 5 since they will still provide a good range even if the user eliminates the extremes.

Likert Scales

Bipolar scaling is sometimes done explicitly making use of what are called Likert-type scales such as:

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

A scale which is **even** in length should be used if you want to prevent the user from being neutral.
Multiple Choice

With multiple choice (and possibly multiple response) questions, the respondent is offered a choice of closed and very specific responses.

How do you most often get help with the system? (tick one)
- [ ] on-line manual
- [ ] paper manual
- [ ] ask a colleague

Which types of software have you used? (tick all that apply)
- [ ] word processor
- [ ] database
- [x] spreadsheet
- [ ] compiler

Ranked

You can ask your participant to rank items in a list by placing an ordering on them.
- This can be useful to indicate a user’s preferences
- However, this is forcing them to choose an order and typically does not allow ties.

Rank the usefulness of these methods of issuing a command where 1 is most useful, 2 is next most useful..., 0 if not used.

[2] context menu selection
[1] drop-down menu selection
[3] control keys
Hybrid

Combining *open-ended* and *closed* question styles can get specific responses while also allowing room for users to express a more general opinion.

It is easy to recover from mistakes:

| disagree | 1 | 2 | 3 | agree | 4 | 5 | comment: the undo facility is really helpful |

--

**Reporting Results**

There are a variety of logistical and ethical questions surrounding qualitative data from surveys. Some are:

– How to display it (bar graph rather than histogram since it’s not continuous).
– How to summarize it (median/mode rather than mean for linear scale results).
– Whether you can/should generalize from it.
Introspection Method: Designer

Typically used with interface design. A design team member tries the system (or prototype) out (doing a walkthrough of the systems screens and features).

- They are looking to determine whether the system “feels right” when being used.
- Is probably still the most common evaluation method…

Potential problems are reliability issues since:

- it is completely subjective
- the “introspector” is a non-typical user
- being so close to the project your intuitions and introspection are often biased and thus wrong…
**Introspection Method: User**

Typically done as a user-centered walkthrough of a system. The idea here is typically one of conceptual model extraction by showing representative users prototypes or even screenshots of a mock-up.

- Can ask the user to explain what each screen element does or represents as well as how they would attempt to perform individual tasks.

This can allow us to gain insight as to a user’s initial perception of our interface and the mental model they might be constructing as they begin to use a system.

NOTE: Since we’re walking them through specific parts as their guide, we won’t really see how a user might explore the system on their own or their learning processes.

---

**Qualitative Evaluation: Direct Observation**
Direct Observation

The evaluator(s) observe and record users interacting with a design/system, either in a lab setting or “field” setting.

– When done in a lab the user is typically asked to complete a set of pre-determined tasks and it might be done in a special instrumented usability lab to facilitate recording.
– When done “in the field” the user might be asked to go through their normal routine, or if they are asked to complete a set of tasks, they are at least doing it in natural setting.

While this can be excellent at identifying gross design/interface problems, the validity and reliability depends on how controlled and/or contrived the situation is...

Direct Observation Approaches

Typically utilized in software design, and there are three general approaches that can be used for direct observations:

– simple observation
– think-aloud
– constructive interaction
Simple Observation

The user is given the task(s) to perform and the evaluator(s) simply watch (and possibly record) what the user does.

Potential problem:
- it is quite possible this does not provide any insight into the user’s decision process or their attitude/feelings while performing the tasks

Think Aloud

A similar setup to simple observation, but the users are asked to say what they are thinking/doing during the tasks.
- what they believe is happening
- what they are trying to do
- why they took an action

This can give insights into what the user is thinking, but there are potential problems
- can be awkward/uncomfortable for subject (thinking aloud is not natural when working alone)
- “thinking” about why they are doing things could alter the way people perform their task
- hard to talk when they are concentrating on problem
Constructive Interaction

Similar to the other two, but here two people work together on the task(s).

- This can lead to a normal conversation between the two users which can then be monitored.
- It should remove the awkwardness of think-aloud but might be less realistic depending on the tasks.

Co-Discovery

A **variant** of constructive interaction is to have co-discovery learning take place, where the pair working together are:

- a semi-knowledgeable “coach”
- a beginner (who is actually using the system)

Ideally, this results in

- the “ naïve” beginner participant asking questions
- the semi-knowledgeable “coach” responding
- insights into thinking process of both beginner and intermediate users
Recording Direct Observations

Make sure you get permission!
Make sure you are mindful of privacy!

Recording Observations: Tools

How do we record user actions during observation for later analysis so that the evaluator doesn’t forget, miss, or misinterpret events?

– paper and pencil are primitive but cheap
  • evaluators record events, interpretations, and extraneous observations
  • challenging to get details (writing is slow) though coding schemes help

– audio recording
  • good for recording dialog produced by think-aloud or constructive-interaction
  • hard to tie into what the user was doing on the screen
  • very hard to search through later

– video recording
  • can see and hear what a user is doing but can be intrusive
  • could use one camera for screen, another for participant(s)
  • hard to search and generates much data to comb through
Example coding scheme...

Tracking a person’s activity in the office with quick notations.

- **s** = start of activity
- **e** = end of activity

<table>
<thead>
<tr>
<th>Time</th>
<th>Desktop Activities</th>
<th>Absences from Desk</th>
<th>Interruptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Working on computer</td>
<td>Working at desk</td>
<td>Using telephone</td>
</tr>
<tr>
<td>9:00</td>
<td>s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:02</td>
<td>e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:13</td>
<td></td>
<td>s</td>
<td></td>
</tr>
</tbody>
</table>

Summary Thoughts?

How do you pick the right approach?