Qualitative Evaluation Techniques

Why evaluation is crucial to interface design

General approaches and tradeoffs with the different approaches to evaluation

The role of ethics

Learning how to quickly debug and evaluate prototypes by observing people using them

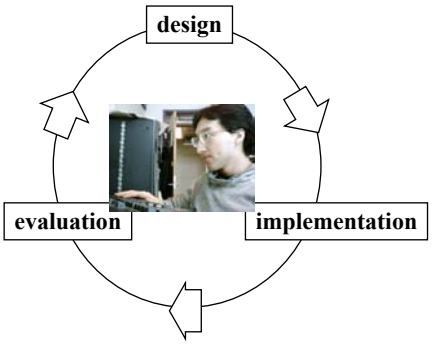
Specific evaluation methods helps you discover people's thoughts and motivations as they are using your system

Typical Arguments Against Evaluation

- **Evaluation takes up too much time and money.**
- **We'll finish the system first and then we'll worry about evaluating it.**

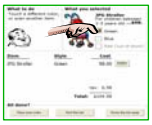
Why Evaluate?

- **(Rationale from the previous section): User-Centered design, account for the needs of the users throughout the design process**
  - One of way of doing this is by building prototypes throughout the design and development phases, and evaluating those designs to see if the user's needs are met.



Why Evaluate? (2)

- **It is typically less expensive to fix problems early than later in development:**
  - e.g., assume $100 per hour per person.

Low Fidelity	Medium Fidelity	High Fidelity
		
Paper sketch Time: 2 hours Cost: $200	**Power point simulation:** Time: 16 hours Cost: $1600	**Complete or nearly completed system:** Time: 100+ hours Cost: $10,000+

Evaluation And The Usability Engineering Lifecycle

Pre-design
- Investing in a new expensive system requires proof of viability

•Initial design stages
- Develop and evaluate initial design ideas with the user to gather requirements

Iterative design
- Verify requirements are being met: Does the system match the user's task?
- Evaluate and improve the design:
  - Are there any specific problems with the design?
  - Can users provide feedback to modify the design

•For these situations it's best to use a low cost, quick to develop prototype
- Low and medium fidelity prototypes are appropriate

Evaluation And The Usability Engineering Lifecycle (2)

Acceptance testing
- Can be used before a project begins
- Rather than specify usability requirements in vague terms:
  - "..the interface should be usable.."
  - "...the system should be user-friendly..."
- Performance criteria can be specified in specific and measurable terms:
  - Time for users to learn specific functions
  - Time to perform specific tasks
  - Error rates by users
  - User retention of commands over time
  - Subjective measures of user satisfaction with the system (assuming an agreed upon measure can be formulated)

Evaluation And The Usability Engineering Lifecycle (3)

Example acceptance test for a food-shopping web site[1]:

The participants will be 35 adults (25-45 years old) with English as their first language, no disabilities, hired from an employment agency. They have moderate web use experience: 1-5 hours/week for at least a year. They will be given a 5 minute demonstration on the basic features. At least 30 of the 35 adults should be able to complete the benchmark tasks, within 30 minutes.

Another test requirement might include:

Special participants in three categories will also be tested: (a) 10 older adults aged 55-65; (b) 10 adults users with varying motor, visual, and auditory disabilities; and (c) 10 adult users who are recent immigrants and use English as a second language.

A third item in the acceptance test might focus on retention:

Ten participants will be recalled after one week, and asked to carry out a new set of benchmark tasks. In 20 minutes, at least 8 of the participants should be able to complete the tasks correctly.

1 From Designing the User Interface (2005) by Shneiderman and Plaisant.

Approaches: Naturalistic

Observation occurs in realistic setting

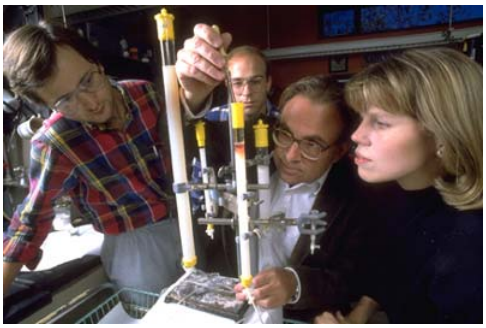
- Real life
- Can be done in a low cost fashion or an extensive study can be performed

- Problems
  - It may be difficult to arrange and to conduct
  - It may not always possible to replicate results

Approaches: Experimental

Experimental

- Classical lab study
- Study relations by manipulating one or more *independent* variables
  - Experimenter controls all environmental factors (nothing else is different)
- Observe effect on one or more *dependent* variables



James Tam

Tradeoffs: Natural Vs. Experimental

Internal validity

- Do you measure what you set out to measure (*correctness*)

External validity

- The degree to which results can be generalized to other situations (*realism*)

	Naturalistic	**Experimental**
Internal validity	Low	High
External validity	High	Low

James Tam

Mitigating External Validity Concerns

Test participants

- Screen participants to ensure that they are representative of the user population.
- Use a pre-test questionnaire to determine things like: education, personality, skill and experience levels which may be relevant to the test.

Tasks performed

- Using approaches like the Task-Centered approach, determine what are the common and important tasks.

Physical environment

- Test in an environment that is similar to the actual place of usage.

Approaches To Setting Up Experiments

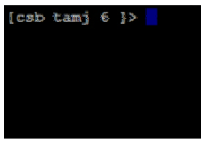
1. **Between subject design**
2. **Within subject design**

Between Subject Design

Test participants are run through only one test condition

Condition #1:

Command line



Participant A

Condition #2:

Menu driven

Participant B

Condition #3:

Touch screen

Participant C

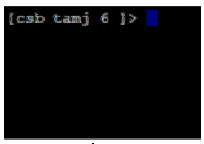
James Tam

Within Subject Design

Test participants are run through multiple (or all test conditions)

Condition #1:

Command line



Participant A

Participant B

Participant C

Condition #2:

Menu driven

Participant A

Participant B

Participant C

Condition #3:

Touch screen

Participant A

Participant B

Participant C

James Tam

Comparing Between And With Subject Design

	Advantage	Disadvantage	Mitigating the disadvantages
Between subject design	• No contamination	• Between subject variability	• Randomize the test condition assigned to participants • Matching
Within subject design	• Controls for individual differences • Fewer participants needed • May allow for comparisons	• Possible carry over effects	• Randomize the order of the tests

James Tam

How Many Participants To Test

The number of participants can have effect on the reliability of the test results.

- Would the same results be achieved if the test were repeated?

Problem: individual differences:

- The best user 10x faster than slowest
- The best 25% of users ~2x faster than slowest 25%

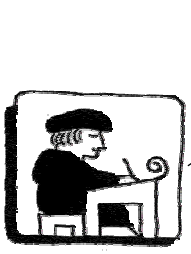
Partial Solution

- Get a reasonable number and range of test participants
- Identify and mitigate the effect of outlier cases (determined via questionnaires)

Images from "The Simpsons" © Fox

James Tam

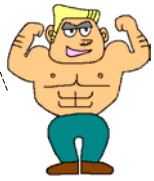
Test Roles



Scribe

Test participant

Test administrator



Security
(optional)

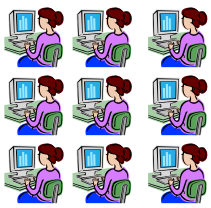
James Tam

Test Procedure

I) Run a pilot study

- "A practice run" of the test
- Purpose: To debug the test
- Results: Used to improve the test
- Participants: Peers or colleagues may be used

II) Run the main test



- Running the test "for real"
- Purpose: To debug the interface
- Results: Used to improve the interface
- Participants: Only use actual members of the user group

James Tam

Test Procedure (2)

1. **Preparation**
2. **Introduction**
3. **Running the system**
4. **Debriefing**

Ethics

Ethics

Testing can be a distressing experience

- People feel pressure to perform so errors are inevitable
- This can result in:
  - Feelings of inadequacy
  - Competition with other test participants

Golden rule

- Test participants should always be treated with respect

Managing Participants In An Ethical Manner

Before the test

- Don't waste the person's time
  - Use pilot tests to debug experiments, questionnaires etc
  - Have everything ready before the participant shows up
  - Try it out yourself one more time
- Make participants feel comfortable
  - Emphasize that it is the system that is being tested, not the person
  - Acknowledge that the software may have problems
  - Let participants know they can stop at any time
- Maintain privacy
  - Tell the participant that individual test results will be kept completely confidential
- Inform the participant
  - Explain any monitoring that is being used
  - Answer all of the person's questions (but avoid biasing them)
- Only use volunteers
  - Typically the test participant must sign an informed consent form

Managing Participants In An Ethical Manner

During the test

- Don't waste the person's time
  - Never have the user perform unnecessary tasks
- Make test participants comfortable
  - Try to give the person an early success experience
  - Keep a relaxed atmosphere in the room
  - Have coffee, breaks, etc
  - Hand out test tasks one at a time
  - Never indicate displeasure with the person's performance
  - Avoid disruptions
  - Stop the test if it becomes too unpleasant

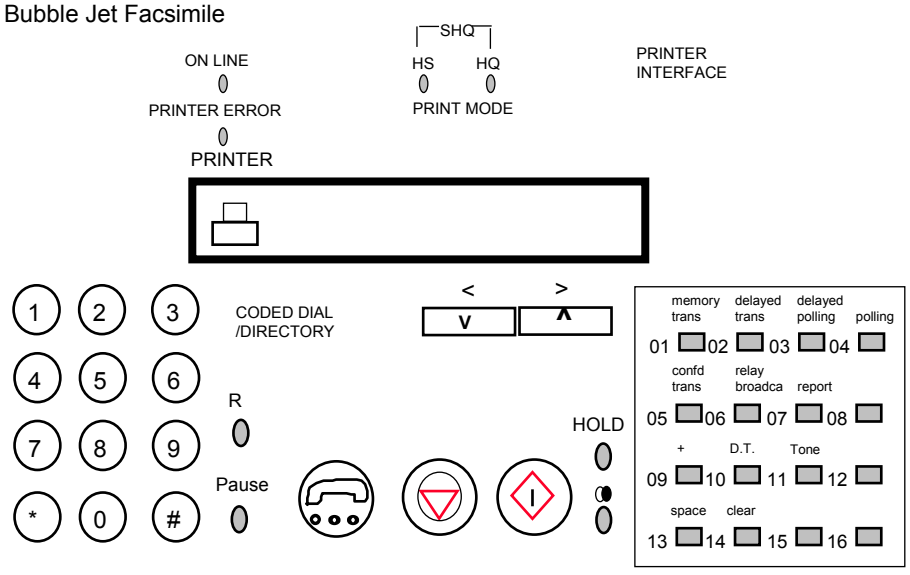
- Maintain privacy
  - This class: Only show test results to people when it is essential (TA and course instructor)
  - Actual practice: Do not allow the participant's management to observe the test

Managing Participants In An Ethical Manner

After the test

- Make the person feel comfortable
  - e.g., state that the participant has helped you find areas of improvement
- Inform the participant
  - Answer particular questions about the study that could have biased the results before
- Maintain privacy
  - Never report results in a way that individuals can be identified e.g., do not use real names in written reports (unless given permission), don't use gender references if it can be traced back to participants (referring to a participant as "she" when there is only one female participant).
  - Keep personal information confidential: Only show test results and other data outside the research group with the participant's permission

Canon
Fax-B320
Bubble Jet Facsimile



James Tam

Discount Usability Evaluation

Low cost methods to gather usability problems

- Approximate: Capture most large and many minor problems

How?

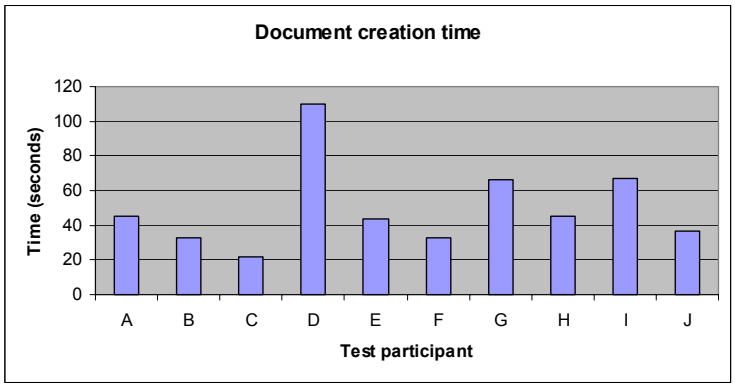
- Quantitative
- Qualitative

James Tam

Quantitative Approach For Usability Evaluation

Description of approach:

- Measure something of interest in user actions
- Count, log, speed, error rate
- A statistical analysis may be performed on the results e.g., analyzing standard deviations, performing T-tests etc.



James Tam

Qualitative Methods For Usability Evaluation

Data gathering

- Observe the actions of the user
- Gather opinions from the user

Produces a description, usually in non-numeric terms

May be quite subjective

Approach:

- Perform the study and gather your data (through questionnaires as well as using techniques to record significant events e.g., note taking).
- After the study is complete go through the data and try to summarize the results into key themes.
  - E.g., when viewing video recordings of the study don't analyze in detail every remark or action of each participant, instead look for critical incidents (such as when the person was obviously stuck).

Star Trek IV: The Voyage Home © Paramount Pictures

James Tam

Qualitative Methods For Usability Evaluation

Techniques

- Inspection
- Extracting the conceptual model
- Direct observation
  - Simple observation
  - Think-aloud
  - Constructive interaction
- Query via interviews and questionnaires
- Continuous evaluation via user feedback and field studies

Star Trek IV: The Voyage Home © Paramount Pictures

The Inspection Method

Designer tries the system (or prototype) out

- Does the system "feel right"?

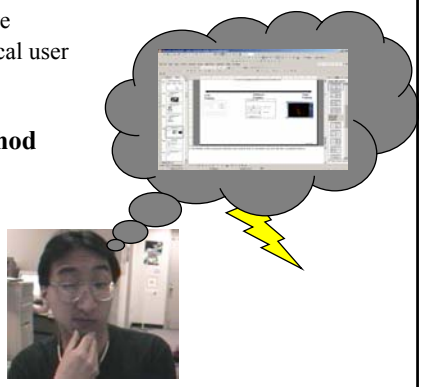
Benefits

- Can probably notice some major problems in early versions during every day use

Problems

- Low reliability rate as it's completely subjective
- Low level of validity as inspector is a non-typical user
- Intuitions and introspections are often wrong

Most widely used informal evaluation method



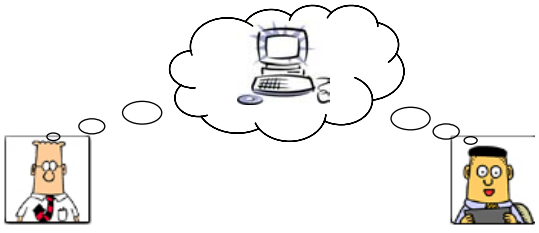
Extracting The Conceptual Model

Show the user static images of:

- The paper prototype *or*
- Screen snapshots *or*
- Actual system screens during use

Have the user try to explain

- What all elements are
- What they would do to perform a particular task
- How they think that the system works



Extracting The Conceptual Model (2)

Initial vs. formative conceptual models

- **Initial:** How person perceives a screen the very first time it is viewed
- **Formative**: The same, except after the system has been used for a while

This approach is:

- Good for eliciting people's understanding before & after use
- Requires active intervention by evaluator, which can get in the way

Direct Observation

Evaluator observes and records users interacting with design/system

- In a lab:
  - User asked to complete a set of pre-determined tasks
  - A specially built and fully instrumented usability lab may be available
- In the field:
  - User goes through normal duties

This approach is:

- Validity/reliability depends on how controlled/contrived the situation is
- Excellent at identifying gross design/interface problems

Three general approaches:

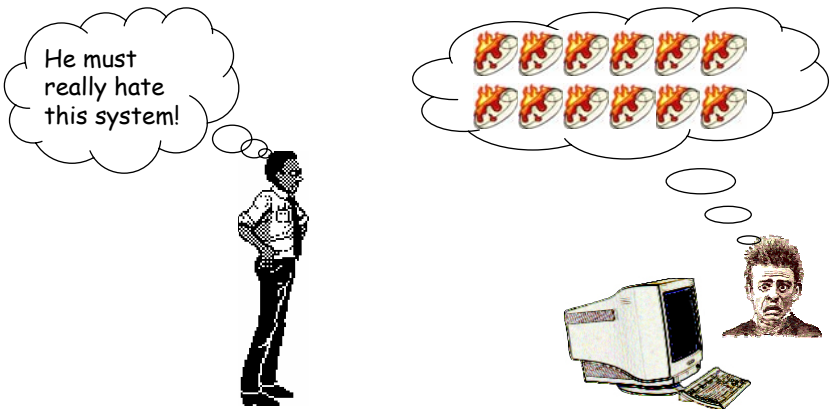
- Simple observation/Silent observer
- Think-aloud
- Constructive interaction

Silent Observer Method

Person is given the task, and the evaluator silently just watches while employing "The Silent Observer" technique.

Problem

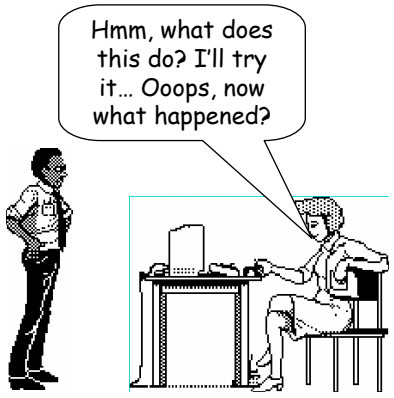
- Does not give insight into the person's decision process or attitude



The Think Aloud Method

Test participants are asked to say what they are thinking/doing

- Gives insight into what the person is thinking
  - What they believe is happening
  - What they are trying to do
  - Why they took an action
- The comments can provide useful quotes to make arguments more convincing.



The Think Aloud Method (2)

Problems

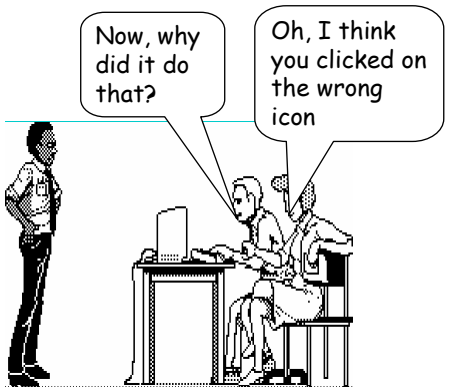
- Awkward/uncomfortable for the person (thinking aloud is not normal!).
- Hard to talk when they are concentrating on a problem.
- "Thinking" about things may alter the way people perform their task (could improve *or* degrade performance).
- Certain situations may prohibit the use of this technique.

Most widely used "formal" evaluation method in industry

The Constructive Interaction Method

Two people work together on a task

- Normal conversation between the two users is monitored
  - Removes the awkwardness of think-aloud
- Try to get participants who already know each other

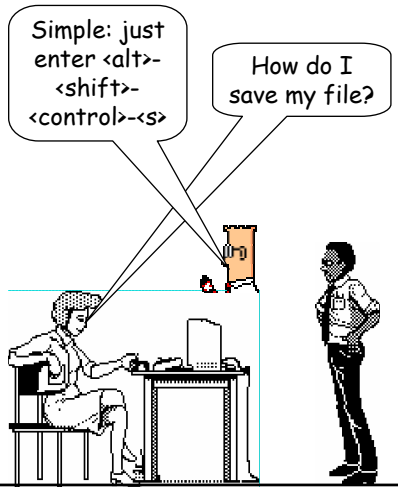


James Tam

Co-Discovery Learning

Variant of Constructive Interaction: Use semi-knowledgeable "coach" and novice user together

- Only the novice uses the interface
- Results in:
  - Novice user asking questions
  - Semi-knowledgeable coach responding
  - Provides insights into the thinking process of both user groups
- Also known as the "Coaching Method"



Dilbert © United Media

Querying People Via Interviews

Use a set of pre-created questions

- Gets things started
- Focuses the interview
- Ensures a base of consistency

Don't slavishly stick to the list!

- Be sure to follow interesting leads rather than bulldozing through a list of questions
- Add additional questions as necessary which could be based on the results of user observations

The degree of structure should be determined by the purpose of the study

- Open-ended/unstructured vs. structured

Querying People Via Interviews (2)

Don't forget

- Balance each question
- Avoid bias
  - Try not to ask leading questions

Querying People Via Interviews (3)

Excellent for pursuing specific issues

- Flexible
  - You can vary questions to suit the context
- Provides a rich depth of data
  - Probe more deeply on interesting issues as they arise
  - Often leads to specific constructive suggestions

Problems:

- Time consuming
- Accounts are subjective
- Requires a skilled and/or experienced interviewer
  - Evaluator can easily bias the interview
- Prone to rationalization of events/thoughts by the person
  - Reconstruction may be wrong

Retrospective Testing

A special type of interviewing technique that was developed in order to address the weaknesses of traditional interviews.

Post-observation interview to clarify events that occurred during system use

Approach:

1. Perform an observational test while recording the session on video

Retrospective Testing (2)

2. Watch the video with the users and encourage them to comment on what they did

Retrospective Testing (3)

Benefits

- Excellent for grounding a post-test interview
- Avoids erroneous reconstruction
- Users often offer more concrete suggestions
- Unlike the silent observer approach it provides insights into what the person is thinking/feeling and it can be used when thinking aloud is not possible.

Drawbacks

- Much like traditional interviews it can be very time consuming

Group Discussions

- **Start with individual discussions to discover different perspectives, and then continue with group discussions.**
- **Be cautious of the pitfalls of using a group!**
- **Increasing the group size may increase the universality of the comments.**
- **May encourage cross discussions.**

Querying People Via Questionnaires And Surveys

Questionnaires / Surveys

- Written queries for usability information

Benefits

- Administration cheap
  - Can reach a wide test group (e.g., mail)
- Administration requires little training
- Anonymous

Drawbacks

- Preparation "expensive" – although this may balanced off by the administrative savings
- Inflexible
- Relies on 'honest' answers

Querying People Via Questionnaires / Surveys (2)

Approach for all types

- Designing questionnaires: Establish the purpose of the questionnaire
  - Determine the audience you want to reach
    Typical survey: random sample of between 50 and 1000 users of the system
  - What information is sought?
  - How would you analyze the results?
  - What would you do with your analysis?
- Pre-test questionnaires:
  - Often used to determine the demographics (e.g., specific and relevant computer experience)
  - Consider questions that only allow for a fixed series of answers
  - Based on the initial answers you may divide participants off into to different categories
- Post-test questionnaires
  - Often used to solicit information about the system being studied
- For both types of questionnaires
  - Do not ask questions whose answers you will not use!
    e.g. How old are you?

Querying People Via Questionnaires / Surveys (3)

- Determine how would you will deliver and collect the questionnaire
  - On-line for computer users (e.g., web site with fill-in forms)
  - Surface mail
    Be sure to include a pre-addressed reply envelope to get a far better response rate
- Again be cautious about biasing the answers with the wording of the questions.
- For general tips on questionnaire design see the url:
  - http://www.cpsc.ucalgary.ca/~tamj/481/assignments/usability/questionnaire_tips.html

Style Of Questions

Open-ended questions

- May provide some unexpected insights
- Good for general subjective information but difficult to analyze rigorously
  - e.g., Can you suggest any improvements to the interface?
- Answers received may be unreliable

Style Of Questions (2)

Closed-ended questions

- Restricts the respondent's responses by supplying alternative answers
- Data is more narrow (less is rich but can be easily analyzed)
- But watch out for hard to interpret (by the reader) responses - *alternative answers should be very specific and should not overlap*
- Example:

(Vague)

Do you use computers at work:

O Often O Sometimes O Rarely

vs.

(Better)

In your typical work day, do you use computers:

O Over 4 hrs a day

O Between 3 and 4 hrs daily

O Between 1and 2 hrs daily

O Less than 1 hr a day

Style Of Questions (3)

•Examples:
(**May be too personal for some respondents**):
How old are you?

(**Overlapping ranges**):
Select the age range that you fall under:
15–20 ___
20–25 ___
25–30 ___

•Whether or not ranges have to be equal depends upon the information that you wish to gather.

(**For a shopping website this may not be a relevant question**)
What operating system do you use:
Windows ___
UNIX ___

•Types of closed-ended questions: scalar, multiple choice, ranked

Closed-Ended Questions: Scalar

Scalar

- Ask the user to judge a specific statement (opinions, attitudes, beliefs) on a numeric scale
- Scale usually corresponds with agreement or disagreement with a statement

Characters on the computer screen are:

Hard to read Easy to read
1 2 3 4 5

Closed-Ended Questions: Multiple Choice

Multi-choice

- Respondent offered a choice of explicit responses

How do you most often get help with the system? (Check only one category)

O On-line manual
✓ Paper manual
O Ask a colleague

Which types of software have you used? (Check all that apply)

✓ Word processor
O Data base
✓ Spreadsheet
O Compiler

Designing Closed-Ended Questions

Determine what features of the system will be evaluated

Create a series of statements about these features

Set the appropriate level of granularity for your questions:

- Too course – middle response tends to be over exaggerated.
- Too fine – makes it too difficult to distinguish between options.
- Rule of thumb: Provide 5 *or* 7 selections
- Try to use good text descriptions rather than simply using numeric values.

Closed-Ended Questions: Ranked

Ranked

- Respondent places an ordering on items in a list
- Useful to indicate a user's preferences
- Forces a choice

Rank the usefulness of the following methods for interacting with a computer
(1 = Most useful, 2 = Next most useful..., 0 = Not used)
__2__ Command line
__1__ Menu selection
__3__ Control key accelerator

Mixing Questionnaire Styles

Combining open-ended and closed-ended questions

- Get a specific response, but allows room for the user's opinion

It is easy to recover from mistakes:

Disagree
1 2 3 (4) 5 Agree

Comment: *The undo facility is really helpful*

Interviews Vs. Questionnaires: Summary Of The Pros And Cons

- **Preparation time**
- **Unanticipated/unexpected events**
- **Depth of information**
- **Analysis time**

Recording Observations

How do we record user actions during observation for later analysis?

- If no record is kept, evaluator may forget, miss, or mis-interpret events
- Regardless of the recording mechanism used make sure that you **separate what happened** from your **interpretations about why** something happened

Mechanisms for data collection

- Notes
- Audio recording
- Video recording

Notes

Evaluators record events, interpretations, and extraneous observations

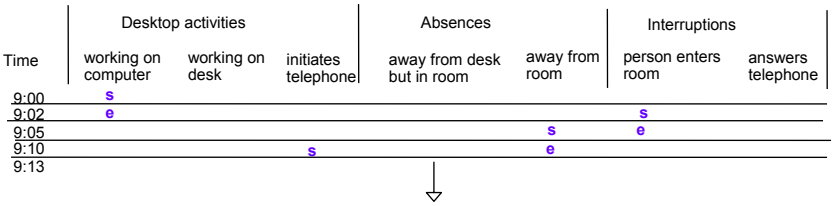
Paper and pencil:

- Cheap and allows for a great deal of flexibility in annotations
- Writing and observing can be tiring and error prone, the scribe records only what he or she deems as important (can be mitigated somewhat by using multiple scribes)
- Hard to get detail (writing is slow)
- The raw notes may have to be transcribed and organized

Notes (2)

- Coding schemes may be helpful:

s = start of activity
e = end of activity



Notes (3)

Electronic note taking:

- Laptops:
  - Information may not have to be transcribed and it's easier to organize.
  - Easy to combine with other recording mechanisms (still images, videos, audio).
  - It may be more cumbersome and obtrusive.
  - Annotations may be more difficult.

Audio Recording

Good for recording the dialog produced by thinking aloud/constructive interaction.

- Unlike note taking the complete dialog is captured.

It may be more obtrusive to test participants than note taking but is generally less obtrusive than video recording

Hard to tie into user actions (i.e., what they are doing on the screen)

Recordings must be transcribed

Video Recording

Can see and hear what a user is doing

- A more complete picture is provided
- It may be more intrusive to test participants than other recording mechanisms.

Multiple views of the study may be captured:

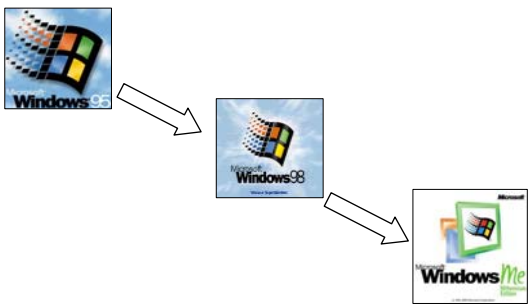
- One camera for screen, another for test user (picture in picture)

Recordings must be transcribed

Continuous Evaluation

1) Developers monitor system while it's actually being used

- Usually done in later stages of development
  - i.e., Beta releases, delivered system
- Good for finding real-world problems
- Problems can be fixed in the next release



Continuous Evaluation (2)

2) Users can provide feedback

- Email
- Special built-in gripe facility e.g., web site
- Telephone hot line
- Help desks
- Suggestion boxes

Best combined with trouble-shooting facility

- Users always get a response (solution?) to their problem

Continuous Evaluation (3)

3) Case/field studies

- Careful study of "system usage" at the site
- Good for seeing "real life" use
- Can be informal or more rigorous qualitative approaches can be attempted

What You Now Know

Evaluation is crucial for designing, debugging, and verifying interfaces

There is a tradeoff in naturalistic *vs.* experimental approaches

- Internal and External validity

The number and range of test participants employed will effect the reliability of your results

Test participants *must* be treated with respect

- The study should be guided by ethical rules of behavior

What You Now Know (2)?

Observing a range of users use your system for specific tasks reveals many successes and problems

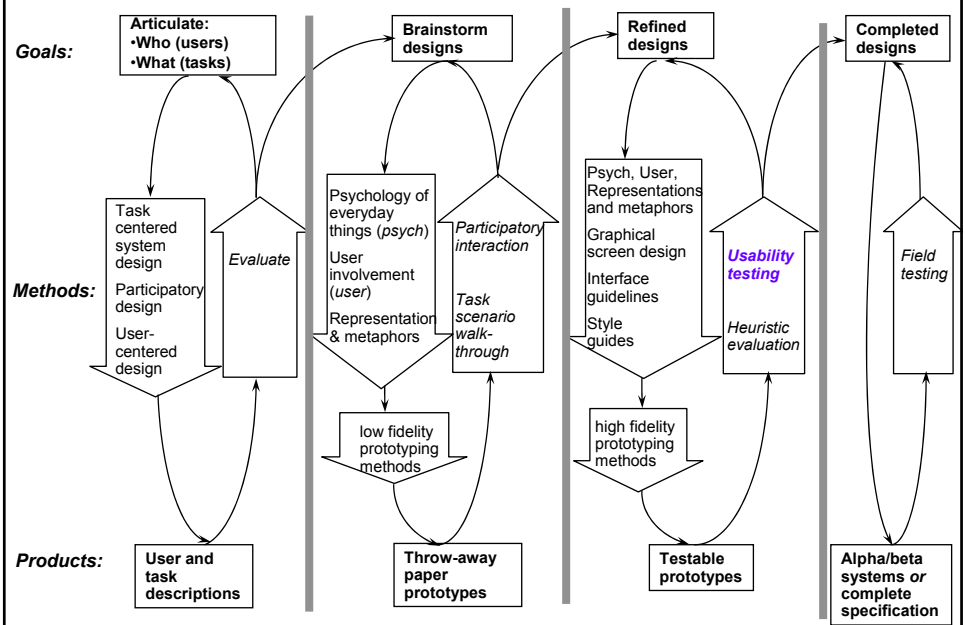
Qualitative observational tests are quick and easy to do

Several methods reveal what is in a person's head as they are doing the test

Particular methods include

- Inspections
- Conceptual model extraction
- Direct observation
  - Simple observation
  - Think-aloud
  - Constructive interaction (Co-discovery learning)
- Query via interviews, retrospective testing and questionnaires
- Continuous evaluation via user feedback and field studies

Interface Design And Usability Engineering



This diagram is a variation of the one presented by Saul Greenberg

James Tam