Evaluating Interfaces with Users

Why evaluation is crucial to interface design

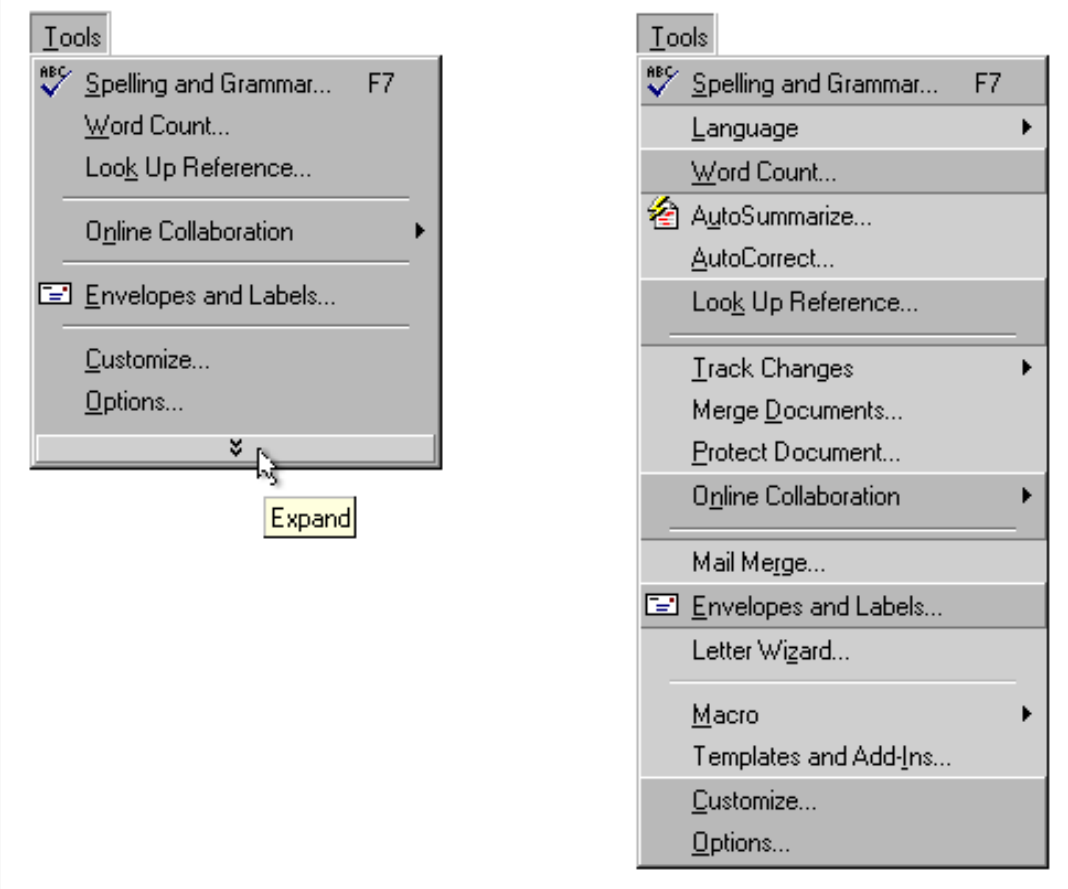
General approaches and tradeoffs in evaluation

How to quickly evaluate prototypes by observing people’s use of them

Methods reveal what a person is thinking about

The role of ethics

Adaptive Menu



User Testing

Test an interface with real users

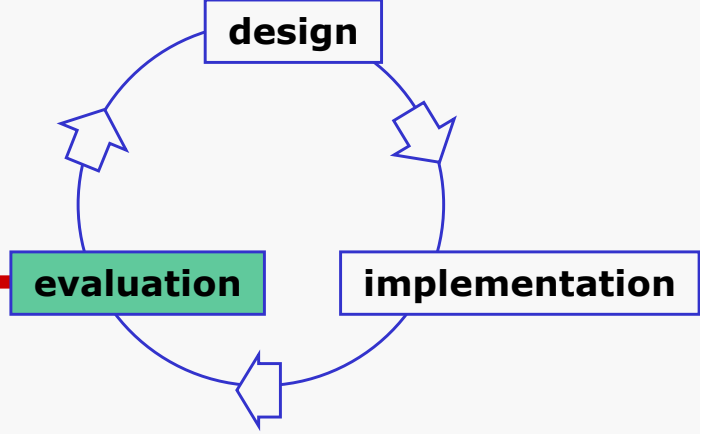
Users are human beings with feelings and rights

- Ethics
- Responsibility of evaluator

Types of User Testing

- Qualitative/Naturalistic
- Quantitative/Experimental
- Field Study

Why bother?



Tied to the usability engineering lifecycle

Pre-design

- investing in new expensive system requires proof of viability

Initial design stages

- develop and evaluate initial design ideas with the user

Iterative design

- does system behavior match the user's task requirements?
- are there specific problems with the design?
- can users provide feedback to modify the design?

Acceptance testing

- verify that human/computer system meets expected user performance criteria
  - Ease of learning, speed of performance, rate of error, retention over time, subjective satisfaction
  - 80% of 1st time customers will take 1-3 minutes to withdraw $50 from the automatic teller

Concerns of User Testing

External validity

- Generalizability of observed results
- confidence that results applies to real situations
- usually good in natural settings

Internal validity

- observed results caused by the independent variables?
- confidence in our explanation of experimental results
- usually good in experimental settings
- watch for "confounding" variables

Reliability

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

Naturalistic/Qualitative approach

Describes an ongoing process as it evolves over time

Observation occurs in realistic setting

- Ecologically valid

Realistic & general, often called "usability study"

Improve whole user interface with overall suggestions

External validity

- Degree to which research results applies to real situations

Considerations on External Validity

Population validity
- How representative are the sampled population?

Ecological validity
- How similar is the testing environment to the real world?

Training validity
- How realistic is the training?

Task validity
- Are the tasks chosen representative of the actual real world tasks?

Experimental/Quantitative approach

Experimenter controls all environmental factors

- So contrived
- study relations by manipulating *independent* variables
- observe effect on one or more *dependent* variables
- nothing else changes (note: confounding variables)

Narrow and specific questions or problems

Internal validity

- confidence that we have in our explanation of experimental results

Considerations on Internal Validity

Ordering effects

- Interface X first or Y first?
- Learning effect
- Get tried or bored

Selection effects

- Assign pre-existing groups to different levels of indep var
- randomization

Experimenter bias

- Biased for desirable result
- rigid experiment protocol

How to counter-balance

- Pure randomization, double-blind experiment

Usability engineering approach

Formative Evaluation

Discount Usability Evaluation

Mostly Qualitative

Observe people using systems in *simulated* settings

- people brought in to artificial setting that simulates aspects of real world setting
- people given specific tasks to do
- Observations made as people do their tasks
- look for problem areas / successes
- good for uncovering 'big effects'

Usability engineering approach

Is the test result relevant to the usability of real products in real use outside of lab?

Problems

- non-typical users tested
- non-typical tasks
- different physical environment
- different social context
  - motivation towards experimenter vs motivation towards boss

Partial Solution

- use real users
- task-centered system design tasks
- environment similar to real situation

Usability engineering approach

How many users should you observe?

- observing many users is expensive
- *but* individual differences matter
  - best user 10x faster than slowest
  - best 25% of users ~2x faster than slowest 25%

partial solution

- reasonable number and range of users tested

  Usability_Problems_Found(i)=N(1-(1-λ)^i), Nielson & Landauer (1993)

  N: total number of usability problems, i: # of test users

  λ: probability for finding any single problem with any single test user

  (typically N=41 and λ =31%, 5 users can find 85% of total usability problems)
- big problems usually detected with handful of users
- small problems / fine measures need many users

Discount usability evaluation

Low cost methods to gather usability problems

- approximate: capture most large and many minor problems

How?

- qualitative:
  - observe user interactions
  - gather user explanations and opinions
  - produces a description, usually in non-numeric terms
  - anecdotes, transcripts, problem areas, critical incidents…
- quantitative
  - count, log, measure something of interest in user actions
  - speed, error rate, counts of activities,

Qualitative methods for usability evaluation

Discount usability evaluation methods

Methods

- introspection
- extracting the conceptual model
- direct observation
  - simple observation
  - think-aloud
  - constructive interaction
- query techniques (interviews and questionnaires)

Introspection Method

Designer tries the system (or prototype) out

- does the system "feel right"?
- Most common evaluation method

- benefits
  - can catch some major problems in early versions

- problems
  - not reliable as completely subjective
  - not valid as introspector is a non-typical user
  - intuitions and introspection are often wrong

Conceptual model extraction

How?

- show the user static images of
  - the prototype *or* screens during use
- ask the user explain
  - the function of each screen element
  - how they would perform a particular task

What?

- **Initial conceptual model**
  - how person perceives a screen the very first time it is viewed
- **Formative conceptual model**
  - How person perceives a screen after its been used for a while

Value?

- good for eliciting people's understanding before & after use
- poor for examining system exploration and learning

Direct observations

Evaluator observes and records users interacting with system

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

Excellent at identifying gross design/interface problems

Validity/reliability depends on how controlled/contrived the situation is

Three general approaches:

- Simple observation
- Think-aloud
- Contructive interaction

Simple observation method

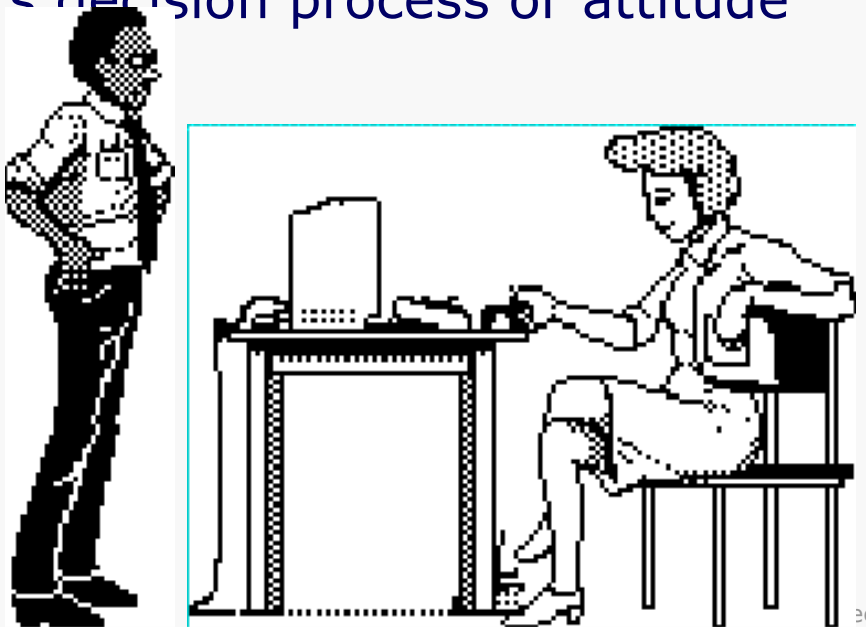
User is given the task

Evaluator just watches the user

- Most realistic

Problem

- does not give insight into the user's decision process or attitude



Think aloud method

Subjects are asked to say what they are thinking/doing

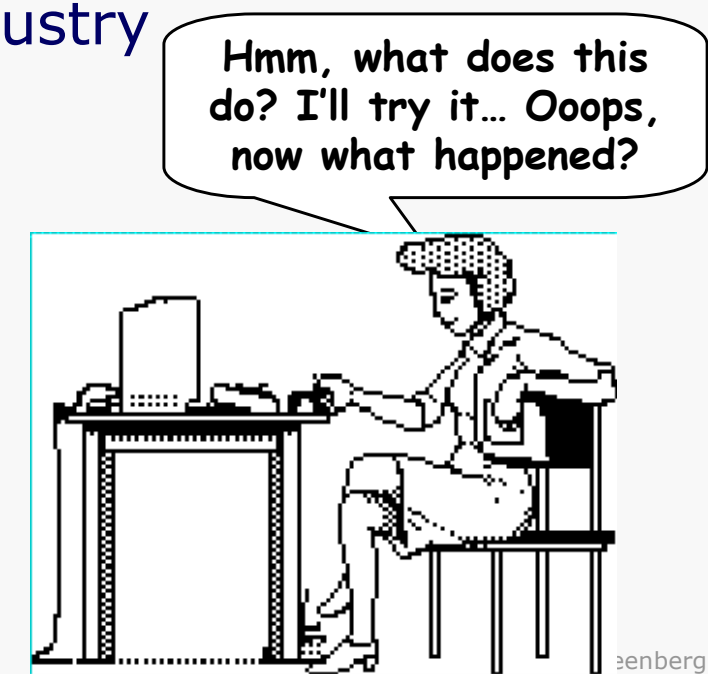
- what they are trying to do
- why they took an action
- how they interpret what the system did

Gives insight into what the user is thinking

Most widely used evaluation method in industry

Problems

- may alter the way users do the task
- unnatural (awkward and uncomfortable)
- hard to talk if they are concentrating



Constructive interaction method

Two people work together on a task

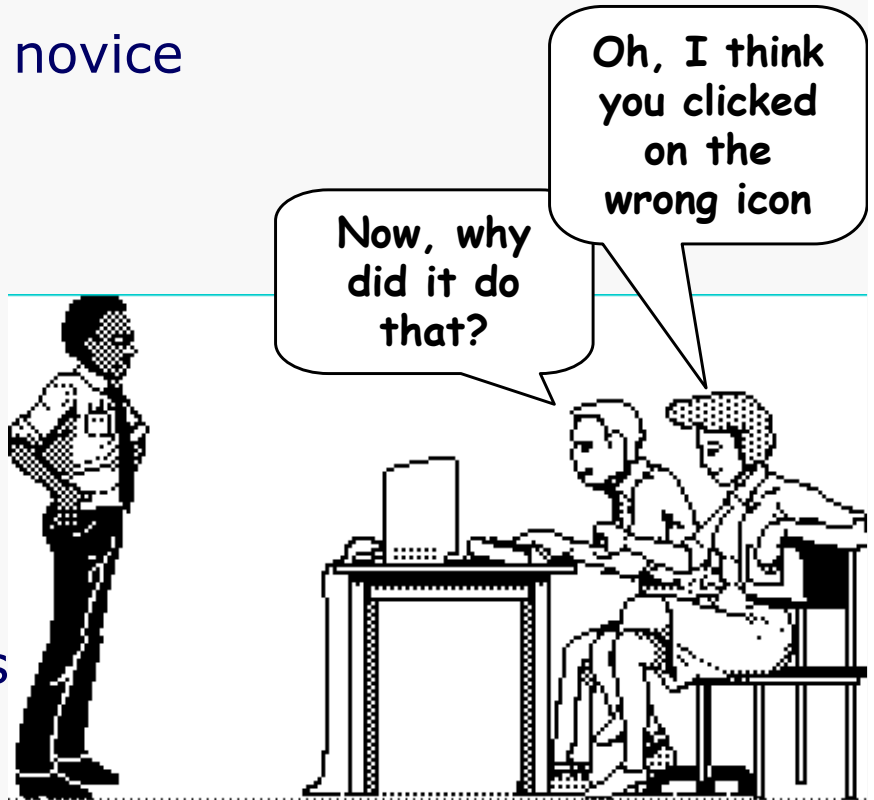
- monitor their normal conversations
- removes awkwardness of think-aloud

Variant: Co-discovery learning

- use semi-knowledgeable "coach" and novice
- make the novice use the interface

Results in

- novice ask questions
- coach responds
- provides insights into the thinking process of both beginner and intermediate users



Recording observations

How do we record user actions for later analysis?

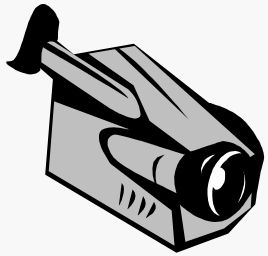
- If no record is kept, evaluator may forget, miss, or misinterpreting events

paper and pencil

- primitive but cheap
- evaluate events, comments, and interpretations
- hard to get detail (writing is slow)
- Code schemes help…

audio recording

- good for recording think aloud talk produced by thinking aloud/constructive interaction
- hard to tie into on-screen user actions
- Hard to search through later

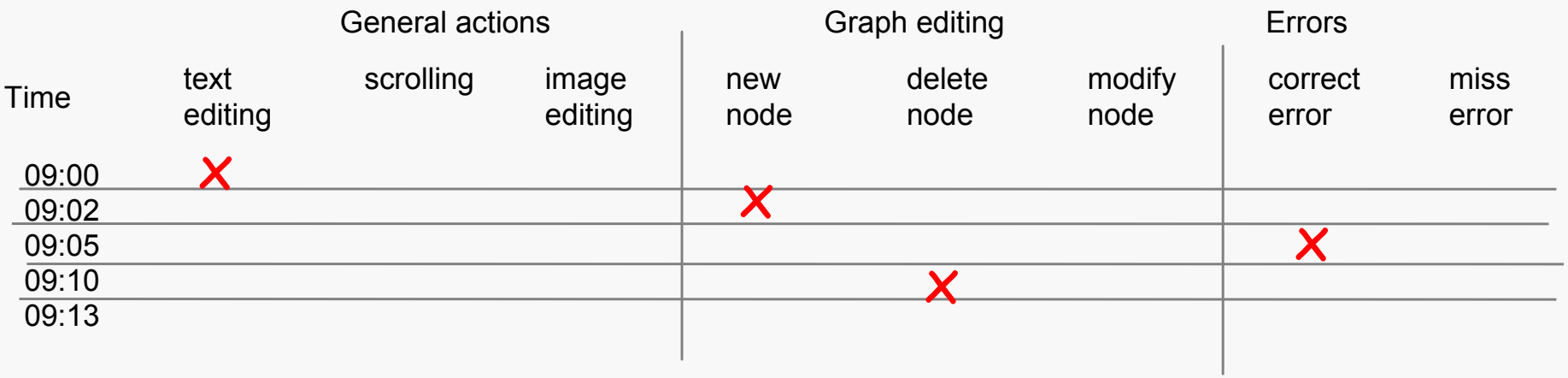


video recording

- can see and hear what a user is doing
- one camera for screen, rear view mirror useful…
- Can be intrusive during initial period of use
- Generate too much data

Coding sheet example...

tracking a person's use of an editor



Querying Users via Interviews

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
- often leads to specific constructive suggestions

Problems:

- accounts are subjective
- time consuming
- evaluator can easily bias the interview
- prone to rationalization of events/thoughts by user
  - user's reconstruction may be wrong

How to Interview

Plan a set of central questions

- Could be based on results of user observations
- a few good questions gets things started
- focuses the interview
- Ensures a base of consistency

Try not to ask leading questions

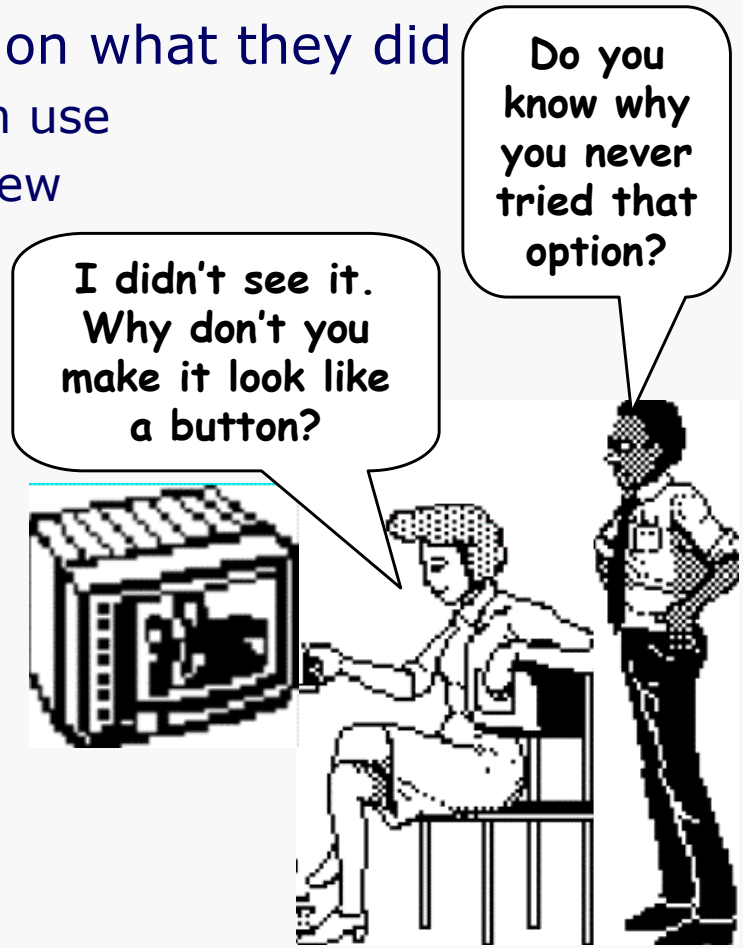
Start with individual discussions to discover different perspectives and continue with group discussions

- The larger the group, the more the universality of comments can be ascertained
- Also encourages discussion between users

Retrospective testing interviews

Post-observation interview to

- perform an observational test
- create a video record of it
- have users view the video and comment on what they did
  - clarify events that occurred during system use
  - excellent for grounding a post-test interview
  - avoids erroneous reconstruction
  - users often offer concrete suggestions

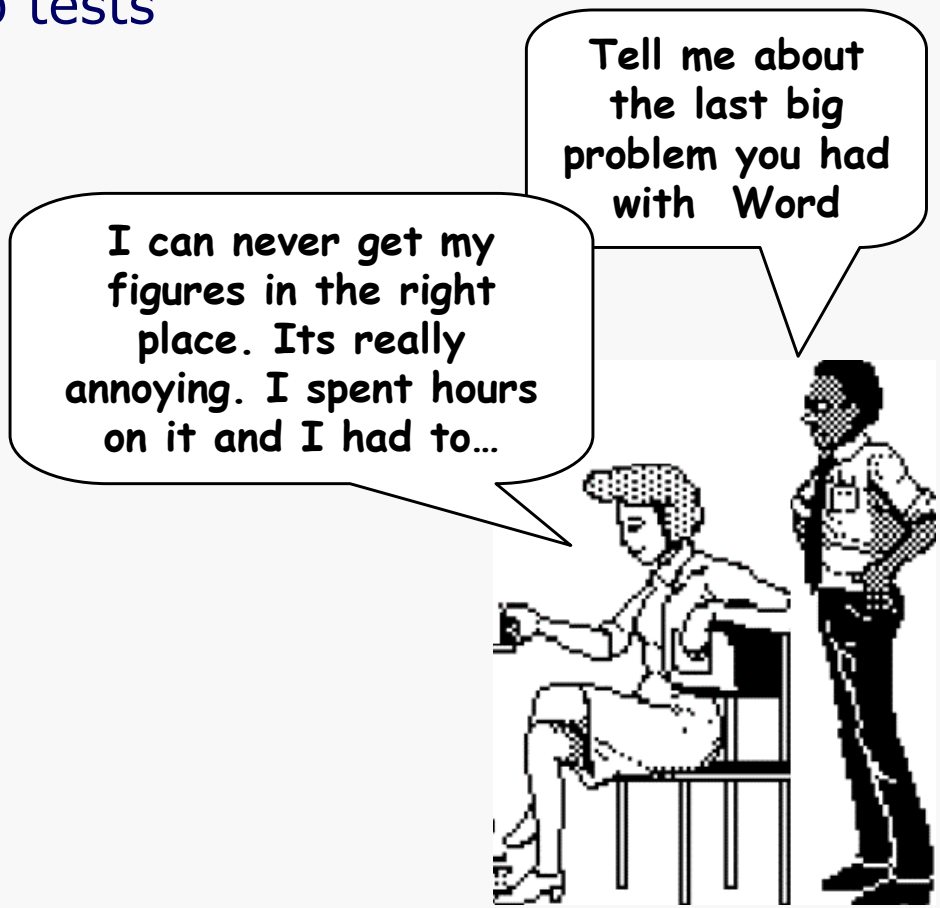


Jinwook Seo/Saul Greenberg

Critical incidence interviews

People talk about incidents that stood out

- usually discuss extremely annoying problems with fervor
- not representative, but important to them
- often raises issues not seen in lab tests



Jinwook Seo/Saul Greenberg

Questionnaires and Surveys

Questionnaires / Surveys

- preparation "expensive," but administration cheap
  - can reach a wide subject group (e.g. mail)
- does not require presence of evaluator
- results can be quantified
- only as good as the questions asked
- Often has low return rate – what's in it for them?

- QUIS – Questionnaire for User Interface Satisfaction

Questionnaires and Surveys

How

- establish the purpose of the questionnaire
  - what information is sought?
  - how would you analyze the results?
  - what would you do with your analysis?
- do not ask questions whose answers you will not use!
  - E.g. how old are you?
- determine the audience you want to reach
  - Typical survey: random sample of between 50 and 1000 users of the product
- determine how would you will deliver / collect the questionnaire
  - on-line for computer users
  - web site with forms
  - surface mail
    - pre-addressed reply envelope gives far better response
- determine the demographics
  - E.g. comptuer experience

Styles of Questions

Open-ended questions

- asks for unprompted opinions
- good for general subjective information
  - but difficult to analyze rigorously

Can you suggest any improvements to the interfaces?

Styles of Questions

Closed questions

- restrict respondent's responses by supplying alternative answers
- makes questionnaires a chore for respondent to fill in
- can be easily analyzed
- watch out for hard to interpret responses!
  - alternative answers should be very specific

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 hrs a day

O between 2 and 4 hrs daily

O between 1and 2 hrs daily

O less than 1 hr a day

Styles of Questions

Scalar

- ask user to judge a specific statement 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

Styles of Questions

Multi-choice

- respondent offered a choice of explicit responses

How do you most often get help with the system? (tick one)

O ✓ on-line manual
O paper manual
O ask a colleague

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

O ✓ word processor
O data base
O ✓ spreadsheet
O compiler

Styles of Questions

Ranked

- respondent places an ordering on items in a list
- useful to indicate a user's preferences
- forced choice

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

Styles of Questions

Combining open-ended and closed questions

- gets specific response, but allows room for user's opinion

It is easy to recover from mistakes:

disagree agree

1 2 3 4 5

comment: *the undo facility is really helpful*

Continuous Evaluation

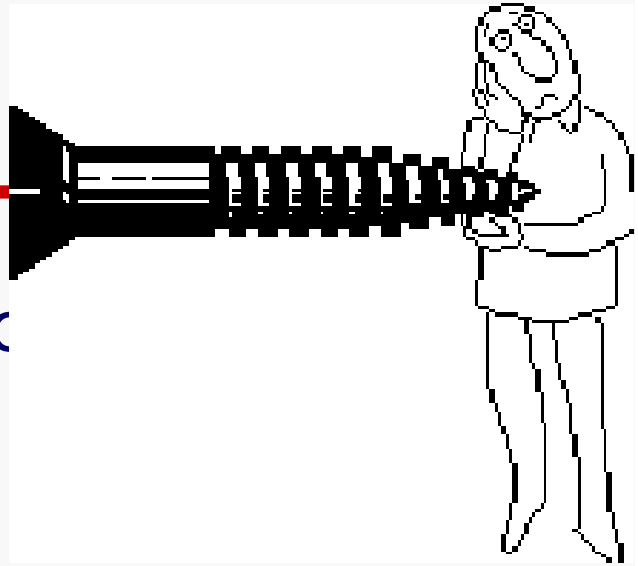
Monitor systems in actual use

- usually late stages of development
  - ie beta releases, delivered system
- fix problems in next release

User feedback via gripe lines

- users can provide feedback to designers while using the system
  - help desks
  - bulletin boards
  - email
  - built-in gripe facility
- best combined with trouble-shooting facility
  - users always get a response (solution?) to their gripes

Ethics



Testing can be a distressing experienc
- pressure to perform
- errors inevitable
- feelings of inadequacy
- Feels like an intelligence test
- competition with other subjects

Golden rule
- subjects should always be treated with respect

Ethics – before the test

Don’t waste the user’s time
- use pilot tests to debug experiments, questionnaires etc
- have everything ready before the user shows up

Make users feel comfortable
- emphasize that it is the system that is being tested, not the user
- acknowledge that the software may have problems
- *let users know they can stop at any time*

Maintain privacy
- tell user that individual test results will be completely confidential

Inform the user
- explain any monitoring that is being used
- answer all user’s questions (but avoid bias)

Only use volunteers
- user must sign an informed consent form

Ethics – during the test

Don’t waste the user’s time

- never have the user perform unnecessary tasks

Make users comfortable

- try to give user an early success experience
- keep a relaxed atmosphere in the room
- coffee, breaks, etc
- hand out test tasks one at a time
- never indicate displeasure with the user’s performance
- avoid disruptions
- stop the test if it becomes too unpleasant

Maintain privacy

- do not allow the user’s management to observe the test

Ethics – after the test

Make the users feel comfortable

- state that the user has helped you find areas of improvement

Inform the user

- answer particular questions about the experiment that could have biased the results before

Maintain privacy

- never report results in a way that individual users can be identified
- only show videotapes outside the research group with the user's permission

You now know

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

Qualitative observational test 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

- conceptual model extraction
- direct observation
  - simple observation
  - think-aloud
  - constructive interaction
- query via interviews, retrospective testing and questionnaires

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
- reliability
- precision
- generalizability

Subjects must be treated with respect

- ethical rules of behavior