

Computer/Human Interaction

Lecture 35

Overview:

- Evaluation methods
 - Empirical methods
- Usability specifications
- Homework 5 discussion

Empirical Evaluation

- Real data from real use
- Main concern is validity
 - Are the users representative
 - Is the test population large/diverse enough
 - Is the test system realistic enough (vs. early prototypes)
 - Does the data reveal real life impact
- Generally, does the investigation genuinely reflect real-world happenings

Empirical Methods

- Field studies - observations of real life
 - + by definition, the tasks are valid and the data is relevant
 - difficult to categorize and summarize data
 - time consuming to set up and conduct
- Interviews - ask about **critical incidents**
 - + collaborative effort between designers and stakeholders
 - memory is biased, tend to reconstruct rather than recall

Controlled Experiments

- Carefully select representative tasks from task analysis and SBD claims
- Control for uninteresting aspects; e.g., location, task order, instructions
- Collect multiple measures of performance (time/errors), output quality, and satisfaction ratings

Controlled Experiments 2

- Define hypothesis in advance. I.e., what is the expected outcome.
 - Independent variable - that which is manipulated; each manipulation method is called a test condition or a **level**. E.g., three different input devices.
 - Dependent variable - the measured experiment outcome. E.g., time to complete task
- Several of each may be included in an experiment. E.g., experience with similar applications (ind.) or accuracy (dep.)

Controlled Experiments 3

- Two kinds of experiment design
 - Within subjects - same participants are exposed to all levels of the independent variable
 - Between subjects - different group for each level
- For either, participant groups must be designed
 - # of people, age range, experience levels, motivation. Often use random assignment. Need at least 10 for statistical validity.

Controlled Experiments 4

- + Control for uninteresting aspects
- + Can collect multiple measures
- Statistical validity is difficult: sample size, sample representativeness
- Between subjects test groups must be matched or pulled from a large enough group that random assignment reduces the effect of the uninteresting aspects
- Expensive and time consuming

“Discount” Evaluation

- Real-world goal of getting the most useful information for guiding redesign at least cost (Nielson)
- Do a little of both analytical and empirical evaluation
 - 3-4 experts find most guideline issues
 - 4-6 users experience most of the actual usage problems
 - Between the two, get a good sense of what to fix

Conducting Experiments

- Recruiting test participants
 - Participatory design users
 - Stipends or other rewards
- Informed consent
 - All tests at UE using human subjects (including surveys) must be approved by the IRB
 - Full disclosure of test procedures, statement of voluntary nature, and participants rights
 - Signature of participant

Surveys

- Good for subjective reactions like satisfaction, usefulness
- Likert scale - strength of agreement to assertion about system or task, usually 5-7 choices converted to a number
- Usually summarized as average and sample size, and compared to usability specification. However, redesign decisions come from the **details** of the test.
- Used pre- and post-test to measure changes in attitude

Usability Specifications

- Describes what needs to be measured and what constitutes satisfactory performance
- Quality objectives must be precise and are managed in parallel with other design specifications
- In SBD, come from scenarios and claims
 - Critical subtasks described by scenarios
 - Claims identify issues and measurable outcomes

Usability Specifications 2

- For each issue/outcome, specify target levels of performance in worst case, planned (average) case, and best case
- E.g. on subtask of uploading a file from a PC with confusion rating from 1 (not confusing at all) to 5 (extremely confusion)
 - Worst case: 3 minutes, 1 error, 3 on confusion
 - Planned case: 30 seconds, 0 errors, 2 on confusion
 - Best case: 10 seconds, 0 errors, 1 on confusion

Homework 5

- Nielsen's heuristic evaluation guidelines applied to www.amazon.com
- Use simple and natural dialog
- Speak the user's language
- Minimize memory load
- Be consistent
- Provide feedback

Homework 5 continued

- Provide clearly marked exits
- Provide shortcuts
- Provide good error messages
- Prevent errors
- Include good help and documentation