CS 350: Computer/Human Interaction
Lecture 21 Overview

- Usability evaluation
- Types of evaluation
- Methods of evaluation
- Usability specifications

**Reminder:** Mid-project Progress Report due on Thursday with peer review forms. No class on Thursday.

**Assignment out:** Homework 5, due Tuesday at beginning of class

CS 350: Computer/Human Interaction
Usability Evaluation

- Any analysis or empirical study of the usability of a prototype or a system
- Goal is to provide feedback in software development. Answer questions like:
  - Is the system sufficiently useful?
  - Is it too difficult to use or learn?
  - Is it satisfying to use?
  - Does it meet the stated goals?
- Understand problems and causes; plan changes to correct them

CS 350: Computer/Human Interaction
Types of Evaluation

- **Formative** evaluation – done during system development. Drives redesign. Often done by asking user to verbalize thoughts during user test of prototypes
- **Summative** evaluation – done at the end of project or other checkpoint. Answers "how well did we do?" Generally done by measuring performance times and error rates.

CS 350: Computer/Human Interaction
Evaluation Methods

- **Analytical** methods – study the design/system characteristics and compare with theory, modeling, or guidelines from experts
- **Empirical** methods – study how users use system via observation, surveys, controlled experiments

Need both analytical and empirical evaluation:
- Is it a bad implementation of a good design or a good implementation of a bad design?
- How to choose what methods to use?
- Which is more expensive?
- Which carries more weight with developers?

SBD claims analysis showing positive and negative impacts

Usability inspection
- Expert walk-through based on guidelines or checklist (more than one if possible)
- Walk-through at different levels or categories
- List problems in each level/category, order by severity
CS 350: Computer/Human Interaction

Analytical Methods

- Cognitive walk-through for walk-up and use systems (e.g. ATMs)
  - Look for affordances, metaphors
  - Careful task selection, answer questions at each step: what comes next, what is assumed, are there any competing goals
- All methods use checklist forms, very popular with industry – generates lots of data for low cost

Heuristic Evaluation

- Nielsen (1994) gives 10 general guidelines for usability inspection
  - Use simple and natural dialog
  - Speak the user’s language
  - Minimize memory load
  - Be consistent
  - Provide feedback

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

For any usability inspection
- Want multiple experts
- Want point of view (POV) of different classes of stakeholders

Model-Based Analysis

- User is modeled as a breakdown of goal identification, steps to achieve goal, implementation of steps, and selection rules
  - Predictive model developed using scientific knowledge of human memory and behavior
  - Use model elements for mental activities
  - Like HTA, also estimate task times for alternatives

GOMS Example

- Goals, Operators, Methods, Selection rules
- Example: how to close Firefox tab

```
GOAL: CLOSE-ACTIVE-TAB
  |-{select GOAL: USE-MENU-METHOD
  |  |  |MOVE-MOUSE-TO-MENU-BAR
  |  |  |DRAG-DOWN-FILEMENU
  |  |  |RELEASE-ON-CLOSE-TAB-OPTION
  |  |GOAL: USE-HANDLE-METHOD
  |  |  |MOVE-MOUSE-TO-INTERNAL-CORNER
  |  |  CLICK-ON-CLOSE-BOX
  |  |GOAL: USE-CONTROL-KEY
  |  |  |PRESS-CTRL-W
```

Tradeoffs

- Usability inspections are fast and cheap
  - BUT
  - Miss details only seen in actual use
  - Doesn’t identify causes of problems
  - Emphasizes problems that are infrequent or atypical in actual use
  - Contributes little to overall HCI theory
CS 350: Computer/Human Interaction

Tradeoffs

- Model-based analysis has scientific foundation and is powerful and credible
  - Limited to the scope of the theory
  - Time-consuming to develop
  - Ignores higher level structures of behavior

CS 350: Computer/Human Interaction

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

CS 350: Computer/Human Interaction

Empirical Evaluation

- 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

CS 350: Computer/Human Interaction

Controlled Experiments

- 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.

CS 350: Computer/Human Interaction

Controlled Experiments

- 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, etc. Often use random assignment. Need at least 10 for statistical validity.
**CS 350: Computer/Human Interaction**

### Controlled Experiments

- 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 (Nielsen)
- 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 must be approved by the IRB, including surveys
  - 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

### 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
**Usability Specifications**

- In SBD, come from scenarios and claims
  - Critical subtasks described by scenarios
  - Claims identify issues and measurable outcomes

- For each issue/outcome, specify target levels of performance in worst case, planned (average) case, and best case

**Example:** 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**

- **Homework 5 due on Tuesday:** apply Nielsen’s guidelines (heuristic evaluation) on page 233 of the textbook to an e-commerce site of your choice; discuss on Tuesday.