Lecture 7: Usability Methods IV

- Usability Testing
  - Verbal Reports
  - Performance Measures (“Usability Metrics”)
  - Questionnaires and Surveys
- Other Methods
  - Experimental Design
  - Storyboards, Scenarios, and Sketching

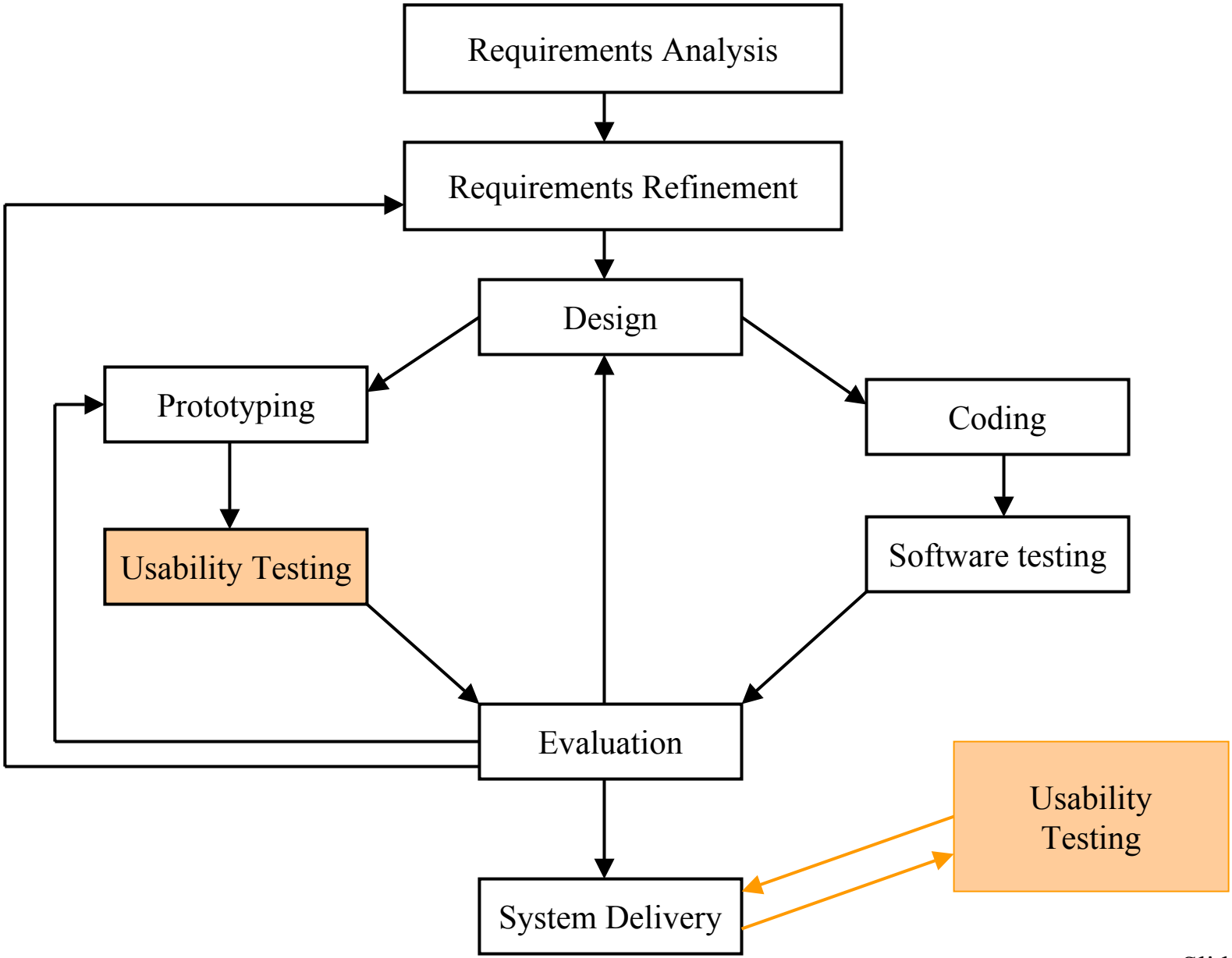
Usability Testing

Usability Testing (1)

- What is it?
  - Users a brought to use the product or a prototype of the product
  - Users' interaction with the system are observed, possibly videotaped
  - Specific tasks and/or goals are defined for the user to accomplish
  - *Usability metrics* are usually collected
    - time it takes to accomplish a task
    - the number of errors in completing a task
    - simple completion of task itself (see Nielsen article handout)
    - ratings (how easy/hard etc. on a ten-point scale, for example)
  - Other observations may be done
    - "think aloud" protocols
    - structured, semi-structured, or unstructured interviews
    - questionnaires

Usability Testing (2)

- When is it done?
  - During the iterative testing phase of design
    - Using prototypes or early, unfinished versions of the system
  - As a final test after system is developed
    - Close to the end stage before product release
  - After a project is in release, as an evaluation
    - Investigate reported problems, e.g. from helplines or reviews
    - As a method of directing new designs


Requirements Analysis
Requirements Refinement
Design
Prototyping
Coding
Usability Testing
Software testing
Evaluation
Usability Testing
System Delivery

Usability Testing Examples

- Voice mail system
- Web site evaluation

The "Data" in Usability Testing

- Observational Data
  - Record observations of users
    - Visual observation
    - Logging by system of key presses, mouse clicks, etc.
  - Audio and Video recording
    - Provides permanent record
    - Can be scorced by different investigators
    - Can provide an effective demonstration tool
      - Video clips from usability testing can persuasively illustrate usability problems to other members of the development team
    - Various scoring methods exist for video recording
      - Task-based analysis
      - Performance-based analysis
- Verbal Protocols
- Questionnaires and Surveys

Verbal Protocols (1)

- Audio record provides users' spoken observations
- Provides insight into users' cognitive processes, their goals, the way they interpret vocabulary, etc.
- *Think-Aloud Protocol*
  - User is instructed to verbalize all their thoughts as they work through the tasks
  - Good solution to getting users to verbalize, solution to long silences, from which you can get no insight
  - Cognitive psychology literature and methodology is extensive on use, validity, and reliability of verbal protocol technique in study of thinking
  - Problem: asks the user to do an extra task while already doing the usability test
    - May add extra stress
    - May bias the results, e.g. ratings of difficulty

Verbal Protocols (2)

- Post event Protocols
  - Ask people to provide observations on what they were doing after they have done the task
  - May do so as narration to viewing themselves on videotape

Usability Metrics

- Users instructed to perform tasks
  - Qualitative observation
  - Usability Metrics
    - Success rate
    - Number of errors
    - Time to complete tasks
    - Users' subjective satisfaction
      - Rating scales
- Advantages of Metrics
  - Track progress of design across re-testing
  - Compare two designs
  - Assess your design against competitor products/services
  - Go/No-go decision → Set Usability Goals
  - Usefully summarizes results

Usability Metrics: Example

- Voice mail usability test
  - Tasks such as login, reply to message, forward message, etc.
- Metrics
  - Time it took to complete task(s)
    - Summarized all tasks into mean & median
  - Success (yes/no)
  - 7-point rating scale hard/easy to complete task & overall

Questionnaires and Surveys

- Open Questions
  - User is free to provide their own answer in their own words
- Closed Questions
  - User selects their answer from a set of alternative replies
  - Simplest version: “Yes / No / Don’t Know”
  - Multiple Choice Questions
  - Rating Scales: Choices are valued or ordered

Rating Scales

- Multi-point scales
  - May or may not be numbered
  - May have meaning of response labels just at end points
  - Each endpoint may be labeled
- Likert scale
  - Strength of agreement with question is measured by associated words on the scale
- Semantic Differential
  - Bipolar adjectives (easy/difficult) at ends of scale, rating is given between the paired adjectives

Multipoint Scale

How easy was it to chose a font in Microsoft Powerpoint?

Easy 1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 Difficult

Likert Scale

Microsoft Powerpoint is easy to use

- Strongly Agree
- Agree
- Slightly Agree
- Neutral
- Slightly Disagree
- Disagree
- Strongly Disagree

Usability Testing: How many participants?

- Nielsen et al.: Discount usability testing
  - 5 users is enough to catch most major usability problems
  - It is more cost effective and useful to run many different tests with 5 users each (iterative design) than to conduct one large study with a large number of users
  - Nielsen, J. (2000) http://www.useit.com/alertbox/20000319.html
  - Nielsen (1993) Usability Engineering, pp. 155-157
- Research:
  - Nielsen & Landauer (1993) “A mathematical model of finding of usability problems.”
  - Virzi (1992) – 80% of usability problems found in first 4-5 users, most significant problems revealed in first 1-3 users.

Nielsen & Landauer (1993)

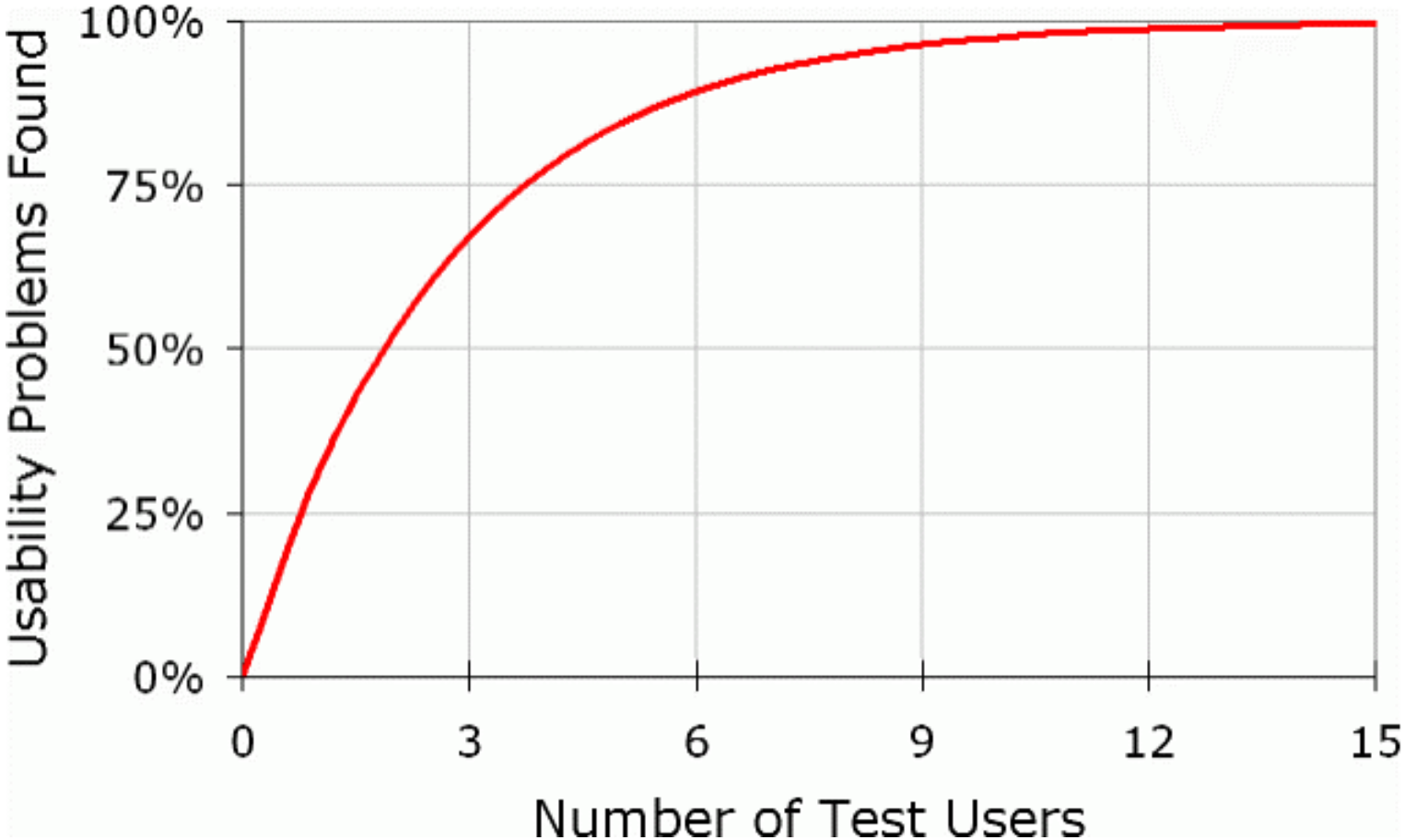
- Number of usability problems found with **n** users is described by

$$N(1–(1–L)^n)$$

- Where:
  - N = total number of usability problems
  - L = proportion of problems discovered on 1 user
  - Typically, L = 31%

- Cost benefit analysis argument
  - Better to test 3 times with different designs, 5 users each, than one large test with 15 users
  - May need more users tested in certain circumstances
    - e.g. have several distinct user problems (young, old, etc.)

Nielsen & Landauer (1993)



Five to Eight Users: Counterarguments

- Five to eight users is insufficient for statistical power
  - Usability metrics may well not be statistically significant
  - Therefore, hard to argue from and interpret
- Single “outlier” can influence interpretation of study
  - But that person may not be representative
- Some testers have observed more usability problems and more stable results as more users are tested

Perfetti & Landesman (www.uie.com)

- Tested usability of web site with 18 users
  - Identified 247 problems throughout testing sessions
  - Found 5 new problems on average with each 5 new users tested
- Interpretation:
  - Web sites are more complex than the simple software tested in Nielsen and Virzi studies
  - User tasks on web sites are more complex
  - Increased complexity requires more users
- Recommendation:
  - Web sites benefit from on-going testing
    - For example, 1-2 users per week
    - After 6 months will have 20 users
    - Can alter web site dynamically over testing period

Conclusions: Number of Users to Test

- Schroeder:
  - Nielsen and Virzi studies don't - and never claimed to - apply to large software projects or websites
- What is "large"?
- Statistics and experimental design:
  - Going beyond qualitative: 3, 4, 6 users tested is insufficient statistically

Other Methods

Other Methods - I

- Planned/controlled experiments
- Surveys and questionnaires
- Focus groups
- User surveys
- Observe statistics from deployed systems
  - System logs
  - Web site hits
  - Completions and hang ups in IVR systems

These methods may be employed at any stage in design. But often are employed near final stages or after release

Controlled Experiments

- What the difference from “Usability testing”
  - Usability testing is observation, single system or one design
  - Experiments test and compare user interfaces under controlled experimental conditions
    - where “user interfaces” are
      - different systems
      - alternative systems
      - alternative designs for a feature or function of a system
- Usually must address *specific* aspects of system under investigation -- “element testing”
- Behavioral science definition of an *experiment*:
  - the investigator “manipulates” some factor(s) of the subject’s environment, and the subject’s behavior is measured
  - If a regular measurable change in behavior is found from the manipulation, the factor can be identified as the “cause”

Experimental Control

- Independent Variables
  - The environmental or subject variable (“factor”) which the experimenter manipulates
  - Takes on two to an infinite number of values or conditions (“levels”, “treatments”)
- Dependent Variable
  - The variable, usually a response, which is influenced by the independent variable -- it is the factor which is measured
- Control Variables (Extraneous variables)
  - All those factors which are kept constant by the experimenter, in order to isolate only the effect of the independent variable, so that causality can be attributed to the independent variable
- Confounding
  - Data are confounded when the influence of variables under study in the experiment cannot be isolated form the influence of extraneous variables

Experimental Design

- Experimental “conditions”
  - Each level or value of the independent variable
  - Ex.: compare reading of text on CRT:
    - Cond 1: Red letters on White
    - Cond 2: White letters on Black
    - Cond 3: Yellow letter on Green
- Between-Subjects Design
  - Different subjects (users) experience each condition
  - Variation: Matched-Subject Design (Matched-Group Design)
    - Subjects paired on organismic variable (e.g. male, female) and assigned randomly to conditions by pairs
    - not often used in user interface design
- Within-Subjects Design (Repeated-measures design)
  - All conditions (treatments) are experienced by all subjects (users)

Between-Subjects Design

- Subjects only experience one condition
- Assignment to condition is random
- Disadvantages
  - Characteristics of subjects may influence (“confound”) your results
  - Example: too many male subjects in color condition 1, may get gender bias effects on color preference -- which you may mistakenly attribute to the independent variable (color combination)
- Advantages
  - Avoids order confounding problems of Between subjects design (next)
  - Often very efficient
  - May be the only way to study some things (esp. if effect of treatment is irreversible, that is, you cannot combine different conditions together)
- Solutions
  - Add more subjects to a between subjects design, action of random assignment should compensate for biases (e.g. like the male/female bias, as more people added, male/female ratios should “even out” between conditions)
  - Balance subjects on variable suspected to be a problem, in a matched subject design

Within-Subjects Design

- Each subject (user) gets each condition
  - Ex.: everyone gets Cond 1, 2, and 3
- Advantage
  - Avoids the effects of subject characteristics, since each subject "contributes" their data to each condition.
    - One subject or set of subjects will not bias a condition
- Disadvantage
  - There may be affects of the order in which you give conditions
  - Examples:
    - Users get tired at end of experiment, last condition given gets bad ratings because of tiredness not because of the treatment
      - Tiredness here is confounded with the treatment
    - Results, knowledge, or learning from the previous conditions will bias subject's responses or answers to later conditions

Within-Subjects Design

- Solutions
  - Manipulate order of conditions
  - Randomly order conditions for each subject
    - Problem: requires many subjects for effect of random assignment to work
  - Counterbalance
    - Determine every combination of orderings of conditions
      - Example: 123, 321, 132, 213, 231, etc.
    - Assign subjects so that an equal number of subjects get each order
    - Effects, for example, of “tiredness” are now equally applied to each condition
    - Differences you see between conditions can now be attributed to the independent variable you manipulated
    - Other counterbalancing methods (Latin squares, etc.)
      - Techniques used when number of orderings is too large

Examples

Experiments as Used in Design

- Compare Voice Mail Systems
- Q/Z Study
- Glare Study
- Upper/Lower Case Letter Study

Other Methods - II

- Storyboards
- Scenarios
- Sketching
  - "Visual brainstorming" -- explore ideas and metaphors in the early process of a design

Generally, these are methods used in early design and planning stages of product development

Scenarios and Storyboards

- Scenario
  - Fictional story of users interacting with products
    - Can assume different environments, events, products, persons, and individual backgrounds as part of the stories
- Storyboards
  - Series of rough sketches of a sequence of actions by user with a product
  - Sketches may be of the user interface, e.g. series of screens on a computer, or may be cartoons

Class Exercise

Preece et al. Ch. 22 Sec. 22.2

- Draw visual metaphor (idea for an icon) for the following programming constructs:
- EDIT
  - Open program in editor for changes
- DEBUG
  - Scan program for syntax errors
- STRING
  - A literal section of text used by the program
- EXECUTE
  - Run the program
- DECLARE
  - Classify a variable in the program as a certain type
- LOOP
  - Iterate the same set of instructions over and over again