Last Class

- Usefulness and usability
- Empirical evaluation, and its use in the design process
- Observing scenarios and prototypes
- User testing with thinking aloud
- Data capture and analysis
- Asking users as well as testing them
- Ethical issues

Outline

- Goals for research in HCI
- McGrath’s taxonomy of research methods
  - Field strategies
  - Experimental strategies
  - Respondent strategies
  - Theroretical strategies
- Demonstrations
- Usability inspection
- Controlled experiments
- Quasi-experiments
- Tradeoffs among empirical methods
- Research methods in the development process

Goals for Research in HCI

- Evaluate or compare existing systems/features/interfaces
- Invent or design new systems/features/interfaces
- Discover and test useful scientific principles
- Establish benchmarks/standards/guidelines

McGrath’s Taxonomy of Research Methods

- Study systems in real use on real tasks in real work environments, i.e., observe under settings with conditions as natural as possible
- Field studies — Study systems in situ, disturbing as little as possible, e.g., with ethnography and interaction analysis (Class 3), contextual inquiry
- Field experiments — Observe impact of changing (ideally) one aspect of a work environment, e.g., in beta testing, studies of technological change and new technology introduction
Quadrant 2 — Experimental Strategies

- Study systems in a lab under controlled conditions, i.e., conditions concocted for research purposes
- Laboratory experiments — Carry out controlled experiments studying impacts of (ideally) one (or two) interface parameter(s) (later this class)
- Experimental simulations — Create in lab for experimental purposes a real system that is used by real users on (usually) artificially simplified tasks, e.g., user testing (last class), usability engineering

Quadrant 3 — Respondent Strategies

- Ask informants to tell us something about themselves and/or their work or about an interface, i.e., where the setting in which questions are asked plays no role
- Judgment studies — Ask respondents about an interface, e.g., in a demonstration (later this class), or with usability inspection (later this class)
- Sample surveys — Ask respondents about themselves and/or their work, i.e., with questionnaires, surveys, interviews (class 3)

Quadrant 4 — Theoretical Strategies

- Ask a theory to tell us something about people’s work and/or about an interface, i.e., no observation of behaviour, experiments, or questions are required
- Formal theory — Use a qualitative theory or some equations, e.g., design theory such as Norman’s 7 stages (classes 6 and 7), or behavioural theory, such as colour vision or Fitts’ Law (next week)
- Computer simulation — Use and run a computer model, e.g., human information processing theory (CSC 428F)

Respondent Strategies

- Judgment studies
  - Demonstrations
  - Usability inspection
    - Heuristic evaluation
    - Cognitive walkthroughs
- “Sample surveys” (Class 3)
  - Questionnaires
  - Surveys
  - Interviews

Demonstrations

- Method
  - Demonstrate system to:
    - Any warm body you can capture
    - Management, potential investors, journalists
    - Potential customers
    - Potential users
    - Potential business partners
  - Take detailed notes
- Role
  - Elicit reactions to user’s model, functionality, interface

Demonstrations

- Advantages
  - Get feedback early in prototype or system construction
  - You’re going to have to give demos anyway — why not learn from them?
- Disadvantages
  - System still rough, which introduces noise into process
- Examples
  - Happens on all projects
Usability Inspection

• Methods
  – Heuristic evaluation — Judgments by a panel of evaluators (e.g., 3 to 5) of the degree to which an interface satisfies a set of usability guidelines, followed by discussion and analysis
  – Cognitive walkthroughs (CSC428)

• Roles
  – Evaluation without users (contrast to usability tests, etc.)
  – Elicit expert opinions re user’s model, functionality, look & feel, etc.

Usability Inspection

• Advantages
  – Structured method of using accumulated wisdom of experts

• Disadvantages
  – Doesn’t take advantage of real insights from real users

• Example — Heuristic evaluation with 10 usability guidelines (Nielsen, BGBG, Fig. 2.7, p. 63)
  – Visibility of system status
  – Match between system and the real world
  – User control and freedom
  – Consistency and standards
  – Error prevention
  – Recognition rather than recall
  – Flexibility and efficiency of use
  – Aesthetic and minimalist design
  – Help users recognize, diagnose, and recover from errors
  – Help and documentation

Questions and Discussion

Controlled Experiments

• Method
  – Manipulate independent variables, system characteristics
  – Control for other variables
  – Measure dependent variables, user behaviour

• Roles
  – Understanding causes of user behaviour
  – Understanding factors influencing interface quality

Controlled Experiments

• Advantages
  – Strong statements about causality
  – Many experimental designs suitable for varying situations

• Disadvantages
  – Requires time, planning, may be expensive
  – Complex designs (more than 3 or 4 independent variables) are often difficult to interpret
  – May legitimize trivial research, and generate results of weak generalization (external validity)
Examples of Real Experiments

- Egan et al. study of searching with print text and electronic text (SuperBook), as a function of whether or not the search term appears in the document heading structure and/or the document text (BGBG, pp. 843-846)
- Myers and Buxton study of impact of using two hands for input on speed of carrying out tasks
- Baecker et al. study of impact of new method of presenting computer program source text on program reading comprehension

Tasks to Design and Run an Experiment

- Design
  - Choose independent variables
  - Choose dependent variables
  - Develop hypothesis
  - Choose design paradigm
  - Choose control procedures
  - Choose a sample size
- Pilot experiment
- Run experiment
- Analyze data
- Interpret results

The Problem: Effectiveness of New Method of Source Code Presentation

- Source code appearance makes inadequate use of capabilities of digital typography
- Potential to make code more readable, more comprehensible
- See book by Baecker and Marcus, Human Factors and Typography for More Readable Programs, Addison-Wesley, 1990
- On following slides, points that refer to an experimental study of our new presentation format indicated by **

Independent Variables

- Definition
  - Factor or treatment
  - The variable manipulated by the experimenter
- Options in experiment design
  - One independent variable: single factor design
  - Two or more independent variables: factorial design
- Characteristics of independent variables
  - Number of levels (2 or more)
  - Quantitative (length of menu) or qualitative (letter vs. number, mouse vs. trackball)
- ** In our example: new typesetting format or traditional presentation format
Dependent Variables

- **Definition**
  - Outcome measure
  - Variable measured by experimenter
  - Variable which should depend on the independent variable
- **Examples**
  - Accuracy
  - Number of subtasks completed in a given time period
- **Criteria for judging**
  - Sensitivity: Responsiveness to changes in independent variable
  - Reliability and consistency: Similar outputs for similar inputs
  - Validity: Measuring what you really want to measure
- **In our example, ability to comprehend program as measured by # of questions answered in given time**

Hypotheses

- Statement of hypothesized relationship between independent and dependent variables
- Statement of how relationship is to be examined or tested
- **Hypothesis in our example: reading comprehension as defined above is improved by new method of source code presentation**
- Typical paradigm for testing the hypothesis
  - Single factor randomized group design with two groups
  - More on next slide

Experimental Design Paradigms

- Between subjects or within subjects manipulation
- **Example: designs with one independent variable**
  - Between subjects (randomized group) design
    - One independent variable with 2 or more levels
    - Subjects randomly assigned to groups
    - Each subject tested under only 1 condition
  - Within subject (repeated measures) design
    - One independent variable with 2 or more levels
    - Each subject tested under all conditions
    - Order of conditions randomized or counterbalanced
- **In our example, within subjects chosen with two conditions, i.e., two sample programs**

Control Procedures

- Goal is to eliminate confound hypothesis, that there are alternative explanation(s), and thereby to increase internal validity
- To do this: Make sure there are no systematic differences between conditions other than the independent variable
- **What to control (next slide)**
- **How to control (slide after next)**
- **In our example, ensure that two sample programs are “identical” in length, complexity, difficulty**

What To Control

- Subject characteristics
  - Gender
  - Ability
  - Experience
- Task variables
  - Instructions
  - Materials used
- Environmental variables
  - Setting
  - Noise, light, etc.
- Order effects
  - Practice
  - Fatigue

How to Control

- Hold constant
  - **Use males only, or students from same class only**
- Randomize
  - **Subjects to groups**
- Balance
  - Same number of novices and experts
- Counterbalance
  - **Half (chosen randomly) get new presentation format first**
- Match
  - Subjects on ability, e.g., programming ability
  - **Materials (programs) on length, difficulty**
- Eliminate
  - **Experience, by using novices only**
Sample Size Selection

- More subjects —> more confidence in results (greater statistical significance)
- But this can be very expensive
- Many methods to reduce the required number of subjects
- Most HCI experiments: 4 to 20 subjects per group
- ** In our example, 44 subjects chosen from a 3rd year programming course

Designing and Running the Experiment and Collecting the Data

- Run pilot studies
  - Check experimental design
  - Test and improve:
    - Task definition
    - Experimental materials (often the most difficult)
    - Instructions
  - Practice tasks
  - Develop experimenter skills
  - Identify and deal with special problems
- Run actual experiment
  - Record data
  - Observe behaviour

** The Presentation Format Experiment

- Within-subjects design, 44 subjects from 3rd year programming course
- Two "similar" short C programs, roughly 200 lines of code, 4 to 5 pages
- 40 minutes to skim first program and attempt to answer 18 questions, half in familiar format and half in new format
- Then each group given other program in other format

Data Analysis and Hypothesis Testing

- Describe data
  - Descriptive statistics (means, medians, standard deviations)
  - Graphs and tables
- Perform statistical analysis of results
  - Are results due to chance? (That is, with what probability)
  - **In our example, mean percentage of correct answers with new format = 44%, with conventional format = 35%
  - **Analysis of variance showed that effect of presentation format in increasing "program readability" was significant, F(1,42)=18.25, p<0.0001.

Interpretation of Results

- Consider plausible causes of differences (internal validity)
  - The independent variable or confounding variables
- Describe limits to generalization (external validity)
  - Variables held constant
    - **200 line programs
  - Task limitations
    - **Skimming programs, answering simple questions
    - Subject characteristics
    - **3rd year computer science students

Quasi-experiments

- Experiments that lack statistical significance (i.e., not enough subjects or individual variability too great for statistical significance) or that lack controls, lacks internal validity
- Typical method
  - Measure change of subjects' behaviour as system changes
  - E.g., study system as it evolves over time, measure performance of group of subjects both before and after experimental treatment such as modification of interface, icons, input devices
Quasi-experiments

• But this is not a controlled experiment
  – Same people used: learning is a confound
  – Subjects know system’s been refined: expectation is a confound
  – Multiple factors changed from v. n to n + 1: these are confounds

• Roles
  – Understanding effects of system change on user behaviour
  – Evaluation at far lower cost than controlled experiments

Examples of Quasi-experiments

• Bewley et al. tests on Star “graphics” (line drawing) functionality (B&B, pp. 662-667)
• Baecker, Small, Mander tests on “animated icons” (BGBG, pp. 444-449) — Confound is learning from test of static icons to test of animated icons
• Perkins et al. iterative design of Freestyle user interface plus tutorial (BGBG, pp. 881-885) — Confound is changing the interface plus the tutorial

Animated Icons (Baecker Small Mander 1991)

• Icons
  – Strengths
    • Compact
    • Quickly recognizable
    • “Universal” because language “not needed”
  – Weaknesses
    • Non-obvious
    • Hard to scale to deal with large numbers
    • Typically, now, both word and image

• Animated icons: a way to improve the comprehensibility of static icons
  – Dynamic visual representations of functions

User Comprehension of Static & Animated Icons
Evaluation of Animated Icons

- 8 subjects with varying degrees of familiarity with paint tools
- Asked to explain static icons, then asked again after viewing animations
- Animations helpful in explaining Selection, Lasso, Paint, Curve, both Polygon tools (where users had trouble with static icons)
- Sound compelling, but not tested

Internal Validity

- Degree of confidence that we’ve found “the” explanation for our results, that is, we know of no other confounding explanations
- Achieve by increasing precision and direct control over the experiment

External Validity

- Degree to which our research applies to other phenomena than just the “experiment”
- Achieve this by increasing range, scope, of phenomena studied

Tradeoffs among Empirical Methods

- Tradeoff between internal validity (soundness) and external validity (generalizability, relevance, realism)
  - Controlled experiments for internal validity
  - Breadth of naturalistic observation for external validity
- “Credible empirical knowledge requires consistency or convergence of evidence across studies based on different methods.” (McGrath, in BGBG, p. 155)
- Different strategies and methods have different advantages and disadvantages — cannot simultaneously maximize:
  - Generalizability of evidence over populations of actors (A)
  - Precision of measurement of the behaviours (B)
  - Realism of the situation or context (C)

McGrath’s Taxonomy of Research Methods

<table>
<thead>
<tr>
<th>DESIGN</th>
<th>IMPLEMENT</th>
<th>ANALYZE AND EVALUATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information collection and requirements analysis</td>
<td>Questionnaires, interviews, observation of potential users</td>
<td>Task analyses, artifact analyses, “day in the life” problem scenarios</td>
</tr>
<tr>
<td>Activity, information &amp; interaction design</td>
<td>Initial design concepts</td>
<td>Design mockups, prototypes, activity scenarios</td>
</tr>
<tr>
<td>Prototyping and prototype system</td>
<td>System functionality and look &amp; feel</td>
<td>&quot;Smoke and mirrors&quot; prototype, partially working system</td>
</tr>
<tr>
<td>Production prototype and its evolution</td>
<td>Complete system, incorporating evaluation insights</td>
<td>Real working system, implemented and installed</td>
</tr>
<tr>
<td>Production system and its evolution</td>
<td>Deliverable system, including monitoring and feedback system</td>
<td>Production system, including monitoring and feedback system</td>
</tr>
</tbody>
</table>
Research and Evaluation Methods in the Design and Development Process

- Information collection
  - Interviews and questionnaires
  - Contextual inquiry
  - Ethnography and interaction analysis
- Concept design
  - Interviews
  - Heuristic evaluation
  - Usability testing
  - Controlled experiments
- Functionality (activity) and interface (information & interaction) design
  - Heuristic evaluation
  - Usability testing
  - Theory-based evaluations
  - Human information processing simulations

Questions and Discussion