Empirical Research Methods: Introduction
Questions Today

Questions

1. What empirical research methods are used in HCI?

2. What are the key differences from related methods in the social sciences?

Why not just of academic interest?

- Sometimes empirical methods from other fields are transferred uncritically, without understanding of important differences
1. What are the goals of empirical studies in HCI?
   - Powerful reasons not to conduct empirical studies

2. Under what conditions are the studies conducted?
   - Special difficulties in generalizing to relevant real conditions

3. What types of data are collected?
   - Need for rich qualitative data

4. How are the data analyzed and interpreted?
   - Emphasis on support for decision making

Choice of examples

- From instructor’s own recent empirical studies
  + Reference to important inside information

1. Purposes of Studies

   Purposes of Studies (1)

A. Development of general principles and models for HCI

Examples

- Predictive mathematical models
  - Example: predicting search time in screen hierarchies (see slide below)
- Validating guidelines and standards

Comments

- In the extreme case, similar to basic psychological research
- To some extent, the points made about other types of study apply
Purposes of Studies

B. Normative evaluation of systems or techniques

Example
- Study of controlled vs. automatic generation of recommendations (see slide below)

C. Discovery of specific problems and possibilities for improvement

Example
- Improving an email client

D. Analyzing user’s requirements

Example
- Web site for a retail chain

Models for Search in Screen Hierarchies

How long will it take a user to select an item within a hierarchy of menus or screens?
Especially simple case (expert user, fast system):
Total time = constant time per screen
* \( \log(\text{number of options}) / \log(\text{branching factor}) \)
Example Evaluation Experiment

Session 1: Acquiring User Models from Multi-Modal User Input (Sun July 15, 9:30 – 11:00 AM)

Harnessing Models of Users’ Goals to Mediate Clarification Dialog in Spoken Language Systems (Add to Hotlist)

Authors: Eric Horvitz, Tim Park

Time: Sun July 15, 9:30 – 10:00 AM

Hotlist Recommender Concepts (with your estimated interest level):
- Natural language dialog (-)
- Decision-theoretic methods (+)

Abstract: Speaker-independent speech recognition systems are being used with increasing frequency for command and control applications. To date, users of such systems must contend with their fragility to subtle changes in language input, and unfamiliar behaviors. We describe an approach to acquire additional information about the user that can be used to improve these systems.

The hotlist of the UM 2001 web site
Issue: automatic vs. controlled presentation of recommendations

Why Studies Tend to Be Skipped

Factors

1. Time pressure
   - Industrial deadlines
   - Rapid change of technology
     Need to understand a technology before it is obsolete
   - Virtually infinite number of questions that could be investigated

2. Excessive trust in intuition and experience

Remedy

- Cost-benefit analyses can demonstrate the value of empirical studies
Two extremes

A. Controlled experiments

• Available subjects
• Assigned tasks
• Controlled, unnatural environment

B. Field studies

• Real users
• Real tasks
• Actual context of use

From: Anna Reiter <anna.reiter2@i-u.de>

Subject: What I’d like to see at UM 2001

For me, the most important methodological approach in the area of user modeling is machine learning. Often, methods from this category are applied in web-based systems, or in systems that select specific news stories for individual users. I’m not interested in these last two types of application of machine learning.

Anything that deals specifically with the improvement of automobile safety would be especially interesting to me.

Another thing I’m interested in is systems that model some type of psychological state of the user, such as emotions or stress.

Best regards,
Anna Reiter
Generalizability Problems

Factors that make generalization of results especially difficult

1. Especially large individual differences among users
   • Domain knowledge, computer knowledge, abilities, motivation, goals, ...

2. Importance of specific context of use
   • Time pressure, distractions, nature of equipment used, ...

Great possibilities for bias to be introduced

• General preference for "new" over "old" techniques
• Difficulty of finding truly representative subjects

3. Types of Data
Types of Data Collected

Quantitative assessment of extent to which criteria have been met

• Success in task completion
• Learning time
• Execution time
• Frequency of errors
• ... (cf. slides from Class 1)

Qualitative observations

• Evidence of specific errors or misunderstandings
• Evidence of concepts, strategies, and methods used
Types of Data Collected

Low-level data

- Records of user behavior
  - Automatically created logs
  - Video recordings
- Eye tracking records
- Physiological indices
  - Example: Heart rate variability as an index of cognitive load

Explicit statements by users

Answers to questionnaire questions

- General satisfaction
- Attitudes and beliefs concerning specific issues

Thinking aloud

Answers to interview-type questions
Data from UM 2001 Experiment

Subjects' desires to make use of hotlist recommendations

Need for Rich Qualitative Data

Why is this need especially great?

- Purpose: analyzing users’ requirements
  - What is needed is a qualitative description of use
  - It is unclear in advance what the key issues will be

- Purpose: discovery of specific problems and possibilities for improvement
  - Problems can be of unforeseen types
  - Users can give valuable hints concerning improvements
4. Analysis and Interpretation

Types of Analysis

Concerning quantitative criteria

Statistical analyses:
- How large and reliable are the observed differences?
- What factors contribute to the observed effects?

Concerning low-level data
- Sometimes used just as a backup for special cases
- Methods for summarizing onto a high level
  - Eye tracking
    - Frequency distributions (see slide with example)
    - Protocol analysis: Attempt to test hypotheses about users’ procedures

Concerning qualitative objective data
- Interpretation in terms of
  - Causes of problems
    - Example: Why do people often forget to attach attachments?
  - Ways of viewing the task
Need to Support Decision Making \(^{(1)}\)

- Why is this need especially great?
  - Almost all of this research is oriented toward practical decision making
    Cf. the "Purposes of research" discussed above

Consequences for qualitative analyses

- Qualitative analyses are more prominent than in purely scientific research
- They tend to involve a good deal of interpretation on the part of the investigator

Need to Support Decision Making \(^{(2)}\)

Consequences for statistical analyses

- Type II errors cannot be viewed as unimportant
  - You can’t just say "We’re not yet sure there’s a difference, so let’s just wait for further studies"
- Effect sizes and confidence intervals are important, not just differences
  - As in application-oriented research more generally
  - "How much improvement can we expect A to bring over B, and what is the standard error of this estimate?"
1. Visual Clarity

4. Is the same type of information (e.g., instructions, menus, messages, titles) displayed:
   a. in the same location on the screen?
   b. in the same layout?

12. Are there standard procedures for carrying out similar, related operations (e.g., updating and deleting information, starting and finishing transactions)?

15. Overall, how would you rate $S$ in terms of visual clarity?

2. Consistency

The way the system looks and works should be consistent at all times

4. Is the same type of information (e.g., instructions, menus, messages, titles) displayed:
   a. in the same location on the screen?
   b. in the same layout?

12. Are there standard procedures for carrying out similar, related operations (e.g., updating and deleting information, starting and finishing transactions)?

15. Overall, how would you rate $S$ in terms of consistency?
3. Compatibility

The way the system looks and works should be compatible with user expectations.

3. Where abbreviations, acronyms, codes, and other alphanumeric information are displayed:
   a. are they easy to recognize and understand?
   b. do they follow conventions where these exist?

11. Is information presented in a way which fits UI's view of the task?

17. Overall, how would you rate $S$ in terms of compatibility?

Usability Laboratories
Typical Usability Labs: Photos (1)
Typical Usability Labs: Photos

Usability labs at Microsoft

Typical Usability Labs: Layout

MicroULab, a Mobile Usability Lab

[Diagram of Usability Lab layout]

[Image of MicroULab setup]

[Image of MicroULab in a suitcase]

[Link: http://www.usabilitylabs.com]

[Link: http://www.oclc.org/oclc/new/n229/ulab2.htm]
MicroULab, a Mobile Usability Lab

User groups
- Those who want to do usability evaluations in a natural context
- Those who are just starting to do usability evaluations and can’t afford a fully equipped usability lab

Features
- Two digital color cameras – for sharp, high resolution images
- Remote control from console – to avoid user interruption
- Three LCD color monitors – to preview video input sources
- Large 10” LCD color monitor – for composite view of images
- 1024x768 scan converter – to capture computer image
- Two audio systems
  - Desk microphone – to record user comments
  - Wireless microphone – to communicate with user

The DRUM Tool for Video Analysis

DRUM Example Screen
DRUM is a software tool which has been developed by close cooperation between Human Factors professionals and software engineers to provide a broad range of support for video-assisted observational studies. It has been in use in commercial organisations since 1992 and has been continuously improved in consultation with NPL’s industrial and research clients.

DRUM directly supports the MUSiC Performance Measurement Method for usability evaluation. Evaluation sessions are recorded on video and subsequently analysed with the help of DRUM.

**DRUM’s role**

DRUM provides support for the usability analyst throughout the process of observation-based usability evaluation. DRUM assists in many aspects of the analyst’s work:

- management of data
- task analysis
- video control
- analysis of data

DRUM greatly speeds up the analysis of video recordings and automates some activities entirely, helps the analyst build up a time-stamped log of each evaluation session, and calculates performance measures and metrics.

DRUM delivers evaluation data in a format compatible with spreadsheets and statistical packages, for further analysis and graphical display.

DRUM also assists the generation and delivery of diagnostic feedback concerning usability defects to a product’s designers.

**Organising evaluations**

The DRUM Evaluation Manager makes it easy to organise data from all stages of usability evaluation. It gives you quick and simple access to evaluation data about:

- users – the people being observed in an evaluation
- tasks – analytic schemes describing the tasks which users perform
- video recordings of evaluation sessions
- logs of user and system activities, created by DRUM
- measures – derived from analysing logged task performance (times, problems, etc.)
- usability metrics – calculated values for individual users and groups
- reports of evaluation findings

DRUM uses text files for data storage, allowing flexible compatibility with word processors, spreadsheets and statistics packages.
Identifying significant events
You may wish to look out for many different kinds of event when studying a video record of an evaluation session. DRUM provides a basic set of event types to support the Performance Measurement Method. With the DRUM Scheme Manager you can define your own event types, and describe the tasks to be performed by users at up to five levels of detail. Each activity is represented on screen as an event button, with associated editable definition and comments.

Analysing video records
The DRUM Recording Logger helps you to build up a time-stamped log marking all the significant events observed on a video. Events are logged with respect to a timecode recorded on the videotape, enabling reliable and efficient access to any part of the video recording at any time. You can add comments to individual logged events.

DRUM gives fully responsive and error-preventative remote-control of the video recorder. As well as the usual video controls, DRUM includes a variable-speed shuttle, and offers automated location and playback of any logged event on the video.

Deriving measures and metrics
The DRUM Log Processor provides automated calculation from any log in the DRUM database of performance measures and performance-based usability metrics, including:

- task time
- snag, help and search times
- effectiveness
- efficiency
- relative efficiency
- productive period

Measures and metrics are presented in tabular and graphical displays. Results for individual users can be grouped, and exported to a spreadsheet or statistics package for further analysis.

Processing analyst-defined events
The DRUM Diagnostic Processor enables you to select any combination of events to generate tables of counts and durations, or comments. The selection mechanism offers great flexibility – allowing events to be selected by type or by time. In addition, any selected subset of the log can be saved as a new log for further processing. Selection templates allow you to define selection criteria to be applied quickly on subsequent logs.

Usability of DRUM
DRUM has an easy to learn graphical user interface, a constantly accessible on-line help system offering context-sensitive and general help, and comprehensive documentation.

What you need to run DRUM – technical information
DRUM requires:

- Apple Macintosh running System 7 or 8
- At least 640 x 480 pixel monitor
- HyperCard 2.1 (or later), allocated at least 2.5 MB RAM

DRUM can at present drive the following video recorders:

- Sony U-Matic VO 7000 and 9000 series, with BKU 701 computer interface
- Sony UVW 1400 series
- Panasonic AG-7350 or AG-7355 with AG-IA232TC adaptor

Other products and services
Additional information leaflets are available for the following:

- Usability Engineering Services
- Usability Services Training Courses
- Usability Context Analysis
- Usability Benefits
- Usability Laboratory Facilities
- Regulations, Standards and Quality
- MUSiC Tools and Methods
- SUMI – Software Usability Measurement Inventory

Usability Services
National Physical Laboratory
Teddington, Middlesex
United Kingdom
TW11 0LW

Tel: 0181-943 7019
Fax: 0181-977 7091
E-mail: usability@npl.co.uk
http://www.npl.co.uk/npl/sections/us/

NPL Helpline 0181-943 6880
Helpline fax 0181-943 6458
**Eye Tracking for Web Site Analysis**

**Eyetracking.com**

Eyetracking.com delivers high quality eye-tracking services to improve your Internet products. We offer you:

- real-time videos that show where users are looking
- detailed graphical and statistical reports
- next-day results

what our clients say:

“Excellent, Excellent, & Excellent! Experiencing EyeTracking testing in person was an extraordinary event for me; we saw a wide range of user ability levels from very novice to expert.... Yes - it was a great SUCCESS! Thanks again for making this critical CBT design evaluation happen.”

Rob Watson, Instructional Designer
Matsushita Avionics Systems Corporation

---

**What is eye tracking?**

- the process of recording by camera & computer
  - where a person looks
  - how his eye changes as he does so

**Products and services**

- Real-time videos that show where users are looking
- Detailed graphical and statistical reports
- Next-day results
- Written summary report with recommendations
Work for Class 13

Reading

Chapter 11 - Evaluation Techniques

- 11.5 Evaluating the implementation, p. 415
  - Exception: Pp. 419–426, concerning statistical measures, do not have to be read
  - On p. 427, start reading again at the beginning of Section 11.5.2
- 11.6 Choosing an evaluation method, p. 436
- 11.7 Summary, p. 440

Chapter 16 - Hypertext, Multimedia and the World Wide Web

- 16.7 Design Focus: Designing for the World Wide Web, p. 602
- 16.8 Summary, p. 609

Design Project Assignment for Class 13 (1)

Part 1: Get the IBIS template

- Following the instructions on the course web site (available from midday on Wednesday, November 28th), acquire a template file for writing an IBIS argumentation

Part 2: Choose design decisions to analyze

- For each member of your group, choose one important issue in the design of your S that you would like to consider carefully
- What are the criteria for a good issue?
  - The best solution should not be obvious
  - Finding a good solution should be really important to the success of S
  - When analyzing it, you should be able to make use of the results of your observation/interview
Design Project Assignment for Class 13

Part 3: Write IBIS analyses of the design decisions

- For each of the issues chosen, write an IBIS argumentation exploring the most important considerations
  - Consider at least two possible answers to the main question
  - In your arguments, refer as much as possible to the results of your interview/observation
  - More generally, take into account the hints on Slides 376 – 382 from Class 10

- An appropriate length will be about 1 to 2 screenfuls of text (using the large font specified in the template) for each of the issues chosen

- To summarize for clarity:
  - If your group has two members, you will submit analyses of two issues, each of which being about 1 to 2 screenfuls long