Evaluating User Interfaces

Lecture slides modified from Eileen Kraemer's HCI teaching material
Department of Computer Science
University of Georgia

Outline

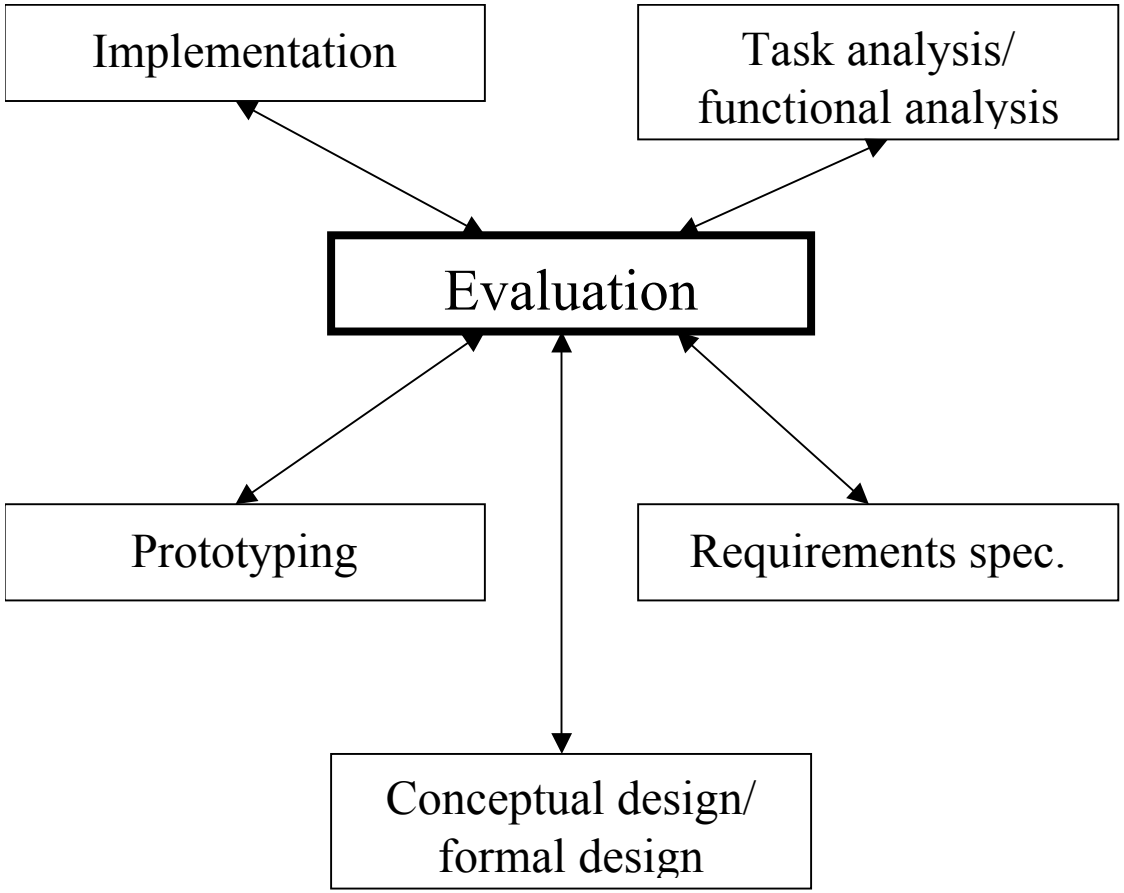
- The Role of Evaluation
- Usage Data: Observations, Monitoring, User's Opinions
- Interpretive Evaluation
- Predictive Evaluation

The Role of Evaluation

In the HCI Design model:

- the design should be **user-centred** and involve users as much as possible
- the design should **integrate** knowledge and expertise from different disciplines
- the design should be highly **iterative** so that testing can be done to check that the design does indeed meet user requirements

The star life cycle



Evaluation

- Evaluation
  - tests usability and functionality of system
  - occurs in laboratory, field and/or in collaboration with users
  - evaluates both design and implementation
  - should be considered at all stages in the design life cycle

Evaluation

- Concerned with gathering data about the usability of
  - a design or product
  - by a specific group of users
  - for a particular activity
  - in a specified environment or work context
- Informal feedback …… controlled lab experiments

Goals of Evaluation

- assess extent of system functionality
- assess effect of interface on user
- identify specific problems

What do you want to know? Why?

- What do users want?
- What problems do they experience?

- **Formative** -- early and often; closely coupled with design, guides the design process
- **Summative** -- judgments about the finished product; near end; have we done well?

Reasons for doing evaluations

- Understanding the real world
  - How employed in workplace?
  - Better fit with work environment?
- Comparing designs
  - compare with competitors or among design options
- Engineering towards a target
  - x% of novice users should be able to print correctly on first try
- Checking conformance to a standard
  - screen legibility, etc.

When and how do you do evaluation?

- Early to
  - Predict usability of product or aspect of product
  - Check design team's understanding of user requirements
  - Test out ideas quickly and informally
- Later to
  - identify user difficulties / fine tune
  - improve an upgrade of product

Case Study: 1984 Olympic Messaging System

- Voice mail for 10,000 athletes in LA -> was successful
- Kiosks placed around Olympic village -- 12 languages
- Approach to design (user-centered design)
  - printed scenarios of UI prepared, comments obtained from designers, management prospective users -> functions altered, dropped
  - produced brief user guides, tested on Olympians, families& friends, 200+ iterations before final form decided
  - early simulations constructed, tested with users --> need ‘undo’
  - toured Olympic villlage sites, early demos, interviews with people involved in Olympics, ex-Olympian on the design team -> early prototype -> more iterations and testing

Case Study: 1984 Olympic Messaging System

- Approach to design (continued)
  - “Hallway” method: -- put prototype in hallway, collect opinions on height and layout from people who walk past
  - “Try to destroy it” method -- CS students invited to test robustness by trying to “crash” it
- Principles of User-Centered Design:
  - focus on users & tasks early in design process
  - measure reactions using prototype manuals, interfaces, simulations
  - design iteratively
  - usability factors must evolve together

Case Study: Forte Travelodge

- System goal: more efficient central room booking
- IBM Usability Evaluation Centre, London
- Evaluation goals:
  - identify and eliminate problems before going live
  - avoid business difficulties during implementation
  - ensure system easy to use by inexperienced staff
  - develop improved training material and documentation

The Usability Lab

- Similar to TV studio: microphones, audio, video, one-way mirror

Particular aspects of interest

- System navigation, speed of use
- screen design: ease of use, clarity, efficiency
- effectiveness of onscreen help and error messages
- complexity of keyboard for computer novices
- effectiveness of training program
- clarity and ease-of-use of documentation

Procedure

- Developed set of 15 common scenarios, enacted by cross-section of staff
- eight half-day sessions, several scenarios per session
- emphasize that evaluation is of system not staff
- video cameras operated by remote control
- debriefing sessions after each testing period, get info about problems and feelings about system and document these

Results:

- Operators and staff had received useful training
- 62 usability failures identified
- Priority given to:
  - speed of navigation through system
  - problems with titles and screen formats
  - operators unable to find key points in doc
  - need to redesign telephone headsets
  - uncomfortable furniture
- New system: higher productivity, low turnover, faster booking, greater customer satisfaction

Evaluation Methods

- **Observing and monitoring usage**
  - field or lab
  - observer takes notes / video
  - keystroke logging / interaction logging
- **Collecting users' opinions**
  - interviews / surveys
- **Experiments and benchmarking**
  - semi-scientific approach (can't control all variables, size of sample)

Evaluation Methods

- **Interpretive Evaluation**
  - informal, try not to disturb user; user participation common
  - includes participatory evaluation, contextual evaluation
- **Predictive Evaluation**
  - predict problems users will encounter without actually testing the system with the users
  - keystroke analysis or expert review based on specification, mock-up, low-level prototype
- Pilot Study for all types!! -- small study before main study to work out problems with experiment itself
- Human Subjects concerns --

Usage Data: Observations, Monitoring, User's Opinions

- Observing users
- Verbal protocols
- Software logging
- Users' opinions: Interviews and Questionnaires

Direct Observation

- Difficulties:
  - people “see what they want to see”
  - “Hawthorne effect” -- users aware that performance is monitored, altering behavior and performance levels
  - single pass / record of observation usually incomplete
- Useful: early, looking for informal feedback, want to know the kinds of things that users do, what they like, what they don’t
- Know exactly what you’re looking for -> checklist/count
- Want permanent record: video, audio, or interaction logging

Eurochange System

- Machine that exchanges one form of European currency for another and also dispenses currency for credit/debit cards -- like an ATM machine
- Intended for installation in airports and railway stations
- Prototype machine installed in Oxford Street
- Your goal: find out how long average transaction takes; note any problems with user's experience
- Problems you might experience?

New school multimedia system

- Being tried out by groups of 13 year olds
- Don’t interfere with children’s activities – note the kinds of things they do and the problems they encounter …
- What difficulties might you encounter?

Indirect Observation: Video recording

- Solves some difficulties of direct observation
- Can be synchronized with keystroke logging or interaction logging
- Problems:
  - effort required to synchronize multiple data sources
  - time required to analyze
  - users aware they're being filmed
    - set up and leave for several days, they get used to it

Analyzing video data

- Task-based analysis
  - determine how users tackled tasks, where major difficulties lie, what can be done
- Performance-based analysis
  - obtain clearly defined performance measures from the data collected (frequency of task completion, task timing, use of commands, frequency of errors, time for cognitive tasks)
  - classification of errors
  - repeatability of study
  - time (5:1) -- tools can help

Verbal protocols

- User's spoken observations, provides info on:
  - what user planned to do
  - user's identification of menu names or icons for controlling the system
  - reactions when things go wrong, tone of voice, subjective feelings about activity
- "Think aloud protocol" -- user says out loud what he is thinking while working on a task or problem-solving
- Post-Event protocols -- users view videos of their actions and provide commentary on what they were trying to do

Think Aloud

- user observed performing task
- user asked to describe what he is doing and why, what he thinks is happening etc.

- Advantages
  - simplicity - requires little expertise
  - can provide useful insight
  - can show how system is actually use
- Disadvantages
  - subjective
  - selective
  - act of describing may alter task performance

Software Logging

- Researcher need not be present
- part of data analysis process automated
- Time-stamped keypresses
- Interaction logging-- recording made in real time and can be replayed in real time so evaluator can see interaction as it happened
- Neal & Simons playback system -- researcher adds own comments to timestamped log
- Remaining problems: expense, volume

Protocol analysis

- paper and pencil – cheap, limited to writing speed
- audio – good for think aloud, difficult to match with other protocols
- video – accurate and realistic, needs special equipment, obtrusive
- computer logging – automatic and unobtrusive, large amounts of data difficult to analyze
- user notebooks – coarse and subjective, useful insights, good for longitudinal studies

- Mixed use in practice.
- audio/video transcription difficult and requires skill.
- Some automatic support tools available

Eye tracking

- head or desk mounted equipment tracks the position of the eye
- eye movement reflects the amount of cognitive processing a display requires
- measurements include
  - fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  - saccades: rapid eye movement from one point of interest to another
  - scan paths: moving straight to a target with a short fixation at the target is optimal

Physiological measurements

- emotional response linked to physical changes
- these may help determine a user's reaction to an interface
- measurements include:
  - heart activity, including blood pressure, volume and pulse.
  - activity of sweat glands: Galvanic Skin Response (GSR)
  - electrical activity in muscle: electromyogram (EMG)
  - electrical activity in brain: electroencephalogram (EEG)
- some difficulty in interpreting these physiological responses - more research needed

Interviews and Questionnaires

- Structured interviews
  - predetermined questions, asked in a set way
  - no exploration of individual attitudes
  - structure useful in comparing responses, claiming statistics
- Flexible interviews
  - some set topics, no set sequence
  - interviewer can follow replies
  - less formal, for requirements gathering

Interviews, continued

- Semistructured interview
  - set of questions available for interviewer to draw on if interviewee digresses or doesn't say much
- Prompted interview
  - draw out more information from interviewee
    - based on screen design or prototype
    - or "… and what do you mean by …"

Example: semi-structured using checklist

- Why do you do this? (To get the user's goal.)
- How do you do it? (To get the subtasks -- ask recursively for each subtask)
- Why not do it this way instead? (Mention alternative -- in order to get rationale for choice of method actually used.)
- What are the preconditions for doing this?
- What are the results of doing this?
- May we see your work product?
- Do errors ever occur when doing this?
- How do you discover and correct these errors?

Variations on interviews

- Card sorting
  - users asked to group or classify cards to answer questions, answers recorded on data collection sheet
- Twenty questions
  - interviewer asks only yes/no questions

Interviews -- summary

- Focus is on style of presentation and flexibility of data gathering
- More structured -> easier to analyze
- Less structured -> richer information
- Good idea: transcribe interviews to permit detailed examination (also true for verbal protocols)

Questionnaires and surveys

- Focus is on preparation of unambiguous questions
- Again, pilot study important
- closed questions:
  - respondent selects from set of alternative replies
  - usually some form of rating scale
- open questions:
  - respondent free to provide own answer

Closed question - simple checklist

Can you use the following text editing commands?

	Yes	No	Maybe
DUPLICATE	[ ]	[ ]	[ ]
PASTE	[ ]	[ ]	[ ]

Closed question -- six-point scale

Rate the usefulness of the DUPLICATE command on the following scale:

very of no

useful |____|____|____|____|____|____| use

Closed question - Likert scale

Computers can simplify complex problems

strongly agree	agree	slightly agree	neutral	slightly disagree	disagree	strongly disagree

Closed question - semantic differential

Rate the Beauxarts drawing package on the following dimensions:

	extremely	quite	slightly	neutral	slightly	quite	extremely	
easy								difficult
clear								confusing
fun								boring

Closed question - ranked order

Place the following commands in order of usefulness (use a scale of 1 to 4 where 1 is the most useful)

___ PASTE

___ DUPLICATE

___ GROUP

___ CLEAR

Questionnaires

- Responses converted to numerical values
- Statistical analysis performed (mean, std_dev, SPSS often used if more statistical detail required)
- Increase chances of respondents completing and returning:
  - short
  - small fee or token
  - send copy of report
  - stamped, self-addressed envelope
- Pre- / post- questionnaires

Questionnaires

- Need careful design
  - what information is required?
  - how are answers to be analyzed?

- Styles of question
  - general
  - open-ended
  - scalar
  - multi-choice
  - ranked

How to write a good survey

- **Write a short questionnaire**
  - what is essential to know? what would be useful to know? what would be unnecessary?
- **Use simple words**
  - Don’t: "What is the frequency of your automotive travel to your parents' residence in the last 30 days?"
  - Do: "About how many times have you driven to your parent's home in the last 30 days?"

How to write a good survey

- **Relax your grammar**
  - if the questions sound too formal.
  - For example, the word "who" is appropriate in many instances when "whom" is technically correct.
- **Assure a common understanding**
  - Write questions that everyone will understand in the same way. Don't assume that everyone has the same understanding of the facts or a common basis of knowledge. Identify even commonly used abbreviations to be certain that everyone understands.

How to write a good survey

- **Start with interesting questions**
  - Start the survey with questions that are likely to sound interesting and attract the respondents' attention.
  - Save the questions that might be difficult or threatening for later.
  - Voicing questions in the third person can be less threatening than questions voiced in the second question.
- **Don't write leading questions**
  - Leading questions demand a specific response. For example: the question "Which day of the month is best for the newly established company-wide monthly meeting?" leads respondents to pick a date without first determining if they even want another meeting.

How to write a good survey

- **Avoid double negatives**
  - Respondents can easily be confused deciphering the meaning of a question that uses two negative words.
- **Balance rating scales**
  - When the question requires respondents to use a rating scale, mediate the scale so that there is room for both extremes.

How to write a good survey

- **Don't make the list of choices too long**
  - If the list of answer categories is long and unfamiliar, it is difficult for respondents to evaluate all of them. Keep the list of choices short.
- **Avoid difficult concepts**
  - Some questions involve concepts that are difficult for many people to understand.

How to write a good survey

- **Avoid difficult recall questions**
  - People's memories are increasingly unreliable as you ask them to recall events farther and farther back in time. You will get more accurate information from people if you ask about the recent past (past month) versus the more distant past (last year).

- **Use Closed-ended questions rather than Open-ended ones**
  - Closed-ended are useful because the respondents know clearly the purpose of the question and are limited to a set of choices where one answer is right for them. Easier to analyze.
  - An open-ended question is a written response. For example: "If you do not want a company picnic, please explain why". .. Can provide new ideas/info.

How to write a good survey

- **Put your questions in a logic order**
  - The issues raised in one question can influence how people think about subsequent questions.
  - It is good to ask a general question and then ask more specific questions..
- **Pre-test your survey**
  - First test to a small number of people.
  - Then brainstorm with them to see if they had problems answering any questions. Have them explain what the question meant to them.

How to write a good survey

- **Name your survey**
  - If you send it out by email, it may be mistaken for “spam”. Also want to pique the interest of the recipients.
  - Here are examples of survey names that might be successful in getting attention:
    - Memo From the Chief Executive Officer
    - Evaluation of Services of the Benefits Office
    - Your Opinion About Financial Services
    - Free T-shirt Win a Trip to Paris
    - Please Respond By Friday
    - Free Subscription
    - Win a notebook computer
  - .. But some of these look like spam to me .. Proceed with caution.

How to write a good survey

- **Cover memo or introduction**
  - If sending by US mail or email, may still need to motivate recipient to complete it.
  - A good cover memo or introduction should be short and includes:
    - Purpose of the survey
    - Why it is important to hear from the respondent
    - What may be done with the results and what possible impacts may occur with the results.
    - Address identification
    - Person to contact for questions about the survey.
    - Due date for response

Interpretive Evaluation

- Contextual inquiry
- Cooperative and participative evaluation
- Ethnography

- rather than emphasizing statement of goals, objective tests, research reports, instead emphasizes usefulness of findings to the people concerned
- good for feasibility study, design feedback, post-implementation review

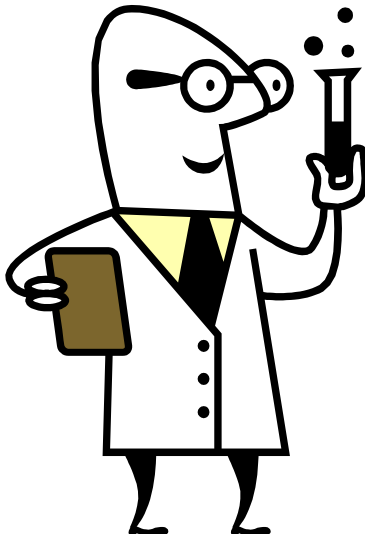
Interpretive Evaluation

- Experimental: Formal and objective
- Interpretive: More subjective
  - Concerned with humans, so no objective reality
  - Sociological, anthropological approach

- Users involved, as opposed to predictive approaches

Beliefs

- Sees limitations in scientific hypothesis testing in closed environment
  - Lab is not real world
  - Can’t control all variables
  - Context is neglected
  - Artificial, short tasks



Contextual Inquiry

- Users and researchers participate to identify and understand usability problems within the normal working environment of the user.
- Makes use of the **contextual interview**.
- *Recommendations to evaluator:*
  - Get as close to work as possible
  - Uncover work practice hidden in words
  - Create interpretations with customers
  - Let customers expand the scope of the discussion

Contextual Inquiry

- Users and researchers participate to identify and understand usability problems within the normal working environment of the user
- Differences from other methods include:
  - work context -- larger tasks
  - time context -- longer times
  - motivational context -- more user control
  - social context -- social support included that is normally lacking in experiments

Why use contextual inquiry?

- Usability issues located that go undetected in laboratory testing.
  - Line counting in word processing
  - unpacking and setting up equipment
- Issues identified by users or by user/evaluator

Contextual interview: topics of interest

- Structure and language used in work
- individual and group actions and intentions
- culture affecting the work
- explicit and implicit aspects of the work

Cooperative evaluation

- A technique to improve a user interface specification by detecting the possible usability problems in an early prototype or partial simulation
- low cost, little training needed
- think aloud protocols collected during evaluation

Cooperative Evaluation

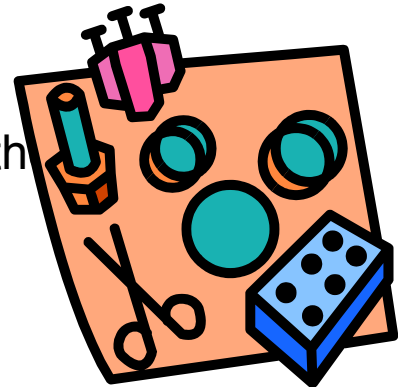
- Typical user(s) recruited
- representative tasks selected
- user verbalizes problems/ evaluator makes notes
- debriefing sessions held
- Summarize and report back to design team

Participative Evaluation

- More open than cooperative evaluation
- subject to greater control by users
- cooperative prototyping, facilitated by
  - focus groups
  - designers work with users to prepare prototypes
  - stable prototypes provided, users evaluate
  - tight feedback loop with designers

Types of Findings

- Can be both
  - Qualitative
    - Observe trends, habits, patterns, …
  - Quantitative
    - How often was something done, what per cent of th
      something occur, how many different …



Predictive Evaluation

- Predict aspects of usage rather than observe and measure
- doesn’t involve users
- cheaper

Why Predictive Evaluation

- User testing is expensive and time consuming, and requires a prototype
- Predictive techniques use expertise of human-computer interaction specialists (in person or via heuristics or models they develop) to identify usability problems without testing or (in some cases) prototypes

Predictive Evaluation Methods

- Inspection Methods
  - Standards inspections
  - Consistency inspection
  - Heuristic evaluation
  - Walkthroughs

- Modeling: The keystroke level model

Standards inspections

- Standards experts inspect the interface for compliance with specified standards
  - e.g., visibility of screen objects
- relatively little task knowledge required

Consistency inspections

- Teams of designers inspect a set of interfaces for a family of products
  - usually one designer from each project

Usage simulations

- Aka - “expert review”, “expert simulation”

- Experts simulate behavior of less-experienced users, try to anticipate usability problems
- more efficient than user trials
- prescriptive feedback

Usage Simulation (Expert Review)

- Pretend you are a novice user; identify usability problems
- Requires
  - Expertise in HCI
  - Expertise in the application area
  - Ability to role play the novice
  - Objectivity (not a developer)
- Problems
  - Bias of experts: use more than one
  - Hard to find experts
  - Real novices do the most unexpected things!

Heuristic evaluation

- Proposed by Nielsen and Molich.

- usability criteria (heuristics) are identified
- design examined by experts to see if these are violated

- Example heuristics
  - system behaviour is predictable
  - system behaviour is consistent
  - feedback is provided

- Heuristic evaluation `debugs' design.

Sample heuristics

- Use simple and natural dialogue
- speak the user's language
- minimize user memory load
- be consistent
- provide feedback
- provide clearly marked exits
- provide shortcuts
- provide good error messages
- prevent errors

Walkthroughs

- Goal - detect problems early on; remove
- construct carefully designed tasks from a system specification or screen mockup
- walk-through the activities required, predict how users would likely behave, determine problems they will encounter

Walkthroughs

- Structured form of usage simulation
  - Identify task, context, and user population
  - Walk through task, predicting user behavior
- Variations:
  - Cognitive walkthrough: simulate cognitive processing of user
  - Pluralistic walkthrough: multiple types of experts

Cognitive Walkthrough

Proposed by Polson *et al.*

- evaluates design on how well it supports user in learning task
- usually performed by expert in cognitive psychology
- expert ‘walks though’ design to identify potential problems using psychological principles
- forms used to guide analysis

Cognitive Walkthrough (ctd)

- For each task walkthrough considers
  - what impact will interaction have on user?
  - what cognitive processes are required?
  - what learning problems may occur?

- Analysis focuses on goals and knowledge: does the design lead the user to generate the correct goals?

Modeling: keystroke level model

- Goal: calculate task performance times for experienced users

- Requires
  - specification of system functionality
  - task analysis, breakdown of each task into its components

Keystroke-level modeling

- Time to execute sum of:
  - Tk - keystroking (0.35 sec)
  - Tp - pointing (1.10)
  - Td - drawing (problem-dependent)
  - Tm - mental (1.35)
  - Th - homing (0.4)
  - Tr - system response (1.2)

Keystroke Modeling Example

Save a file in application using mouse and pull down menu

1. Initial homing to mouse T_H = 0.4
2. Move cursor to file menu T_P + T_M = 1.35 + 1.10 = 2.33
3. Select “save as” in file menu (click, move, click): T_M + T_K + T_P + T_K = 0.35 + 1.35 + 1.10 + 0.35 = 7.05
4. Application prompts for file name T_R = 1.2; user types 8 characters: T_R + T_M + T_K*8 + T_K for return = 1.2 + 1.35 + 0.35*8 + 1.35 + 0.35 = 7.05

Total = 13.05

Choosing an Evaluation Method

when in process:	design vs. implementation
style of evaluation:	laboratory vs. field
how objective:	subjective vs. objective
type of measures:	qualitative vs. quantitative
level of information:	high level vs. low level
level of interference:	obtrusive vs. unobtrusive
resources available:	time, subjects, equipment, expertise

Example: Star Workstation, text selection

- Goal: evaluate methods for selecting text, using 1-3 mouse buttons
- Operations:
  - Point (between characters, target of move,copy, or insert)
  - Select text (character, word, sentence, par, doc)
  - Extend selection to include more text

Selection Schemes

	A	B	C	D	E	F	G
Button1	Point	Point	Point C Drwthru	Point C, W, S, P, D Drwthru	Point C, W, S, P, D, Drwthru	Point C Dthru	Point C, W, S, P, D
Button2	C Drwthru	C, W, S, P, D Drwthru	W, S, P, D Drwthru		Adjust	Adjust	Adjust
Button3	W, S, P, D Drwthru						

Methodology

- Between-subjects paradigm
- six groups, 4 subjects per group
- in each group: 2 experienced w/mouse, 2 not
- each subject first trained in use of mouse and in editing techniques in Star w.p. system
- Assigned scheme taught
- Each subject performs 10 text-editing tasks, 6 times each

Results: selection time

Time:

Scheme A :12.25 s

Scheme B: 15.19 s

Scheme C: 13.41 s

Scheme D: 13.44 s

Scheme E: 12.85 s

Scheme F: 9.89 s

Results: Selection Errors

- Average: 1 selection error per four tasks
- 65% of errors were drawthrough errors, same across all selection schemes
- 20% of errors were “too many clicks” , schemes with less clicking better
- 15% of errors were ‘click wrong mouse button”, schemes with fewer buttons better

Selection scheme: test 2

- Results of test 1 lead to conclusion to avoid:
  - drawthroughs
  - three buttons
  - multiple clicking
- Scheme “G” introduced -- avoids drawthrough, uses only 2 buttons
- New test, but test groups were 3:1 experienced w/mouse to not

Results of test 2

- Mean selection time: 7.96s for scheme G, frequency of “too many clicks” stayed about the same
- Conclusion: scheme G acceptable
  - selection time shorter
  - advantage of quick selection balances moderate error rate of multi-clicking

Experimental design - concerns

- What to change? What to keep constant? What to measure?
- Hypothesis, stated in a way that can be tested.
- Statistical tests: which ones, why?

Selecting subjects - avoiding bias

- Age bias -- Cover target age range
- Gender bias -- equal numbers of male/female
- Experience bias -- similar level of experience with computers
- etc. ...

Experimental Designs

- Independent subject design
  - single group of subjects allocated randomly to each of the experimental conditions
- Matched subject design
  - subjects matched in pairs, pairs allocated randomly to each of the experimental conditions
- Repeated measures design
  - all subjects appear in all experimental conditions
  - Concerns: order of tasks, learning effects
- Single subject design
  - in-depth experiments on just one subject

Critical review of experimental procedure

- User preparation
  - adequate instructions and training?
- Impact of variables
  - how do changes in independent variables affect users
- Structure of the tasks
  - were tasks complex enough, did users know aim?
- Time taken
  - fatigue or boredom?

Critical review of experimental results

- Size of effect
  - statistically significant? Practically significant?
- Alternative interpretations
  - other possible causes for results found?
- Consistency between dependent variables
  - task completion and error scores versus user preferences and learning scores
- Generalization of results
  - to other tasks,users, working environments?

Assignment #5

- Reading:
  - **Electronic Voting System Usability Issues by Bederson et al. CHI 2003**
- **Quiz on May 3 class**