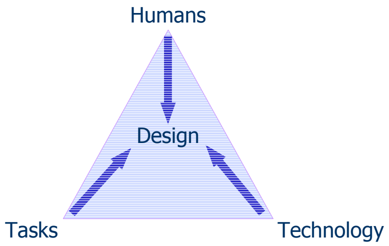
CS 590
Research Methods in HCI

Acknowledgement: Some of the material in these lectures is based on material prepared for similar courses by Saul Greenberg (University of Calgary), Ravin Balakrishnan (University of Toronto), James Landay (University of California at Berkeley), monica schraefel (University of Toronto), and Colin Ware (University of New Hampshire). Used with the permission of the respective original authors.

What is HCI?

Design, *Implementation* and *Evaluation* of interactive systems for HUMAN use.



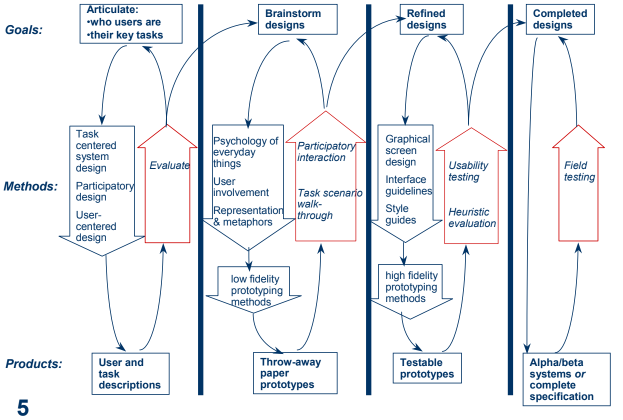
Examples of HCI Innovations

- mouse [Englebart, ’65]
- direct manipulation [Sutherland, ’63]
- desktop metaphor [Xerox Star, ’81]
- spreadsheet [VisiCalc, Fankston & Bricklin, ’77]

HCI – a multidisciplinary field

- Computer Science
- Psychology
- Sociology
- Education
- Anthropology
- Library Science
- Mechanical Engineering
- Industrial Engineering
- …

Interface Design and Usability Engineering



Goals:

Articulate:
•who users are
•their key tasks

Brainstorm designs

Refined designs

Completed designs

Methods:

Task centered system design

Participatory design

User-centered design

Evaluate

Psychology of everyday things

User involvement

Representation & metaphors

Participatory interaction

Task scenario walk-through

Graphical screen design

Interface guidelines

Style guides

Usability testing

Heuristic evaluation

Field testing

low fidelity prototyping methods

high fidelity prototyping methods

Products:

User and task descriptions

Throw-away paper prototypes

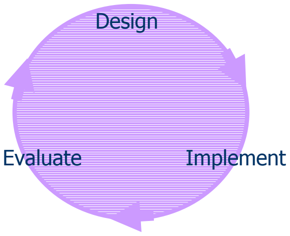
Testable prototypes

Alpha/beta systems or complete specification

Gould's article…

- How are the four principles of the usability design process reflected in the diagram?
- And the usability design phases?

Highest level abstraction

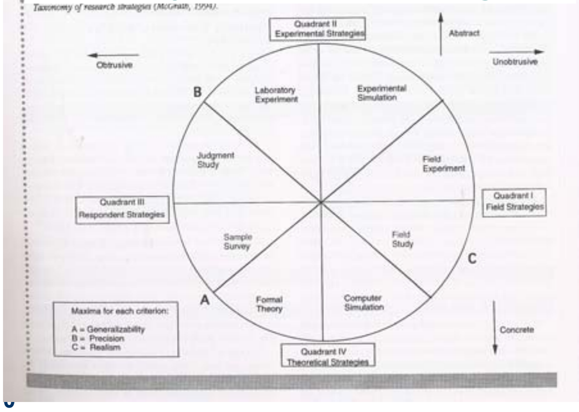


What do we learn from McGrath?

The “right” method – no such thing!

- Methods enable but also limit evidence
- All methods are valuable, but all have weaknesses or limitations
- You can offset the different weaknesses of various methods by using multiple methods
- You can choose such multiple methods so that they have patterned diversity, i.e., so the strengths of some methods offset the weaknesses of others

Taxonomy of Research Strategies



Taxonomy of research strategies (McGrath, 1994).

Maximization of 3 desirable features

A. **Generalizability** of the evidence over the populations of Actors

B. **Precision** of measurement of the behaviours that are being studied (and precision of control over extraneous factor that are not being studied)

C. **Realism** of the situation or context within which the evidence is gathered, in relation to the contexts to which you want your evidence to apply

Although you always want to maximize all three of these criteria, A, B, and C simultaneously, you *cannot* do so. This is the fundamental dilemma of the research process. Therefore, each study must be interpreted in relation to other evidence bearing on the same questions.

Quadrant I: Field Strategies

- Field Study
  - direct observations of “natural”, ongoing systems
  - minimal intrusion/disturbance of systems
  - e.g., cultural anthropology, “case studies”
- Field Experiment
  - within an ongoing natural system
  - some intrusion: one or more features of system manipulated
  - e.g., Hawthorne studies (vary lighting in organization)

Quadrant II: Experimental Strategies

- Lab experiment
  - concocted situation, rules of operation, individuals or groups engage in behaviours specified by rules
  - extraneous factors eliminated (which may or may not be relevant)
  - considerable precision
  - more obtrusive, reduced realism, less generalizable
  - e.g., unnatural task in a lab setting (target acquisition)
- Lab simulation
  - to gain some realism concocted situation made to seam natural
  - e.g., giving a natural task in a lab setting

Quadrant III: Respondent Strategies

- Sample Survey
  - evidence obtained to estimate the distribution of some variables, or relationships among them, within a specified population
  - careful sampling from that population
  - e.g., public opinion surveys
- Judgment Study
  - obtain information about the properties of a certain set of stimulus materials
  - focus is set of properties of stimulus materials, rather than attributes of the respondents
  - e.g., psychophysics studies (systematic relations between properties of the physical stimulus world and the psychological perception of those stimuli)

Quadrant IV: Theoretical Strategies

- Formal Theory
  - does not involve the gathering of any empirical observations
  - general relations among a number of variables of interest
  - based on earlier empirical evidence
  - e.g., model human processor, Fits Law
- Computer Simulation
  - complete and closed system that models the operation of the concrete system without any behaviour by any system participants
  - e.g., physics simulator

Comparison Techniques

- Baserates
  - must know how often Y occurs in the general case, to know if Y is some particular case is (not) notable
  - e.g., users can set up a network connection in less than 5 minutes in WinXP (is this an improvement?)
- Correlation
  - how the values of property X vary in relation to the values in property Y
  - not necessarily causal
  - e.g., number of files and time spent in Windows Explorer

Basic Experimental Design

- Independent variables
  - Factors that are manipulated in the experiment (e.g., *W, A* in Fitts' Law)
- Dependent variables
  - Factors that *may* depend on the independent variables (e.g., performance time)
- Wide range of independent variables
  - E.g. Fitts' law expt:
    - *W* 's range from character size (10) to icons (40) pixels
    - *D* 's from short (50) to large (screen size ~800 pixels)

Other experimental examples

	large screen	small screen
blue font	10	10
black font	10	10

reading task, dependent variable: reading performance

	Mac users	PC users
easy	15	15
medium	15	15
hard	15	15

formatting task, dependent variables: speed and accuracy

Randomization and "true experiments"

- can only control a small number of variables, what do you do with the others?
- have to do *something else* with all other factors
- randomization: random assignment procedure allocating "cases" to "conditions"
- does not guarantee an equal distribution of the extraneous factors, but makes an unequal distribution of any one factor highly unlikely
- statistical inference – selection and allocation of cases to conditions require random component to the procedure

Validity of Findings

- Internal validity
  - presence of X (or variations in level of X) caused the altered level of Y values
  - need to rule out plausible rival hypotheses
  - e.g., study comparing readability on small and large screens that finds small screen slows reading, when in fact it was the glare of the screen that caused the difference in performance
- Construct validity
  - the extent to which the methods used are in agreement with the theoretical concept (construct) of interest

- External validity
  - findings will be replicable (repeatable)
  - generalizable to intended population
  - no one study has external validity
  - typical threats:
    - non-representative users evaluated
    - non-representative tasks
    - non-representative environment (quiet lab vs. noisy office)

Measures and Manipulations

- record made by: actor, investigator, uninvolved third party
- degree to which actors aware of being observed impacts naturalness of behaviour
- Self-reports: participants knowingly report their own behaviour
- Observations: participants behaviour recorded by investigator or tool (visible vs. non-visible)
- Archival records: data recorded independent of study (public vs. private)
- Trace measures: records of behaviour without actors' awareness

Strengths and Weaknesses

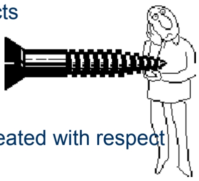
- Self-reports
  - questionnaires, interviews, rating scales, paper and pencil tests
  - frequently-used, very versatile, relatively cheap
  - potentially reactive
- Observations
  - by visible observer, potentially reactive
  - vulnerable to observer errors
  - can only be used on overt behaviour, not thoughts
  - versatile, costly
- Strength of one measure can compensate and offset weakness of another. Unlike study designs, investigator can and should use multiple measures.

Manipulating Variables

- Selection: select cases to be alike on a certain variable (e.g., Mac users vs. PC users)
  - not a true experiment, because not random
- Direct intervention: force the independent variable (e.g., small vs. large screen)
  - true experiment, but not always possible
- Inductions: less direct intervention
  - 3 ways: misleading instructions, false feedback, experimental confederates

Ethics in treatment of subjects

- Testing can be a distressing experience
  - pressure to perform, errors inevitable
  - feelings of inadequacy
  - competition with other subjects



- Golden rule
  - subjects should always be treated with respect

Managing subjects in an ethical manner

- Before the test
  - don’t waste the user’s time
    - use pilot tests to debug experiments, questionnaires etc
    - have everything ready before the user shows up
  - make users feel comfortable
    - emphasize that it is the system that is being tested, not the user
    - acknowledge that the software may have problems
    - let users know they can stop at any time
  - maintain privacy
    - tell user that individual test results will be kept completely confidential
  - inform the user
    - explain any monitoring that is being used
    - answer all user’s questions (but avoid bias)
  - only use volunteers
    - user must sign an informed consent form

Managing subjects in an ethical manner

- During the test
  - don’t waste the user’s time
    - never have the user perform unnecessary tasks
  - make users comfortable
    - try to give user an early success experience
    - keep a relaxed atmosphere in the room
    - coffee, breaks, etc
    - hand out test tasks one at a time
    - never indicate displeasure with the user’s performance
    - avoid disruptions
    - stop the test if it becomes too unpleasant
  - maintain privacy
    - do not allow the user’s management to observe the test

Managing subjects in an ethical manner

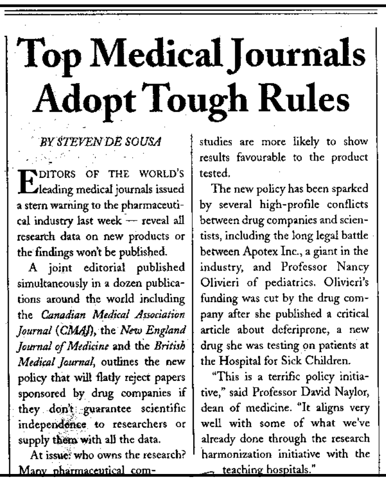
- After the test
  - make the users feel comfortable
    - state that the user has helped you find areas of improvement
  - inform the user
    - answer particular questions about the experiment that could have biased the results before
  - maintain privacy
    - never report results in a way that individual users can be identified
    - only show videotapes outside the research group with the user’s permission

University Involvement in Ethics

- Document evaluation protocol (strategy, methods, measures, number of subjects, subject recruitment, consent form, etc.)
- Document purpose of evaluation
- Submitted to Office of Research Studies (ORS)
- Reviewed by a committee (different committees for different kinds of evaluation)
- Usually 2 – 8 weeks for approval

Ethics in reporting

UofT Bulletin
24 Sept 2001
(also covered in
The Economist)



Top Medical Journals Adopt Tough Rules

BY STEVEN DE SOUSA

EDITORS OF THE WORLD'S leading medical journals issued a stern warning to the pharmaceutical industry last week — reveal all research data on new products or the findings won't be published.

A joint editorial published simultaneously in a dozen publications around the world including the *Canadian Medical Association Journal* (*CMAJ*), the *New England Journal of Medicine* and the *British Medical Journal*, outlines the new policy that will flatly reject papers sponsored by drug companies if they don't guarantee scientific independence to researchers or supply them with all the data.

At issue: who owns the research? Many pharmaceutical com-

studies are more likely to show results favourable to the product tested.

The new policy has been sparked by several high-profile conflicts between drug companies and scientists, including the long legal battle between Apotex Inc., a giant in the industry, and Professor Nancy Olivieri of pediatrics. Olivieri's funding was cut by the drug company after she published a critical article about deferiprone, a new drug she was testing on patients at the Hospital for Sick Children.

"This is a terrific policy initiative," said Professor David Naylor, dean of medicine. "It aligns very well with some of what we've already done through the research harmonization initiative with the teaching hospitals."

More on Observation

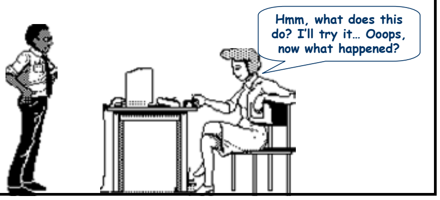
- Three general approaches:
  - simple observation
  - think-aloud
  - co-discovery learning

Simple Observation

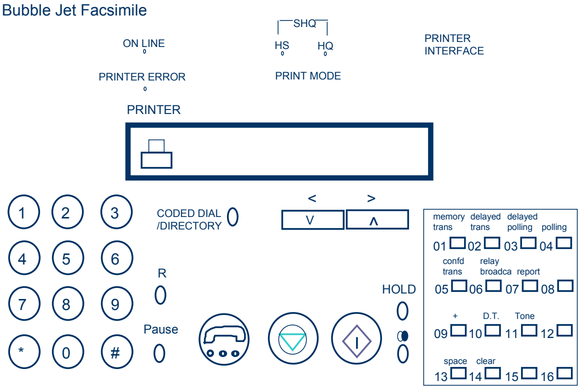
- User is given the task (or not), and evaluator just watches the user
- Problem
  - does not give insight into the user's decision process or attitude

The Think Aloud Method

- Subjects are asked to say what they are thinking/doing
    - what they believe is happening
    - what they are trying to do
    - why they took an action
  - – Gives insight into what the user is thinking
- Problems
    - awkward/uncomfortable for subject (thinking aloud is not normal!)
    - “thinking” about it may alter the way people perform their task
    - hard to talk when they are concentrating on problem
- Most widely used evaluation method in industry

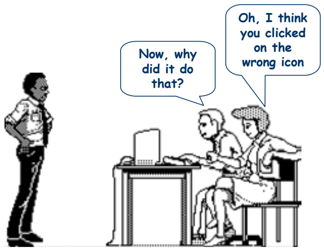

Hmm, what does this do? I’ll try it… Ooops, now what happened?

Canon
Fax-B320
Bubble Jet Facsimile



Co-discovery Learning

- Two people work together on a task
  - – normal conversation between the two users is monitored
    - removes awkwardness of think-aloud, more natural
    - provides insights into thinking process of both users



Recoding observations

- How do we record user actions during observation for later analysis?
    - if no record is kept, evaluator may forget, miss, or mis-interpret events
  - – paper and pencil
    - primitive but cheap
    - evaluators record events, interpretations, and extraneous observations
    - hard to get detail (writing is slow)
    - coding schemes or forms that just need to be ticked off
  - – audio recording
    - good for recording talk produced by thinking aloud/co-discovery interaction
    - hard to tie into user actions (i.e., what they are doing on the screen)
  - – video recording
    - can see and hear what a user is doing
    - one camera for screen, another for subject (picture in picture)
    - can be intrusive during initial period of use
    - Companies often build “usability labs” with one-way mirrors, video cams, etc.
  - – ideally have a system that synchronizes all these different records together

Querying Users via Interviews

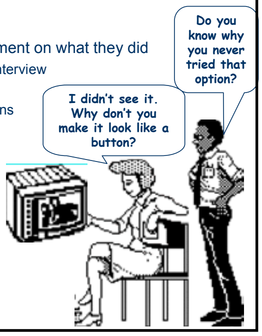
- Excellent for pursuing specific issues
  - vary questions to suit the context
  - probe more deeply on interesting issues as they arise
  - good for exploratory studies via open-ended questioning
  - often leads to specific constructive suggestions
- Problems:
  - accounts are subjective
  - time consuming
  - evaluator can easily bias the interview
  - prone to rationalization of events/thoughts by user
    - user's reconstruction may be wrong

How to interview

- Plan a set of central questions
  - could be based on results of user observations
  - gets things started
  - focuses the interview
  - ensures a base of consistency
- Structured interview – only ask planned questions
- Semi-structured interview – allow new questions to follow from answers to planned questions
- Try not to ask leading questions
- Start with individual discussions to discover different perspectives, and continue with group discussions
  - the larger the group, the more the universality of comments can be ascertained
  - also encourages discussion between users

Retrospective Interview

- Post-observation interview to clarify events that occurred during system use
  - perform an observational test
  - create a video record of it
  - have users view the video and comment on what they did
    - excellent for grounding a post-test interview
    - avoids erroneous reconstruction
    - users often offer concrete suggestions


Do you know why you never tried that option?
I didn't see it. Why don't you make it look like a button?

Querying users via Questionnaires and Surveys

- Questionnaires / Surveys
  - preparation “expensive,” but administration cheap
    - can reach a wide subject group (e.g. mail)
  - does not require presence of evaluator
  - results can be quantified
  - only as good as the questions asked

Querying Users via Questionnaires / Surveys

- establish the purpose of the questionnaire
  - what information is sought?
  - how would you analyze the results?
  - what would you do with your analysis?
- do not ask questions whose answers you will not use!
  - e.g. how old are you?
- determine the audience you want to reach
  - typical survey: random sample of between 50 and 1000 users of the product
- determine how would you will deliver and collect the questionnaire
  - on-line for computer users
  - web site with forms
  - surface mail
    - including a pre-addressed reply envelope gives far better response
- determine the demographics
  - e.g. computer experience

Styles of questions

- Open-ended questions
  - asks for unprompted opinions
  - good for general subjective information
    - but difficult to analyze rigorously

E.g., Can you suggest any improvements to the interfaces?

Styles of questions

- Closed questions
  - restricts the respondent's responses by supplying alternative answers
  - can be easily analyzed
  - but watch out for hard to interpret responses!
    - alternative answers should be very specific

Do you use computers at work:
[✓ often] O sometimes O rarely

vs

In your typical work day, do you use computers:
O over 4 hrs a day
[✓ between 2 and 4 hrs daily]
O between 1 and 2 hrs daily
O less than 1 hr a day

Styles of questions

- Scalar
  - ask user to judge a specific statement on a numeric scale
  - scale usually corresponds with agreement or disagreement with a statement

Characters on the computer screen are:

hard to read easy to read
1 2 [✓ 3] 4 5

Scale usually has an uneven length – why?

Styles of questions

- Multi-choice
  - respondent offered a choice of explicit responses

How do you most often get help with the system? (tick one)

○ on-line manual
☑ paper manual
○ ask a colleague

Which types of software have you used? (tick all that apply)

☑ word processor
○ data base
☑ spreadsheet
○ compiler

Styles of questions

- Ranked
  - respondent places an ordering on items in a list
  - useful to indicate a user's preferences
  - forced choice

Rank the usefulness of these methods of issuing a command
(1 most useful, 2 next most useful..., 0 if not used
__2__ command line
__1__ menu selection
__3__ control key accelerator

Styles of questions

- Combining open-ended and closed questions
  - gets specific response, but allows room for user's opinion

It is easy to recover from mistakes:

disagree agree comment: *the undo facility is really helpful*

1 2 3 [✓ 4] 5

Assessing any evaluation…

- What strategy, method, measures were used?
- What are the inherent weaknesses/strengths of the strategies, methods, measures?
- How (if at all) did the investigators mitigate/address the weaknesses? (Did they acknowledge the weaknesses?)
- Key: think of these questions when you are planning your own evaluation!

WRT last Friday’s readings…

- Which research strategies were used?
- Which methods were used?
- Internal, Construct, External validity?
- Which of the three desirable features (generalizability, precision, realism) were *least* achieved?
- What study design would increase that feature?

Readings

- McGrath, J. (1994). Methodology matters: Doing research in the behavioural and social sciences. (BGBG 152-169)
- Gould, J. (1988). How to Design Usable Systems, In Helander (Ed.), Handbook of Human-Computer Interaction. North-Holland: Elsevier, 1988, 757-789. (Excerpt reprinted with some additions in BGBG, p. 93 - 121)