Evaluation 1

John Stasko

Spring 2007

This material has been developed by Georgia Tech HCI faculty, and continues to evolve. Contributors include Gregory Abowd, Al Badre, Jim Foley, Elizabeth Mynatt, Jeff Pierce, Colin Potts, Chris Shaw, John Stasko, and Bruce Walker. Permission is granted to use with acknowledgement for non-profit purposes. Last revision: January 2007.

Agenda (for next 3 lectures)

- Evaluation overview
- Designing an experiment
  - Hypotheses
  - Variables
  - Designs & paradigms
- Participants, IRB, & ethics
- Gathering data
  - Objective; Subjective data
- Analyzing & interpreting results
- Using the results in your design

Evaluation, Part 1

- Evaluation overview
- Designing an experiment
  - Hypotheses
  - Variables
  - Designs & paradigms
- Participants, IRB, & ethics

Project Part 4

- All about evaluation
  - Use what you learn in next 3 classes

Why Evaluate?

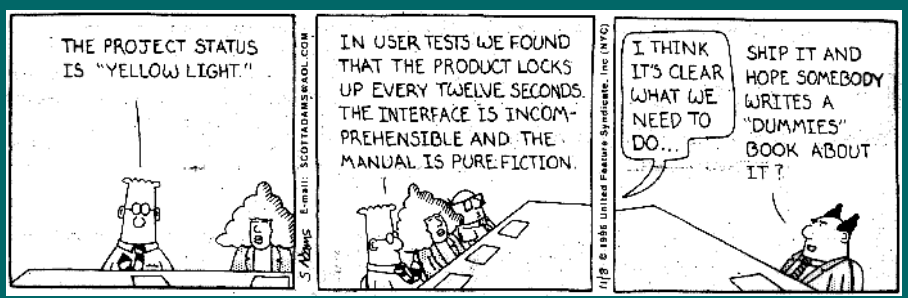
Recall:

- Users and their tasks were identified
- Needs and requirements were specified
- Interface was designed, prototype built

- *But is it any good? Does the system support the users in their tasks? Is it better than what was there before (if anything)?*

One Model

Evaluation can help your design…



Types of Evaluation

- Interpretive and Predictive (a reminder)
  - Heuristic evaluation, cognitive walkthroughs, ethnography, GOMS, …
- Summative vs. Formative
  - What were they, again?

Now With Users Involved

- Interpretive (naturalistic) vs. Empirical:
- Naturalistic
  - In realistic setting, usually includes some detached observation, careful study of users
- Empirical
  - People use system, manipulate independent variables and observe dependent ones

Why Gather Data?

- Design the experiment to collect the data to test the hypotheses to evaluate the interface to refine the design

- Information gathered can be: *objective* or *subjective*
- Information also can be: *qualitative* or *quantitative*

Which are tougher to measure?

Conducting an Experiment

- Determine the TASK
- Determine the performance measures
- Develop the experiment
- IRB approval
- Recruit participants
- Collect the data
- Inspect & analyze the data
- Draw conclusions to resolve design problems
- Redesign and implement the revised interface

The Task

- Benchmark tasks - gather quantitative data
- Representative tasks - add breadth, can help understand process
- Tell them what to do, not how to do it
- Issues:
  - Lab testing vs. field testing
  - Validity - typical users; typical tasks; typical setting?
  - Run pilot versions to shake out the bugs

"Benchmark" Tasks

- Specific, clearly stated task for users to carry out
- Example: Email handler
  - "Find the message from Mary and reply with a response of 'Tuesday morning at 11'."
- Users perform these under a variety of conditions and you measure performance

Defining Performance

- Based on the task
- Specific, objective measures/metrics
- Examples:
  - Speed (reaction time, time to complete)
  - Accuracy (errors, hits/misses)
  - Production (number of files processed)
  - Score (number of points earned)
  - …others…?

Types of Variables

- Independent
  - What you're studying, what you intentionally vary (e.g., interface feature, interaction device, selection technique)
- Dependent
  - Performance measures you record or examine (e.g., time, number of errors)

"Controlling" Variables

- Prevent a variable from affecting the results in any systematic way
- Methods of controlling for a variable:
  - Don't allow it to vary
    - e.g., all males
  - Allow it to vary randomly
    - e.g., randomly assign participants to different groups
  - Counterbalance - systematically vary it
    - e.g., equal number of males, females in each group
  - The appropriate option depends on circumstances

Hypotheses

- What you predict will happen
- More specifically, the way you predict the dependent variable (i.e., accuracy) will depend on the independent variable(s)
- "Null" hypothesis (H_o)
  - Stating that there will be no effect
  - e.g., "There will be no difference in performance between the two groups"
  - Data used to try to disprove this null hypothesis

Example

- Do people complete operations faster with a black-and-white display or a color one?
  - Independent - display type (color or b/w)
  - Dependent - time to complete task (minutes)
  - Controlled variables - same number of males and females in each group
  - Hypothesis: Time to complete the task will be shorter for users with color display
  - H_o: $Time_{color} = Time_{b/w}$
  - Note: Within/between design issues, next

Experimental Designs

- Within Subjects Design
  - Every participant provides a score for all levels or conditions

	Color	B/W
P1	12 secs.	17 secs.
P2	19 secs.	15 secs.
P3	13 secs.	21 secs.
...		

Experimental Designs

- Between Subjects
  - Each participant provides results for only one condition

	Color		B/W
P1	12 secs.	P2	17 secs.
P3	19 secs.	P5	15 secs.
P4	13 secs.	P6	21 secs.
...			

Within vs. Between

- What are the advantages and disadvantages of the two techniques?

Within Subjects Designs

- More efficient:
  - Each subject gives you more data - they complete more "blocks" or "sessions"
- More statistical "power":
  - Each person is their own control
- Therefore, can require fewer participants
- May mean more complicated design to avoid "order effects"
  - e.g. seeing color then b/w may be different from seeing b/w then color

Between Subjects Designs

- Fewer order effects
  - Participant may learn from first condition
  - Fatigue may make second performance worse
- Simpler design & analysis
- Easier to recruit participants (only one session)
- Less efficient

Now What…?

- You've got your task, performance measures, experimental design, etc.
- You have hypotheses about what will happen in the experiment
- Now you need to gather the data
- …So you need… PARTICIPANTS

IRB, Participants, & Ethics

- Institutional Review Board (IRB)
  - http://www.osp.gatech.edu/compliance.htm
- Reviews all research involving human (or animal) participants
- Safeguarding the participants, and thereby the researcher and university
- Not a science review (i.e., not to asess your research ideas); only safety & ethics
- Complete Web-based forms, submit research summary, sample consent forms, etc.
- All experimenters must complete NIH online history/ethics course prior to submitting

Recruiting Participants

- Various "subject pools"
  - Volunteers
  - Paid participants
  - Students (e.g., psych undergrads) for course credit
  - Friends, acquaintances, family, lab members
  - "Public space" participants - e.g., observing people walking through a museum
- Must fit user population (validity)
- Motivation is a big factor - not only $$ but also explaining the importance of the research
- Note: Ethics, IRB, Consent apply to *all* participants, including friends & "pilot subjects"

Ethics

- Testing can be arduous
- Each participant should consent to be in experiment (informal or formal)
  - Know what experiment involves, what to expect, what the potential risks are
- Must be able to stop without danger or penalty
- All participants to be treated with respect

Consent

- Why important?
  - People can be sensitive about this process and issues
  - Errors will likely be made, participant may feel inadequate
  - May be mentally or physically strenuous
- What are the potential risks (there are always risks)?
  - Examples?
- "Vulnerable" populations need special care & consideration (& IRB review)
  - Children; disabled; pregnant; students (why?)

Before Study

- Be well prepared so participant's time is not wasted
- Make sure they know you are testing software, not them
  - (Usability testing, not User testing)
- Maintain privacy
- Explain procedures without compromising results
- Can quit anytime
- Administer signed consent form

During Study

- Make sure participant is comfortable
- Session should not be too long
- Maintain relaxed atmosphere
- Never indicate displeasure or anger

After Study

- State how session will help you improve system ("debriefing")
- Show participant how to perform failed tasks
- Don't compromise privacy (never identify people, only show videos with explicit permission)
- Data to be stored anonymously, securely, and/or destroyed

Attribution Theory

- Studies why people believe that they succeeded or failed--themselves or outside factors (gender, age differences)
- Explain how errors or failures are not participant's problem---places where interface needs to be improved

Project

- IRB approval?
- P3 due Thursday after break
  - Prototype description
  - Evaluation plan & usability specs

Midterm Exam

- Grades
- Review

Upcoming

- Thursday – No class
  - Project work day
- More on evaluation (after break)
  - Gathering data
    - Recording, measuring, observing
    - Objective data
    - Subjective data, questionnaires
  - Analyzing Data, Interpreting Results
  - Usability specifications