Evaluation Paradigms, Associated Techniques, and Biases

Modified from materials originally created by Prof. Jamie Ruiz
Evaluation paradigms

There are 4 evaluation paradigms. These are ways to approach evaluation, not actual methods or techniques to perform evaluation.

A. “Quick and dirty” evaluations
B. Usability testing
C. Field studies / naturalistic studies
D. Analytical evaluations

Example techniques(s) of each paradigm are discussed after the paradigm.
Paradigm A. Quick and Dirty - 1

• What it is:
  – Informal feedback: low cost, fast, qualitative data
  – Feeds back into design process
  – Quick “reality checks” to get feedback quickly to help push design forward
  – Can quickly uncover misconceptions in your own head

• Typical “metrics”:
  – Does user understand system? How much do you need to explain for them to understand it?
  – Do they seem to quickly adopt to system? Does it seem to mesh with their expectations?
  – Do they look puzzled? Or do they react with enthusiasm?

• How to do it:
  – Show scenarios, storyboards, sketches, Woz the system for them …
  – …or simply tell them your idea, and sketch it out for them on the fly
Quick and Dirty - 2

• Who:
  – Real users
  – Consultants
  – HCI experts
  – Person two offices down

• Keep in mind:
  – *Any* feedback is better than no feedback
  – The sooner evaluation is performed, the cheaper it is to fix problems
  – In general, the less familiar the person is with project, the better their feedback will be
    • Will not bring preconceived notions

• Pros and Cons:
  – Positives: fast, cheap, can do it at ANY time, little planning needed, do not need complete or functional system for this testing
  – Negatives: strictly qualitative data, “spot checks” may not have a lot of generalizability
Quick & Dirty Paradigm Technique: Think Aloud

• What it is/what it provides:
  – A “quick and dirty” methodology: user thinks aloud as they use an application or prototype to accomplish a particular task
  – Results are qualitative data that provide insights into how user perceives and interprets the application
  – Data is provided/collection immediately: not a retrospective account. This increases the chance that timely, “accurate” data is given

• How to do it: PILOT THE TASK!
  – Decide your question(s): What do you want to learn? ...
  – Create a specific task for subject to perform
    • Structured tasks are much better than extremely vague tasks
    • You want the participant to focus on using the interface, not on figuring out what type of problems to solve

• Instructing the Subject:
  – Ask participant to say whatever comes to mind as they use the application
  – Tell them that anything that may be confusing is not their fault, but the fault of the application’s design
    • You want them to feel comfortable criticizing the application
    • Shift blame from internal to external attribution (participant to designer)
  – Be prepared to prompt subject to remind them to continue thinking aloud
    • “What are you thinking about now?”
Paradigm B. Usability Testing - 1

• What it is:
  – Carefully measure performance on prepared tasks typical for system’s intended uses
  – Emphasizes collection of *quantitative* data (qualitative data may also be collected e.g., video)
  – Performed in laboratory-like conditions (highly controlled environment)
  – User is given very specific, pre-determined tasks
  – Often includes detailed logging (video, keystroke, mouse, command use, software logging)

• Typical “metrics”:
  – Number of errors (error rate), time to complete task, number of tasks performed
  – Satisfaction, cognitive load (NASA TLX), learning time

• How to do it:
  – *Very formal process*
  – Usability test planned, piloted
  – Script prepared to ensure consistency for each test
  – Subjects recruited, consent gained
  – Subjects informed about usability test (using script)
  – Subjects perform usability test (using script)
  – Subjects debriefed on process (using script)
  – Data analyzed
Paradigm B. Usability Testing - 2

• Who:
  – Representative end-users; you can use other subjects, but ....

• Keep in mind:
  – Choice of subject significantly affects the validity of the results
    • If test with experts, but will be used by novices, results may be meaningless

• Pros and Cons:
  – Positives:
    • Quantitative data that lends itself to statistical analyses,
    • Can focus on very specific questions
  – Negatives:
    • Significant start-up costs (detailed plan, pilot testing, recruitment...)
    • Controlled environment strips out “reality” – lose the influence of peripheral, real-world activities
    • Need a functional system for subject to use (additional time to implement and test the evaluation system)
    • Care must be taken to develop authentic, representative tasks
Experiment Motivations

“Results from our tests indicate it took the user 30 seconds, on average, to input a new entry into the database.”

– If we’re trying to improve an interface, what does a result like this tell us?

“To test our interface, subjects first performed 10 tasks using the current interface. Subjects then performed the same 10 tasks using our new interface. Results indicate that our new interface improved task performance by 70%.”

– How much faith can we have in these results?

“We added shortcut keys, switched to a multiple-document GUI, ported it to the Mac, eliminated the need for a mouse to be used, and added type-ahead capabilities to our application. Our users are 50% faster because of the type-ahead feature.”

– What can you say about this claim?
Determining Cause and Effect

• To determine cause-effect relationships, we must carefully observe differences between when suspected cause is, and is not, present.

• Making tightly controlled *comparisons* is only way we can know how one particular change affects user.

• Experimental process provides framework for establishing cause-effect relationships.
Usability Testing Technique: Experimental Process

1. Formulate a hypothesis
2. Identify independent, dependent variables
3. Design a controlled experiment
4. Check for:
   – Validity
   – Reliability
   – Confounds
5. Pilot the study
6. Select representative subjects
7. Randomly assign to conditions
8. Analyze results
1. Formulate Hypothesis

- A suggested explanation of a phenomenon
  - “If I change A, then B will change thusly”
- In experimentation, we want the hypothesis to be as specific as possible: this makes it easier to test
- Hypothesis must be falsifiable; we should be able to show it is not true
  - “Using two mice, we believe users will be able to more quickly create rectangles of a given size than when using one mouse”
  - we just have to show 1 case where this isn’t true to disprove the hypothesis
- To test hypothesis, must identify what variables we think will lead to expected outcome
- Must identify how manipulating these variables will result in expected outcome
  - “If I provide keyboard shortcuts, users will be able to complete the tasks faster”
  - “If I use pie menus rather than vertically-oriented context menus, users will be able to select items faster”
- Clearly identify which variables will influence what outcomes, and how
2. Identify Independent, Dependent Variables

• Independent variables:
  – Variables that are manipulated by experimenter
    • Everything else held constant
  – *Levels* are the number of different conditions tested with an independent variable
    • Keyboard shortcuts vs. no shortcuts = 2 levels

• Dependent variables:
  – Variables that are measured
  – Are presumed to be *dependent* on manipulations of the independent variables
  – Examples
    • Time to complete task, number of errors made, user preference, quality of response
Variable Relationships

• Independent variables are assumed to produce an effect on dependent variables’ values when manipulated
  – “If I use pie menus rather than vertically-oriented context menus, users will be able to select items faster”
  – Pie menus vs. context menus (independent variables)
  – Item selection (dependent variable)

• If we only manipulate independent variables, this increases our confidence that any observed changes in dependent variables are due to changes in independent variables
Hypothesis Testing

• In testing hypothesis, we are seeking to disprove the *null hypothesis*

• Null hypothesis
  – There exists no relationship between manipulating the independent variables and the resultant changes in the dependent variables
3. Design the Experiment

- Need at least two conditions
  - Control condition
    - No experimental manipulations performed
  - Experimental condition
    - Experimental variable is manipulated
- Results are compared between two conditions
- Subject pool important
  - Do they represent the intended user population?
- Modify only one thing between conditions
- Ensure experimenter is not biasing subject
  - Blind vs. double-blind studies
- Random assignment to conditions and/or counterbalancing assignment to conditions essential
  - Want to control for learning, ordering effects
Main Experimental Study Designs

• Between-subjects
  – Also known as “randomized” design
  – Each subject exposed to only one condition
  – Positives: no learning effects
  – Negatives: problems possible with variations in subjects, need more people

• Within-subjects
  – Every subject experiences every condition (control and experimental)
  – Order of conditions usually counterbalanced to counter learning effects
  – Positives, need fewer subjects than between-subjects design
  – Negatives, must be sure to control for ordering/learning effects
4. Experiment Design Checks

• Validity
  – Are we measuring what we say we are measuring?

• Reliability
  – If we run the experiment several times, do we get the same results every time?

• Confounds
  – Are there variables we didn’t control for which may be influencing the results we’re obtaining?

• Examples
  – Testing new features after subjects do control condition; what’s the issue?
Running the Study

• The more subjects the better: increases confidence of findings
• For large effects, fewer subjects needed to notice effect is statistically significant: the smaller the effect, the more subjects needed
• Interpreting the results: consult your nearest, friendly statistics expert...
Applying Experimental Concepts

• You will most likely not have enough time to perform experimental studies for this class...

• But the concepts are applicable to understanding the quality of the data you collect

• Think about quantitative data you can collect as part of your evaluation and add it to your plan; small survey, task completion, task time experiments may be feasible, but make sure to address design checks (slide 47).

• Be sensitive to all of the potential confounds affecting the data you collect
  – Are people trying to be kind to you?
  – Is your instruction of how to use interface biasing people to like it more? To learn it more quickly?
  – What other explanations might apply to the results you observe?
Paradigm C. Field / Naturalistic Studies

• What it is:
  – Same qualitative methods you used to initially study users...
    • Interviews and Observations
  – Emphasis is on observing use of system in a natural environment
  – Subjects use system for real or representative tasks

• Typical “metrics”:
  – Does system mesh with desired work practices?
  – What parts of system do not seem to work well?
  – How do users actually adopt and use the system?

• How to do it:
  – Like Contextual Inquiry, identify users to study, identify tasks you want to observe, perform observations of them using the system
Paradigm C. Field / Naturalistic Studies

• Who:
  – Actual end users in their actual work environment

• Keep in mind:
  – Choosing tasks to observe is critical; choose work that has been re-designed

• Pros and Cons:
  – Positives:
    • A “true,” real-world test of system, complete with all peripheral activities and distractions expected of real-world use
  – Negatives:
    • May require a fully functional system
    • Data does not easily reduce to quantitative measures
    • Access to subjects can be difficult since you will likely be disrupting typical work practices
Field / Naturalistic Studies Technique: Observations

• What it is/what it provides:
  – Field/Naturalistic method
  – Rich qualitative data

• How to do it:
  – Identify specific work you want to observe. Usually having to do with the re-designed work. The task needs to be specific.
  – Identify users and arrange time (probably less than an hour this time)
  – Observe users as they work through the task; you provide the Wizard of Oz actions
Paradigm D. Analytical Evaluation

• What it is:
  – Evaluation techniques that make use of theoretical models of users and their capabilities
  – Heuristic evaluation, cognitive walkthrough, GOMS...
    • GOMS: Goals, Operators, Methods, Selection

• Typical “metrics”:
  – Depends on the technique/method used

• How to do it:
  – Techniques/methods prescribe a set of questions to ask of system
    • (Different techniques/methods ask different questions)
Paradigm D. Analytical Evaluation

• Who:
  – No subjects required

• Keep in mind:
  – Need a description of the interface, task, and actions required to perform the task

• Pros and Cons:
  – Positives:
    • No functioning system, no subjects required
    • Can uncover many significant problems relatively quickly
  – Negatives:
    • Lacks context of real-world situation
    • Real-world users may have different mental models than experts who developed the techniques/methods
Analytical Evaluation Technique 1: Heuristic Evaluation

• What it is/what it provides:
  – Analytical method
  – Structured critique of system using set of heuristics
  – Performed by expert evaluators on design specs or completed applications
  – Developed by Jakob Nielsen and Rolf Molich: Nielsen claims that 75% of overall usability problems discovered with 5 evaluators

• How to do it:
  – Evaluators assess application according to heuristics and any violations noted
  – Severity rated according to following criteria:
    • How common is problem?
    • How easy is it for user to overcome?
    • Will it be a one-off problem or a persistent problem?
    • How seriously will problem be perceived?
  – Problems rated according to following severity scale:
    • 0: I don’t agree this is a usability problem
    • 1: Cosmetic problem only
    • 2: Minor usability problem: Fixing it should be low priority
    • 3: Major usability problem: High priority to fix
    • 4: Usability catastrophe: Must fix before releasing software
Heuristic Evaluation Using Nielsen’s Usability Heuristics

1. Visibility of system status
   – Can user determine state of system at any point in time?
2. Match between system and real world
   – Does system communicate in language of user?
3. User control and freedom
   – Can users always return to a previous state if they make a mistake or if they explore?
4. Consistency and standards
   – Are terms and conventions used consistently within application and within platform?
5. Error prevention
   – Are constraints used to prevent errors from even occurring?
6. Recognition rather than recall
   – Make objects, actions, and options visible
7. Flexibility and efficiency of use
   – Provide accelerators for frequent actions
8. Aesthetic and minimalist design
   – Provide only the information necessary at any point in time
9. Help users recognize, diagnose, and recover from errors
   – Precisely indicate problems with possible solutions
10. Help and documentation
    – Provide easily accessible help with concrete examples

http://www.nngroup.com/articles/ten-usability-heuristics/
Analytical Evaluation Technique 2: Cognitive Walkthrough

• What it is/what it provides:
  – Analytical method
  – Experts analyze application or its design
  – Evaluation method intended to assess learnability of system

• How to do it:
  – Start with a detailed spec, prototype, or functioning system
  – Create a detailed description of a representative task the user is to perform
  – Develop a detailed, written list of actions needed to complete the task
  – Indicate who the users are, their experience level, and knowledge of application
  – Provide evaluators with the above information
  – Evaluators answer the following questions for each step of action sequence
Checks in a Cognitive Walkthrough

1. Is the effect of the action the same as the user’s goal at that point?
   – Example: Choosing animation for slides can switch out of outline mode in PowerPoint

2. Will users see that the action is available?
   – Question of visibility of system state with respect to what users want to do at a particular point in time

3. Once users have found the correct action, will they know it is the one they need?
   – Will they understand that the option offered by system is the option they need to complete the task?

4. After the action is taken, will users understand the feedback they get?
   – Does system communicate promptly, in the language of the user?
Experimental Biases in the Real World

• Hawthorne effect/John Henry effect

• Experimenter effect/Observer-expectancy effect

• Placebo effect

• Novelty effect
Hawthorne Effect

• Named after the Hawthorne Works factory in Chicago
  – Original experiment asked whether lighting changes would improve worker productivity
  – Found that anything they did improved productivity, even changing the variable back to the original level
  – Benefits stopped though; when the studying stopped, the productivity increase went away

• Why?
  – Motivational effect of interest being shown in workers

• Also, the flip side, the John Henry effect
  – Realization that you are in a control group makes you work harder
Experimenter Effect

• A researcher’s bias influences what they see
• Example from Wikipedia: music backmasking
  – From Wikipedia: “**Backmasking** is a recording technique in which a sound or message is recorded backward onto a track that is meant to be played forward. Backmasking is a deliberate process, whereas a message found through phonetic reversal may be unintentional.

  Backmasking was popularised by the Beatles, who used backward instrumentation on their 1966 album *Revolver*. Artists have since used backmasking for artistic, comedic and satiric effect, on both analogue and digital recordings. The technique has also been used to censor words or phrases for "clean" releases of explicit songs.”
  – **Once the subliminal lyrics are pointed out, they become obvious**

• Dowsing
  – Not more likely than chance

• The issue:
  – If you expect to see something, maybe something in that expectation leads you to see it

• Solved via double-blind studies
  – Neither subject nor researchers know who is receiving the control condition versus the experimental condition
Placebo Effect

• Subject expectancy
  – If you think the treatment, condition, etc., has some benefit, then it may

• Placebo-based anti-depressants, muscle relaxants, etc.
  – In computing, an improved GUI, a better device, etc.
Novelty Effect

• Typically with technology
• Performance improves when technology is instituted because people have increased interest in new technology
• Examples: Computer-Assisted instruction in secondary schools, computers in the classroom in general, etc.
How do you control for biases?

• Cannot fully
  – More an awareness issue
• Approach any test data with some skepticism
• Assume subjects are trying to be helpful, so any errors must be pretty serious
• Aggressively seek contradictory data