Goals of Usability Evaluation

Formative Evaluation: *What and how to re-design?*

Summative Evaluation: *How well did we do?*

Formative and Summative Goals

Formative: *during* development, guides process
Summative: *after* development, or at a *checkpoint*

- What sorts of “test data” might aid formative goals? What about summative goals?

- SBD relies on *mediated* usability evaluation
  - claims analysis documents design features of concern
  - users’ performance & reactions tied to these features, establishing a *usability specification*
Analytic and Empirical Methods

Analytic: theory, modeling, guidelines (from experts)
Empirical: observations, surveys (from users)

“If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axeman.”

- Which is more expensive? Why?
- Which carries more weight with developers? Why?

Usability Inspection

- Expert walk-through based on usability guidelines, often working from a checklist
  - generally want more than one expert (if affordable!)
- Guidelines (and walkthrough) can be at many levels
  - e.g., screen layout, detailed analysis of cognitive states
- May or may not use a standard set of tasks
  - depends on how comparable you want judgements to be
- Summarize by listing problems identified in each category, also often rating them for severity
Heuristic Evaluation

- Use simple and natural language
- Speak the users’ language
- Minimize memory load
- Be consistent
- Provide feedback
- Provide clearly marked exits
- Provide shortcuts
- Provide good error messages
- Prevent errors
- Include good help and documentation

Multiple experts review with respect to these issues

Can also include different classes of stakeholders, e.g. developers, users

Cognitive Walkthrough

- Form-based inspection of system or prototype
  - for walk up and use systems (affordances, metaphors)
  - careful task selection, then answer questions at each step
  - e.g., how obvious is the next action? what knowledge is assumed? what if any competing goals must be ignored?
- Check-list approach is attractive to practitioners
  - concerns with how to select the tasks for analysis, i.e. complexity/realism trades off with cost of evaluation
  - in practice, often can be used in more lightweight fashion, more of a “tuning” to issues
## Excerpt from Cognitive Walkthrough Form

...  
**Step [B] Choosing the Next Correct Action:**  
**[B.1] Correct Action:** Describe the action that the user should take at this step.  
**[B.2] Knowledge Checkpoint:** If you have assumed user knowledge or experience, update the USER ASSUMPTION FORM.  
**[B.3] System State Checkpoint:** If the system state may influence the user, update the SYSTEM STATE FORM.  
**[B.4] Action Availability:** Is it obvious to the user that this action is a possible choice here? If not, indicate why.  
**[B.5] Action Identifiability:**  
**[B.5.a] Identifier Location, Type, Wording, and Meaning:**  

| No identifier is provided. (Skip to subpart [B.5.d])  
| Identifier type: Label | Prompt | Description | Other (Explain) | Identifier wording:  

| Is the identifier’s location obvious? If not, indicate why.  
**[B.5.b] Link Between Identifier and Action:** Is the identifier clearly linked with this action? If not, indicate why.  
**[B.5.c] Link Between Identifier and Goal:** Is the identifier clearly linked with an active goal? If not indicate why.  
...  

## GOMS Analysis

- Build *predictive* model using scientific knowledge about human memory and behavior  
  - like HTA, can analyze for complexity, consistency  
  - or build computational version, to estimate task times for different design alternatives  
  - if successful, can provide huge benefit...why?  
- Extends general techniques of HTA  
  - goals, subgoals, plans, actions  
  - BUT adds model elements for *mental activities* such as goal creation and selection, memory retrieval, etc.
a GOMS snippet

GOAL: CLOSE-ACTIVE-WINDOW
   [select GOAL: USE-MENU-METHOD
      MOVE-MOUSE-TO-MENU-BAR
      DRAG-DOWN-FILEMENU
      RELEASE-ON-CLOSE-OPTION
      GOAL: USE-HANDLE-METHOD
      MOVE-MOUSE-TO-CORNER
      CLICK-ON-CLOSE-BOX
      GOAL: USE-CONTROL-KEY
      PRESS-APPLE+W]

Downsides of Analytic Methods

- Usability inspections are rapid, relatively cheap
  - but may miss details only seen in realistic use contexts involving real users, say little about what caused the problems, or expected impact

- Model-based approaches have good scientific foundation, are credible, can be very powerful
  - but current theories have limited scope, and developing the models takes time/expertise
Empirical Evaluations: Validity

- Conclusions based on actual use, BUT...
  - are the users representative?
  - is the test population large, diverse enough?
  - is the test system realistic (versus early prototype)?
  - do the tasks match what happens in real use?
  - do the data (measures) reveal real life impacts?
- These are the general concerns of “ecological validity”, the extent to which an investigation is a genuine reflection of real-world happenings

Field Studies

- Variants of ethnographic methods we discussed during requirements analysis
  - observation of realistic tasks, interviews, data files, etc.
  - avoids the problem of ecological validity
- Summarize data through content classification
  - e.g., problem categories, as in themes analysis
  - can also sort by severity, based on observed impacts
- Field data collection and analysis time-consuming
  - also, much of the data simply reveals successful use!
Interviews, User Introspection

- Ask users to report what stands out in memory
  - more efficient access to data of interest: *critical incidents*
  - can enhance by making this collaborative, a discussion among usability personnel and multiple stakeholders
- BUT, human memory is biased by many things
  - wanting things to make sense, assuming things work as they always have

Usability Evaluation in the Lab

- Carefully selected set of representative tasks
  - e.g., based on task analysis of the system, design goals
  - in SBD, claims are used to guide task selection
- Control aspects of situation that are uninteresting
  - e.g., experimenter, location, task order, instructions
- Collect multiple measures of usability impacts
  - performance (time and errors), output quality
  - satisfaction ratings or other subjective measures
- Interpretation comes back to validity of the test
  - both ecological (realism) and internal (controls)
a simulated office environment, with as many “normal” details as is reasonable, possible

observational data collected via a one-way mirror, video cameras, screen capture, and possible event logging (sometimes observer in room, too)

Controlled Experiments

- If asking a specific question, making a choice
- Operationalize independent and dependent variables
  - what is manipulated, what outcomes are measured
- Define hypotheses in advance of the test
  - causal relation of independent and dependent variables
  - testing these requires the use of inferential statistics
- Construct an effective and practical design
  - within-subjects or between-subjects testing conditions
  - how many people to test, how assign to conditions
Some Variations

- Usability testing with *think-aloud* instructions
  - users comment as they work on their current goals, expectations, and reactions
  - BUT, thinking aloud takes capacity, changes task itself
  - very useful in supporting formative evaluation goals
- *Storefront testing*: bring the prototype into the hall!
  - fast, easy, quick cycle...but no control of users, tasks
- All of these can (should!) be supplemented with interviews and/or user reaction surveys
  - objective measures of behavior are not always correlated with subjective measures of experience or satisfaction

“Discount” Usability Evaluation

- Goal: get the most useful information for guiding re-design with the least cost
  - pioneered by Jacob Nielsen (heuristic inspection)
- Do a little bit of each (analytic and empirical)
  - 3-4 experts find most of the guidelines issues
  - 4-6 users experience most of the actual use problems
  - so between the two, get a good sense of what to fix
- Not surprisingly, a popular strategy, pretty much what you find in practice