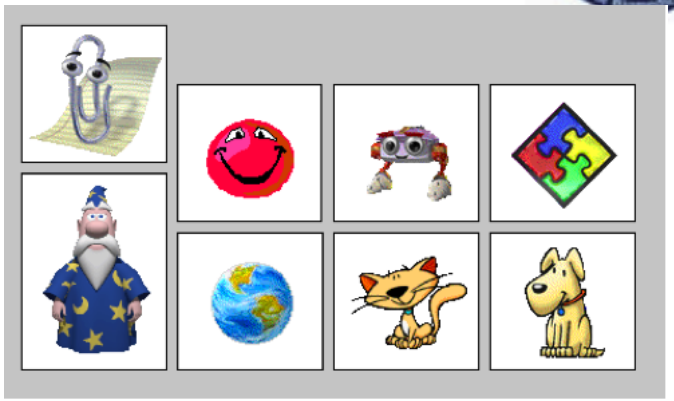
General Overview of Usability Assessment and Testing Methods

Harry Hochheiser
Biomedical Informatics
harryh@pitt.edu

What is Usability?

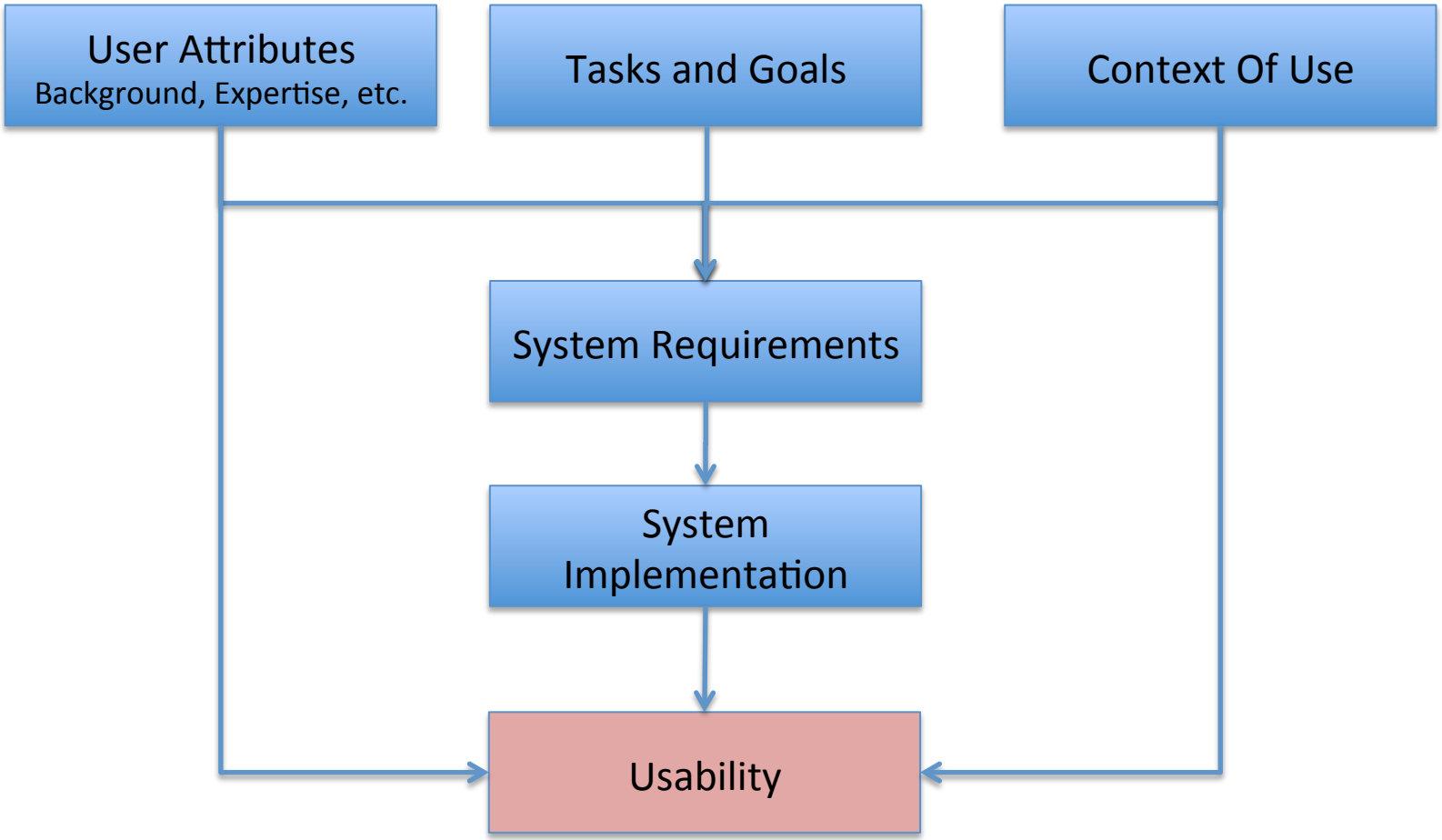
- Not "user-friendly"?



- A usable helps people accomplish their goals and complete their tasks.

- An unusable system is one that interferes with these goals

Determinants of Usability



Understanding User Requirements

- Understanding work in context – goals, motivations, priorities, behavior, difficulties, etc. is necessary for building better systems
  - Most go beyond just talking about computer systems to address bigger picture questions
- Successful implementations may require work redesign
  - Translating the same old methods and procedures to computers may not help much,
  - But reference to the familiar can be helpful

Requirements vs. goals

Goal - where you want to end up?

Requirements
what you must do to get there?

Stakeholder Analysis

Rosson & Carroll 2002

Identify stakeholder groups

Background

Expectations

Needs

Preferences

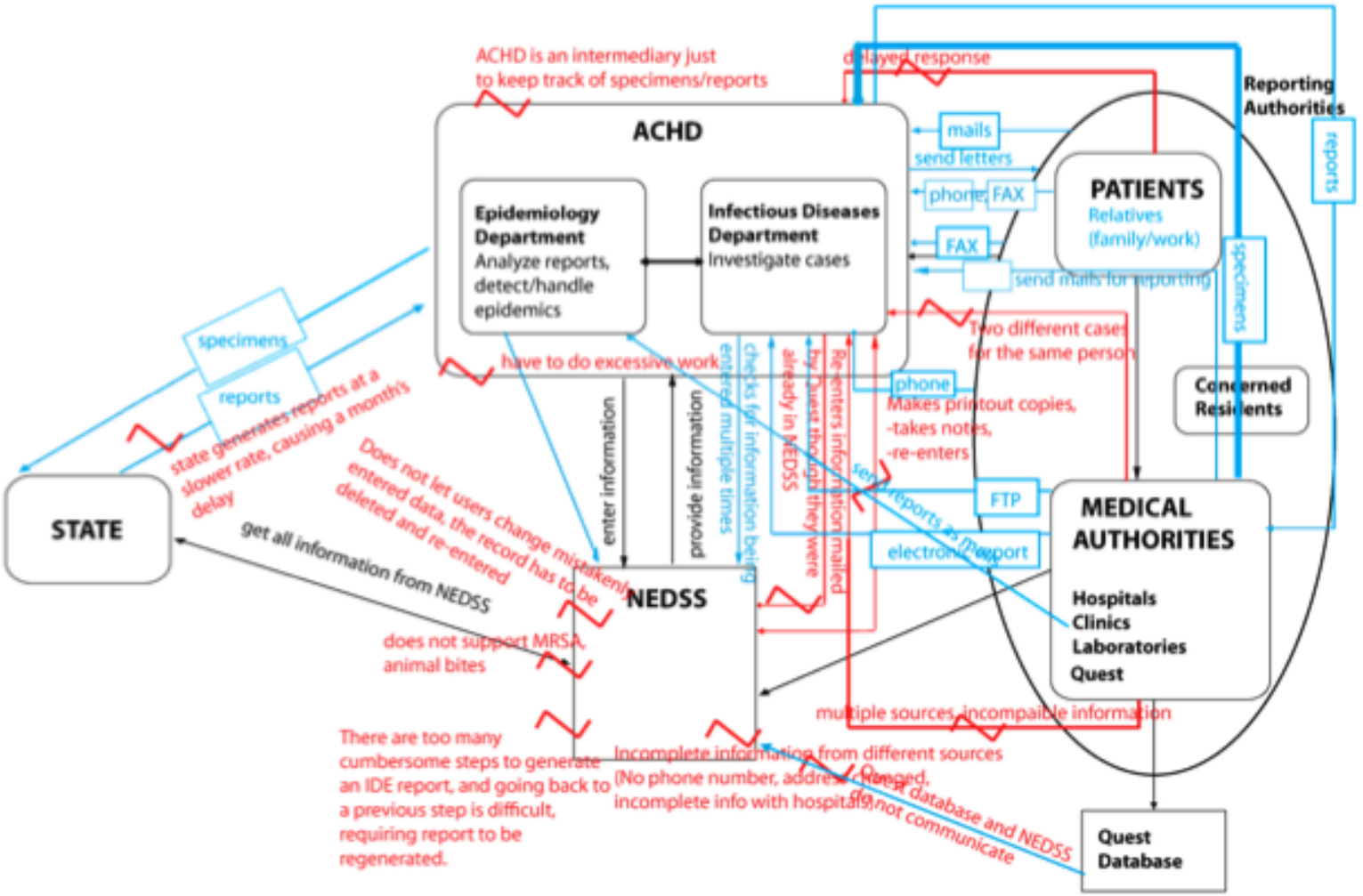
Concerns

Values

An important, but often overlooked step

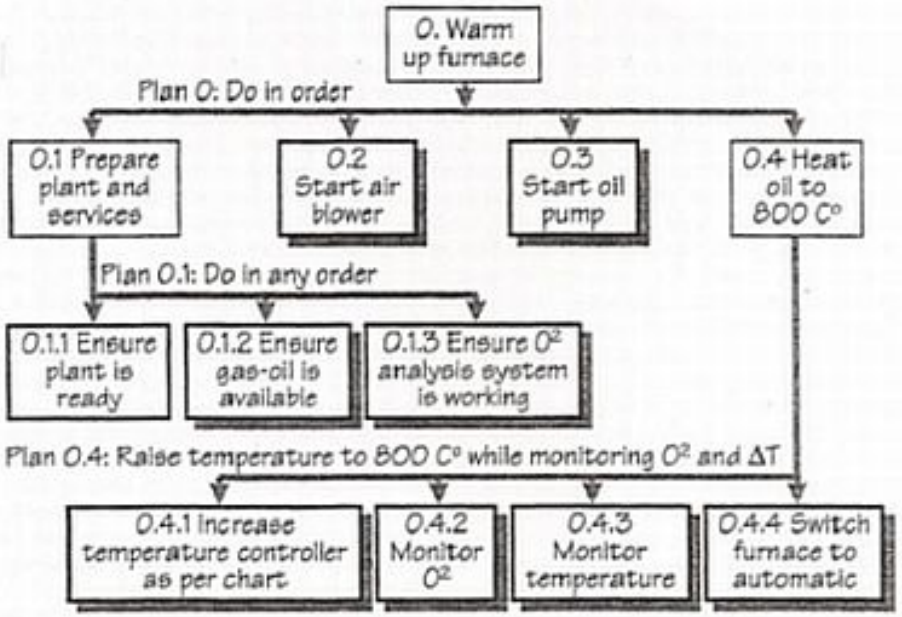
Flow Model

thanks to M. Wagner and A. Dey

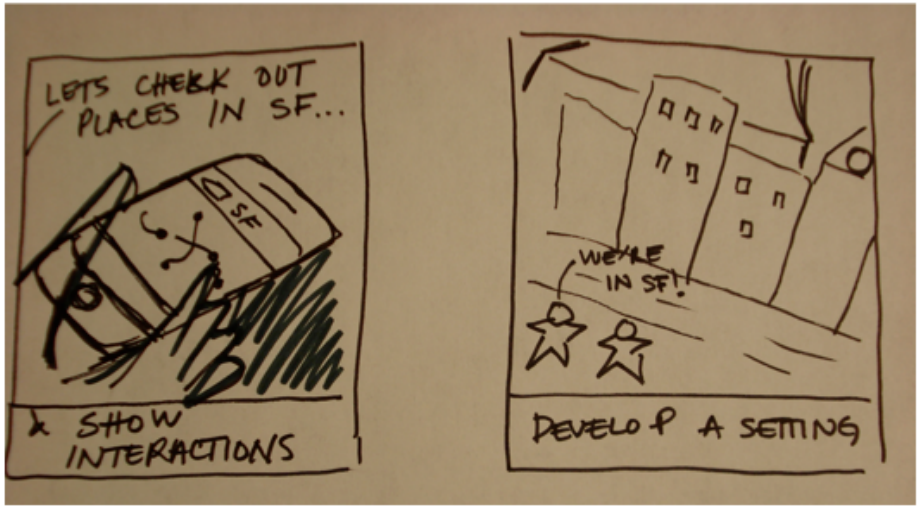


Hierarchical Task Analysis

http://en.wikipedia.org/wiki/File:Hierarchical_Task_Analysis.jpg



Storyboards



- Amal Dar Aziz – Guide to storyboarding http://hci.stanford.edu/courses/cs147/assignments/storyboard_notes.pdf

Dimensions of Usability

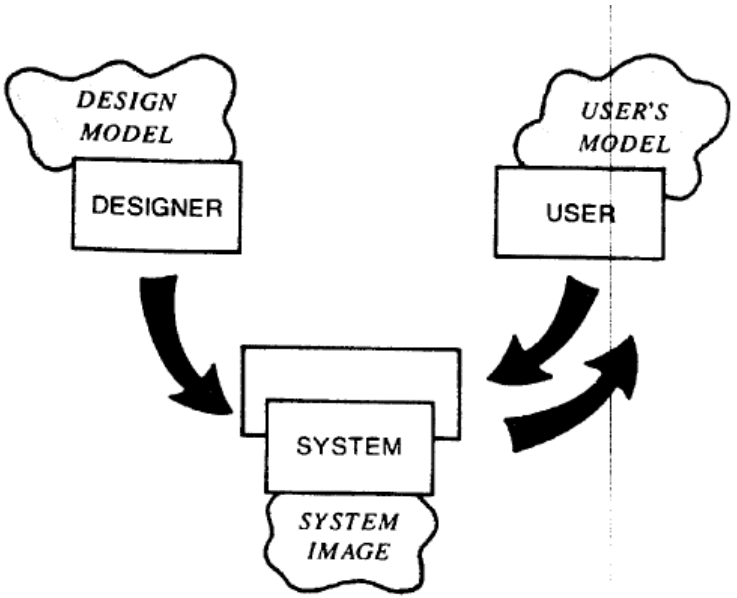
- Efficiency
- Learnability
- Memorability
- Error-Handling/Prevention
- Satisfaction

Dimensions of Usability

- Efficiency
  - Task Completion time? # of operations/movement
- Learnability
  - How quickly can a novice learn tool? What help is given?
- Memorability
  - Retention of proficiency over time? Cognitive load?
- Error-Handling/Prevention
  - Error rate? Slips vs. Mistakes? Error Prevention
- Satisfaction

Mental Models

- Cognitive representation of observed phenomena
  - What you think is going on.
- "Deep" vs. "Shallow"
  - Deep models based in understanding of underlying mechanisms
- Designer model vs. user's model?



Slips vs. Mistakes

- Slip – you know what to do, but you do the wrong thing.
  - Click the wrong button.
  - Generally less serious

- Mistake – you don't know what to do; Don't know which menu to look under
  - Potential indicator of mismatch between system model and user model.

Goal of Usability Assessments

- Can users complete tasks?
- Appropriateness of mental models
- Comparative efficiency
- Subjective satisfaction
- How do we assess?

Spectrum of Methods

Inspection Methods
- Heuristic Evaluation
- Cognitive Walkthrough

Lab Usability Studies
- "think-aloud"

Comparative Empirical Studies

in situ evaluation
- Quantitative and qualitative

Low Cost
Low Fidelity

High Cost
High Fidelity

Usability Inspections

- "clean-room" static examination of usability
- Methodically scrutinize interfaces in search of potential problems

- Pros:
  - Inexpensive – no users, relatively easy
  - Identify major issues at a relatively early stage
- Cons:
  - May miss problems: generally find < 50%
  - All results are hypothetical – don't' know which problems might really lead to errors

Two Broad classes of inspections

Heuristic Evaluations: How well does an interface conform to guidelines for interface design?

Walkthrough: Analytic examination of interface and interaction requirements, usually informed by some model of the user

Many variants...

Who inspects?

Heuristic inspection

- Usability experts
- Domain experts
- Combination? (Double experts)
- Users should participate as users when possible

3-5 experts? (Nielsen)

Or more...

Work alone, or in teams..

Walkthroughs

- May require more cognitive background

Domain expert feedback helpful

Conducted by a team?

Tasks?

Heuristic inspections

- Set tasks
- Open-ended exploration

Walkthroughs

- Generally, specific tasks

How to interpret?

Use severity judgments to prioritize fixes

- Frequency of problem
- Impact of problem
- Persistence – will users be repeatedly bothered?
- Multiple independent raters increase reliability

Bigger questions – does this design work at all?

As with usability studies, try to generalize

- Don't solve lots of small problems if the design is inherently problematic

Nielsen's Heuristics

Nielsen, 1994 - http://www.useit.com/papers/heuristic/

- Visibility of system status
- Match between system and real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

Heuristic Evaluation: Procedure

Evaluators work alone

(except for when they work in teams)

Optional observer can help explain confusing issues and to record issues.

Go through interface several times

overview and specific

Heuristic focus or task focus

Note discrepancies between interface and heuristic

Individual evaluators meet to aggregate results

Heuristic Evaluation Procedure, cont.

List of heuristics is not exhaustive – use other principles as needed

Develop specific heuristics for particular classes of
- Tool
- Users
- Contexts

Use specific scenario and/or open-ended exploration

Multiple investigators provide greater coverage

Heuristic Evaluation Output

List of usability problems

Reference to principles that were violated

List all violations – even if multiple problems with a single interface element

Suggest fixes if possible?

1994 Case study. $10,500 heuristic evaluation led to expected benefits of $500,000

Case Study

- Clinical trials registry
  - Clinicaltrials.gov
  - https://www.researchmatch.org/
  - https://www.researchregistry.pitt.edu/

Cognitive Walkthrough
(Wharton, et al, 1994, Spencer 2000)

Evaluate software for learning by exploration
    Preferred mode of learning for many users

Conducted with respect to one or more specific tasks

Consider, in sequence, user actions needed to complete the task

Tell a story about interactions

Ask what user would be trying to do and what interface affords

Successful interfaces will lead user to correct the appropriate action
    and provide clear feedback that progress is being made

Procedural details

Focus on ease of learning might bias results
Narrowly-focused method

Evaluators – group or individual

Developers, designers, marketing people, interface experts

Give people roles – contribute specific expertise.

Inputs to the Walkthrough

- Who will be the users? - Be specific – background, experience, knowledge
- What task (or tasks) will be evaluated? - Reasonable but representative set of benchmark tasks
- What is the correct action sequence for each task and how is it described? - describe at same level as a good tutorial
- How is interface defined? Provide detail relevant for presumed user and context – don't bother with information that can be assumed.

The walkthrough

(Wharton, et al. 1994)

Examine each action in solution path

Attempt to tell a credible story as to why the expected users would choose the correct action

Based on user's background and goals

Critical features – those that link task description and correct action

Four questions for the walkthrough

(Wharton, et al. 1994)

. Will the users try to achieve the right effect?
. Will the user notice that the correct action is available?
. Will the user associate the correct action with the effect?
. If the correct action is performed, will the user see that progress is being made toward solution of the task?

If all four questions can be answered yes – success

Any single "no" – failure story

Information to capture
(Wharton, et al, 1994)

User knowledge requirements

Assumptions about user population

Notes about side issues

Design changes?

Three displays
- Key points of group story
- Information about each class of user
- Side issues and design changes

Case Study

- Clinical trials registry
  - Clinicaltrials.gov
  - https://www.researchmatch.org/
  - https://www.researchregistry.pitt.edu/

- Users?
- Tasks?
- Correct Sequences

Usability Studies: Goals

- Understand if the system supports completion of intended tasks
- Be specific –
  - Users
  - Tasks – detailed scenarios
  - Define success
- User Satisfaction?
  - Do users like the tool?

Formative Usability Studies: Conditions

- Usability Lab
  - Two-way mirrors/separate rooms
- Workspace
- Online?
- Often video and/or audio-recorded
- Screen-capture
- Logs and instrumented software
- Goal: Ecological Validity

Usability Studies: Measures

- Key question "can users complete tasks"?
- Lists of usability problems
  - Description of difficulty
  - Severity
- Task completion times – depending on methods
- Error rates?
- User Satisfaction
- Quantitative results for measuring success
  - Not comparative

Usability Studies: Methodology

- Define Scope
- Users complete tasks
- Researchers observe process
- What happens?
- What goes right? What goes wrong?
- Note difficulties, confusions?
- Record – audio/video, screen capture, Techsmith Morae

Usability Studies: Participants

- Somewhat representative of likely users
- Willing guinea-pigs
- Need folks who are patient, willing to deal with problems
- Well-motivated
    - Compensated
    - Eager to use the tool
- Small numbers – repeat until diminishing returns

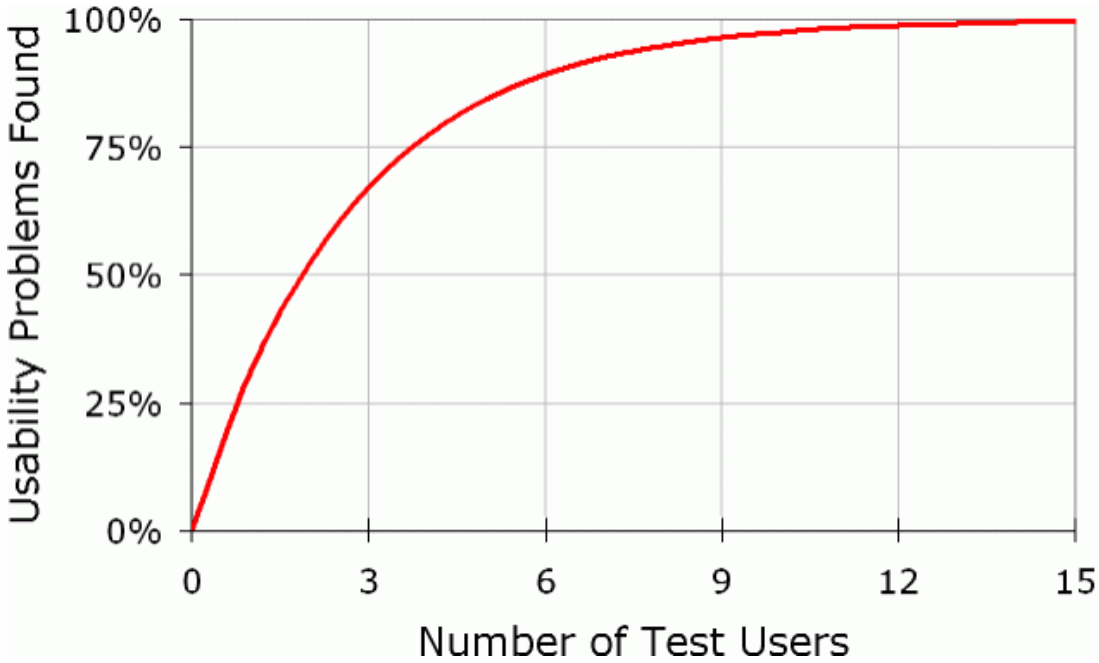
… But how many?

Only 5 users – or maybe not

Nielsen – why you only need to test with 5 users

http://www.useit.com/alertbox/20000319.html

Hwang & Salvendy (2010) – maybe need 10 +/- 2



Two approaches

- Observation
  - Subject performs tasks, researchers observe
  - Ecological validity, but no insight into users
- "Think aloud"
  - User describes mental state and goals

Think-Aloud Protocols

User describes what they are doing and why as they try to complete a task http://www.youtube.com/watch?v=l-OC1_QxIdw

- Describe both goals and steps taken to achieve those goals.
- Observe
  - Confusions – when steps taken don't lead to expected results
  - Misinterpretations – when choices don't lead to expected outcomes
- Goal: identify both micro- and macro-level usability concerns

Caveats

- Think-aloud is harder than it might sound
- What is the role of the investigator?
  - How much feedback to provide?
  - What (if anything) do you say when the user runs into problems?
  - What if it's a system that you built?
- How to identify/describe a usability problem?

Reporting Usability Problems

adapted from Mack & Montaniz, 1994

- Look for Breakdowns in goal-directed behavior
  - Correct action, noticeable effort
    - To find
    - To execute
  - Confused by consequence
    - Correct action, confusing outcome
    - Incorrect action requires recovery
    - Problem tangles
- Qualitative analysis by interface interactions
  - Objects and actions
  - Higher-level categorization of interface interactions

Reporting Usability Problems

adapted from Mack & Montaniz, 1994

- Inferring possible causes of problems
- Problem reports
  - Design-relevant descriptions
  - Quantitative analysis of problems by severity

Analysis

- Challenge – identify problems at the right level of granularity?
    - When does a series of related difficulties lead to a need for redesign?
    - What if these difficulties come from different tasks?
- When appropriate, relate usability observations back to contextual inquiry or other earlier investigations
    - Does the implementation fail to line up with the needs?
        - Perhaps in some unforeseen manner?

Completion – Summative User Studies

- Demonstrate successful execution of system
- With respect to
  - Alternative system – even if straw man
  - Stated performance goals – Acceptance Tests
- User studies
- Generally empirical

Completion – Summative Studies of systems in use

- Case studies
  - Descriptions of individual deployments
  - Qualitative
- Longitudinal study of ongoing use
  - Collect data regarding impact
  - Similar to case studies, but potentially more quantitative.

Acceptance Tests

Usability tests aimed at measuring success

Does the tool do what the client wants
95% task completion rate within 3 minutes, etc.?

Client has clearer idea – not just "user friendly"

What: Empirical Studies

- Quantitative measure of some aspect of successful system use
  - Task completion time (faster is better)
  - Error rate
  - Learnability
  - Retention
  - User satisfaction…
- Quality of output?

Tension in empirical studies

- Metrics that are easy to measure may not be most interesting
    - Task completion time
    - Error rate
- Great for repetitive data entry tasks, less so for complex tasks
    - Analytics, writing…
- Powerball vs. smallball

Design

- Controlled experiments from cognitive psychology
- State a testable/falsifiable hypothesis
- Identify a small number of independent variables to manipulate
    - hold all else constant
    - choose dependent variables
    - assign users to groups
    - collect data
    - statistically analyze & model

Independent Variables

- What are you going to test?
- Condition that is "independent" of results
  - independent of user's behaviors
  - independent of what you're measuring.
  - one of 2 (or 3 or 4) things you're comparing.
  - can arise from subjects being classified into groups
- Examples
  - Galileo: dropping a feather vs. bowling ball
  - Menu structures – broad/shallow vs. narrow/deep

Dependent variable

- Values that hypothesis test
  - falling time
  - task performance time, etc.
- May have more than one
- Goal: show that changes in independent variable lead to measurable, reliable changes in dependent variables.
- With multiple independent variables, look for interactions
  - Differences between interfaces increase with differences in task complexity

Controls

- In order to reliably say that independent variables are responsible for changes in dependent variables, we must **control** for possible **confounds**
- **Control** – keep other possible factors constant for each condition/value of independent variables
- **confound** – uncontrolled factor that could lead to an alternate explanation for the result.

Between-Groups vs. Within-Groups Design

- How do you assign participants to conditions?
- All people do all tasks/cells?
    - Within-groups – compare within groups of individuals.
    - one group of test participants
    - Fewer participants, but learning effects
- Certain people for certain cells?
    - between groups – compare between groups of individuals
    - 2 or more groups
- Mixed models

Analysis

- Plan your analysis in advance
- Necessary for determining number of participants
- Consult a statistician

Other Challenges

- Ordering tasks?
- How many?
    - Want to avoid fatigue, boredom, and expense of long sessions
- Variability among subjects
    - May be unforeseen.
    - Bi-modal distribution of education or computer experience?
- Training materials
- Run a pilot

Longitudinal use

- Lab studies are artificial
- Many tools used over time.
  - use and understanding evolve
- Longitudinal studies look at usage over time
- Expensive, but better data
- Techniques
  - Interviews, usability tests with multiple sessions, continuous data logging, Instrumented software, Diaries

Case Studies

- In-depth work with small number of users
    - Multiple sessions
    - Describe scenarios
- Illustrate use of tool to accomplish goals
- Good for novel designs, expert users
- Formative evaluation – can be used to gather requirements
- Summative – show validity of idea
- Possibly less compelling than usability evaluations.

Informed Consent

- Research must be done in a way that protects participants
- Principles
  - Respect for persons
  - Beneficence – minimize possible harms, maximize possible benefits
  - Justice – costs and benefits should not be limited to certain populations
- Institutional Review Board (IRB) – revies and approves experiments
  - www.irb.pitt.edu
- Crucial for responsible research

Shameless plugs..
For more information

- BIOINF 2121 – Human-Computer Interaction and Evaluation
  http://faculty.dbmi.pitt.edu/harryh/classes/2013/2121/