EVALUATING USER INTERFACES

15 & 17 January 2014
Is this any good?
Goal

• You Should Be Able To
  • Explain common **approaches** to evaluating interfaces
  • Explain **when** to evaluate interfaces
  • Identify common **metrics** and **questionnaires** used in evaluating interfaces
Introduction

• Why evaluate?
  • Designers become too entranced
    • What I like - become tied to design and lose focus of user
  • Experienced designers know extensive testing is required
• The determinants of the evaluation plan include
  • stage of design (early, middle, late)
  • novelty of project (well defined vs. exploratory)
  • number of expected users
  • criticality of the interface (life-critical medical system vs. museum exhibit support)
  • costs of product and finances allocated for testing
  • time available
  • experience of the design and evaluation team
Introduction

• How do you test…
  • a web site?
  • an air traffic control system?
What does testing **not** do?

- Guarantee perfection
- Hard to “finish” testing
- Difficult to test unusual situations
  - Military attack
  - Heavy load (e.g. voting)
- Case study: Project ORCA
- Simulate accurate situations
  - E.g. driving games, military games, medical sims
Project ORCA

- Web app for 37,000 volunteers at polls, 800 central trackers to track voters – who still needed to vote
  - Training
    - 90 minute training for 800 central volunteers
    - No training for field volunteers
  - Testing – local, not “real world”
    - System crash during actual load usage
  - Usage
    - Users’ first usage was 6 AM on election day
    - Confusion about “app” (where would you look for it)?
    - http vs https: no redirect to the secure site
    - Wrong pins and passwords
    - VoIP phones incorrectly configured – volunteers unable to provide support
  - Automated vs. Human testing
  - Difference in FL, OH, VA, CO was 520,000 votes
Expert Review

• Colleagues or Customers
  • Informal
  • Natural starting point
  • Ask for opinions

• Expert reviews
  • Formal
  • Considerations
    • What is an expert? User or designer?
    • How long? Half day to week
  • Different review methods from which to choose
Heuristic Evaluation

- Experts evaluate interface on design heuristics
  - Shneiderman’s Eight Golden Rules

- Specific to application area
  - Heuristics for gaming (Pinelle et al 2008)
    - Provide consistent responses to user’s actions (1)
    - Allow users to customize video and audio setting, difficulty, and game speed (2)
    - Provide users with information on game status (3)
    - …etc
Eight Golden Rules of Interface Design

1. Strive for consistency
2. Cater to universal usability
3. Offer informative feedback (e.g. button clicked)
4. Design dialogs to yield closure (e.g. confirmation page)
5. Prevent errors (e.g. gray out menu items)
6. Permit easy reversal of actions
7. Support internal locus of control
8. Reduce short-term memory load (e.g. 7 ± 2)
Cognitive Walkthrough

• Experts “simulate” being users going through the interface
• Tasks are ordered by frequency
• Good for interfaces that can be learned by “exploratory browsing” (Wharton 1994) [novices]
• Usually walkthrough by themselves, then report their experiences (written, video) to designers
• Useful if application is geared for a group the designers might not be familiar with:
  • Military, Assistive Technologies
Formal Usability Inspection

- Experts hold courtroom-style meeting
- Each side gives arguments (in an adversarial format)
  - Merits and weaknesses of interface
- There is a judge or moderator
- Good for novice designers and managers
- Extensive and expensive
  - Requires time to prepare and more personnel than other approaches
- An example of formal usability inspections in practice at HP Company
  - http://www.sigchi.org/chi95/proceedings/intpost/cg_bdy.htm
Expert Reviews

• Can be conducted at several times in the design process
• Focus on comprehensiveness over specific features or random improvements
• Rank recommendations by importance
  • HIGH: Shorten log in procedure – because users are over busy
  • MIDDLE: Reordering sequence of displays, removing nonessential actions, providing feedback.
• Less vital fixes and novel features for future releases
Expert Review

- Placed in situation similar to user
  - Take training courses
  - Read documentation
  - Take tutorials
  - Try the interface in a realistic work environment (complete with noise and distractions)

- Bird’s eye view
  - Studying a full set of printed screens laid on the floor or pinned to the walls
  - Easier to evaluate topics such as consistency

- Software tools to speed analyses
  - WebTango
Usability Testing and Labs

- 1980s, testing was luxury (but deadlines crept up)
- Usability testing was incentive for deadlines
  - Sped up projects*
  - Cost savings*
- Labs are different than academia
  - Less general theory
  - More practical studies
Usability Labs

- IBM early leader
- Microsoft next (>25 labs)
- Now hundreds of companies

From http://www.ergosign.de/
Staff

- Expertise in testing (Psych, HCI, CS)
- 10 to 15 projects per year
- Meet with UI architect to plan testing (Figure 4.2)
- Participate in early task analysis and design reviews
- t – 2-6 weeks, creates study design and test plan
  - Task lists
  - Participants – number, types and sources (current customers, in company staff, temp agencies, advertisements)
- t – 1 week, pilot test (procedures, tasks, and questionnaires)
  - 1-3 participants
Participants

• Labs categorize users based on:
  • Computing background
  • Experience with task
  • Motivation
  • Education
  • Ability with the language used in the interface

• Controls for
  • Physical concerns (e.g. eyesight, handedness, age)
  • Experimental conditions (e.g. time of day, physical surroundings, noise, temperature, distractions)
Recording Participants

- Logging is important, yet tedious
  - Software to help (Live Logger, Morae, Spectator)
  - Powerful to see people use your interface
- New approaches: eye tracking
  - Tobii (infrared to generate reflection patterns on corneas)
Thinking Aloud

• Concurrent think aloud
  • Invite users to *think aloud*
  • Nothing they say is wrong
  • Don’t interrupt, let the user talk (sometimes need to prompt)
  • Spontaneous, encourages positive suggestions
  • Can be done in teams of participants

• Retrospective think aloud
  • Asks people afterwards what they were thinking
  • Issues with accuracy (remembering)
  • Does not interrupt users (timings are more accurate)
Types of Usability Testing

- Paper mockups and prototyping
  - Inexpensive, rapid, very productive
  - Low fidelity is sometimes better (Synder, 2003)

http://user.meduni-graz.at/andreas.holzinger/holzinger/papers%20en/
Types of Usability Testing

• Discount usability testing
  • Test early and often (with 3 to 6 testers)
  • Pros: Most serious problems can be found with few testers. Good for formative evaluation (early)
  • Cons: Complex systems can’t be tested this way. Not good for summative evaluation (late)

• Competitive usability testing
  • Compares new interface against prior or competitor’s versions
  • Experimenter bias, be careful to not “prime the user”
  • Counterbalancing order of presentation of interfaces important
  • Within-subjects is preferred because can make comparisons
Types of Usability Testing

• Universal usability testing
  • Test with highly diverse…
    • users (experience levels, ability, etc.)
    • platforms (mac, pc, linux)
    • hardware (old (how old is old?) -> latest)
    • networks (dial-up -> broadband)

• Field tests and portable labs
  • Tests UI in realistic environments
  • Use of logging software an advantage
  • Beta tests – 10s or 1000s of users evaluate beta versions and asked to comment
Types of Usability Testing

• Remote usability testing (via web)
  • Recruited via online communities, email
  • Pros: large \( n \), adds to realism
  • Difficulty in logging, validating data
  • Software can help (NetMeeting, WebEx, Sametime)

• Can You Break this Test
  • Challenge testers to break a system
  • Games, security, public displays
Limitations

- Focuses on first-time users
- Limited coverage of interface features
  - Testing only 1-3 hours – can’t know performance after week or month
    - Rarely used features
- Difficult to simulate realistic conditions
  - Ex: Military, first response
    - Stress in real usage situation can lead to inability to process info (Rahman 2007)
- Yet formal studies on usability testing have identified
  - Cost savings
  - Return on investment (Sherman 2006; Bias & Mayhew 2005)
Survey Instruments

- Questionnaires
  - Paper or online (e.g. Qualtrics, SurveyMonkey)
  - Easy to grasp for many people
  - The power of many can be shown
    - 80% of the 500 users who tried the system liked Option A
    - 3 out of the 4 experts like Option B

- Success depends on
  - Clear goals in advance
  - Focused items

- Advantage of pilot testing – define and refine the survey questions!!
Designing survey questions

• Ideally
  • Based on existing questions
  • Reviewed by colleagues
  • Pilot tested with small sample
  • Have built in methods to verify respondents represent population
    • Age
    • Gender
    • Experience
    • Etc…
Likert Scales

• Most common methodology
  • Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
• Typically 5, 7, 9-point scales
• Examples
  • Improves my performance in book searching and buying
  • Enables me to search and by books faster
  • Makes it easier to search for and purchase books
Most used Likert scales

- **Questionnaire for User Interaction Satisfaction**
- Example questions
  - How long have you worked on this system?
  - Learning to operate the system (Difficult 1 2 3 4 5 6 7 8 9 Easy)
- **System Usability Scale (SUS)** – Brooke 1996
  - 10 questions: 5 pos worded, 5 neg worded (reverse coding)
- Post-Study System Usability Questionnaire
- Computer System Usability Questionnaire
- Software usability Measurement Inventory
- Website Analysis and MeasureMent Inventory
- Mobile Phone Usability Questionnaire
- Questionnaire websites
  - [Gary Perlman’s website](#), [Jurek Kirakowski’s website](#)
Acceptance Tests

- Set goals for performance
  - Objective
  - Measurable
- Requirements document
  - We want the software to be “user friendly”
    - Vague and misleading criterion
- How could we rephrase it?
  - Use a metric such as Shneiderman’s goals for interface design
    - Time for users to learn specific functions
    - Speed of Task performance
    - Rate of Errors
    - User retention of commands over time
    - Subjective satisfaction
Examples: food shopping website

Test A
- The participants will be
  - 35 adults (25-45 years old)
  - Native speakers with no disabilities
  - Hired from an employment agency
  - Moderate web-use experience (1-5 hours/week) for at least one year
- ≥30 of the 35 should complete the benchmark tests within 30 minutes

Test B
- 10 participants will be recalled after one week
- Carry out new set of benchmark tests
- In 20 minutes, at least 8 should be able to complete tasks correctly
Acceptance Tests

- Completing the acceptance tests
  - Can be part of contractual fulfillment
  - Demonstrate objectivity

- Different than usability tests
  - More adversarial
  - Neutral party should conduct that

- Central goal
  - To verify requirements adhered to
  - Not to detect flaws
Evaluation during active use

• Interviews with individual users
  • Pro: Get very detailed on specific concerns
  • Cons: Costly and time-consuming

• Focus group discussions
  • Pro: identify patterns of usage or hidden problems
  • Con: Certain people can dominate or sway opinion

• Targeted interviews and focus groups
  • Experienced or long-term users
  • Identify different issues than novice users
Evaluation during active use

- Continuous logging
  - The system itself logs user usage
    - Video game example
  - Other examples
    - Track frequency of errors** (gives an ordered list of what to address via tutorials, training, text changes, etc.)
    - Speed of performance
    - Track which features are used and which are not
    - Web Analytics

- Ethical considerations
  - Privacy? What gets logged? Opt-in/out?
  - What about companies?
Online and Telephone Help

- Users feel reassured having people ready to help (real-time chat online or via telephone)
- E.g. (2008) Netflix has 8.4 million customers, how many telephone customer service reps?
  - 375
  - Expensive, but higher customer satisfaction
- Cheaper versions are bug report systems
  - Windows, Chrome, Bugzilla