# Conducting Evaluations

# User Testing - Previously

- Give Users Task
- Roles
  - Greeter
  - Facilitator
  - Computer (paper prototyping)
  - Observers
- Take notes, make changes, test again

# Beyond “Do you like my interface?”

- “How much do you like my interface?”
- “This is a useful interface: agree/disagree”

# Usability Evaluation

- What's the comparison?
  - Design 1, 10 users failed task 9
  - Design 2, 2 users failed task 9
- Evaluate with Concrete **Tasks**
- Independent vs. Dependent Variables
- Formative vs. Summative Studies
- Qualitative vs. Quantitative Data

# Independent vs. Dependent Variables

- Independent
  - Manipulate this
  - Different Designs: 1 & 2
- Dependent variables
  - Measure this
  - Task Success: 1, 2, 3, 4, …, n

# Formative vs. Summative Research Studies

- empirical data – observations and/or measurements collected
- Formative – Refinement
  - informal
  - understand why interface element x does get used by participants
- Summative – Evaluation
  - formal, comparative study
  - Is design x better than design y?

# Formative Evaluation

- Refine your design
  - Watching your users fail to use your design, will help you understand why it doesn't work and how to fix it.
- Usually tasks based
- Informal – interrupt and asks questions or clarify why they did something
  - Think aloud protocol helpful

# Summative Evaluation

- Evaluate Hypothesis (Independent Variable(s))
  - Is X better than Y?
    - Compare Design 1 against Design 2
    - Compare System 1 against System 2
- Formal – Do not interrupt
- Gather empirical data
  - observations/measurements

# When to use?

- Formative First
  - refine your design
  - get a feel for how users use it
  - is there evidence of an effect?
- Pilot Study
  - Evaluate your study design
- Then Summative
  - Evaluate whether your design works

# Collecting Data

- notes
- surveys – validated
  - Likert Scale (i.e 1 to 7 )
- logging/instrument your software
- recordings
  - video
  - audio
  - screen

# Qualitative vs. Quantitative Data

- Measure dependent variable(s)
  - Formative & Summative
- Qualitative
  - data gathered not in numerical form; often descriptive
    - think aloud protocol
    - interviews (probably semi-structured)
- Quantitative
  - data gathered in numerical form; often measurements
    - task success
    - near transfer tasks
    - number of clicks via logging

# Study Design – Task Ordering

- “Learning Effects”
  - If I always do task 1 first, then I may do better on tasks 2..n.
  - Sometimes you want this; other times you don't
- Randomization
  - user completes tasks in random order
- Latin Squares (Balanced/Williams Design)
  - user 1: tasks 1, 2, 3
  - user 2: tasks 2, 3, 1
  - user 3: tasks 3, 1, 2

# Study Design – Subjects

- Between-subjects design
  - Random assignment
  - one half of participants use design/system 1
  - other half of participants use design/system 2
- Within-subjects design
  - all participants use design/system 1 & 2

# Analyzing Qualitative Data

- Grounded Theory Approach
  - Pull themes from qualitative data
- Affinity Diagramming
  - useful for finding themes
- Interrator agreement/reliability
  - 2 people rate random sample of 20% of data set.
  - Cohen's Kappa > .8

# Analyzing Quantitative Data

- Explore your Data
  - Plot it!
- Aggregate Statistics
  - X% of users did this, Y% of users did this
  - mean + standard deviation
- Statistical Tests
  - p-values (t-test, chi-squared, ANOVA)
    - likelihood result is by chance ($p < .05$)
    - skewed by number of participants
  - effect size
    - tries to be independent of population/sample size

# Before you Start

- Consent
- Introduction to study
  - “we are testing the software; we are not testing you”
- Training/Training Tasks
  - At the beginning of your study, provide any training that your users might need
  - Remove variance due to “this is new” vs. independent variable

# Pilot Studies

- Practice run of you summative study
- Treat it like the **real** thing
- “prototype of your experimental design”
  - Fix problems with your study protocol
  - Fix errors that can derail your results

# Testing Environment

- Usability Lab
  - Testing Room
  - Observation Room
- In the Field

http://www.museumsandtheweb.com/mw2009/papers/dowden/dowden.html

# Field Study

- Study your system in the field
  - timeline: hours to months
- Diaries
  - people often forget; not very accurate
- Direct Observations & Interviews
  - Observe your system in the field
  - Interview if necessary
- Caution: field studies can change user behavior

# Field Study Example

- Evaluating cell phones in different world markets
- Examine cross-cultural differences
- Nokia evaluated phones in the field in emerging markets
  - e.g. China, India, Brazil

# In the Wild

- Evaluate system “in situ”
  - natural setting
- Release your system
- Observe Remotely

# Lab vs. Field/Wild

- usability issues – lab
- use aspects – field
  - how users approach new technology
  - benefits derived from it
  - how it folds into everyday activities
  - sustained use over time

# A/B Testing

- Test two versions of a design in the wild
- Release A and B, see which *performs* better
- Good for small changes, especially web pages
  - It's scope is limited, but important
- It's cheap

# A/B Testing Details

- Version A
  - current version of website
- Version B
  - move checkout button to top of page
- Show 50% of your visitors A, the other 50% B
- Compare your analytics, to check performance (~ 1%)
- Why didn't B work?
  - WHO KNOWS!?

# Summary

- Usability
  - formative – refinement
  - pilot – study design evaluation
  - summative – performance evaluation
- Field
  - direct observations
- Wild
  - observe remotely
- A/B Testing
  - small improvements