Pick Project Groups

Choose your group partner by Friday.
Use your PhD/MS/MEng project, if you have one.
Otherwise: incorporate Kinect or talk to me during office hours.
HW3 in, HW4 out

HW3 not graded yet.

HW4 Goal: become familiar with running a user study & analyzing study data.

Short deadline: Due Monday before class.
HW4 Details

Evaluate either HW1 (online) or HW2 (in person).

Gather quantitative and qualitative data; perform simple statistical analysis; report results with figures.
Genres of assessment

Automated: Usability measures computed by software

Inspection: Based on skills, and experience of evaluators

Formal: Models and formulas to calculate measures

Empirical: Usability assessed by testing with real users
Purpose of Assessment

**Formative:** to inform the next design

**Summative:** to characterize a design, once it is complete
Empirical Testing is Costly

User studies are very expensive – you need to schedule (and normally pay) many subjects.

User studies may take many hours of the evaluation team’s time.

A user test can easily cost $10k’s
“Discount” Usability Techniques

Cheap
No special labs or equipment needed
The more careful you are, the better it gets

Fast
On order of 1 day to apply
(Standard usability testing may take a week)
“Discount” Usability Techniques

Heuristic Evaluation
Assess interface based on a predetermined list of criteria

Cognitive Walkthroughs
Put yourself in the shoes of a user
Like a code walkthrough

Other, non-inspection techniques are on the rise
e.g., online remote experiments with Mechanical Turk
Cognitive Walkthrough

Given an interface prototype or specification, need:

A detailed task with a concrete goal, ideally motivated by a scenario
Action sequences for user to complete the task

Ask the following questions for each step:

Will the users know what to do?
Will the user notice that the correct action is available?
Will the user interpret the application feedback correctly?

Record: what would cause problems, and why?

From: Preece, Rogers, Sharp – Interaction Design
Heuristic Evaluation

Developed by Jakob Nielsen (1994)

Can be performed on working UI or on sketches

Small set (3-5) of evaluators (experts) examine UI

Evaluators check compliance with usability heuristics

Different evaluators will find different problems

Evaluators only communicate afterwards to aggregate findings

Designers use violations to redesign/fix problems
Nielsen’s Ten Heuristics

**H2-1**: Visibility of system status
**H2-2**: Match system and real world
**H2-3**: User control and freedom
**H2-4**: Consistency and standards
**H2-5**: Error prevention
**H2-6**: Recognition rather than recall
**H2-7**: Flexibility and efficiency of use
**H2-8**: Aesthetic and minimalist design
**H2-9**: Help users recognize, diagnose, recover from errors
**H2-10**: Help and documentation
Empirical Assessment: Qualitative

**Contextual Inquiry**: try to understand user’s tasks and conceptual model

**Usability Studies**: look for critical incidents in interface

Qualitative methods help:
Understand what is going on
Look for problems
Roughly evaluate usability of interface
Empirical: Quantitative Studies

Quantitative
Use to reliably measure some aspect of interface
Compare two or more designs on a measurable aspect
Contribute to theory of Human-Computer Interaction

Approaches
Collect and analyze user events that occur in natural use
Controlled experiments

Examples of measures
Time to complete a task, Average number of errors on a task, Users’ ratings of an interface
Compare & Contrast?
Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. Identify variables
   (independent, dependent, control, random)
3. Design the experimental protocol
4. Choose user population
5. Apply for human subjects protocol review
6. Run pilot studies
7. Run the experiment
8. Perform statistical analysis
9. Draw conclusions
Example: Bubble Cursor
Lucid, Testable Hypothesis

H1: Users will acquire targets faster with the Bubble cursor (their movement time will be lower).

H2: Users will have a lower error rate with the Bubble cursor.

Other hypotheses?
Experiment Design

Testable hypothesis
Precise statement of expected outcome

Independent variables (factors)
Attributes we manipulate/vary in each condition
Levels – values for independent variables

Dependent variables (response variables)
Outcome of experiment (measurements)
Usually measure user performance
Experiment Design

Control variables
Attributes that will be fixed throughout experiment
Confound – attribute that varied and was not accounted for

Problem: Confound rather than IV could have caused change in DVs
Confounds make it difficult/impossible to draw conclusions

Random variables
Attributes that are randomly sampled
 Increases generalizability
Common Metrics in HCI

Performance metrics:

- Task success (binary or multi-level)
- Task completion time
- Errors (slips, mistakes) per task
- Efficiency (cognitive & physical effort)
- Learnability

Satisfaction metrics:

- Self-report on ease of use, frustration, etc.
Performance Metric: Errors
Performance Metric: Lostness

Smith 1996:
N: # of different pages visited
S: # of total pages visited, incl. revisits
R: minimum # of pages to accomplish task

Lostness = \sqrt{(N/S-1)^2 + (R/N-1)^2)

Optimum number of steps (three) to accomplish a task that involves finding a target item on Product Page C1 starting from the homepage.

Actual number of steps a participant took in getting to the target item on Product Page C1. Note that each revisit to the same page is counted, giving a total of eight steps.
Satisfaction Metric: Likert

Respondents rate their level of agreement to a statement.

Likert data is ordinal, not continuous (matters for analysis)!

“Overall, I am satisfied with the ease of completing the tasks in this scenario”

1: Strongly Disagree
2: Disagree
3: Neither agree nor disagree
4: Agree
5: Strongly agree
Variables for the Bubble Cursor

Independent variables

Dependent variables

Control variables

Random variables
Variables

**Independent variables**
- Cursor type (bubble, normal, area?)
- Target Distance
- Target Width (Effective vs. Actual?)

**Dependent variables**
- Movement Time
- Error Rate
- User Satisfaction

**Control variables**
- Color scheme, input device, screen size

**Random variables**
- Location, environment,
- Attributes of subjects
- Age, gender, handedness, …

Conducting studies online vs. in person strongly influences which variables are controlled and which are random.
Goals

**Internal validity**
Manipulation of IV is cause of change in DV
Requires eliminating confounding variables (turn them into IVs or RVs)
Requires that experiment is replicable

**External validity**
Results are generalizable to other experimental settings
*Ecological validity* – results generalizable to real-world settings

**Confidence in results**
Statistics
Experimental Protocol

What is the task? (must reflect hypothesis!)
What are all the combinations of conditions?
How often to repeat each combination of conditions?
Between subjects or within subjects
Avoid bias (instructions, ordering, …)
Between vs. Within Subjects

**Between subjects**
Each participant uses one condition
+/- Participants cannot compare conditions
+ Can collect more data for a given condition
- Need more participants

**Within subjects**
All participants try all conditions
+ Compare one person across conditions to isolate effects of individual diffs
+ Requires fewer participants
- Fatigue effects
- Bias due to ordering/learning effects
Within Subjects: Ordering Effects

In within-subjects designs ordering of conditions is a variable that can confound results. Why?
Run the Experiment

Always pilot it first!
Reveals unexpected problems
Can't change experiment design after starting it

Always follow same steps – use a checklist

Get consent from subjects

Debrief subjects afterwards
Are the Results Meaningful?

**Hypothesis testing**
Hypothesis: Manipulation of IV effects DV in some way
Null hypothesis: Manipulation of IV has no effect on DV
Null hypothesis assumed true unless statistics allow us to reject it

**Statistical significance (p value)**
Likelihood that results are due to chance variation
p < 0.05 usually considered significant (Sometimes p < 0.01)
Means that < 5% chance that null hypothesis is true

**Statistical tests**
T-test (1 factor, 2 levels)
Correlation
ANOVA (1 factor; > 2 levels, multiple factors)
MANOVA ( > 1 dependent variable)
Significance ≠ Relevance
There are people out there that are better at statistics than we are. Feel free to ask them for help.

E.g. Berkeley Statistics Department Consulting Service

http://www.stat.berkeley.edu/43
What are common evaluation designs?