Last Chances…

1. Check out a camera and tripod after class
2. Bring official DSP letter for special midterm accommodations to us

Charles Thacker wins Turing Award!

"Thacker created and collaborated on what would become the fundamental building blocks of the PC business. The Alto computer, developed in 1974, incorporated bitmap (TV-like) displays which enable modern graphical user interfaces (GUIs), including What You See Is What You Get (WYSIWYG) editors. These components have dominated computing during the last two decades."

Topics

1. Managing study participants (qualitative and quantitative studies)
2. Why do we conduct quantitative studies?
3. Designing controlled experiments
Managing Study Participants

The Participants' Standpoint

Testing is a distressing experience
Pressure to perform
Feeling of inadequacy
Looking like a fool in front of your peers, your boss,…
(from “Paper Prototyping” by Snyder)

The Three Belmont Principles

Respect for Persons
Have a meaningful consent process; give information, and let prospective subjects freely choose to participate

Beneficence
Minimize the risk of harm to subjects, maximize potential benefits

Justice
Use fair procedures to select subjects (balance burdens & benefits)

To ensure adherence to principles, most schools require Institutional Review Board approval of research involving human subjects.

Ethics: Stanford Prison Experiment

1971 Experiment by Phil Zimbardo at Stanford
24 Participants – half prisoners, half guards ($15 a day)
Basement of Stanford Psychology bldg turned into mock prison
Guards given batons, military style uniform, mirror glasses,…
Prisoners wore smocks (no underwear), thong sandals, pantyhose caps

Experiment quickly got out of hand
Prisoners suffered and accepted sadistic treatment
Prison became unsanitary/in hospitable
Prisoner riot put down with use of fire extinguishers
Guards volunteered to work extra hours

Zimbardo terminated experiment early
Grad student Christina Maslach objected to experiment
Important to check protocol with ethics review boards
(from Wikipedia)
Ethics

Was it useful?
“...that’s the most valuable kind of information that you can have - and that certainly a society needs it” (Zimbardo)

Was it ethical?
Could we have gathered this knowledge by other means?

http://www.prisonexp.org/slide-42.htm

Ethics (more recently)

“In 2001, a faculty member from the business school of a major university designed a study to see how restaurants would respond to complaints from putative customers. As part of the project, the researcher sent letters to restaurants falsely claiming that he and/or his wife had suffered food poisoning that ruined their anniversary celebration. The letters disclaimed any intention of contacting regulatory agencies and stated that the only intent was to convey to the owner what had occurred “in anticipation that you will respond accordingly.” Restaurant owners were understandably upset and some employees lost their jobs before it was revealed that the letter was a hoax.”

CITI Human Subject Training Material

Beneficience: Example

MERL DiamondTouch:
User capacitively coupled to table through seating pad.
No danger for normal users, but possibly increased risk for participants with pacemakers.
Inform subjects in consent!


Privacy and Confidentiality

Privacy: having control over the extent, timing, and circumstances of sharing oneself with others.
Confidentiality: the treatment of information that an individual has disclosed with the expectation that it will not be divulged

Examples where privacy could be violated or confidentiality may be breached in HCI studies!
### Treating Subjects With Respect

<table>
<thead>
<tr>
<th>Follow human subject protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual test results will be kept confidential</td>
</tr>
<tr>
<td>Users can stop the test at any time</td>
</tr>
<tr>
<td>Users are aware (and understand) the monitoring technique(s)</td>
</tr>
<tr>
<td>Their performance will not have implications on their life</td>
</tr>
<tr>
<td>Records will be made anonymous</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use standard informed consent form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Especially for quantitative tests</td>
</tr>
<tr>
<td>Be aware of legal requirements</td>
</tr>
</tbody>
</table>

### Conducting the Experiment

<table>
<thead>
<tr>
<th>Before the experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have them read and sign the consent form</td>
</tr>
<tr>
<td>Explain the goal of the experiment in a way accessible to users</td>
</tr>
<tr>
<td>Be careful about the demand characteristic</td>
</tr>
<tr>
<td>(Participants biased towards experimenter’s hypothesis)</td>
</tr>
<tr>
<td>Answer questions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>During the experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stay neutral</td>
</tr>
<tr>
<td>Never indicate displeasure with users performance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>After the experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debrief users (Inform users about the goal of the experiment)</td>
</tr>
<tr>
<td>Answer any questions they have</td>
</tr>
</tbody>
</table>

### Managing Subjects

<table>
<thead>
<tr>
<th>Don’t waste users time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use pilot tests to debug experiments, questionnaires, etc…</td>
</tr>
<tr>
<td>Have everything ready before users show up</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Make users comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep a relaxed atmosphere</td>
</tr>
<tr>
<td>Allow for breaks</td>
</tr>
<tr>
<td>Pace tasks correctly</td>
</tr>
<tr>
<td>Stop the test if it becomes too unpleasant</td>
</tr>
</tbody>
</table>

### If you want to learn more…


- The Belmont Report: Ethical Principles and Guidelines for the protection of human subjects of research
  - 1979 Government report that describes the basic ethical principles that should underly the conduct of research involving human subjects
**Why Quantitative Studies?**

**Qualitative Studies**
- **Qualitative:** What we've been doing so far
  - **Contextual Inquiry:** try to understand user’s tasks and conceptual model
  - **Usability Studies:** look for critical incidents in interface

**Qualitative methods help us:**
- Understand what is going on
- Look for problems
- Roughly evaluate usability of interface

**Quantitative Studies**
- **Use** to reliably measure some aspect of interface
- Compare two or more designs on a measurable aspect

**Approaches**
- Collect and analyze user events that occur in natural use
  - mouse clicks, key presses
- Controlled experiments

**Examples of measures**
- Time to complete a task
- Average number of errors on a task
- Users’ ratings of an interface *
  - Ease of use, elegance, performance, robustness, speed, ...

* You could argue that users’ perception of speed, error rates etc is more important than their actual values

**Comparison**
- **Qualitative studies**
  - Faster, less expensive → esp. useful in early stages of design cycle
  - In real-world design, quantitative study not always necessary

- **Quantitative studies**
  - Reliable, repeatable result → scientific method
  - Best studies produce generalizable results
Pilot User Study Assignment (after midterm)

You will conduct a **qualitative** study
We don’t have enough time or subjects for quantitative studies
But you should do a little quantitative analysis
What are your measures?
Compute summary statistics (mean, stdev)
Do you have independent, dependent, and control variables?

Designing Controlled Experiments

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: Chart Perception

![Chart Perception Examples](chart1.png)
Example: Chart Perception

Lucid, Testable Hypothesis

H1: Because users must mentally combine bars in Grouped Bar, comparisons will be slower than in other groups.

H2: Comparisons will be less accurate in Grouped Bar for non-leaf comparisons.

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
Variable Types

Nominal: categories with labels, no order

Ordinal: categories with rank order

Continuous:
  interval (w/o zero point), ratio (w/ zero point)

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

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

\[
\text{Lostness} = \sqrt{(N/S-1)^2 + (R/N-1)^2}
\]
Satisfaction Metric: Likert Scales

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

Independent variables
- Chart type
- Leaf Node vs Non-Leaf Node Comparison
- Data Density (# of Leaf Nodes)

Dependent variables
- Response Time
- Estimation Error
- User Satisfaction

Control variables
- Color scheme, rendering style

Random variables
- Location, environment, Attributes of subjects
  - Age, sex, ...

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, …)

Number of Conditions

Consider all combinations to isolate effects of each IV
(factorial design)
(2 chart types) * (3 leaf/non-leaf combinations) * (3 densities) = 27 combinations

Adding levels or factors can yield lots of combinations!

Reducing Num. of Conditions

Vary only one independent variable leaving others fixed

Problem: ?

Reducing Num. of Conditions

Vary only one independent variable leaving others fixed

Problem: Will miss effects of interactions
### Other Reduction Strategies

**Run a few independent variables at a time**
- If strong effect, include variable in future studies
- Otherwise pick fixed control value for it

**Fractional factorial design**
- Procedures for choosing subset of independent variables to vary in each experiment

### Choosing Subjects

**Pick balanced sample reflecting intended user population**
- Novices, experts
- Age group
- Sex
- ...

**Example**
- 12 non-colorblind right-handed adults (male & female)

**Population group can also be an IV or a controlled variable**

**What is the disadvantage of making population a controlled var?**

---

### Between Subjects Design

- Wilma and Betty use one interface
- Dino and Fred use the other

---

### Within Subjects Design

- Everyone uses both interfaces
**Between vs. Within Subjects**

**Between subjects**
- Each participant uses one condition
  + 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 differences
  - 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?*

- Turn it into a random variable
  - Randomize order of conditions across subjects
  - Counterbalancing (ensure all orderings are covered)
  - Latin square (partial counterbalancing)

**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

**Results: Statistical Analysis**

**Descriptive Statistics**
- Continuous data:
  - Central tendency (mean, median, mode),
  - Dispersion,
  - Shape of distribution
- Categorical data:
  - Frequency distributions
Descriptive Statistics: Error

What's missing from this bar chart?

Descriptive Statistics: Error

What is going on?

Descriptive Statistics: Time

Error bars indicate 95% confidence interval

Avg. Absolute Error (percentage points)

Grouped Bar 5.6
Line Icicle 15.54
Treemap 3.74

Avg. Absolute Error (percentage points)

Grouped Bar 7796
Line Icicle 8002
Treemap 11440
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)

Outliers

Q: What percentage is the SMALLER value of the LARGER value?

0 10 90 760-860

T-test

Compare means of 2 groups
- Null hypothesis: No difference between means

Assumptions
- Samples are normally distributed
- Very robust in practice
- Population variances are equal (between subjects tests)
- Reasonably robust for differing variances
- Individual observations in samples are independent
- Extremely important!

Correlation

Measure extent to which two variables are related
- Does not imply cause and effect
- Example: Ice cream eating and drowning
- Need a large enough sample size

Regression
- Compute the “best fit”
- linear
- logistic
ANOVA

Single factor analysis of variance (ANOVA)
Compare means for 3 or more levels of a single independent variable

Multi-Way Analysis of variance (n-Way ANOVA)
Compare more than one independent variable
Can find interactions between independent variables

Repeated measures analysis of variance (RM-ANOVA)
Use when > 1 observation per subject (within subjects exp.)

Multi-variate analysis of variance (MANOVA)
Compare between more than one dependent var.

ANOVA tests whether means differ, but does not tell us which means differ – for this we must perform pairwise t-tests

Which should we use for the menu selection example?

Our example
ANOVA: means for estimation error were significantly different
(F(2,185) = 5.32, p < .001)
But how do the means differ?
Use pairwise t-tests!
Diff in means for (Treemap, Grouped Bar) was NOT significant (t(122) = 1.26, p=0.05)
Diff in means for other pairs WERE significant.

Our Example
ANOVA: means for estimation time were not significantly different
(F(2,185)=1.05, p=0.35)
Takeaway: it's hard to measure response time when focus on task is not controlled.
What are some ways to achieve robustness?
Interactions

Are we done with our analysis? No!
Multiple IVs effect DV non-additively!
We had 3 IVs (chart type, density, node/leaf combo),
so we should investigate interaction effects!

Descriptive Statistics: Error

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>GB</th>
<th>LI</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>4.47</td>
<td>5.19</td>
<td>3.07</td>
<td>3.67</td>
</tr>
<tr>
<td>64</td>
<td>7.65</td>
<td>7.06</td>
<td>3.71</td>
<td>3.79</td>
</tr>
<tr>
<td>128</td>
<td>2.84</td>
<td>25.14</td>
<td>2.07</td>
<td>4.27</td>
</tr>
</tbody>
</table>

Descriptive Statistics: Time

<table>
<thead>
<tr>
<th></th>
<th>SD: GroupBar:</th>
<th>GB: 1479</th>
<th>LN: 2306</th>
<th>NN: 1630</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line icicle:</td>
<td>LL: 4704</td>
<td>LN: 3423</td>
<td>NN: 1917</td>
<td></td>
</tr>
<tr>
<td>Treemap:</td>
<td>LL: 1692</td>
<td>LN: 5404</td>
<td>NN: 1615</td>
<td></td>
</tr>
</tbody>
</table>
Draw Conclusions
What is the scope of the finding?
Does the experiment reflect real use?
External validity
Are there other parameters at play?
Internal validity

Summary
Quantitative evaluations
Repeatable, reliable evaluation of interface elements
To control properly, usually limited to low-level issues
Menu selection method A faster than method B

Pros/Cons
Objective measurements
Good internal validity \(\rightarrow\) repeatability
But, real-world implications may be difficult to foresee
Significant results doesn’t imply real-world importance
3.05s versus 3.00s for menu selection

Next Time
Midterm review!
No new readings – revisit old readings.
Heuristic Evaluation & Low-Fi Prototypes due!