**CMPE 235: User Evaluation of Technology**

**Lecture 1: User Study 101**

Course contact: srikur@soe.ucsc.edu

---

### Types of Study: Zikmund vs. Purists

- **Causal Study** (Study that looks for cause & effect)
- **Descriptive Study** (Study that describes)
- **Comparative Study** (Study that compares)
- **Associational Study** (Study that associates)

---

#### Exploratory Study

- Initial study conducted to explore (clarify and define) the nature of a problem
- Does not provide reasoning behind findings so subsequent study is expected
  - Helps to diagnose a situation
  - Screen alternatives
  - Discover new ideas
- Example: Do students procrastinate?

---

#### Descriptive Study

- Describes characteristics of a population or phenomenon
- Some understanding of the nature of the problem
- Deals with the who, what, where, when, how...but not the why.
- Weight Watchers average customer
  - Woman about 40 years old
  - Household income of about $50,000
  - At least some college education
  - Trying to juggle children/family and a job

---

#### Study Process

- **Problem Discovery and Definition**
- **Study Design**
- **Sampling**
- **Data Gathering**
- **Data Processing and Analysis**

- **Conclusions and Report**

- "The formulation of the problem is often more essential than its solution"
Problem Definition

- Decide on the overarching objectives
- Understand background of the problem
- Isolate/identify the problem, not symptoms
- Determine unit of analysis
- Determine relevant variables
- State study questions and objectives

Study Design

- Experimental
  - Quasi-experimental
  - Time series
  - Predictive
  - Correlational
  - Comparative
  - Descriptive
  - Surveys
  - Multivariate
  - Longitudinal

- Pilot Study
  - Initial study using a small amount of people to test a study question

- Exploratory
  - Small to moderate scale study exploring the study question

- Large Scale Study
  - Large scale study using many people to answer study questions

- Focus groups
  - Ethnographies
  - Interviews
  - Case study
  - Grounded theory
  - Narrative
  - Historical
  - Phenomenological
  - Triangulation

Methods

- Attitude Behavior
- Qualitative Quantitative
- Formative evaluation
- Summative evaluation
- Ethnographic research
- Participatory design
- Expert Evaluation
- Cognitive models
- Eye tracking
- Surveys
- Heuristic Evaluation
- Cognitive Walkthrough
- Cognitive models
- Card sorting
- Attitude

Cross-Sectional vs. Longitudinal

- A **cross-sectional** study is one that takes place at a single point in time → prone to chance results
- A **longitudinal** study is one that takes place over time
  - repeated measures - two or a few waves of measurement (pre- and post-tests included)
  - time series – many (20+) waves of measurement over time
  - But many things can affect validity
    - Mortality threat: dropouts
    - Individual difference in memory retention
    - Testing threat: need to make sure the instrumentation and setup are kept constant
    - Maturation threat: as individuals mature, their changes might affect the later tests

Experimental Designs

- Between subject design
  - each subject is allocated randomly to each of the experimental conditions (for experiment = treatment, control vs. experiment group)
  - concerns: need more people, prone to individual difference
- Matched subject design
  - subjects matched in pairs, pairs allocated randomly to each of the experimental conditions
- Within subject design
  - all subjects appear in all experimental conditions
  - concerns: fatigue, learning effects
- Single subject design
  - in-depth experiments on just one subject
  - must have a strong reason of why only 1 subject

Types of experiments

- True experiment – random assignment to groups
- Quasi experiment – no random assignment, but has a control group or multiple measures
- Non-experiment – no random assignment, no control, no multiple measures

Decision Tree:

1. Is random assignment used?
   - Yes: True experiment
   - No: Are there multiple groups or measures?
     - Yes: Quasi experiment
     - No: Non-experiment
Sampling

- Population
  - theoretical population: group you wish to generalize to
  - accessible population: subset of that population that is accessible to the experimenter
- sampling frame: list of the accessible population you’ll draw your sample from
- sample: group selected to be in your study

- External validity
  - the degree to which study conclusions would hold for other experimenters in other similar studies

- Internal validity
  - the approximate truth about inferences regarding cause-effect (causal) relationships

Sampling types

- Simple random sampling
  - based on random number generation
- Stratified random sampling
  - divide pop into homogenous subgroups, then simple random sample w/in
- Systematic random sampling
  - select every kth individual
- Cluster (area) random sampling
  - randomly select clusters, sample all units w/in cluster
- Multistage sampling
  - combination of methods

Sampling concept

- sampling distribution - the distribution of an infinite number of samples of the same characteristics as the sample in our study
- standard deviation: the spread of scores around the average in a single sample
- standard error, sampling error
  - the spread of averages around the average of averages in a sample distribution
  - indicates the precision of our statistical estimate
  - calculated based on standard deviation of sample
  - larger sample size → smaller standard error (statistical power)
- confidence interval: the interval associated with the probability the population will fall into

An example:

- Confidence interval
  - $\hat{p}(3.725<\mu<3.775) = 0.68$
  - $\hat{p}(3.700<\mu<3.800) = 0.95$
  - $\hat{p}(3.675<\mu<3.825) = 0.99$

- In general, for bell-shaped distribution
  - 68% of cases fall w/in 1 S.D.
  - 95% fall w/in 2 S.D.
  - 99% fall w/in 3 S.D.

  That’s why outlier is everything > 3 S.D.

The rest

- Data analysis: turning raw data into information
  - Ranges from simple frequency distributions to complex multivariate data analysis techniques.
- Conclusions and Report
  - Executive Summary/Abstract
  - Introduction, Motivation
  - Problem Definition, Hypotheses, Questions
  - Relevant Work
  - Methodology
  - Results, Findings, Discussion, Conclusions

Hypothesis

- Hypothesis: (unproven) proposition or question formulated for empirical testing → specific, measurable/testable
- Educated guess/interpolation
  - Ex: Flashing ads increase sales by 10% annually (as large display ads on Times Square do)

- Null hypothesis (H_0) vs. Alternative hypothesis (H_1)
  - No effect, e.g., “There will be no difference in completion times between the two groups” vs. there is an effect

- One-tailed vs. two-tailed
  - The prediction specifies a direction, e.g., “Increased icon size decreases click time” vs. “Icon size affects click time”

- The goal of testing and analysis is usually to see if there is enough evidence to reject H_0.
Type I and II errors

- Type I: mistake of rejecting Ho when it is true and should not be rejected
  - The probability of making this error = \( \alpha \) (or p value)
  - Usually chosen at \( p=0.05, 0.01 \) or 0.001
- Type II: mistake of not rejecting Ho when it is false and should be rejected; the probability of making this error is called \( \beta \)
- Statistical power = 1 - \( \beta \)

<table>
<thead>
<tr>
<th>Reality</th>
<th>Type I (( \alpha ))</th>
<th>Type II (( \beta ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>No difference</td>
<td>Mouse is faster than joystick</td>
<td></td>
</tr>
</tbody>
</table>

Study conclusion

No difference

Mouse is faster than joystick

Variables

- Variable: any entity that can take on different values.
- Attribute: a specific value of a variable.
- Independent variable: things you vary (mouse vs. joystick)
- Dependent variable: things of interest (click time)
- Control: things you don’t want to interfere (task order, gender or age or skill composition between IV)
  - Randomize the composition (random generator, dice, coin)
  - Balance (almost equal composition)
  - Keep constant (20 min time limit, all >55 years)

Good variable characteristics

- exhaustive – include all possible responses (all possible values of Windows icon sizes)
- mutually exclusive - no response should be able to have two attributes simultaneously (2-4” and 4-6”)

Choosing measurement scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Characteristics</th>
<th>e.g.</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Numbers identify &amp; classify objects</td>
<td>SSN, # sport players</td>
<td>Percentages, mode</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Relative positions, uneven distances</td>
<td>Rankings of teams in a tournament</td>
<td>Percentile, median</td>
</tr>
<tr>
<td>Interval</td>
<td>Even distances, arbitrary zero</td>
<td>Temperatures (Celsius)</td>
<td>Range, mean, S.D.</td>
</tr>
<tr>
<td>Ratio</td>
<td>Fixed zero, even distances</td>
<td>Length, weight</td>
<td>Range, mean, S.D.</td>
</tr>
</tbody>
</table>

More detailed statistical tests

<table>
<thead>
<tr>
<th>Level of Measurement</th>
<th>Sample Characteristics</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sample</td>
<td>2 Samples</td>
<td>K Samples (&gt;2)</td>
</tr>
<tr>
<td>Independent</td>
<td>Independent</td>
<td>Independent</td>
</tr>
<tr>
<td>Categor/ Nominal</td>
<td>X^2 or binom</td>
<td>X^2</td>
</tr>
<tr>
<td>Rank or Ordinal</td>
<td>Mann Whitney U</td>
<td>Matched Pairs Signed Ranks</td>
</tr>
<tr>
<td>Parametric (Interval &amp; Ratio)</td>
<td>z/t test between groups</td>
<td>t test between groups</td>
</tr>
<tr>
<td>Factorial (2 way) ANOVA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parametric and Nonparametric Tests

- Parametric tests estimate at least one parameter (in t-test it is population mean)
  - Usually for normal distributions and when the dependent variable is interval/ratio
  - More likely to inflate \( \alpha \)
  - Prone to violation to normality of data
- Nonparametric tests do not test hypothesis about specific population parameters
  - Distribution-free tests
  - Although appropriate for all levels of measurement most frequently applied for nominal or ordinal measures
  - Easier to compute and have less restrictive assumptions
A Classification of Scaling Techniques

Comparative Scaling
- Direct comparison of stimulus objects.
- Data must be interpreted in relative terms and have only ordinal or rank order properties.
- Disadvantages
  - Ordinal nature of the data
  - Inability to generalize beyond the stimulus objects scaled.
- Advantages
  - Small differences between stimulus objects can be detected.
  - Same known reference points for all respondents.
  - Easily understood and can be applied.
  - Involve fewer theoretical assumptions.
  - Tend to reduce halo or carryover effects from one judgment to another.

Comparative Scaling
- Paired comparison
  - Kittenwar!
- Rank order
  - See all alternatives, rank them (can’t have the same rank)
- Constant sum
  - Give users 100 pt, ask to distribute among alternatives
- Q-sort
  - Divide and conquer

Noncomparative Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Characteristics</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Mark on a continuous line</td>
<td>Easy to construct</td>
<td>Subjective analysis</td>
</tr>
<tr>
<td>Likert</td>
<td>Degrees of agreement</td>
<td>Easy to construct, administer, understand</td>
<td>Time consuming</td>
</tr>
<tr>
<td>Semantic</td>
<td>7-pt scale on bipolar</td>
<td>Versatile</td>
<td>Controversy as to whether data are</td>
</tr>
<tr>
<td>differential</td>
<td>labels</td>
<td></td>
<td>interval</td>
</tr>
<tr>
<td>Stapel</td>
<td>Unipolar 10-pt scale (-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>to +5) without zero</td>
<td>Easy to construct and administer</td>
<td>Can be confusing to respondents</td>
</tr>
</tbody>
</table>

Scale Evaluation

Measurement Accuracy

The true score model provides a framework for understanding the accuracy of measurement.

\[ X_O = X_T + X_S + X_R \]

where
- \( X_O \) = the observed score or measurement
- \( X_T \) = the true score of the characteristic
- \( X_S \) = systematic error
- \( X_R \) = random error
Potential Sources of Measurement Errors

1) Relatively stable characteristics of the individual, such as intelligence, social desirability, and education.
2) Short-term or transient personal factors, such as health, emotions, and fatigue.
3) Situational factors, such as the presence of other people, noise, and distractions.
4) Sampling of items included in the scale: addition, deletion, or changes in the scale items.
5) Lack of clarity of the scale, including the instructions or the items themselves.
6) Mechanical factors, such as poor printing, overcrowding items in the questionnaire, and poor design.
7) Administration of the scale, such as differences among interviewers.
8) Analysis factors, such as differences in scoring and statistical analysis.

Reliability

- **Reliability** can be defined as the extent to which measures are free from random error, $X_R$. $X_R = 0$ → perfectly reliable.
- In **test-retest reliability**, identical sets of scale items at two different times and the degree of similarity between the two measurements is determined.
- In **alternative-forms reliability**, two equivalent forms of the scale are constructed and the same respondents are measured at two different times.
- **Internal consistency reliability** determines the extent to which different parts of a summated scale are consistent in what they indicate about the characteristic being measured.
- In **split-half reliability**, the items on the scale are divided into two halves and the resulting half scores are correlated.
- The Cronbach's alpha, is the average of all possible split-half coefficients (different ways of splitting the items).

Validity

- The **validity** of a scale = the extent to which differences in scale scores reflect true differences among objects being measured. Perfect validity = no measurement error ($X_O = X_T$).
- **Content validity** is how well the content of a scale represents the measurement task at hand.
- **Criterion validity** = whether a scale performs as expected in relation to other variables selected as meaningful criteria.
- **Construct validity** = what construct or characteristic the scale is, in fact, measuring.
- **Convergent validity** is how much the scale correlates positively with other measures of the same construct.
- **Discriminant validity** is the idea that the scale must not correlate with scales from other constructs.
- **Nomological validity** is how much the scale correlates in theoretically predicted ways with measures of different but related constructs.

Relationship Between Reliability and Validity

- If a measure is perfectly valid, it is also perfectly reliable. In this case $X_O = X_T$, $X_R = 0$, and $X_S = 0$.
- If a measure is unreliable, it cannot be perfectly valid, since at a minimum $X_O = X_T + X_R$. Furthermore, systematic error may also be present, i.e., $X_S ≠ 0$. Thus, unreliability implies invalidity.
- If a measure is perfectly reliable, it may or may not be perfectly valid, because systematic error may still be present ($X_O = X_T + X_S$).
- Reliability is a necessary, but not sufficient, condition for validity.