HUMAN-COMPUTER INTERACTION

MEASURING IN HCI RESEARCH

Professor Bilge Mutlu

Computer Sciences, Psychology, & Industrial and Systems Engineering
University of Wisconsin–Madison

CS/Psych-770 Human-Computer Interaction
LAST WEEK
REVIEW

What is the most basic form of a research question?
Independent vs. dependent variables?
What is a factor? What is a level?
Fixed vs. random factors?
How are hypotheses determined?
What is an interaction effect?
Why follow a factorial design?
Demand characteristics, transfer effects; when might they be problematic?

What is a bias vs. power tradeoff? Why do we have to choose?

What is a mixed design?

What is an example of a control condition?

What are consideration in you identifying participants?
MEASUREMENT
WHAT DO WE MEASURE?
VARIABLES

Variables are things that change
E.g., gender, preference, performance

Attributes qualify variables
E.g., male vs. female, high-performance, low-performance

Quantitative measurements describe the degree of an attribute
E.g., an IQ of 110, an under-three-hour marathon runner

Qualitative measurements describe subjective observations
E.g., “the first customer was a tall man”
TYPES OF VARIABLES

Nominal data are names of groups or categories
   E.g., males vs. females, American vs. Japanese

Ordinal data is a rank-ordering of measurements
   E.g., very satisfied, satisfied, neutral, unsatisfied, very unsatisfied

Interval data are measurements along a scale with no real zero
   E.g., happiness in a scale of 1 to 7

Ratio data are measurements along a scale with a real zero
   E.g., a person’s weight
TYPES OF VARIABLES

- Ratio
- Interval
- Ordinal
- Nominal
## Types of Variables

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinctiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Rank ordering</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Equal intervals</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Absolute zero</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
TYPES OF VARIABLES

Descriptive
E.g., “a tall man”

Categorical
E.g., male vs. female, high vs. mid. vs. low

Numeric
E.g., age

Discrete
E.g., subjective ratings of an interface from 1 to 7

Continuous
E.g., performance measures
QUESTIONS?
TYPES OF MEASUREMENTS
TYPES OF MEASUREMENTS

**Objective** measurements

Data directly measured from participants, comparable across participants
E.g., performance in a knowledge test

**Behavioral** measurements

Data on the actions and behaviors of participants
E.g., how much eye-contact participants maintain with a robot

**Subjective** measurements

Data participants evaluate subjectively; comparisons across participants are less meaningful
E.g., preferences, personality

**Physiological** measurements

Data measured directly from participants’ bodies
E.g., body temperature, GSR, EEG, EMG, fMRI
WHAT MAKES MEASUREMENTS GOOD?
FACTORS AFFECTING GOODNESS

Measurement quality

Reliability of measurements, task, context, analyses, etc.

Validity of measurements, task, context, analyses, etc.
MEASUREMENT ERROR
MEASUREMENT ERROR
(OBSERVATIONAL ERROR)

The difference between the measurement and the true quantity of the variable

\[ X = T + e_r + e_s \]

Observed measurement is *what is recorded*

True measurement is *what the true value is*

Measurement error are *distortions* that case the observed measurement to be different from true quantities
RANDOM ERROR

Inherent in any measure that randomly varies

E.g., the baseline “mood” participants might be in when they come in

Only affects variance, not the mean of measurements

Also called “noise”
SYSTEMATIC ERROR

Caused by external factors
E.g., delay in coding, noise during test, camera delay

Consistently affects the mean

Also called “bias”
HOW TO REDUCE ERRORS?

Pilot test instruments
If coders are used, train them, measure reliability
If data entered manually, repeat data entry
Use statistical methods to measure error
Use multiple measures
QUESTIONS?
RELIABILITY
DEFINITIONS

The reliability of a measure defines its **consistency** across repeated measurements and judgments.

E.g., more robot gaze leads to better information recall; could we replicate this result with a second set of subjects or with the same subject another time?

The more error there is the less reliable the measure is.
ESTIMATING RELIABILITY

Reliability can be estimated using statistical methods

\[ R = \frac{\nu_{\text{true}}}{\nu_{\text{true}} + \nu_{\text{error}}} \]

Provides a value between 0 and 1.

Rule of thumb

A reliability of .70 and higher is acceptable.
HOW TO ENSURE RELIABILITY

Test-retest reliability

The same test is repeated with the same group at another time

Alternative-form method

A second test with similar measures given to the same set of people

Split-half technique

The test is split into half and the results from the two are correlated
# PROS & CONS

<table>
<thead>
<tr>
<th>Reliability method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Test-retest</em></td>
<td>Uses the same test items</td>
<td>First testing may contaminate the second</td>
</tr>
<tr>
<td></td>
<td>Simple to administer</td>
<td>Respondent may change with time</td>
</tr>
<tr>
<td><em>Alternative-form</em></td>
<td>Minimizes repeat-item contamination</td>
<td>Use of different items lowers reliability</td>
</tr>
<tr>
<td></td>
<td>Little time passes before retesting</td>
<td>Requires a longer test</td>
</tr>
<tr>
<td></td>
<td>Useful for pre/post-testing</td>
<td></td>
</tr>
<tr>
<td><em>Split-half</em></td>
<td>Minimizes repeat-item contamination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No time passes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Done at a single session</td>
<td></td>
</tr>
</tbody>
</table>
INTERNAL RELIABILITY

Inter-item correlation

Mean of all pairwise correlations across items of a measure

Split-half correlation

Correlations between two randomly-split halves of the measure

Cronbach’s alpha

Iterative calculation of inter-item correlations across randomly-selected subsets of the measure

Rule of thumb

An $\alpha$ of .70 or above is acceptable.
INTER-CODER RELIABILITY

The extent to which independent coders evaluate a behavior to reach the same conclusion

Not very easy — requires a lengthy, rigorous process
INTER-CODER RELIABILITY MEASURES

Agreement between raters

How much two rates agree — too coarse of a measure

Cohen’s kappa

Takes into account agreement occurring by chance

Rule of thumb

A \( \kappa \) of .80 and above indicates substantial agreement

Fishers’ kappa, Krippendorff’s alpha

Alternative methods
PROCESS

Select one or more appropriate indices

Obtain the necessary tools to calculate the index or indices selected

Select an appropriate minimum acceptable level of reliability for the index or indices to be used

Assess reliability informally during coder training

Assess reliability formally in a pilot test

Assess reliability formally during coding of the full sample

Select and follow an appropriate procedure for incorporating the coding of the reliability sample into the coding of the full sample

Report inter-coder reliability in a careful, clear, and detailed manner
QUESTIONS?
VALIDITY
MEASURE VALIDITY

Whether we are measuring what we want to measure
E.g., measuring aggression in children

Measure the amount of time children play with

  Aggressive toys (guns, swords, tanks)

  Non-aggressive toys (trucks, tools, dolls)

Challenges

  They might be playing with toys that they are more familiar with — they see guns and tanks on TV all the time

  Children might play with trucks and dolls in aggressive ways as well
FACE VALIDITY

How much a measure “appears” to be measuring what it intends to measure

Not statistical, involves judgment

Face validity ≠ Validity

A measure with face validity might not be valid overall

A measure without face validity might be valid overall
CONSTRUCT VALIDITY

How much conceptual constructs relate to they intend to measure

A construct is conceptual formulation of a high-level phenomena of interest for measurement

Measures with high construct validity should relate appropriately with other measures

E.g., a measure of self-esteem should correlate positively with optimism
# A CONSTRUCT OF SELF-ESTEEM

<table>
<thead>
<tr>
<th>+ keyed</th>
<th>– keyed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feel comfortable with myself.</td>
<td>Dislike myself.</td>
</tr>
<tr>
<td>Just know that I will be a success.</td>
<td>Am less capable than most people.</td>
</tr>
<tr>
<td>Seldom feel blue.</td>
<td>Feel that my life lacks direction.</td>
</tr>
<tr>
<td>Like to take responsibility for making decisions.</td>
<td>Question my ability to do my work properly.</td>
</tr>
<tr>
<td>Know my strengths.</td>
<td>Feel that I'm unable to deal with things.</td>
</tr>
</tbody>
</table>
EMPIRICAL VALIDITY
(CRITERION-RELATED VALIDITY)

How much results from a measure relate to more established measures of the phenomena

Concurrent validity

A valid measure should correlate with other, existing measures

E.g., PhD qualifier scores should correlate with CGPA in area courses

Predictive validity

A valid measure should predict future actions

E.g., SAT scores should predict college graduation CGPAs
CONVERGENT & DISCRIMINANT VALIDITY

Convergent validity

A measure is *correlated* with another measure assessing the same phenomena

E.g., intellect and competence should correlate

Discriminant validity

A measure *not correlating* with measures that it should not correlate with

E.g., intellect and control should *not* correlate
CONTENT VALIDITY

How much a construct captures the full extent of the phenomena of interest for measurement

E.g., the GREs verbal test captures vocabulary but not grammar, understanding, or communication
ECOLOGICAL VALIDITY

How much the measurements represent real-world phenomena of interest

E.g., the ability to perceive dots on a screen fast might not help with detecting cars in traffic
RELIABILITY IS A NECESSARY CONDITION FOR VALIDITY BUT NOT VICE VERSA
QUESTIONS?
THANKS!

Professor Bilge Mutlu
Computer Sciences, Psychology, & Industrial and Systems Engineering
University of Wisconsin–Madison

CS/Psych-770 Human-Computer Interaction