University of Colorado Boulder

Human-Robot Interaction

Measuring in HRI Research I

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder

Video of the Day

Atlas, The Next Generation

Boston Dynamics

Previously…

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?

Review

Demand characteristics, transfer effects

- What are they
- 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 considerations in identifying participants?

Measurement

What do we measure?

Variables

Variables are things that change or take on different values

E.g., gender, preference, performance

Attributes qualify variables

E.g., male/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, Americans vs Japanese

Ordinal data represents a rank-ordering of measurements

E.g., very satisfied, satisfied, neutral, unsatisfied, very unsatisfied

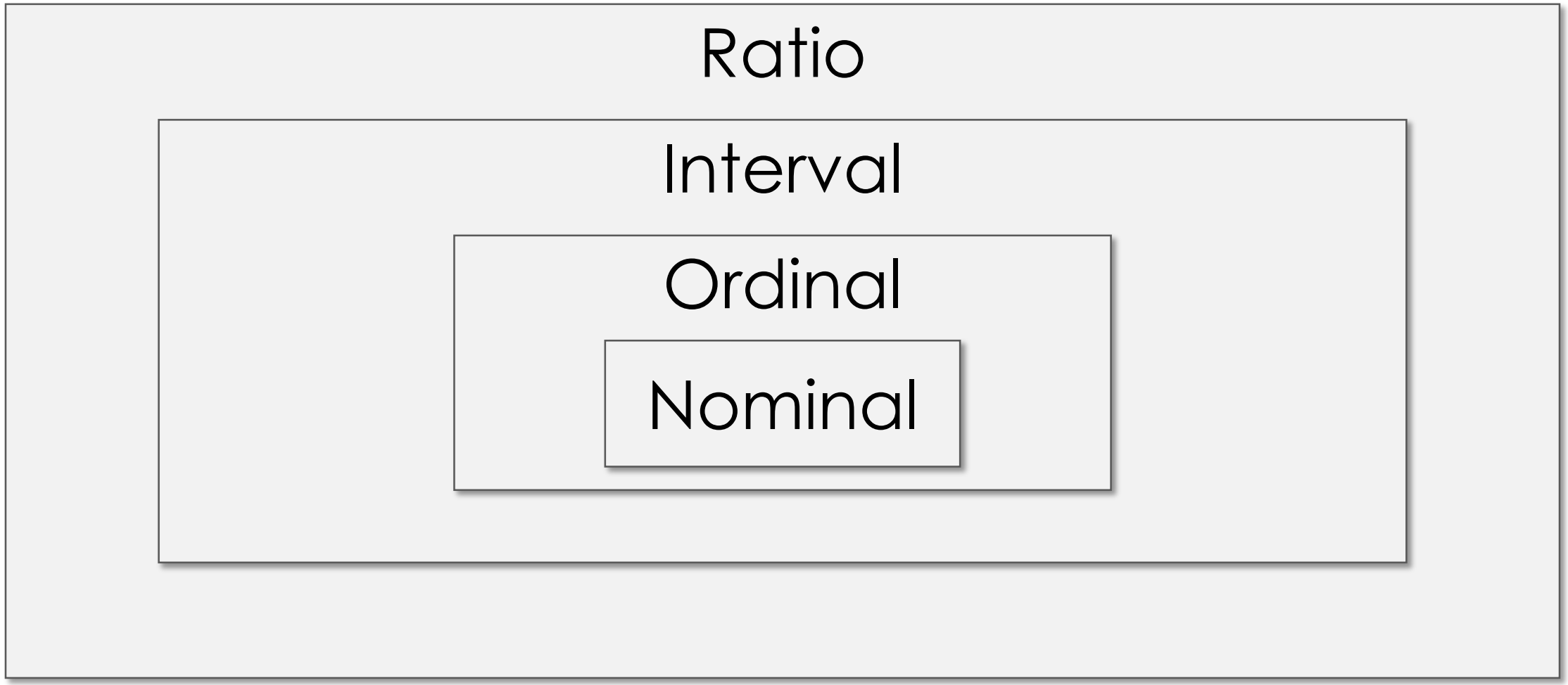
Interval data are measurements along a scale with no real zero

E.g., a slider for inputting "happiness"; IQ; temperature in Celsius

Ratio data are measurements along a scale with a real zero

E.g., a person's weight

Types of Variables



Types of Variables

	Nominal	Ordinal	Interval	Ratio
Distinctiveness	*Yes*	*Yes*	*Yes*	*Yes*
Rank ordering	*No*	*Yes*	*Yes*	*Yes*
Equal intervals	*No*	*No*	*Yes*	*Yes*
Absolute zero	*No*	*No*	*No*	*Yes*

Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔

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

Data directly measured from participants

Comparable across participants

E.g., performance in a knowledge test

Subjective

Data participants evaluate subjectively

Comparisons across participants are less meaningful

E.g., preferences, personality

Behavioral

Data on the actions and behaviors of participants

E.g., how much eye-contact participants maintain with a robot

Physiological

Data measured directly from participants' bodies

E.g., body temperature, EEG, EMG, fMRI, fNIRS, EDA

What makes measurements good?

Factors affecting “Goodness”

Measurement quality

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

Validity of measurements, task, context, analysis, etc.

Measurement Error

Measurement Error
(observational error)

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

$$\boldsymbol{X = T + e_r + e_s}$$

Observed measurement is *what is recorded*

True measurement is *what the true value is*

Measurement error is the *distortions* that cause the observed measurement to be different from true quantities

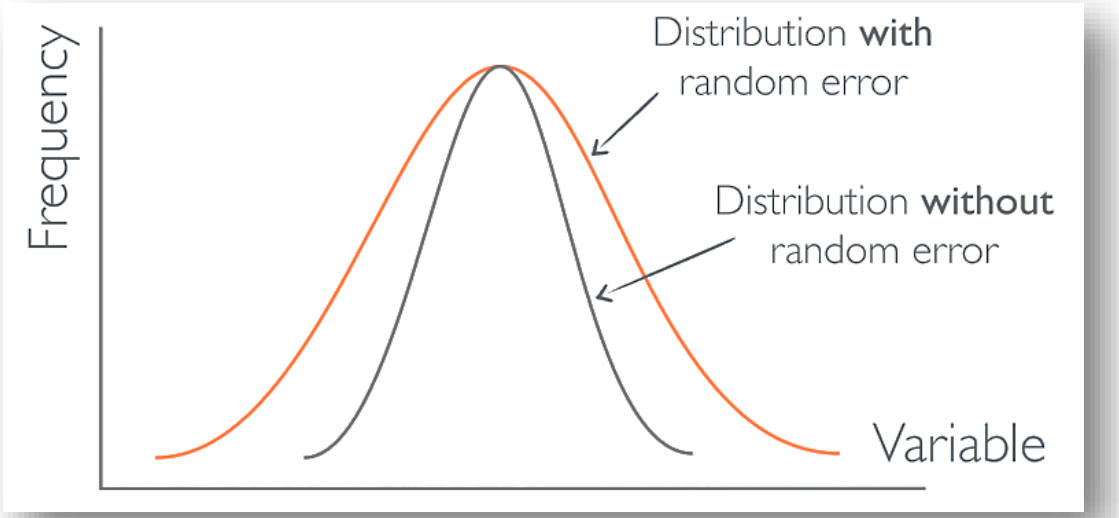
Random Error

Inherent in any measure that randomly varies

- E.g., the baseline "mood" participants might be in when they arrive

Often affects variance of measurements (not mean)

Also called "noise"



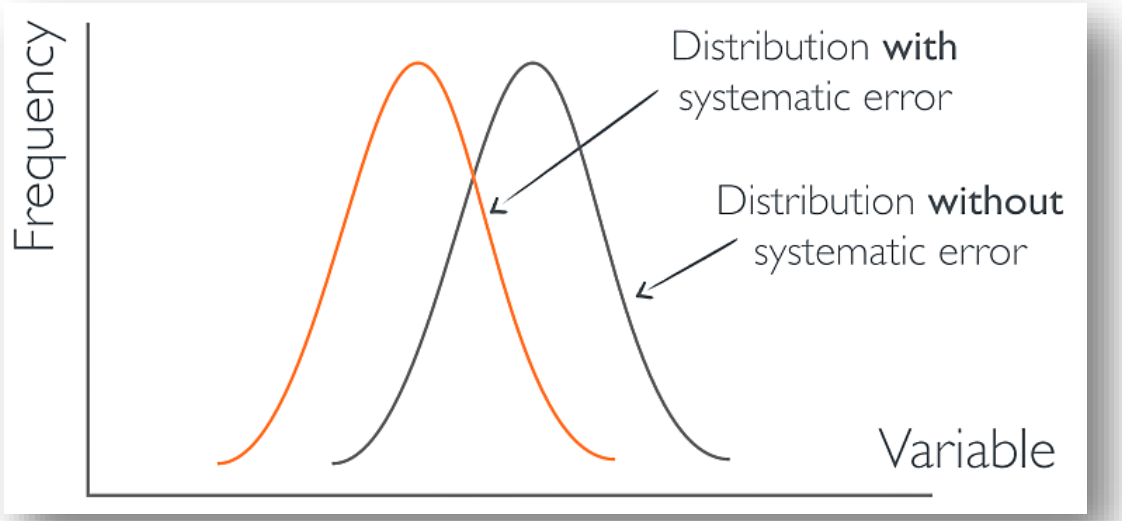
Systematic Error

Caused by external factors

- E.g., delay in coding, noise during test, camera delay, slow robot actuators

Consistently affects the mean (not the variance)

Also called “bias”



How to Reduce Errors?

Pilot test instruments

If coders are used, train them and measure reliability

If data is entered manually, repeat data entry

Use statistical measures to measure error

Use multiple measures

Questions?

Reliability

Definitions

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

- E.g., we find that 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

Types of Reliability

External Reliability ("test-retest reliability")

Extent to which a measure varies from one use to the next

Internal Reliability ("internal consistency reliability")

Extent to which a measure is consistent with itself

Inter-rater Reliability

Degree of agreement between two or more raters

Estimating Reliability

Reliability can be estimated with statistical methods

$$\boldsymbol{R = \frac{v_{true}}{(v_{true} + v_{error})}}$$

Provides a value between 0 and 1

Rule of thumb

- Reliability of **.70** and higher is acceptable

Reliability Methods

Test-retest reliability

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

Alternative-form method (“parallel-forms”)

- 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

Reliability method	Pros	Cons
Test-retest	Uses the same test items Simple to administer	First testing may contaminate the second Respondent may change with time
Alternative-form	Minimizes repeat-item contamination Little time passes before retesting Useful for pre/post-testing	Must develop second, equivalent test May be impossible to truly validate
Split-half	Minimizes repeat-item contamination No time passes Done at a single session	Can't be used for all applications (e.g., learning)

Estimates of Internal Reliability

Inter-item correlation

Split-half correlation

Cronbach's alpha

Estimates of Internal Reliability

Inter-item correlation

Mean of all pairwise correlations across items of a measure

Table 2. Correlation coefficients between the items (initial version - 11 questions).

Q	1	2	3	4	5	6	7	8	9	10	11
1	1										
2	0.869	1									
3	0.851	0.775	1								
4	0.445	0.487	0.513	1							
5	0.440	0.435	0.415	0.786	1						
6	0.273	0.288	0.273	0.495	0.564	1					
7	0.394	0.373	0.296	0.633	0.747	0.784	1				
8	0.459	0.542	0.359	0.459	0.271	0.153	0.323	1			
9	0.467	0.576	0.382	0.288	0.113	-0.112	0.121	0.780	1		
10	0.705	0.779	0.500	0.289	0.124	0.199	0.280	0.662	0.672	1	
11	0.252	0.306	0.098	0.327	0.373	0.031	0.428	0.396	0.368	0.249	1

Average value of the inter-item correlations. R=0.425.

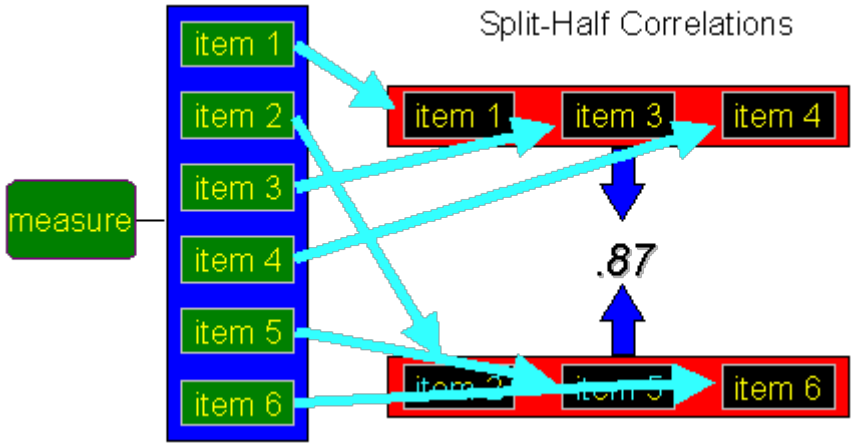
Estimates of Internal Reliability

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



Estimates of 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 of **α** of **.70** or above is acceptable

Inter-rater Reliability
(inter-coder reliability)

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

Not easy – requires a lengthy, rigorous process

Tradeoffs

Inter-rater Reliability Measures

- Percent agreement between raters
  - How much two raters agree – often too coarse of a measure
- Cohen's kappa: $p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$
  - Takes into account agreement occurring by chance
  - Rule of thumb
    - A of **κ** of **.80** and above indicates substantial agreement
- Alternative methods
  - Fishers' kappa, Krippendorff's alpha

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

C. Analysis & Findings

Video recordings of each interaction were analyzed to identify high-level patterns of user behaviors (Figure 3). After a period of iteration examining and refining observed behaviors, eight dominant patterns emerged for classifying participant activities. Two coders annotated video data from each interaction with these eight classifications, which could consist of a single utterance (often seen in sequential commands) or a series of back-and-forth utterances between the participant and confederate (often seen in refinement behaviors). Data was divided evenly between the coders, with an overlap of 12.5% of the data coded by both. Inter-rater reliability analysis revealed substantial agreement between the raters (Cohen's $\kappa = .72$)

Questions?

Validity

Measure Validity
(internal 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 – might see guns and tanks more often on TV
  - Children might play with trucks and dolls in an aggressive manner as well

Face Validity

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

Not statistical, involves judgement

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 what they intend to measure

A construct is a 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

+ keyed
  - Feel comfortable with myself
  - Just know that I will be a success
  - Seldom feel blue
  - Like to take responsibility for making decisions
  - Know my strengths

- keyed
  - Dislike myself
  - Am less capable than most people
  - Feel that my life lacks direction
  - Question my ability to do my work properly
  - Feel that I'm unable to deal with things

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., SAT scores should correlate with ACT scores

Predictive validity

- A valid measure should predict future actions
- E.g., SAT scores should predict college graduation GPAs

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 height should *not* correlate

Content Validity

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

- E.g., the GRE verbal test captures vocabulary but not grammar, understanding, or communication

Ecological Validity
(external validity)

How much the measurements represent real-world phenomena

- E.g., the ability to quickly perceive dots on a screen may not help with detecting cars in traffic

Reliability is a necessary condition for validity but not vice versa

Questions?

Next

Assignment #2

Due midnight tonight Monday 3/4

See guidelines and submission template on Moodle

Project

Work with your groups to determine experimental design (if appropriate) as in assignment 2

Upcoming IRB assignment (to be posted on Moodle)

University of Colorado
Boulder

THANKS!

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder