CS/Psych/EdPsych 770 Human-Computer Interaction

Measurement Basics

Yuhang Zhao
Computer Sciences, UW-Madison

Today's Agenda

- Topic overview: *Measurement Basics*
- Hands-on Activity: *Objective Measures*

What do we measure when we measure?

- **Definition**: Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events.

But it's not just numbers! We can measure:

- **Quantitative measurements** describe the degree of an attribute
  - e.g., an under-three-hour marathon runner, someone who scores 1600 in SAT
- **Qualitative measurements** describe subjective observations
  - e.g., "the first customer was a tall man"

What are different types of variables

- **Nominal** (categorical) 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, temperature, Likert-scale: happiness in a scale of 1 to 7
- **Ratio** data are measurements along a scale with a real zero
  - E.g., the weight of something

	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

Other items you might hear…

- **Descriptive**, e.g., “a tall man”
- **Categorical**, e.g., novice vs. expert, high vs. mid vs. low
- **Numerical**, e.g., age
- **Discrete**, e.g., subjective ratings of an interface from 1 to 7
- **Continuous**, e.g., performance measures

What are different kinds of measurements we can take?

- **Objective**: Measurement from participants against an objective standard, e.g., performance in a test
- **Behavioral**: Measurement of the actions and behaviors of participants, e.g., how much eye-contact participants maintain with a robot
- **Subjective**: Measurement of self-report data on subjective evaluations, e.g., preferences, personality
- **Physiological**: Measurements taken directly from participants' bodies, e.g., body temperature, GSR, EEG (electroencephalogram), EMG, fMRI

What makes measurements good?

- Validity
- Reliability
- Quality

What is validity?

- **Definition**: Validity is the extent to which a concept, conclusion, or measurement is well-founded and likely corresponds accurately to the real world
- In other words, are we measuring what we want to measure?

What is an example of validity problem?

- Consider wanting to measure **aggression in children**
- We can measure the amount of time children play with: *aggressive toys* (guns, swords, tanks) vs. *non-aggressive toys* (trucks, tools, dolls)
- What are threats to validity?
  - Children might be playing with toys that they are more familiar with, e.g., they see guns and tanks on TV all the time
  - Children can also play with trucks and dolls in aggressive ways

What are different forms of validity

- Face validity
- Construct validity
- Criterion validity
- Content validity
- Ecological validity

Face validity

- **Definition**: The extent to which a measure appears valid
- Based on logical reasoning and judgement, not statistical

Construct validity

- **Definition**: The extent to which the method of measurement matches the conceptual *construct* to be measured.

What is a construct?

- **Definition**: A psychological construct is a label for a cluster or domain of covarying behaviors.
- A high construct validity means that the measurement sufficiently captures and covers the construct.

A measure of **introversion**

Positively keyed

- Don’t like to draw attention to myself
- Keep in the background
- Dislike being the center of attention
- Don’t talk a lot

Negatively keyed

- Don’t mind being the center of attention
- Take charge
- Want to be in charge
- Am the life of the party
- Can talk others into doing things

Two types of construct validity

- **Convergent validity**: The extent to which the measure is associated with things it should be associated with, e.g., *depression* and *self-confidence* should correlate.
- **Discriminant validity**: The extent to which the measure is not associated with things it should not be associated with, *depression* and *height* should not correlate.

Criterion validity (empirical validity)

- **Definition**: The extent to which the results from the measure relate to existing, well-established measures.

- Two forms of criterion validity
  - **Concurrent validity**: the extent to which the measure correlates with other measures of the same construct taken at the same time
  - **Predictive validity**: the extent to which the measure can predict other measures of the same construct that will be taken in the future

Content validity

- **Definition**: The extent to which the test is representative of all aspects of the construct.
- E.g., GRE's verbal test captures vocabulary but not grammar, understanding, or communication

Ecological validity

- **Definition**: The extent to which research results can be applied to real-life situations
- E.g., the ability to perceive dots on a screen fast might not help with detecting cars in traffic

What is reliability?

- **Definition**: Reliability in statistics and psychometrics defines the consistency of a measure across repeated measurements and judgements
- E.g., more robot gaze leads to better information recall; could we replicate this results with a second set of subjects or with the same subject another time?
- High reliability indicates that the measure produces similar results under consistent conditions

How do we ensure reliability?

- Over time: test-retest reliability
- Across items: internal reliability
- Across researchers: inter-coder reliability

Test-retest

- **Definition**: Test-retest reliability evaluates the consistency of a measure over time
- Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the *same* group of people at a later time, and then looking at *test-retest correlation* between the two sets of scores.

Internal reliability

- **Definition**: A measure of whether several items that propose to measure the same general construct produce similar scores.

What are measures 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 α**: Conceptually, α is the mean of all possible split-half correlations for a set of items (α > .7 desirable)

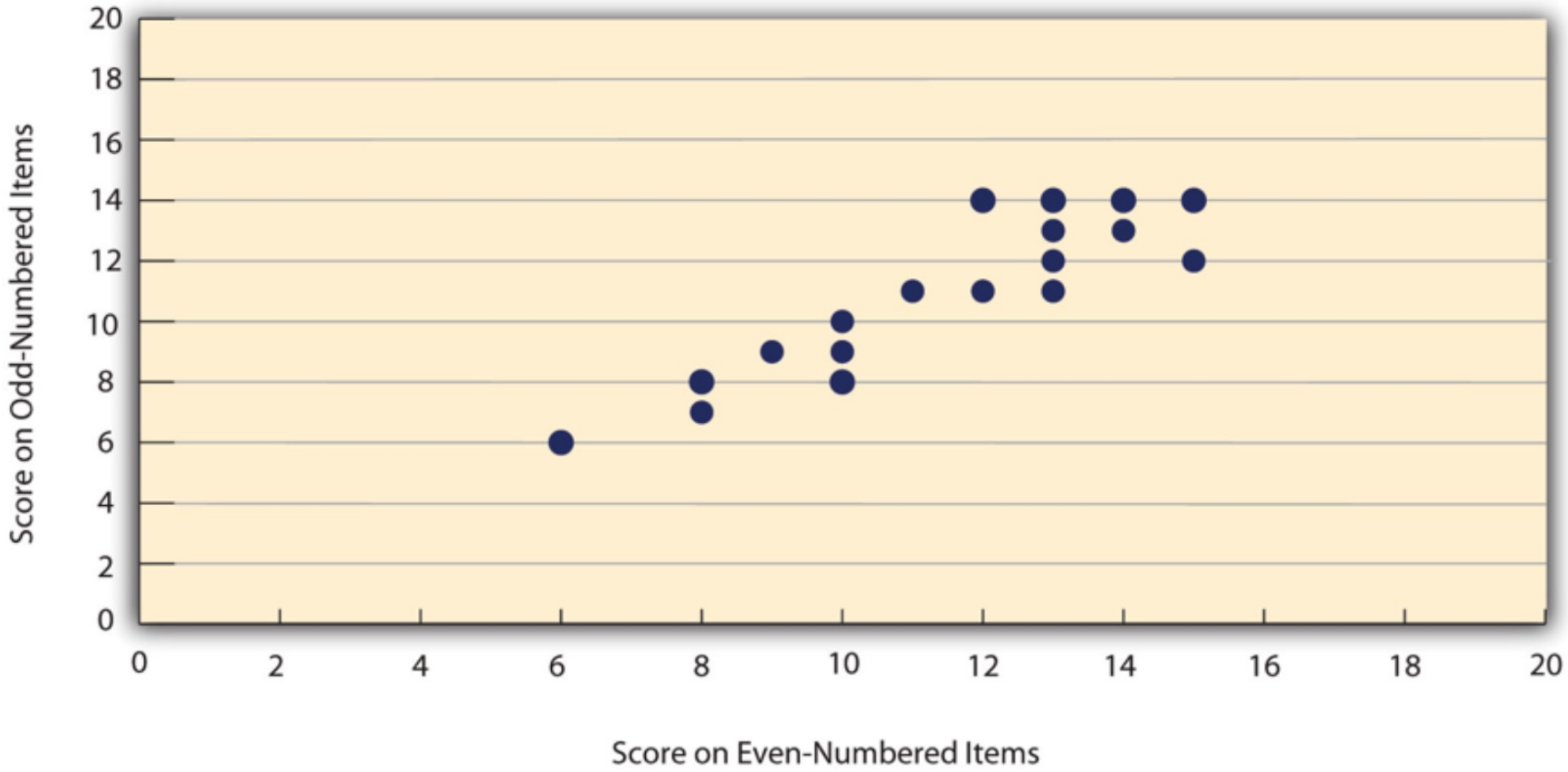


Image source

Inter-coder reliability

- **Definition**: The degree of agreement among raters of the same phenomenon

Measures of inter-coder reliability

- Percent agreement

$$\%\text{agreement} = \frac{\text{the number of cases coded the same way by multiple coders}}{\text{the total number of cases}}$$

- Cohen's κ

$$\kappa = \frac{P_a - P_c}{1 - P_c}$$

- Other measures: Fisher's κ, Kripperndorff's α

What does data quality mean?

- Data quality is affected by **measurement error** (or observational error)
- **Definition**: The difference between the measurement (what is recorded) and the true quantity of the variable, i.e., distortion that cause the observed measurement to be different from the true quantities.

What are different types of error?

- Random error
- Systematic error

$$X = T + e_r + e_s$$

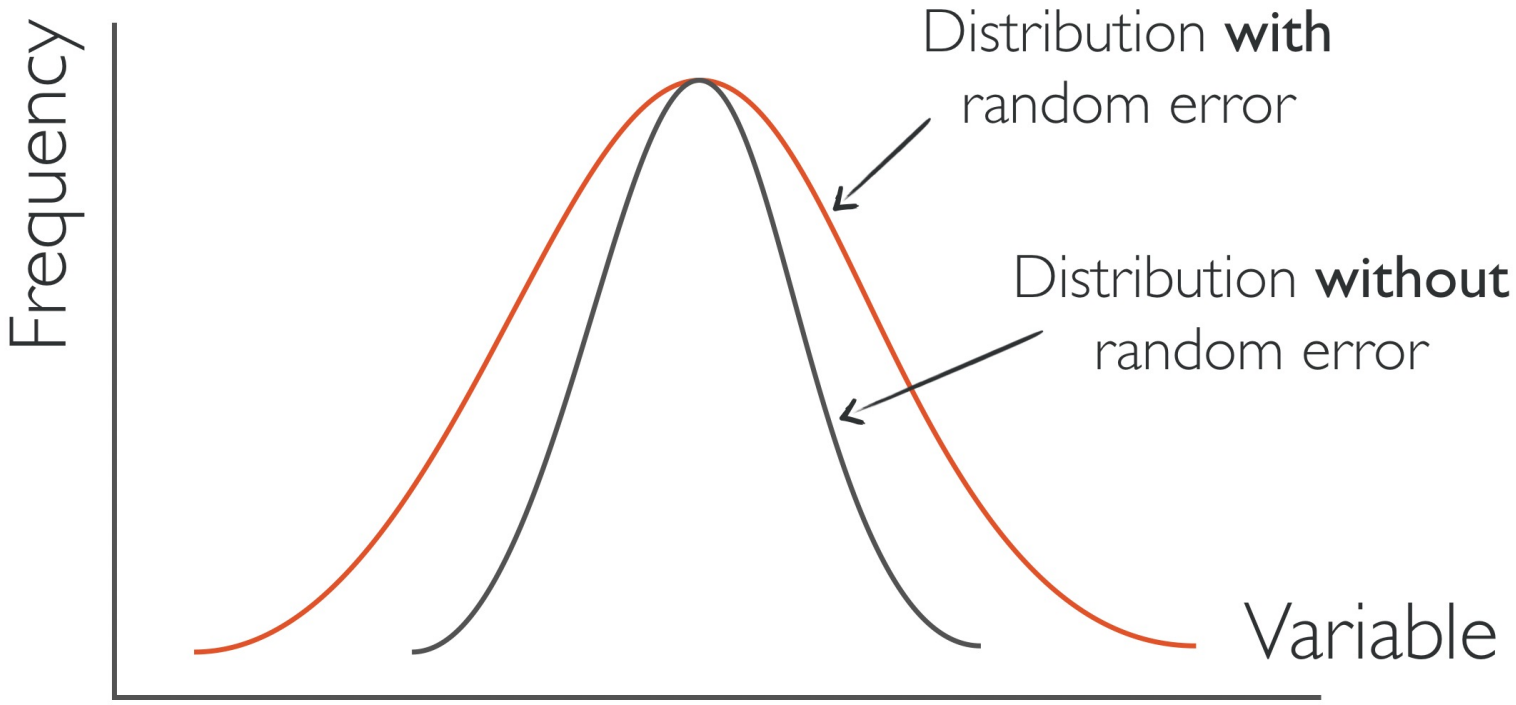
What is random error?

- **Definition**: Random errors are errors in measurement that lead to measurable values being inconsistent when repeated measurements of a constant attribute or quantity are taken.
- Random error is also called ***noise***.

Session 1: 46 words per minute
Session 2: 52 words per minute
Session 3: 47 words per minute
Session 4: 51 words per minute
Session 5: 53 words per minute

What causes random error?

- Inherent in any measure, randomly varies, and affects variance, not the mean.

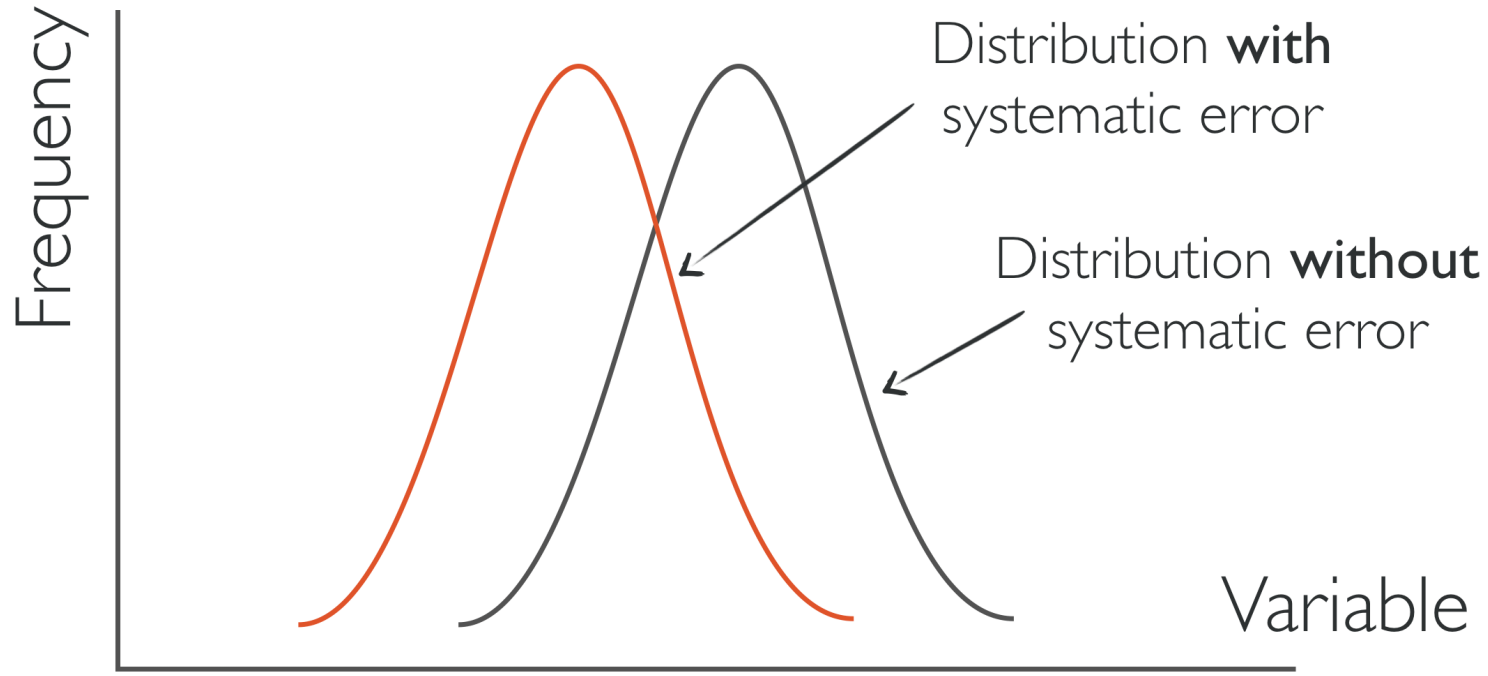


What is systematic error?

- **Definition**: Systematic errors are errors that are not determined by chance but are introduced by an inaccuracy (involving either the observation or measurement process) inherent to the system.
- Systematic error is also called ***bias***.

What causes systematic errors?

- Caused by external factors (e.g., equipment delay, experimental procedures) and affects the mean.



How do we increase measurement quality?

- Piloting experimental instruments
- Testing reliability of coders, retaining
- Using reliable software-driven instruments
- Triangulation
- …

Hands-on Activity