

Human-Computer Interaction

# Measurement

# Basics

Professor Bilge Mutlu

# 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.<sup>1</sup>

*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”

---

<sup>1</sup>[Wikipedia](#)

*How do we represent measurements?*

**Variables** are things that change, e.g., gender, preference, performance

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

*What are different 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

*Ratio*

*Interval*

*Ordinal*

*Nominal*

	<b>Nominal</b>	<b>Ordinal</b>	<b>Interval</b>	<b>Ratio</b>
<b>Distinctiveness</b>	Yes	Yes	Yes	Yes
<b>Rank ordering</b>	No	Yes	Yes	Yes
<b>Equal intervals</b>	No	No	Yes	Yes
<b>Absolute zero</b>	No	No	No	Yes

*Other terms you might hear...*

- >> **Descriptive**, e.g., “a tall man”
- >> **Categorical**, e.g., novice vs. expert, 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



## *What are different kinds of measurements we can take?*

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

## *What makes measurements good?*

1. Validity
2. Reliability
3. 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.<sup>2</sup>

In other words, are we measuring what we want to measure?

---

<sup>2</sup>Wikipedia)

*What is an example 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?*

1. Face validity
2. Construct validity
3. Empirical validity
4. Content validity
5. Ecological validity

## *Face validity*

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

## *Construct validity*

**Definition:** The extent to which conceptual *constructs* relate to what they intend to measure.

### *Two types of construct validity*

*Convergent validity:* The extent to which the measure is associated with things it should be associated with, e.g., *intellect* and *competence* should correlate

*Discriminant validity:* The extent to which the measure is *not* associated with things it should *not* be associated with, *intellect* and *control* should not correlate

*What is a construct?*

**Definition:** A psychological construct is a label for a cluster or domain of covarying behaviours.<sup>3</sup>

*What happens when we have low or high construct validity?*

A high construct validity means that the measurement sufficiently captures and covers the construct.

---

<sup>3</sup>Brittanica



*A measure of **introversion** ( $\alpha = .85$ )<sup>4</sup>*

*Positive keyed*

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

*Negative 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.

---

<sup>4</sup> IPIP

*Empirical validity (criterion validity)*

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

*What are different forms of empirical 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 sample covers a representative sample of the behavior that is measured

E.g., GREs 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 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?

High reliability indicates that the measure produces similar results under consistent conditions.

Reliability is decreased by error.

*How do we calculate reliability?*

We can't—we can only *estimate* it using statistical methods.

$$R = \frac{v_{measurement}}{v_{true} + v_{error}}, R \in [0, 1]$$

*Pro Tip:* A reliability of .70 or higher is acceptable

## *How do we ensure reliability?*

1. **Test-retest reliability:** Involves repeating the same measurement with the same population another time
2. **Alternative form method:** Involves administering a second measure with similar measures with the same population
3. **Split-half technique:** Involves splitting data from the measure into half and correlating the results

Reliability method	Pros	Cons
<i>Test-retest</i>	<ul style="list-style-type: none"> <li>- Uses the same test items</li> <li>- Simple to administer</li> </ul>	<ul style="list-style-type: none"> <li>- First testing may contaminate the second</li> <li>- Respondent may change with time</li> </ul>
<i>Alternative-form</i>	<ul style="list-style-type: none"> <li>- Minimizes repeat-item contamination</li> <li>- Little time passes before retesting</li> <li>- Useful for pre/post-testing</li> </ul>	<ul style="list-style-type: none"> <li>- Use of different items lowers reliability</li> <li>- Requires a longer test</li> </ul>
<i>Split-half</i>	<ul style="list-style-type: none"> <li>- Minimizes repeat-item contamination</li> <li>- No time passes</li> <li>- Done at a single session</li> </ul>	



*What are different forms of reliability?*

1. Internal reliability
2. Inter-coder reliability

## *Internal reliability*

**Definition:** A measure of whether several items that propose to measure the same general construct produce similar scores.<sup>5</sup>

*What are measures of internal reliability?*

1. **Inter-item correlation:** mean of all pairwise correlations across items of a measure
2. **Split-half correlation:** correlations between two randomly split halves of the measure
3. **Cronbach's  $\alpha$ :** inter-item correlations that are iteratively calculated across randomly selected subsets of the measure ( $\alpha > .70$  desirable)

---

<sup>5</sup>[Wikipedia](#)

## *Inter-coder reliability*

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

## *Measures of inter-coder reliability*

- >> Percent agreement
- >> Cohen's  $\kappa$ , Fisher's  $\kappa$ , Krippendorff's  $\alpha$

*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., distortions that cause the observed measurement to be different from the true quantities.

$$X = T + e_r + e_s$$

*What are different types of error?*

1. Random error
2. Systematic error

*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.<sup>6</sup>

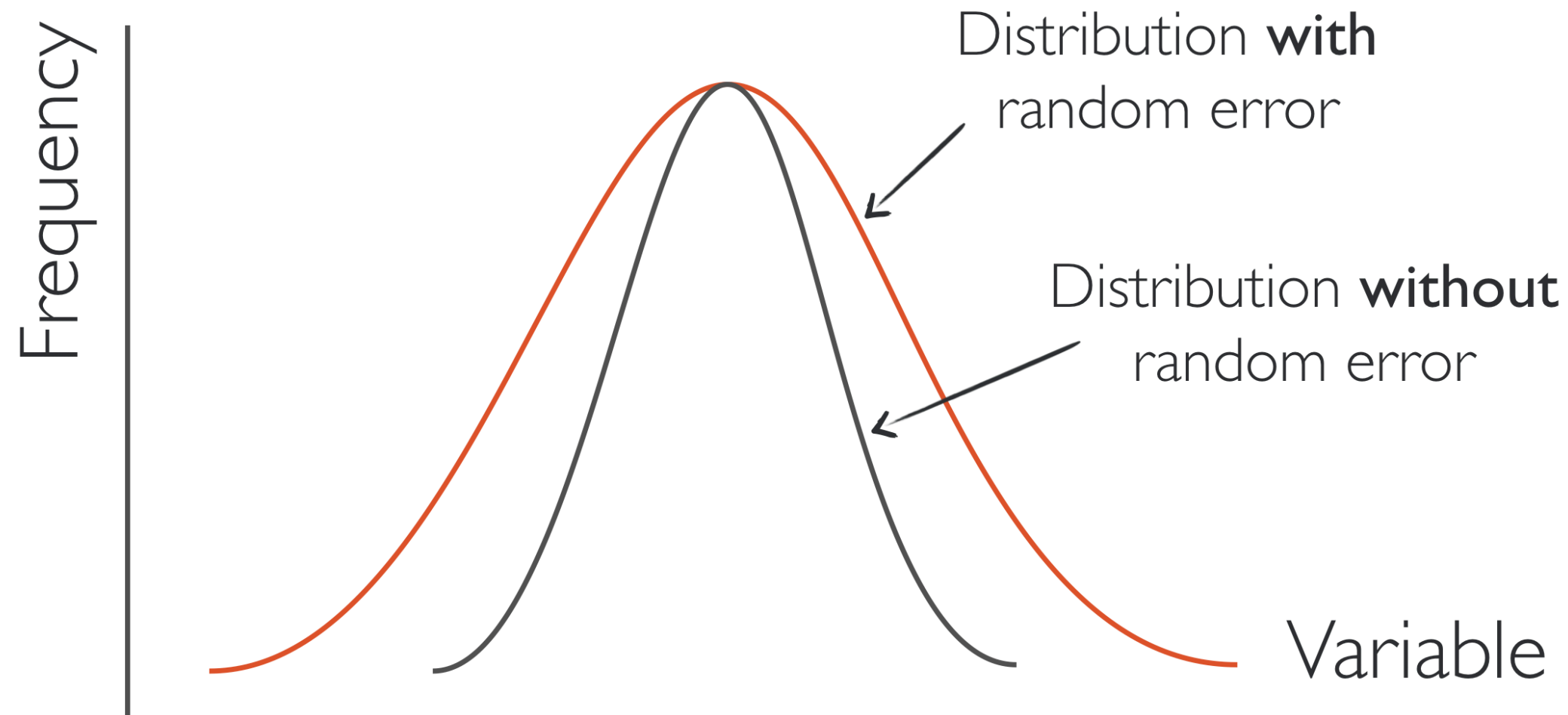
Random error is also called *noise*.

---

<sup>6</sup>[Wikipedia](#)

## What causes random error?

Inherent in any measure that randomly varies (e.g., participant prior experience with robots) 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.<sup>6</sup>

Systematic error is also called *bias*.

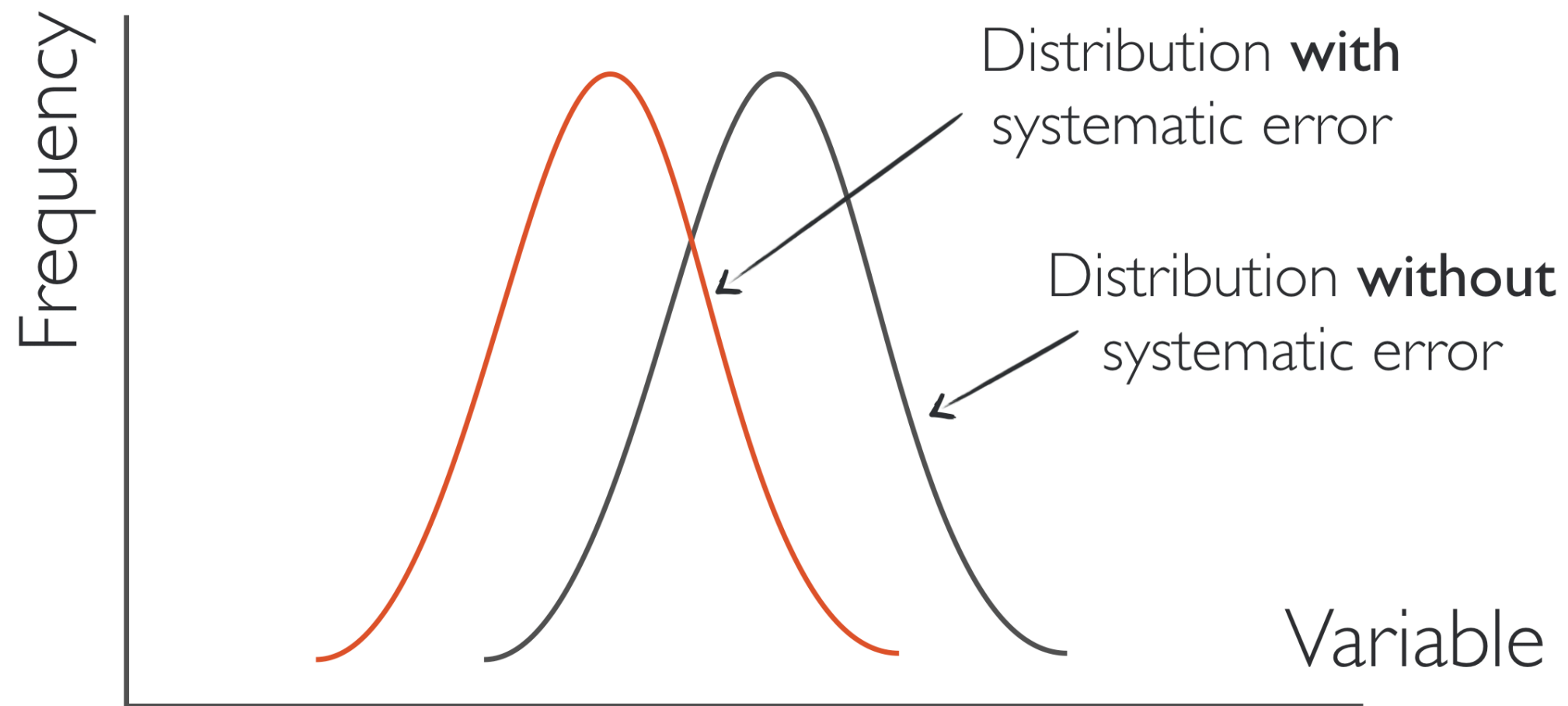
---

<sup>6</sup>[Wikipedia](#)



## What causes systematic errors?

Caused by external factors (e.g., measurement noise, equipment delay, coder error) and affects the mean.



## *How do we increase measurement quality?*

- >> Piloting experimental instruments
- >> Testing reliability of coders, retraining
- >> Repeating manual data entry
- >> Measuring error using statistical methods
- >> Triangulation