



---

# **Week 5 Lecture - Measurement**

Undergraduate Research Methods in Psychology

---

Quinton Quagliano, M.S., C.S.P

Department of Psychology

## Table of Contents

<b>1</b>	<b>Chapter Overview</b>	<b>2</b>
1.1	Introduction to Measurement . . . . .	2
1.2	Visual Representation . . . . .	2
<b>2</b>	<b>How to Measure Something?</b>	<b>4</b>
2.1	Overview . . . . .	4
2.2	Constructs vs. Observed Variables (Again) . . . . .	4
2.3	Three Types of Measure . . . . .	4
2.3.1	Self-Report . . . . .	5
2.3.2	Observational . . . . .	5
2.3.3	Physiological . . . . .	5
<b>3</b>	<b>Reliability (Consistency)</b>	<b>6</b>
3.1	Overview . . . . .	6
3.2	Three Types of Reliability . . . . .	6
3.3	Scatterplot Visualization . . . . .	6
3.3.1	Perfect Positive Relationship . . . . .	7
3.3.2	Imperfect Positive Relationship . . . . .	7
3.4	Correlation Coefficient . . . . .	7
3.4.1	Perfect Positive Relationship . . . . .	8
3.4.2	Imperfect Positive Relationship . . . . .	8
3.4.3	Internal Reliability . . . . .	8
3.5	Reading About Reliability . . . . .	9
<b>4</b>	<b>Measurement Validity (Accuracy)</b>	<b>9</b>
4.1	Overview . . . . .	9
4.2	Measurement Validity . . . . .	9
4.3	Face & Content Validity . . . . .	10
4.4	Criterion Validity . . . . .	10
4.4.1	Correlative Methods (for Continuous Behavior) . . . . .	10
4.4.2	Known Groups Methods (for Categorical Behavior Groups) . . . . .	11
4.5	Convergent & Divergent Validity . . . . .	11
4.5.1	Convergence with Related Measures . . . . .	11
4.5.2	Divergence with Unrelated Measures . . . . .	11
4.6	Relationship Between Reliability and Validity . . . . .	12

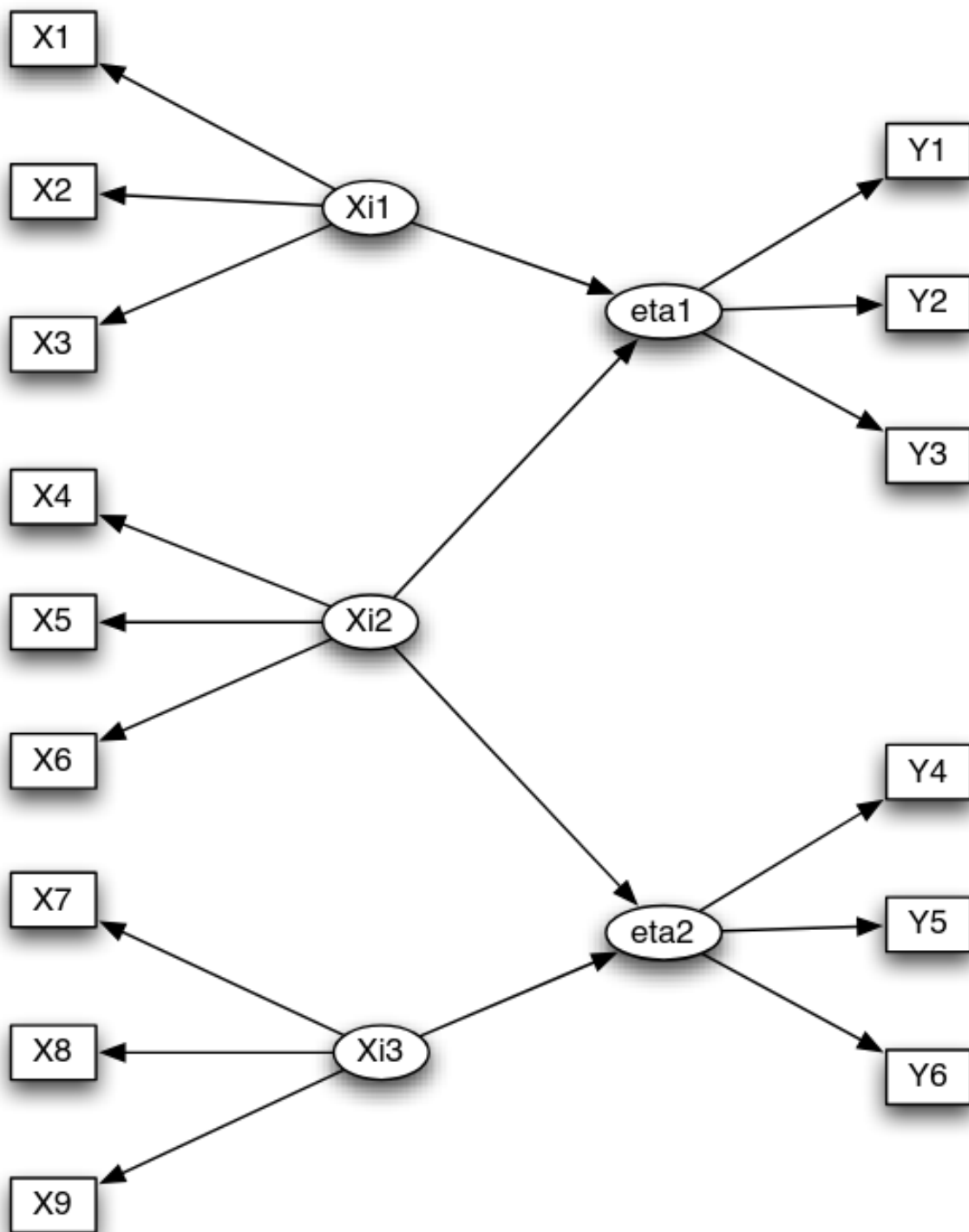
# 1 Chapter Overview

## 1.1 Introduction to Measurement

- Valid and reliable \_\_\_\_\_ is an essential part of any good quantitative research - without it, it is difficult to test for differences, associations, \_\_\_\_\_, or frequencies.
- We must be *systemic, rigorous*, and \_\_\_\_\_ in our measurement, and report (through our writing) thoroughly on the methods we use to capture phenomena and experiences.
- Psychological constructs are, in some ways, more difficult to measure than phenomena explored in other \_\_\_\_\_. For example:
  - A chemist is able to \_\_\_\_\_ parts of solutions with pH, graduated cylinders, etc.
  - Physicists can measure weight, mass, speed, velocity with \_\_\_\_\_ and scales
  - Biologists can measure \_\_\_\_\_ of animals or number of times a certain trait appears in a creature
  - In psychology, we cannot \_\_\_\_\_ measure cognitive traits that are *internal* to people, and even in behaviors things may be complicated...
- Remember that for \_\_\_\_\_ validity: we must make operational variables from latent/conceptual/construct variables, and we must do this well!

## 1.2 Visual Representation

- Rectangles -> Observed Variables (Our Measurements)
  - Ellipses -> Latent Variables (Our Constructs)
  - We want to use strategies that \_\_\_\_\_ the link between the two
-



## 2 How to Measure Something?

### 2.1 Overview

- There are many decisions to be made on how to operationalize, which will have a direct \_\_\_\_\_ on the construct validity of a study.
- There are also different mediums for measurement:
  - \_\_\_\_\_-report
  - \_\_\_\_\_
  - \_\_\_\_\_

### 2.2 Constructs vs. Observed Variables (Again)

- Expanding on what we already know from lecture 3:
  - For any construct under study, we must come up with some conceptual \_\_\_\_\_, that is, some theoretical description of a construct. This usually involves having a reasonable knowledge of \_\_\_\_\_ and theoretical work in a certain topic area
  - Then, we must link that conceptual definition to an operational measure or tool that fully \_\_\_\_\_ that meaning.
  - Note: different measures for the “same” \_\_\_\_\_ may have very different underlying conceptual definitions! Understand the \_\_\_\_\_ that your tool makes before you use it.
- Example: take the concept of “\_\_\_\_\_” - what even is intelligence?
  - Depends on whom we ask: Weschler says \_\_\_\_\_ from Binet says different from ...
  - We also may ask: is cognitive intelligence different from emotional intelligence different from \_\_\_\_\_ intelligence
  - Do we take into account age? \_\_\_\_\_ level? Race? Socioeconomic background?
  - This is why having a clear literature review and background for a tool can help readers understand a \_\_\_\_\_ description which goes into the measure of choice

### 2.3 Three Types of Measure

- All three types have drawbacks and biases which will be discussed more in week 6
-

### 2.3.1 Self-Report

- This is a \_\_\_\_\_ completed by the person it is measuring, often requiring some amount of introspection
- *Example:* Ever been to a doctor's office and have to fill out a bunch of paperwork? We would call that self-reported \_\_\_\_\_
- This can be either through a paper form or through a \_\_\_\_\_ questionnaire
- Related: in some cases we may use what is called a \_\_\_\_\_ report which involves a third-party (e.g., parent, teacher, friend) providing their \_\_\_\_\_ of another person

### 2.3.2 Observational

- This is derived from a third party \_\_\_\_\_ a person's behavior/actions and \_\_\_\_\_ how many times a certain behavior occurs or in what manner the behavior occurs.
- What I do every day in clinic is technically an observational measure: I present a person with some task or stimuli and I *observe* their response or \_\_\_\_\_ on a test

### 2.3.3 Physiological

- This is some sort of measurement of \_\_\_\_\_ characteristics of a person, tends to be much more of a "concrete" measurement than the other two described
  - A lot of physiological measures enjoy some associations with the types of measures above
    - Example: a person who reports a high level of anxiety may also show a specific pattern of \_\_\_\_\_ in the brain during an fMRI
  - Examples:
    - Brain scans: CT, MRI, fMRI, PET, EEG
    - Facial movement: EMG
  - Ideally, we may choose to use all 3 \_\_\_\_\_ of measures or some combination of two of them, to provide multiple operationalizations for the same construct, and they should all be \_\_\_\_\_ (correlated) with one another.
-

## 3 Reliability (Consistency)

### 3.1 Overview

- \_\_\_\_\_ is all about how *consistent* a certain scale or measurement is across different raters, times, and contexts.
- We want a measure to be reliable, otherwise, we have a tool that may very well tell us a \_\_\_\_\_ answer every time we take a measurement!

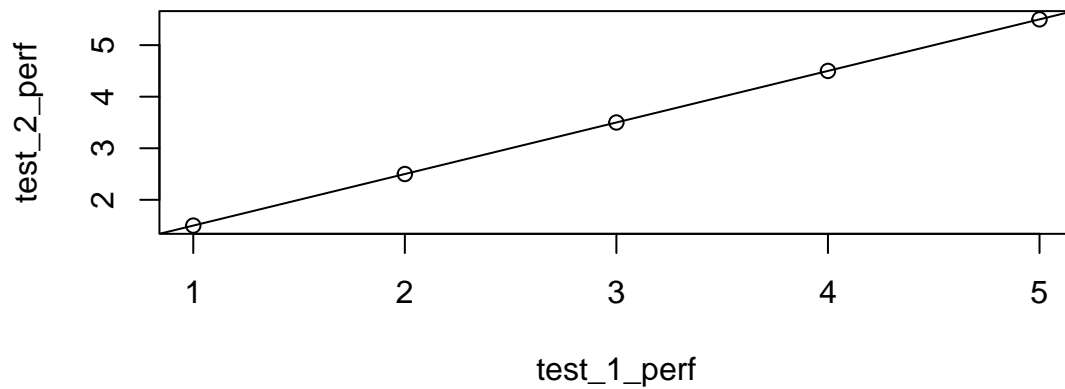
### 3.2 Three Types of Reliability

- There are generally 3 types of reliability:
  - Test-retest: Between different \_\_\_\_\_ points
  - Interrater: Between different observers/raters - how often are they rating something the \_\_\_\_\_?
  - Internal: Between items on the \_\_\_\_\_ measure - how well are related questions regarding the same construct co-varying with one another?

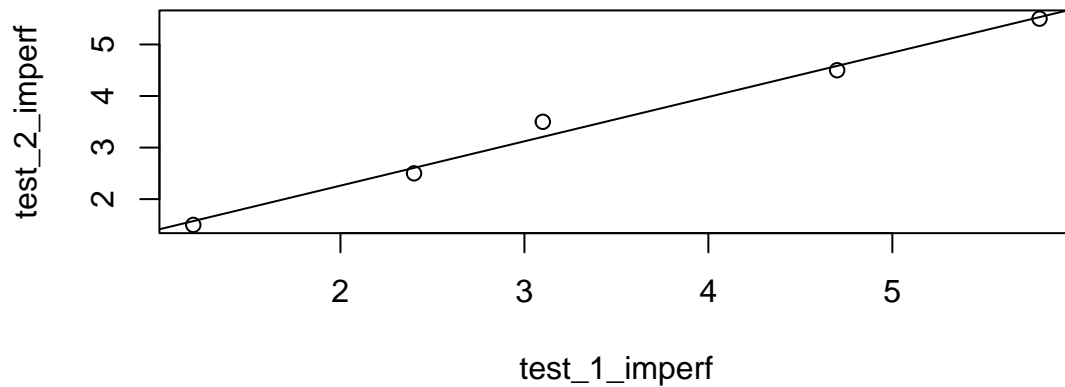
### 3.3 Scatterplot Visualization

- We may logically approach questions of reliability similar to any other claim of \_\_\_\_\_
    - In test-retest, we claim that the measure scores at two different times are \_\_\_\_\_
    - In \_\_\_\_\_, we claim that the measure scores, as recorded by each observer, are *associated* with one another
    - In internal, we claim that two or more \_\_\_\_\_ on the same measure, for the same construct, *covary* with one another
  - Graphically, we may use a \_\_\_\_\_ when we have two sets of continuous data, e.g., two sets of scores of any of the above 3 types
  - The more the points sit close to the line of best fit (which is usually an OLS linear regression line), the stronger the \_\_\_\_\_ between the two measures.
  - In the case of reliability, we would usually like to see a positive relationship between the two sets of data, which would be represented by a line of best fit traveling up and to the \_\_\_\_\_. Put another way, as one set of data increases, the other generally does as well.
-

### 3.3.1 Perfect Positive Relationship



### 3.3.2 Imperfect Positive Relationship



## 3.4 Correlation Coefficient

- To mathematically \_\_\_\_\_ the direction and strength of relationship between two variables we may use  $r$ . The type we will be talking about now is technically called Pearson's product-moment correlation coefficient  $r$ .



- $r$  will always be between -1 and 1, with 0 representing \_\_\_\_\_ correlation or relationship and -1/1 representing a perfectly strong relationship between the two. In practice, you will never get exactly 0 or -1/1, but likely some number in \_\_\_\_\_.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- This correlation coefficient is technically only applicable to continuous data, but is \_\_\_\_\_ to ordinal data by using Spearman's  $\rho$  (rho) or Kendall's  $\tau - B$  (tau-B).

### 3.4.1 Perfect Positive Relationship

[1] 1

### 3.4.2 Imperfect Positive Relationship

[1] 0.9944883

- The \_\_\_\_\_ process is fairly straightforward for inter-rater or test-retest reliability -> higher  $r$  = greater consistency/reliability.
- Another popular statistic for inter-rater reliability is called Cohen's  $\kappa$  (kappa), but that is only \_\_\_\_\_ when raters are grouping/classifying objects or people. It's interpretation is the same as  $r$ .

### 3.4.3 Internal Reliability

- The simplest way to arrange this is to use a correlation \_\_\_\_\_ of all the items of a measure, and calculate the  $r$  between each two items.
- In such a table, we are looking to make sure construct-related items are positive, highly \_\_\_\_\_ and that theoretical unrelated items are negative or weakly correlated.
- You may also calculate an \_\_\_\_\_ inter-item correlation which is just an average correlation across the entire matrix (only recommended if all items *should* be related)
  - Side note: I wouldn't recommend using the acronym AIC for this - most statisticians use that more often for the Akaike information criterion, used often in regression
  - We want \_\_\_\_\_ between 0.15 and 0.50 for this to be "reliable"

- Finally, you can take \_\_\_\_\_'s  $\alpha$  (alpha) which is taken from the average inter-item correlation and number of items on a scale.
  - This is probably the “preferred” statistic for \_\_\_\_\_ reliability
  - Above 0.80 is good for self-reports, we want as close to 1 as possible

### 3.5 Reading About Reliability

- The most important things to know are the \_\_\_\_\_ of each value for reliability and a basic \_\_\_\_\_ of each one. You do not have to know how to calculate these values by hand (for this class), but you should be familiar with the conceptual definition of each.

## 4 Measurement Validity (Accuracy)

### 4.1 Overview

- This is where our \_\_\_\_\_ get confusing, because we already talked about the 4 validities for investigating claims. For the sake of clarity I will use the terms “*claim validity*” and “*measurement validity*” to separate the two terms.
- Essentially, \_\_\_\_\_ validity is the second major component of construct validity, alongside measurement reliability. They both are individual steps in establishing construct validity.
- A “good” measure (i.e., one with good construct validity) will have evidence of \_\_\_\_\_ the following measurement validities, usually across different studies

### 4.2 Measurement Validity

- Validity is all concerned with how well we represent the construct with an \_\_\_\_\_ tool. It is multifaceted and often quite complicated - usually the measurement validity of any given tool has to be well-established across \_\_\_\_\_ studies.
  - But remember, just like any claim or evidence, we never *prove* something as completely and flawlessly valid - rather the \_\_\_\_\_ of evidence is for or against its validity.
-

### 4.3 Face & Content Validity

- Both of these are more \_\_\_\_\_ validities which relate to whether it *seems* like a certain measurement captures the concept well. However, they are a little \_\_\_\_\_ and tend to be more valued when a measurement scale is first proposed, vs. less so when a scale is more established.
- **Face validity** is largely just an assessment of “well, does it seem like this would work?” - may be \_\_\_\_\_ by the general public or by experts in a domain
- **Content validity** asks whether it appears a measurement would capture *all* components of a theoretical construct. This is usually \_\_\_\_\_ assessed by domain experts which have a strong knowledge of the theory underlying a certain construct.
- For a more empirical, albeit subjective approach, some studies will gather a panel of experts, calculating  $r$  or  $\kappa$  for a measurement or the individual items of a measure across the experts. This could be taken as evidence of somewhat decent expert \_\_\_\_\_ on a tool, but it strays close to an appeal to authority (i.e., the expertise of the judges). For more than 2 judges we would use extensions of those statistics that are applicable to multivariate data (e.g., multivariate regression, G-study, D-study, etc.)

### 4.4 Criterion Validity

- **Criterion validity** focuses on whether a measurement is positively \_\_\_\_\_ with behaviors that are also said to be representative of the construct. Now those selected \_\_\_\_\_ are also a subjective choice - but this measurement validity can help establish that a measure is related to what behaviors we normally associate with a trait.

#### 4.4.1 Correlative Methods (for Continuous Behavior)

- Once again, we are able to use \_\_\_\_\_ methods and scatterplots, like those previous discussed.
  - In the scatterplot, we would place the \_\_\_\_\_ or magnitude of behavior on one axis and the measure on the other axis.
  - A strong, positive relationship between the two would be \_\_\_\_\_ of good criterion validity
- For example, consider we are developing a collateral-report measure for temper in children where a parent reports how often a child engages in disruptive behavior. Now, we sit in a classroom with the child and measure the number of \_\_\_\_\_

they engage in disruptive behavior. A higher score on the measure should be associated with a higher number of occurrences of disruptive behavior.

#### 4.4.2 Known Groups Methods (for Categorical Behavior Groups)

- We can also assess whether a measure is able to discern differences between some *known-groups* by some \_\_\_\_\_ standard.
  - This means that, prior to testing the new measure, we must have some “source of truth” for whether a person belongs to a certain group.
- Example: We have two groups of people, those diagnosed with schizophrenia and those not diagnosed with schizophrenia. We have a continuous measure designed to detect psychotic disorders. Are those individuals with schizophrenia and those without scoring different on this measure?
- In this method, we could use between-groups \_\_\_\_\_, such as t-tests and ANOVA to decide whether scores on the measure are significantly different between the known-group members.
  - In modern research, you are more likely to see techniques like logistic regression, receiver operating characteristic (ROC), and area-under-the-curve (AUC) analysis - as these methods are able to detect appropriate “cut-off” scores for the measure that discriminate between the two groups with ideal sensitivity (detection of true positives) and specificity (detection of true negatives)

### 4.5 Convergent & Divergent Validity

- We can also determine how a new measure relates to existing measures for the same/different \_\_\_\_\_. Ideally, we want a new measure to strongly correlate with measures for the same construct (convergence) and have little to no relationship with measures for other constructs (divergence).
- Just like the other measurement validities, correlation is our analysis of choice for these

#### 4.5.1 Convergence with Related Measures

- For a valid measure, we want *high convergence (i.e., strong positive correlation)* with tools that are meant to measure the *same* construct.

#### 4.5.2 Divergence with Unrelated Measures

- For a valid measure, we want *high divergence (i.e., no/negative correlation)* with
-

tools that are meant to measure a *different* construct.

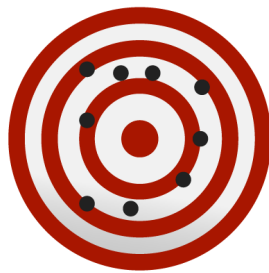
## 4.6 Relationship Between Reliability and Validity

- Measurement validity and reliability are **not** \_\_\_\_\_ terms, though they are related.
  - Reliability is a core necessity for a tool to be considered valid, but just because something is reliable, does not make it valid.
  - “A valid tool is reliable”
  - “A reliable tool is not necessarily valid”

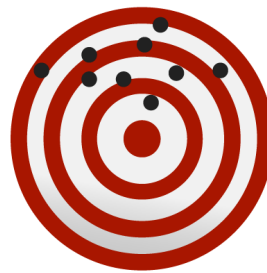
### Reliability and Validity



Reliable  
Not valid



Low validity  
Low reliability



Not reliable  
Not valid



Both reliable  
and valid