Chapter 5 - Measurement

PSY-300: Research Methods in Psychology

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

Department of Psychology

Table of Contents

1 Last Week Review

1.1 Announcements and Due Dates

- Keep turning in reading evidence if you are reading the textbook and taking good notes (which you should be doing either way!)
- I will have office hours again at 2:00pm - 5:00pm EST in AuSable 1307 on Friday 09/27/2024. Come stop by!
- GVSU Annual Undergraduate Research Fair on Oct 1st, 2024 at 2250 Kirkhof Center (Grand River) from 5:00pm - 7:00pm EST!
  - This does fall during class time, so you'd only be able to attend the first bit
  - Please see my class announcement on Blackboard for more information

1.2 Last Week Content

- Historical examples of ethical violations in medical and psychological studies and the later ramifications of these studies (i.e., the Belmont Report and the APA Code of Ethics)
- Descriptions of the principles found in the Belmont Report and APA Code of Ethics and the practical implications of each of these principles for research design
- Case study of a more modern example of ethical violations in research, an exploration of the professional consequences of poor practice

2 Quiz 3 Review

2.1 Areas for Review

- In Mak et al. (2023):
  - When it comes to good, transparent representation of results, we look mostly to two things: the completeness of the statistical tests reporting, and any graphical representations of effects. In the case of this article, I specifically commented that the graphs gave an excellent visual aid to understand the tightly clustered nature of the different conditions.
  - Whenever you are looking for something related to internal validity, we are often going to be focused on words like "confounds", "mediators", "moderator", "cause" - because that validity is focused on the like between the manipulated (experimental) and the measured (outcome) variables

  - A randomized controlled trial (RCT) is a clinical research design term that is, effectively, a type of experiment - which IS a valid design for investigating causal claims. The "randomized" refers to random assignment (a critical part of experiments) and the "controlled" alludes to the tight control for confounds that experiments should have.

- *Re: this statement: "Among patients who have suffered a traumatic brain injury (TBI), most patients see a logarithmic return of cognitive skills post-injury"*
  - This is a **frequency claim**, which is indicated by the word "most". I am saying that a greater proportion of people have this trend of recovery.

- *Re: this statement: "Time (in seconds) to complete a 100 m dash"*
  - Will be a continuous/interval/ratio type data, as it has a clear, consistent, known distance between points - we know how long is in between two seconds. That alone makes it continuous in nature. It is not ordinal, because that would imply we *don't* know the distance between two values in the scale.

3 Quiz 4

3.1 Quiz Content

- Covers all content from 09/17 class meeting, including but not limited to:
  - Chapter 4 of Morling Textbook
  - Lecture on Chapter 4
  - Ethics case study (from the New Yorker, regarding Prof. Gino)
- *Any last minute questions?*

3.2 Quiz Rules

- *From the Syllabus:*
  - Each quiz is 10 multiple-choice questions, 1 point for each question
  - Quizzes will be taken at the start of the class period on the Blackboard LMS
  - Quizzes will be on content covered in the previous lecture and the associated reading for that lecture
  - Quizzes are timed, 23 minutes only (previously was 15 minutes). If you finish before time is up, please remain in class and find another activity to work on quietly.
  - Quizzes are open-note and open-book, that is, you are allowed to use those resources during the quizzes. Thus, they reward good structure in thoughtfulness in your notes and preparation

  - You may not collaborate with others during the quizzes, or discuss questions with other students after the quiz. You cannot use AI tools or the internet to help you during the quiz.
  - Quizzes and exam will be ended early if all students are clearly finished and content with their answers
  - Quizzes will be graded promptly and reviewed the following week

4 Learning Objectives

4.1 Textbook Objectives

- Interrogate the construct validity of a study's measured variables.
- Describe the kinds of evidence that support the construct validity of a measured variable.

4.2 Professor's Objectives

- Describe the problems and difficulty in psychological measurement compared to other scientific fields
- Be able to describe and identify the different forms or mediums of measurement: self-report, observed, physiological
- Differentiate between reliability and measurement validity of a scale, and also be able to describe the similarities in those regards.

5 Chapter Overview

5.1 Introduction to Measurement

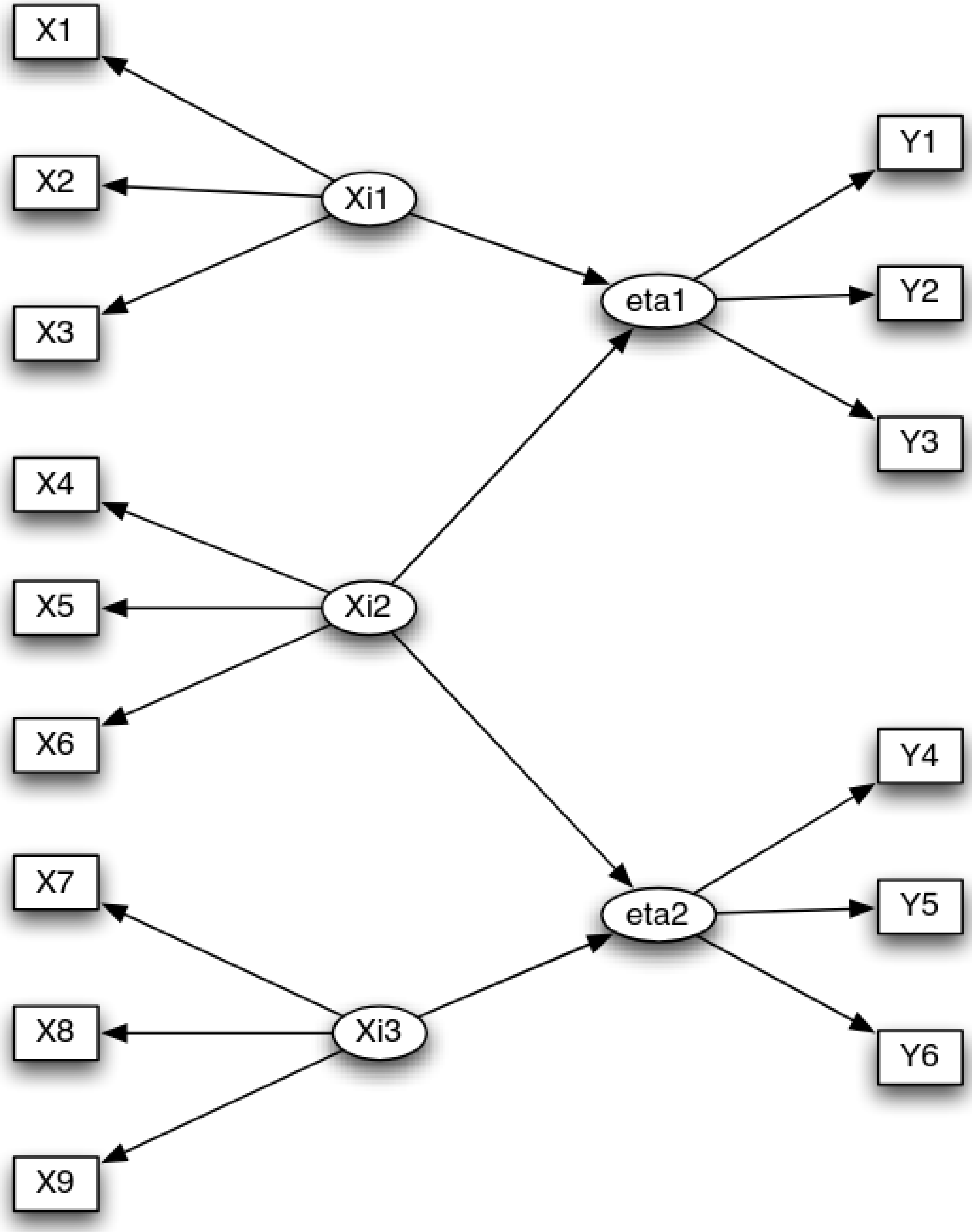
- Valid and reliable measurement is an essential part of any good quantitative research - without it, it is difficult to test for differences, associations, or frequencies.
- We must be *systemic*, *rigorous*, and *consistent* in our measurement, and report (through our writing) thoroughly on the methods we use to capture phenomena and experiences.
- Psychological constructs are, in some ways, more difficult to measure than phenomena explored in other sciences. For example:
  - A chemist is able to measure parts of solutions with pH, graduated cylinders, etc.
  - Physicists can measure weight, mass, speed, velocity with time and scales

  - Biologists can measure numbers of animals or number of times a certain trait appears in a creature
  - In psychology, we cannot directly measure cognitive traits that are *internal* to people, and even in behaviors things may be complicated...

- Remember that construct validity: we must make operational variables from latent/conceptual/construct variables, and we must do this well!

5.2 Visual Representation

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


X1
X2
X3
X4
X5
X6
X7
X8
X9
Xi1
Xi2
Xi3
eta1
eta2
Y1
Y2
Y3
Y4
Y5
Y6

6 How to Measure Something?

6.1 Overview

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

6.2 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 definition**, that is, some theoretical description of a construct. This usually involves having a reasonable knowledge of empirical and theoretical work in a certain topic area
  - Then, we must link that conceptual definition to an operation measure or tool that fully captures that meaning.
  - Ex: Pg 288. describes operationalizing "happiness" two different ways
  - Note: different measures for the "same" construct may have very different underlying conceptual definitions! Understand the assumptions that your tool makes before you use it.
- Lecture example: take the concept of "intelligence" - what even is intelligence?
  - Depends on whom we ask: Weschler says different from Binet says different from ...
  - We also may ask: is cognitive intelligence different from emotional intelligence different from social intelligence
  - Do we take into account age? Education level? Race? Socioeconomic background?
  - This is why having a clear literature review and background for a tool can help readers understand a theoretical description which goes into the measure of choice
- Good examples of types of measures for different constructs in table 5.1

6.3 Three Types of Measure

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

Self-Report

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

Observational

- This is derived from a third party observing a person's behavior/actions and recording how many times a certain 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 success on a test

Physiological

- This is some sort of measurement of physical characteristics of a person, tends to be much more of a "concrete" measurement that than other two described
- A lot of physiological measures enjoy some associations with the types of measures above
  - For example: a person who reports a high level of anxiety may also show a specific pattern of activation 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 types of measures or some combination of two of them, to provide multiple operationalizations for the same construct, and they should all be associated (correlated) with one another.

6.4 Scales of Measurement (Again)

- This topic was covered in the previous chapter 3 lecture, as I felt it more relevant to that discussion. Please make sure you review and understand this topic, as it will be very helpful to your choice of methodology and to later courses, like PSY-350.

7 Reliability (Consistency)

7.1 Overview

- **Reliability** 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 different answer every time we take a measurement!

7.2 Three Types of Reliability

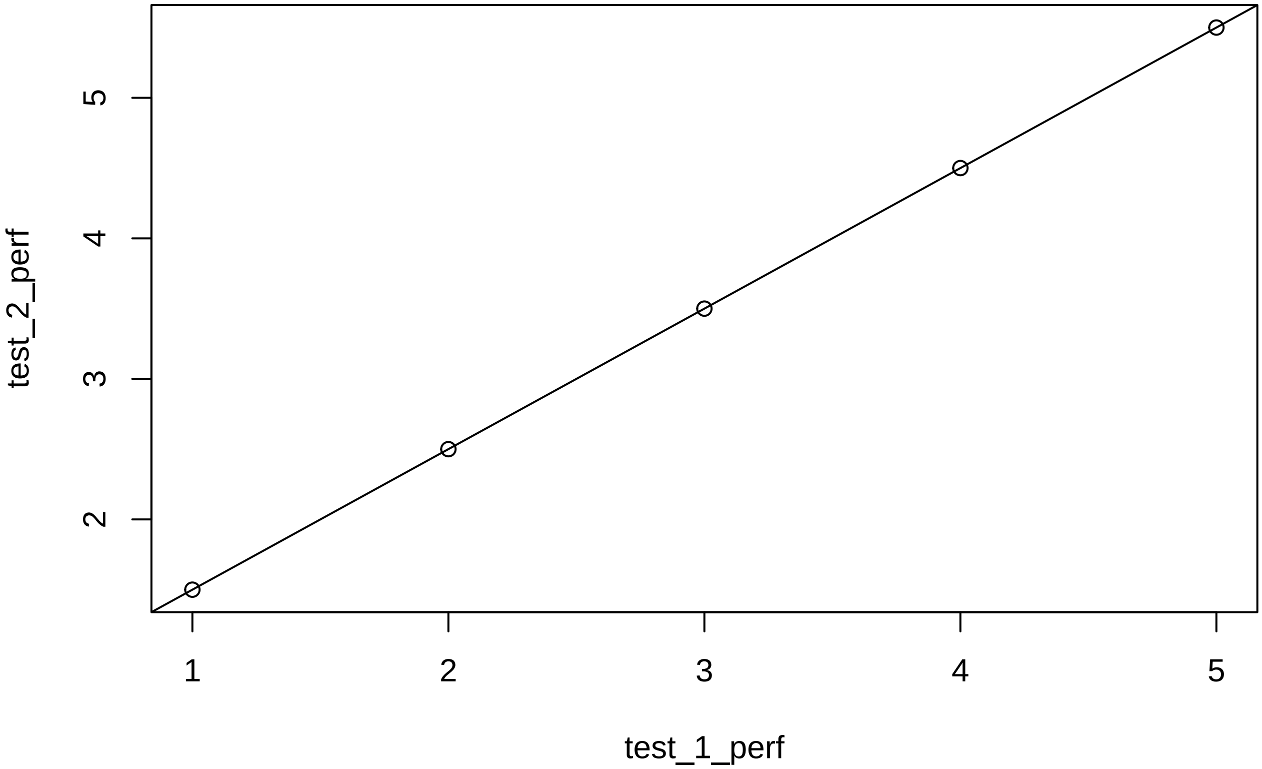
- There are generally 3 types of reliability:
  - Test-retest: Between different time points
  - Interrater: Between different observers/raters - how often are they rating something the same?
  - Internal: Between items on the same measure - how well are questions regarding the same construct co-varying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

7.3 Scatterplot Visualization

- We may logically approach questions of reliability similar to any other claim of association
  - In test-retest, we claim that the measure scores at two different times are *linked*
  - In interrater, we claim that the measure scores, as recorded by each observer, are *associated* with one another
  - In internal, we claim that two or more items on the same measure, for the same construct, *covary* with one another
- Graphically, we may use a scatterplot 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 relationship 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 right. Put another way, as one set of data increases, the other generally does as well.

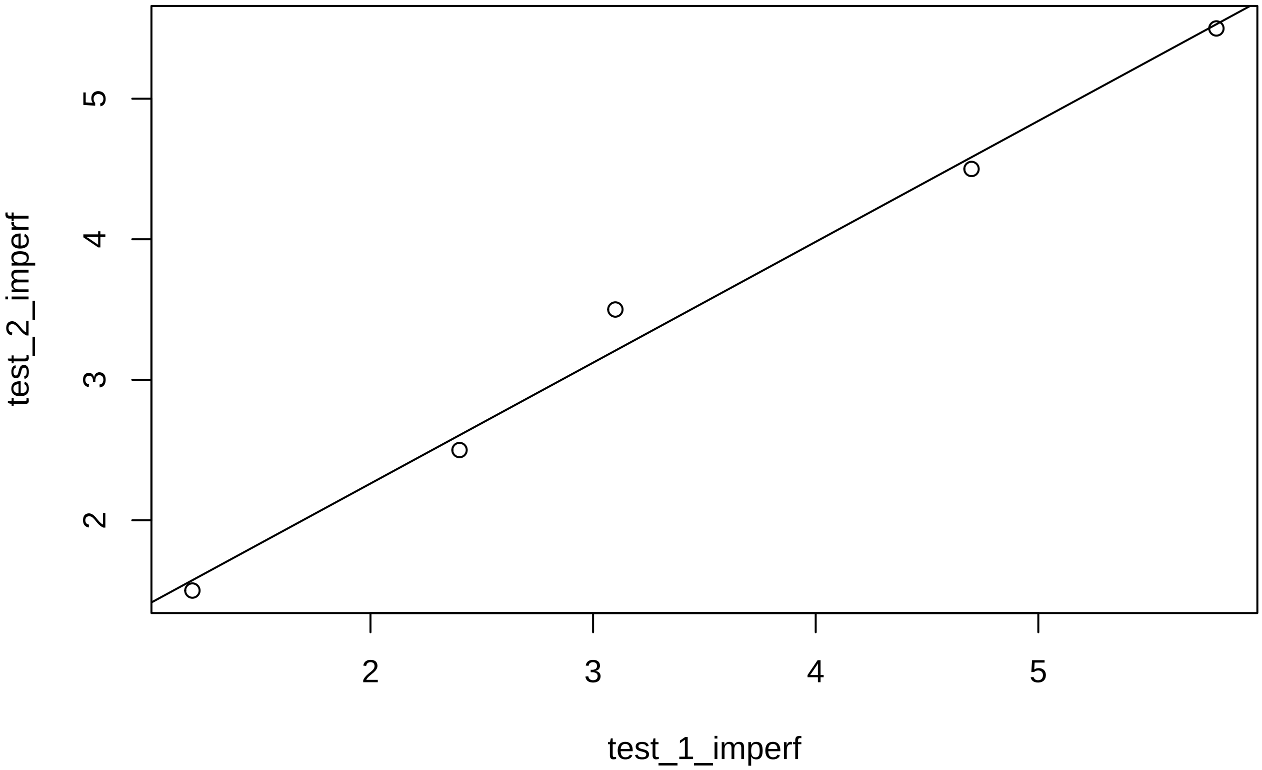
Perfect Positive Relationship

```r
test_1_perf <- c(1, 2, 3, 4, 5)
test_2_perf <- c(1.5, 2.5, 3.5, 4.5, 5.5)

plot(test_1_perf, test_2_perf)
abline(lm(test_2_perf ~ test_1_perf))
```



Imperfect Positive Relationship

```r
test_1_imperf <- c(1.2, 2.4, 3.1, 4.7, 5.8)
test_2_imperf <- c(1.5, 2.5, 3.5, 4.5, 5.5)

plot(test_1_imperf, test_2_imperf)
abline(lm(test_2_imperf ~ test_1_imperf))
```



7.4 Correlation Coefficient

- To mathematically calculate 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 no 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 between.

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

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

Perfect Positive Relationship

```
cor(test_1_perf, test_2_perf)
```

```
[1] 1
```

Imperfect Positive Relationship

```
cor(test_1_imperf, test_2_imperf)
```

```
[1] 0.9944883
```

- The interpretation 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 κ (kappa), but that is only applicable when raters are grouping objects or people. It's interpretation is the same as r.

Internal Reliability

- The simplest way to arrange this is to use a correlation matrix of all the items of a measure, and calculate the r between each two items. See table 5.3 on in your book (pg 311) as an example.
- In such a table, we are looking to make sure construct-related items are positive, highly correlated and that theoretical unrelated items are negative or weakly correlated.
- You may also calculate an average 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 roughly between 0.15 and 0.50 for this to be "reliable"
- Finally, you can take Cronbach's α (alpha) which is taken from the average inter-item correlation and number of items on a scale.
  - This is probably the "preferred" statistic for internal reliability
  - Above 0.80 is good for self-reports, we want as close to 1 as possible

7.5 Reading About Reliability

- The book provides a brief example of reading for reliability values. The most important things to know are the names of each value for reliability and a basic interpretation 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.

8 Measurement Validity (Accuracy)

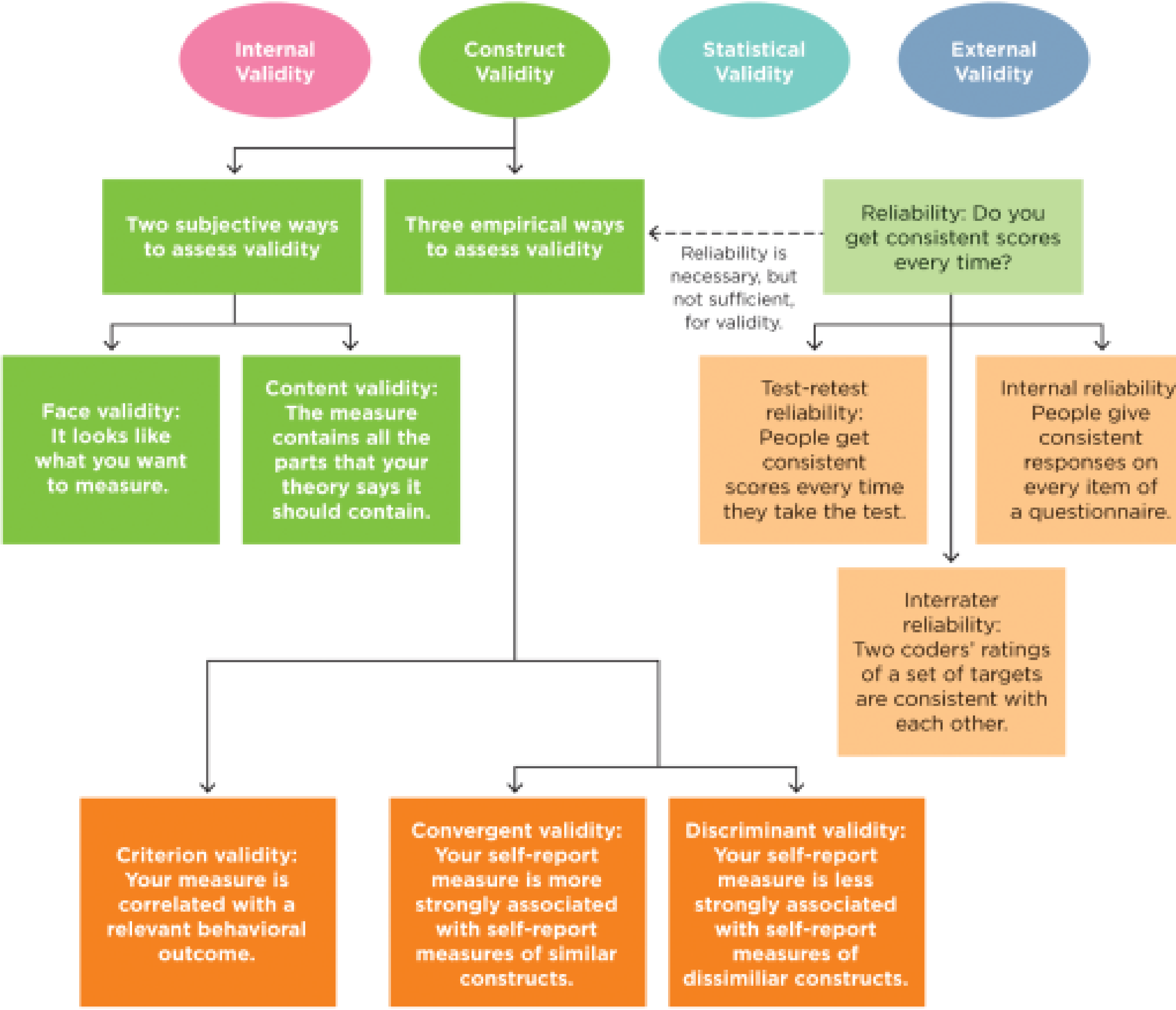
8.1 Overview

- This is where our terms 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, measurement 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 all the following measurement validities, usually across different studies

8.2 Measurement Validity

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

Figure 5.8 in the book:


Internal Validity
Construct Validity
Statistical Validity
External Validity
Two subjective ways to assess validity
Three empirical ways to assess validity
Reliability: Do you get consistent scores every time?
Reliability is necessary, but not sufficient, for validity.
Face validity: It looks like what you want to measure.
Content validity: The measure contains all the parts that your theory says it should contain.
Test-retest reliability: People get consistent scores every time they take the test.
Internal reliability People give consistent responses on every item of a questionnaire.
Interrater reliability: Two coders' ratings of a set of targets are consistent with each other.
Criterion validity: Your measure is correlated with a relevant behavioral outcome.
Convergent validity: Your self-report measure is more strongly associated with self-report measures of similar constructs.
Discriminant validity: Your self-report measure is less strongly associated with self-report measures of dissimiliar constructs.

8.3 Face & Content Validity

- Both of these are more subjective validities which relate to whether it *seems* like a certain measurement captures the concept well. However, they are a little superficial 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 evaluated 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 best 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 κ for a measurement or the individual items of a measure across the experts. This could be taken as evidence of somewhat decent expert consensus 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.)

8.4 Criterion Validity

- **Criterion validity** focuses on whether a measurement is positively associated with behaviors that are also said to be representative of the construct. Now those selected behaviors 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.

Correlative Methods (for Continuous Behavior)

- Once again, we are able to use correlative methods and scatterplots, like those previous discussed.
  - In the scatterplot, we would place the number or magnitude of behavior on one axis and the measure on the other axis.
  - A strong, positive relationship between the two would be indicative 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 times they engage in disruptive behavior. A higher score on the measure should be associated with a higher number of occurrences of disruptive behavior.

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 established 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 statistics, 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)

8.5 Convergent & Divergent Validity

- We can also determine how a new measure relates to existing measures for the same/different constructs. 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

Convergence with Related Measures

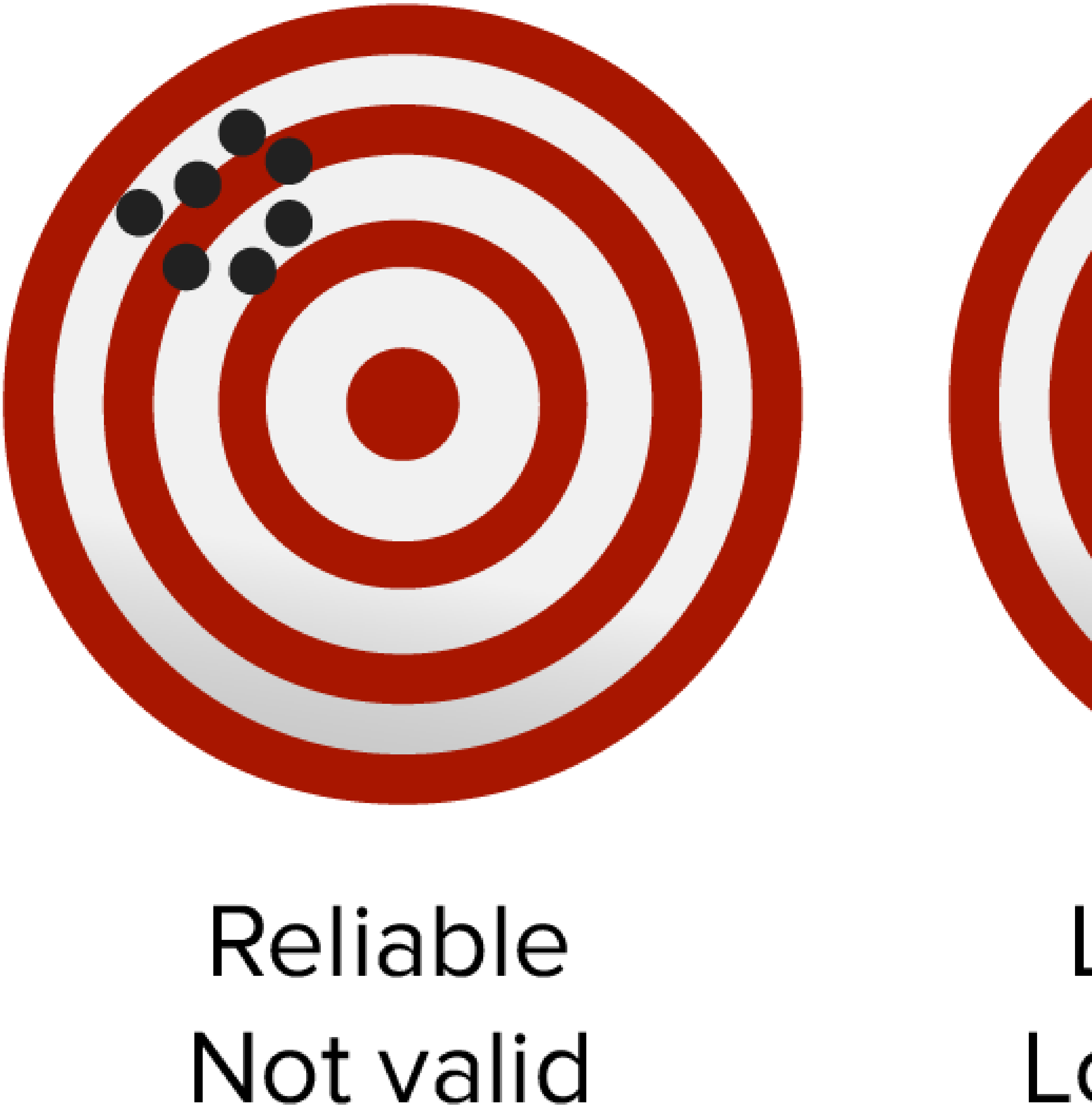
- For a valid measure, we want *high convergence (i.e., strong positive correlation)* with tools that are meant to measure the *same* construct.

Divergence with Unrelated Measures

- For a valid measure, we want *high divergence (i.e., no correlation)* with tools that are meant to measure a *different* construct.

8.6 Relationship Between Reliability and Validity

- Measurement validity and reliability are **not** interchangeable 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”


Reliable
Not valid
L
Lo