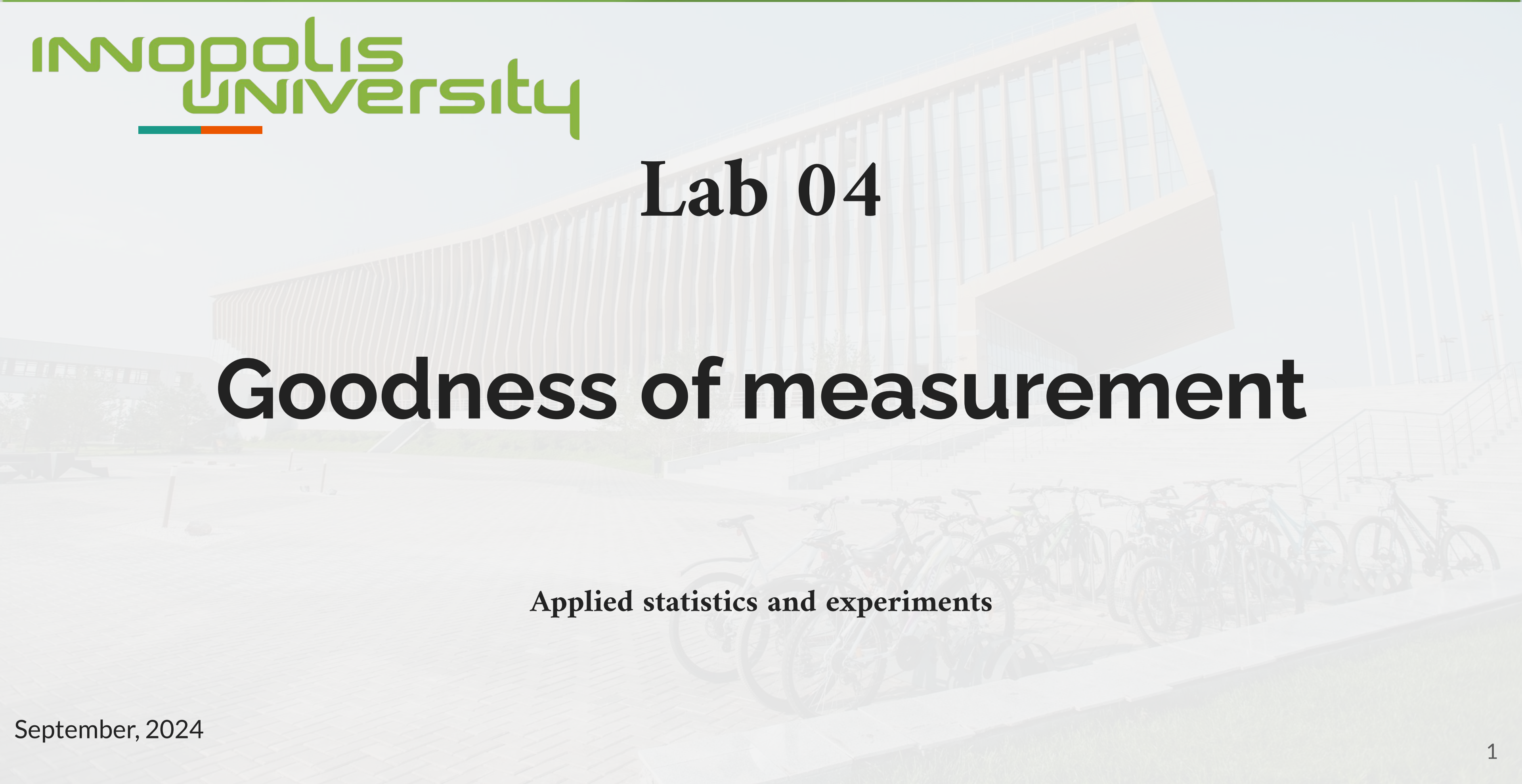


Lab 04

Goodness of measurement

Applied statistics and experiments

September, 2024

Agenda

1. Characteristics of a good measurement
2. Sensitivity
3. Validity
4. Reliability
5. Measurement errors

Lecture Recap

https://quizizz.com/join

Join and enter game code

Research construct

- An abstraction that researchers use to represent a phenomenon that’s not directly measurable
  - **Job satisfaction:** a social construct reflecting the degree to which employees feel content with their work environment and overall experience in their workplace.
  - **Quality of life**: a complex multi-dimensional construct encompassing various aspects of an individual’s well-being such as physical health, emotional stability, social relationships, and economic status.

Metric
Measure
Measurement
Measurement scale
Representational theory of measurement
Representation condition

Research construct

- An abstraction that researchers use to represent a phenomenon that’s not directly measurable
- Not directly measurable
  - inferred from other measurable variables, which are gathered through **observation**.
- For example, the construct of intelligence can be inferred based on a combination of measurable indicators such as problem-solving skills and language proficiency.

Metric
Measure
Measurement
Measurement scale
Representational theory of measurement
Representation condition

Measurement

- Measuring a construct involves assigning scores to represent an attribute.

Quality of your measurement?

Metric
Measure
Measurement
Measurement scale
Representational theory of measurement
Representation condition

Measurement

- Measuring a construct involves assigning scores to represent an attribute.
- This process creates the data that we analyze. However, to provide meaningful research results, that data must be good. And not all data are good!

Measurement

- Researches can **not** just **assume** they have **good** measurements. Typically, researchers need to collect data using an instrument and evaluate the quality of the measurements.
- In other words, they **conduct** a measurement **assessment** before the primary research with respect to its **reliability** and **validity**.

Goodness of measurement

For data to be good enough to allow you to draw meaningful conclusions from a research study, they must be **reliable and valid**.

What are the properties of good measurements?

- In a nutshell, reliability relates to the consistency of measures, and validity addresses whether the measurements are quantifying the correct attribute.
- Reliability and validity are criteria by which researchers assess measurement quality.

Goodness of measurement

- The extent to which a measurement accurately reflects the concept it is intended to measure.
- An important aspect of research design, as inaccurate or unreliable measurements can lead to flawed conclusions and invalid results.
- Need to assess the “goodness” of measuring instrument.
- A good measurement is valid, reliable, and free from bias, ensuring that the data collected is accurate and meaningful for analysis and interpretation.
- Characteristics of a good measurement:
  - Validity
  - Reliability
- Sensitivity – accuracy of the measuring instrument.

Sensitivity of the measuring tool

- Instrument’s ability to accurately measure variability in responses.
- For example:
  - A dichotomous response category, such as “agree or disagree” does not allow the recording of subtle attitude changes.
  - A sensitive measure, with numerous items on the scale, may be needed. For example: Increase items. Increase response categories. (Strongly agree, agree, neutral, disagree, strongly disagree). It will increases a scale’s sensitivity.
- Another example:
  - Thermometers which are used to measure temperature. A good thermometer should be sensitive enough to detect small changes in temperature, as even slight variations can affect the accuracy of the measurement.

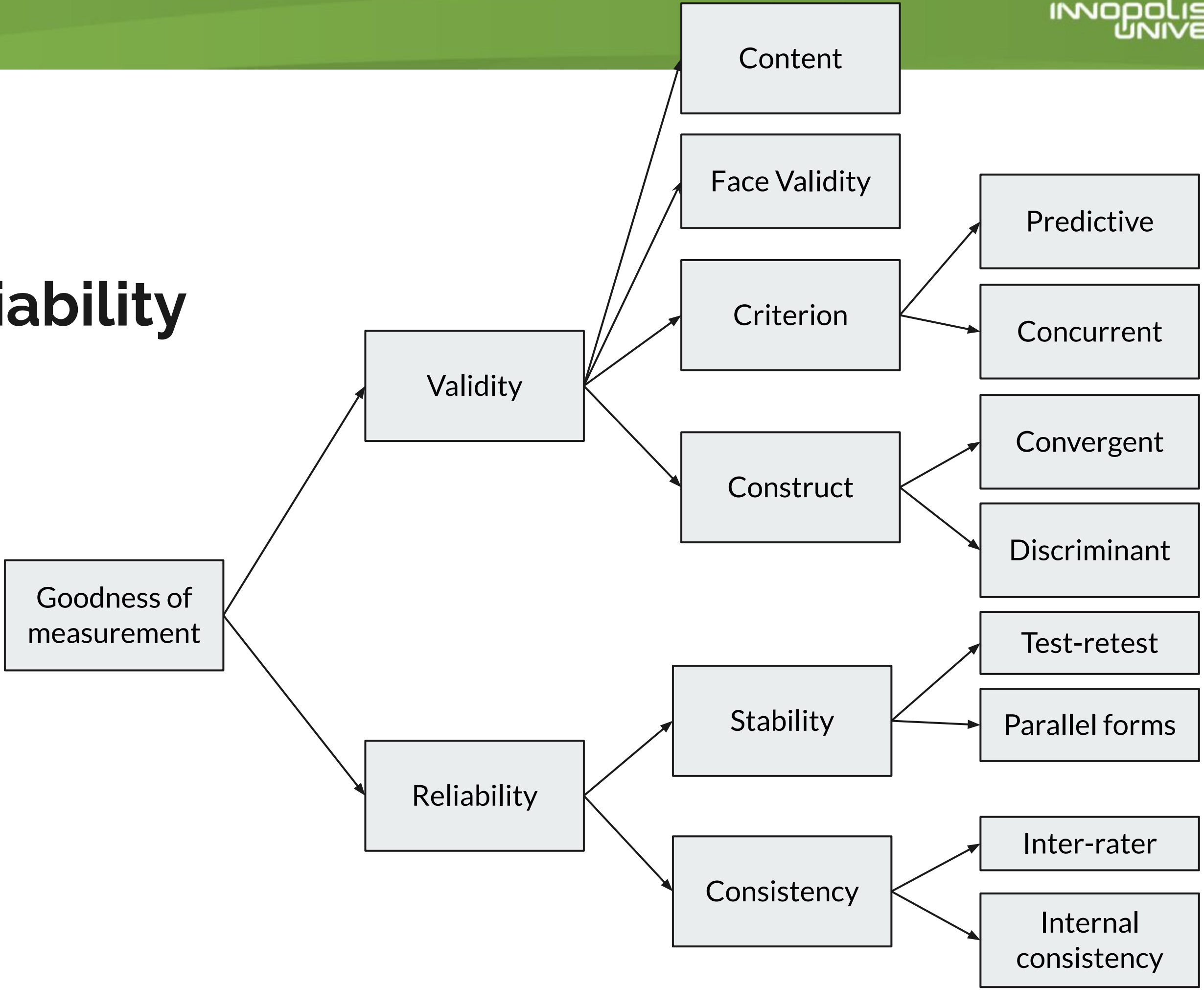
Validity & Reliability

Validity

- Content
- Criterion
- Construct

Reliability

- Stability
- Consistency

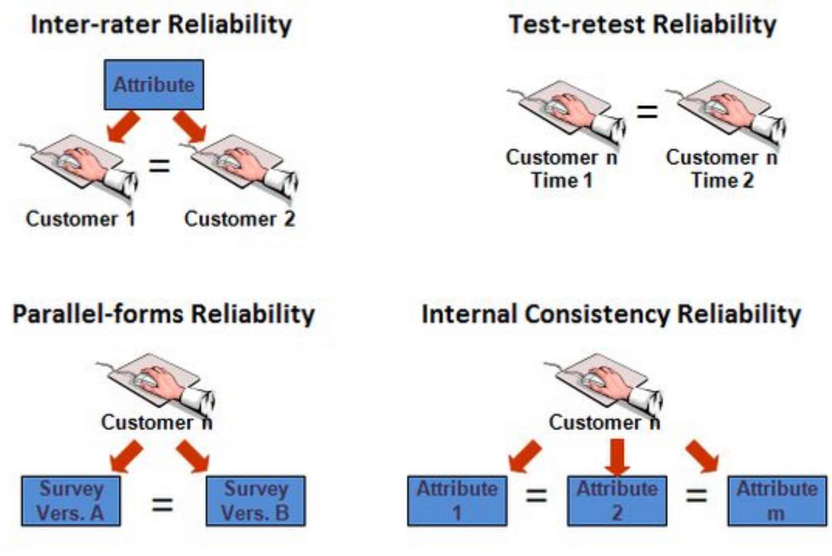


Reliability

- The degree to which measures are free from random error and therefore yield consistent and stable results.
- Stability and consistency with which the instrument measures the concept and helps to assess the goodness of a measure.
- Maintains stability over time in the measurement of a concept
- Two important dimensions of reliability:
  - stability
  - consistency

Types of Reliability

- ➢ Inter-rater (consistency over raters)
- ➢ Test-retest & intra-rater (stability over time)
- ➢ Internal consistency (consistency over different items)
- ➢ Parallel forms reliability (stability over different forms)



Inter-rater Reliability

- is a measure of reliability used to assess the degree to which different judges or raters agree in their assessment decisions.
- is useful because human observers will not necessarily interpret answers the same way;
  - raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed.

Inter-rater Reliability

Example

- Different judges are evaluating the degree to which art portfolios meet certain standards. Inter-rater reliability is especially useful when judgments can be considered relatively subjective.
- Thus, the use of this type of reliability would probably be more likely when evaluating artwork as opposed to math problems.

Used for only nominal measures

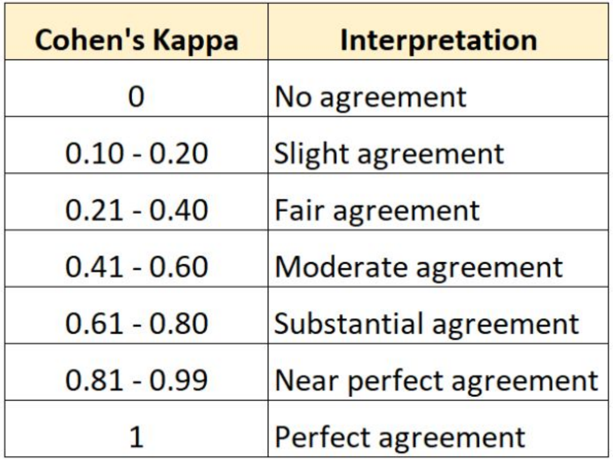
Cohen’s kappa coefficient (*κ*)

- is a statistic that is used to measure inter-rater reliability between **two raters** for qualitative (categorical) items.
- Cohen's kappa measures the agreement between **two raters** who each classify N items into C mutually exclusive categories. The definition of **κ** is:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

- p_0 = relative observed agreement among raters.
- p_e = the hypothetical probability of chance agreement.

Statistics are skipped in these slides



Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Fleiss’ kappa coefficient (*κ*)

- **More than two raters.**

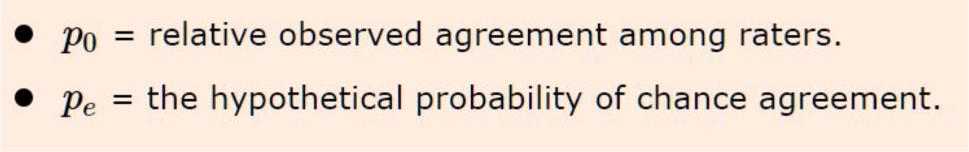
Cohen’s kappa coefficient (*κ*)

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Given the answers of the readers in the following table, calculate Cohen's kappa coefficient.

Statistics are skipped in these slides

		reader B	
		Yes	No
reader A	Yes	20	5
	No	10	15

Cohen’s kappa coefficient (*κ*)



- p_0 = relative observed agreement among raters.
- p_e = the hypothetical probability of chance agreement.

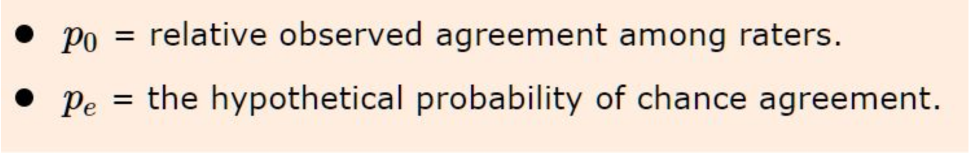
Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Given the answers of the readers in the following table, calculate Cohen's kappa coefficient.

Statistics are skipped in these slides

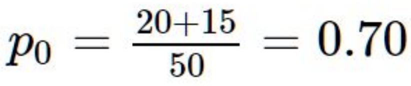
		reader B	
		Yes	No
reader A	Yes	20	5
	No	10	15

$$p_0 = \frac{20+15}{50} = 0.70$$

Cohen’s kappa coefficient (*κ*)



- p_0 = relative observed agreement among raters.
- p_e = the hypothetical probability of chance agreement.

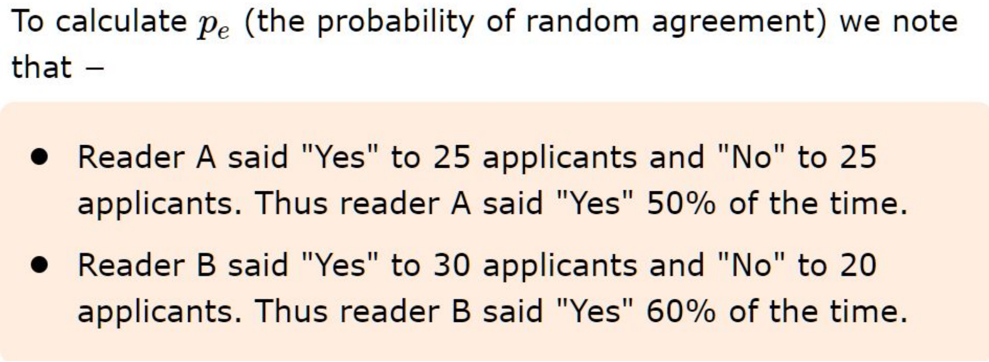


$$p_0 = \frac{20+15}{50} = 0.70$$

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Given the answers of the readers in the following table, calculate Cohen's kappa coefficient.

Statistics are skipped in these slides

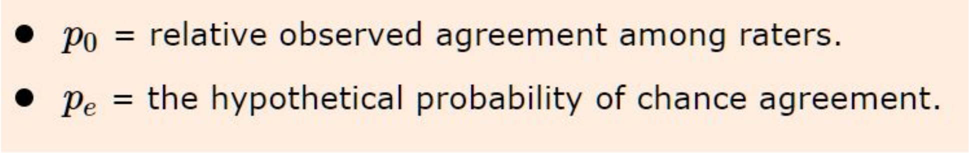
		reader B	
		Yes	No
reader A	Yes	20	5
	No	10	15



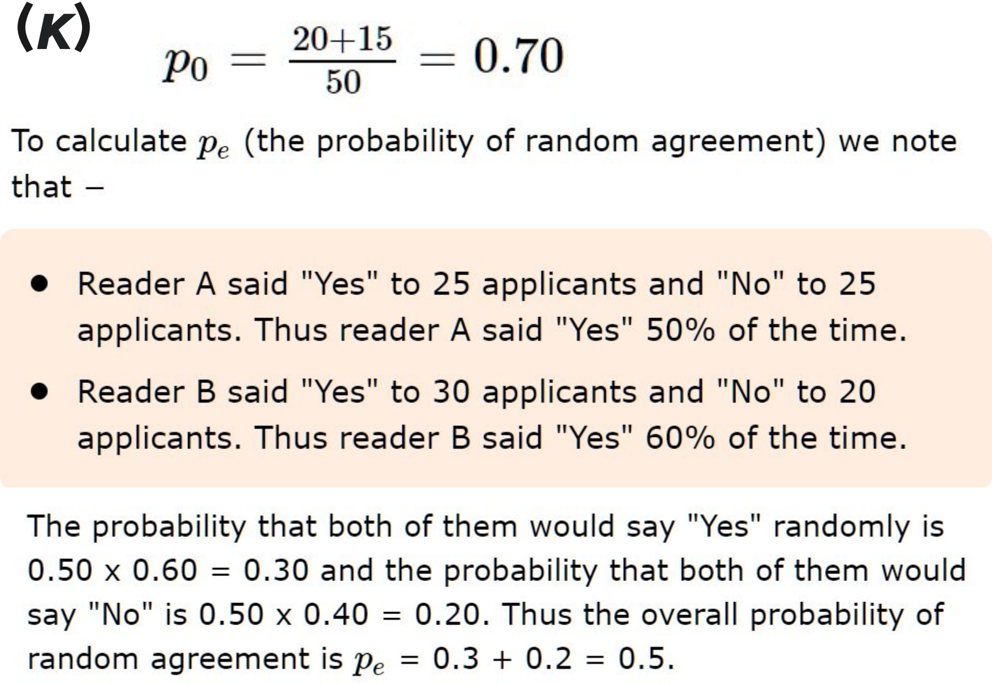
To calculate p_e (the probability of random agreement) we note that –

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

Cohen’s kappa coefficient (*κ*)



- p_0 = relative observed agreement among raters.
- p_e = the hypothetical probability of chance agreement.



$$p_0 = \frac{20+15}{50} = 0.70$$

To calculate p_e (the probability of random agreement) we note that –

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

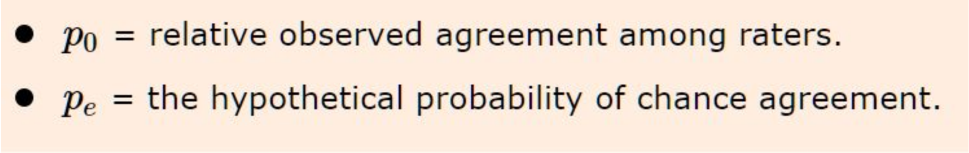
The probability that both of them would say "Yes" randomly is 0.50 x 0.60 = 0.30 and the probability that both of them would say "No" is 0.50 x 0.40 = 0.20. Thus the overall probability of random agreement is p_e = 0.3 + 0.2 = 0.5.

Statistics are skipped in these slides

		reader B	
		Yes	No
reader A	Yes	20	5
	No	10	15

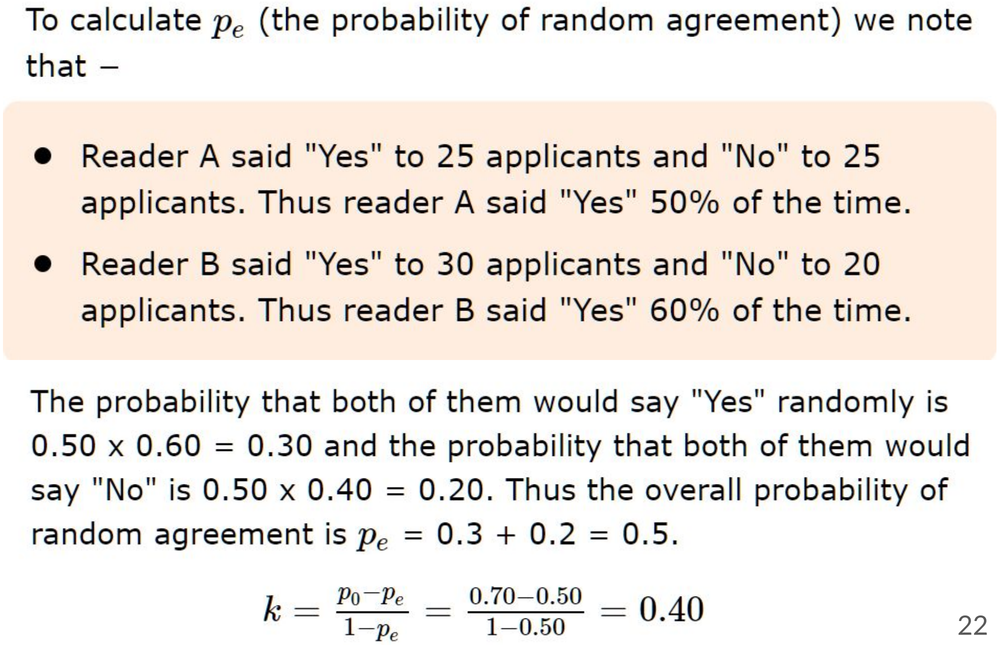
Measuring inter-rater reliability in R
https://colab.research.google.com/drive/15MXI-XO2qCyVpls-u-vHI7D-9MRDRjOa?usp=sharing

Cohen’s kappa coefficient (*κ*)



- p_0 = relative observed agreement among raters.
- p_e = the hypothetical probability of chance agreement.

$$p_0 = \frac{20+15}{50} = 0.70$$

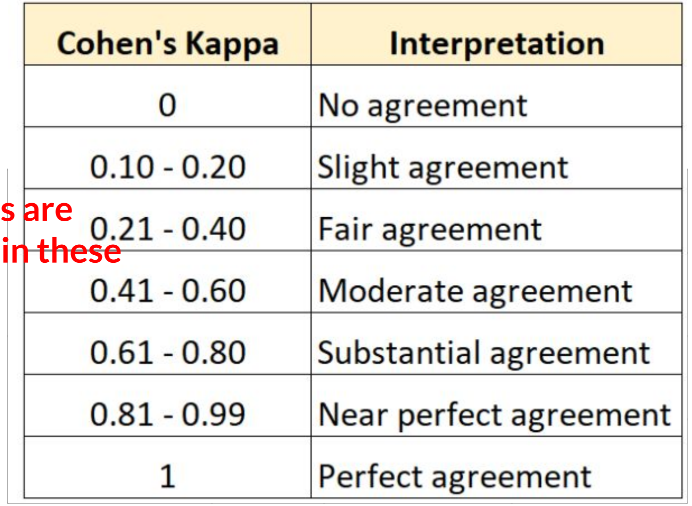


To calculate p_e (the probability of random agreement) we note that –

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

The probability that both of them would say "Yes" randomly is 0.50 x 0.60 = 0.30 and the probability that both of them would say "No" is 0.50 x 0.40 = 0.20. Thus the overall probability of random agreement is p_e = 0.3 + 0.2 = 0.5.

$$k = \frac{p_0 - p_e}{1 - p_e} = \frac{0.70-0.50}{1-0.50} = 0.40$$



Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Statistics are skipped in these slides

Used for only nominal measures

Fleiss’ kappa coefficient (κ)

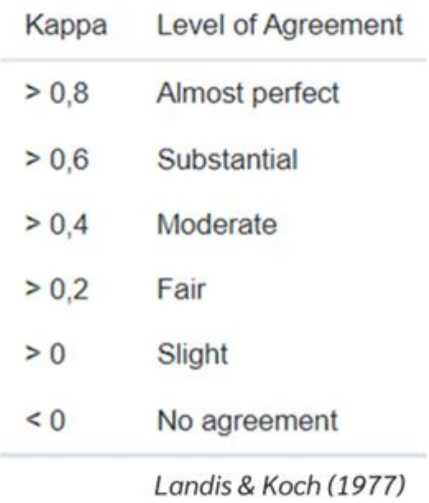
- is another statistic for evaluating the inter rater consistency among **more than two raters** when assigning categorical ratings to a number of items or classifying items.
- Let N be the total number of subjects, let n be the number of ratings per subject, and let k be the number of categories into which assignments are made.
- It is given by the following formula (similar to Cohen’s formula)

Statistics are skipped in these slides

Observed agreement

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Expected agreement if random judgment



Kappa	Level of Agreement
> 0,8	Almost perfect
> 0,6	Substantial
> 0,4	Moderate
> 0,2	Fair
> 0	Slight
< 0	No agreement

Landis & Koch (1977)

Measuring inter-rater reliability in R
https://colab.research.google.com/drive/15MXI-XO2qCyVpls-u-vHI7D-9MRDRjOa?usp=sharing

Fleiss' kappa coefficient (κ)

- Let's say we have 7 patients and three raters. Each patient has been assessed by each rater.
- Here, we simply count how many times a patient was judged to be depressed and how many times they were judged not to be depressed.
- 8/21=38% of the patients are rated as not depressed by the raters.
- 13/21=62% of the patients are rated as depressed by the raters.

Statistics are skipped in these slides

Patient	Rater 1	Rater 2	Rater 2	☺	☹
1	☹	☹	☹	0	3
2	☹	☺	☹	1	2
3	☺	☺	☺	3	0
4	☺	☹	☹	1	2
5	☹	☹	☹	0	3
6	☺	☺	☹	2	1
7	☹	☹	☺	1	2

Σ 8 13 21

$$\frac{8}{21} = 0.38 \qquad \frac{13}{21} = 0.62$$

$$p_e = \sum p_j^2 = 0.38^2 + 0.62^2 = 0.53$$

Measuring inter-rater reliability in R

https://colab.research.google.com/drive/15MXI-XO2qCyVpls-u-vHI7D-9MRDRjOa?usp=sharing

Fleiss' kappa coefficient (κ)

- Let's say we have 7 patients and three raters. Each patient has been assessed by each rater.
- Here, we simply count how many times a patient was judged to be depressed and how many times they were judged not to be depressed.
- N=7 patients, n=3 raters, k=2 categories (depressed, not depressed).
- nij is the number of votes of category j for the patient i.

Statistics are skipped in these slides

$$p_e = \sum p_j^2 = 0.38^2 + 0.62^2 =$$

Patient	Rater 1	Rater 2	Rater 2	🙂	☹
1	☹	☹	☹	0	3
2	☹	🙂	☹	1	2
3	🙂	🙂	🙂	3	0
4	🙂	☹	☹	1	2
5	☹	☹	☹	0	3
6	🙂	🙂	☹	2	1
7	☹	☹	🙂	1	2

$$p_0 = \underbrace{\frac{1}{N \cdot n \cdot (n-1)}}_{\frac{1}{7 \cdot 3 \cdot (3-1)} = 0.024} \left(\underbrace{\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2}_{0^2 + 3^2 + \cdots 1^2 + 2^2 = 47} - \underbrace{N \cdot n}_{7 \cdot 3 = 21} \right)$$

$$p_0 = 0.024 \cdot (47 - 21) = 0.624$$

Measuring inter-rater reliability in R
https://colab.research.google.com/drive/15MXI-XO2qCyVpls-u-vHI7D-9MRDRjOa?usp=sharing

Fleiss' kappa coefficient (κ)

- Let's say we have 7 patients and three raters. Each patient has been assessed by each rater.
- Here, we simply count how many times a patient was judged to be depressed and how many times they were judged not to be depressed.
- N=7 patients, n=3 raters, k=2 categories (depressed, not depressed).
- nij is the number of votes of category j for the patient i.

Statistics are skipped in these slides

$$p_e = \sum p_j^2 = 0.38^2 + 0.62^2 =$$

$$p_0 = 0.024 \cdot (47 - 21) = 0.624$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$= \frac{0.624 - 0.53}{1 - 0.53} = 0.19$$

slight agreement

Test-Retest (& Intra rater) Reliability

- It is obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time.
- Procedure:
  - Give a test (make a rating -the rater as the instrument)
  - Allow time to pass.
  - Give another test (make another rating)
  - Analyze the correlation of the two test scores (ratings).

Test-Retest (& Intra rater) Reliability

Example

- A test designed to assess student learning in psychology could be given to a group of students twice, with the second administration perhaps coming a week after the first. The obtained correlation coefficient would indicate the stability of the scores.
- Pearson Correlation Coefficient
- Intra-class correlation coefficient

Statistics are skipped in these slides

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Internal Consistency reliability

is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results.

Example:

For example, if designing a test on geometry, then all questions on the test should be about geometry.

Used for only measurement scales which has the order property (ordinal, interval, and ratio)

Cronbach’s Alpha coefficient α

is a way to measure the internal consistency of a measure (questionnaire or survey).

Statistics are skipped in these slides

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right)$$

k represents the number of items in the measure

$\sigma_{y_i}^2$ the variance associated with each item *i*

σ_y^2 the variance associated with the total scores $\left(y = \sum_{i=1}^{k} y_i\right)$

Cronbach’s Alpha	Internal consistency
0.9 ≤ α	Excellent
0.8 ≤ α < 0.9	Good
0.7 ≤ α < 0.8	Acceptable
0.6 ≤ α < 0.7	Questionable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

Statistics are skipped in these slides

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right)$$

Cronbach's Alpha coefficient α

Suppose a restaurant manager wants to measure overall satisfaction among customers. She decides to send out a survey to 10 customers who can rate the restaurant on a scale of 1 to 3 for various categories.

Respondent	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
Q1	3	1	1	3	2	2	3	1	2	3
Q2	2	1	2	3	2	2	2	2	3	2
Q3	3	1	2	3	2	2	2	3	2	3

https://docs.google.com/spreadsheets/d/1W5uHzPZj14xZH1aX7LBGIOsArQ3_S62hMeHXoroiGk4/edit?usp=sharing

Statistics are skipped in these slides

Var(total=y) = 2.94
SUM(variance) = 1.54
Alpha = 0.71

Acceptable

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right)$$

Cronbach’s Alpha coefficient α

Suppose a restaurant manager wants to measure overall satisfaction among customers. She decides to send out a survey to 10 customers who can rate the restaurant on a scale of 1 to 3 for various categories.

Respondent	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Variance
Q1	3	1	1	3	2	2	3	1	2	3	0.77
Q2	2	1	2	3	2	2	2	2	3	2	0.32
Q3	3	1	2	3	2	2	2	3	2	3	0.46
Total	**8**	**3**	**5**	**9**	**6**	**6**	**7**	**6**	**7**	**8**	

https://docs.google.com/spreadsheets/d/1W5uHzPZj14xZH1aX7LBGIOsArQ3_S62hMeHXoroiGk4/edit?usp=sharing

Statistics are skipped in these slides

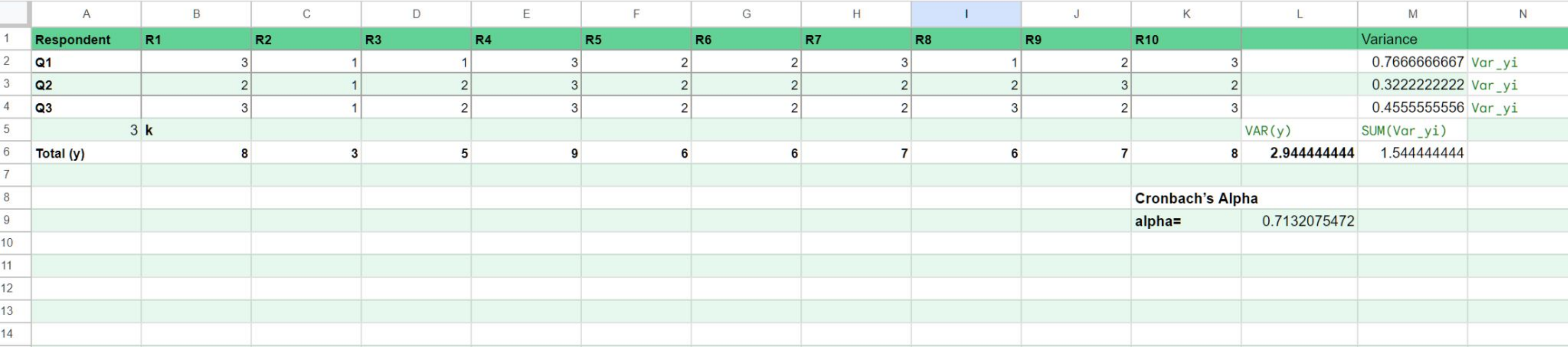
Var(total=y) = 2.94
SUM(variance) = 1.54
Alpha = 0.71

Acceptable

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_y^2}\right)$$

Cronbach’s Alpha coefficient α

Suppose a restaurant manager wants to measure overall satisfaction among customers. She decides to send out a survey to 10 customers who can rate the restaurant on a scale of 1 to 3 for various categories.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Respondent	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10		Variance	
2	Q1	3	1	1	3	2	2	3	1	2	3		0.7666666667	Var_yi
3	Q2	2	1	2	3	2	2	2	2	3	2		0.3222222222	Var_yi
4	Q3	3	1	2	3	2	2	2	3	2	3		0.4555555556	Var_yi
5	3	k										VAR(y)	SUM(Var_yi)	
6	Total (y)	8	3	5	9	6	6	7	6	7	8	2.944444444	1.544444444	
7														
8											Cronbach’s Alpha			
9											alpha=	0.7132075472		
10														
11														
12														
13														
14														

Parallel forms reliability

- is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals.
- The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternative versions

Parallel forms reliability

Example

- If you wanted to evaluate the reliability of a critical thinking assessment, you might create a large set of items that all pertain to critical thinking and then randomly split the questions up into two sets, which would represent the parallel forms.

Validity

- **Reliability** is necessary, it is alone not sufficient.

- For example, if your scale is off by 50 g, it reads your weight every day with an excess of 50 g. The scale is reliable because it consistently reports the same weight every day, but it is not valid because it adds 50 g to your true weight. It is not a valid measure of your weight.

Validity

- It refers to how well a test measures what it is purported to measure.
- Measures are valid for a single purpose (the thermometer measures temperature and nothing else)
- Four types of validity:
  - Face – as judged by others or by logic
  - Content – captures the entire meaning of the experience
  - Construct – it measures the construct adequately.
  - Criterion – agrees with a validated, reliable external source:
    - Concurrent, agrees with a preexisting measure
    - Predictive, agrees with a future behavior or outcome

Face Validity

- ascertains that the measure appears to be assessing the intended construct under study.
- The stakeholders can easily assess face validity.
- Although this is **not a very “scientific”** type of validity, it may be an essential component in enlisting motivation of stakeholders.
- If the stakeholders do not believe the measure is an accurate assessment of the ability, they may become disengaged with the task.

Face Validity

Example:

- ➢ A survey that asks people about their job satisfaction might have high face validity if the questions appear to be directly related to job satisfaction and the words are easy to understand

Construct Validity

- is used to ensure that the test is actually measuring what it is intended to measure (i.e. the construct), and not other variables. It refers to whether a test measures the construct adequately.

Example:

➢ A women's studies program may design a cumulative assessment of learning throughout the major. The questions are written with complicated wording and phrasing. This can cause the test unintentionally becoming a test of reading comprehension, rather than a test of women's studies. It is important that the measure is actually assessing the intended construct, rather than an extraneous factor.

Construct Validity (Convergent validity)

➢ helps to establish construct validity when you use two different measurement methods (e.g., participant observation and a survey) in your research to collect data about a construct (e.g., anger, depression, motivation, task performance).

➢ The extent to which convergent validity has been demonstrated is established by the strength of the relationship between the scores that are obtained from the two different measurement methods

➢ The idea is that if these scores converge, despite the fact that we use two different measurement methods, we must be measuring the same construct.

Construct Validity (Convergent validity)

- **Construct:** Sleep quality
- Study the relationship between fitness level and sleep quality;
  - the impact that exercise has on how well people sleep
- When participants in the study wake up in the morning, they record their sleep quality using a self-completed survey (i.e., they fill in a questionnaire)
  - This gives us insight into how well the participants felt they slept.
- Is this a reliable measurement procedure to measure the construct, sleep quality?

Construct Validity (Convergent validity)

- **Construct:** Sleep quality
- Study the relationship between fitness level and sleep quality;
  - the impact that exercise has on how well people sleep
- Self-completed measurement procedures can be prone to certain biases.
- Therefore, we also observe the participants whilst they are sleeping using a video camera to monitor their sleeping patterns.
  - When making the observations, we score the participants' sleep quality.

Construct Validity (Convergent validity)

- **Construct:** Sleep quality
- Self-completed survey
- Participant observation
- We hope that by using two different research methods to assess sleep quality, we will have a more reliable measurement procedure for the construct we are interested in.
- Two different sets of scores from the two different measurement procedures used under the two research methods

Construct Validity (Convergent validity)

- **Construct:** Sleep quality
- Self-completed survey
- Participant observation
- If there is a strong relationship between the two scores
  - convergent validity
  - We can be more confident that the measurement procedures that we are using to measure sleep quality are a valid measure of the construct, sleep quality.

Construct Validity (Divergent validity)

- ➢ helps to establish construct validity by demonstrating that the construct you are interested in (e.g., anger) is different from other constructs that might be present in your study (e.g., depression)
- ➢ To assess construct validity in your research, you should first establish convergent validity, before testing for divergent validity.
- ➢ The extent to which divergent validity has been demonstrated is establish by the **strength** of the relationship between the scores that are obtained from the two different measurement procedures and research methods that you have used to collect data about the **two constructs** you are interested in.
- ➢ We are interested in the extent to which the scores diverge (i.e., we want to see little or no relationship between the two scores from the two constructs)

Construct Validity (Divergent validity)

- This is a two-step process:
  - **Establish convergent validity:** A strong relationship should be established between the two scores for each of the two constructs (e.g., a strong relationship for anger and a strong relationship for depression).
  - **Establish divergent validity:** Little or no relationship should be found between the two scores between the two constructs (e.g., little or no relationship between anger and depression) when comparing the same methods used to collect the data (e.g., comparing anger and depression from the observational scores, and comparing anger and depression from the survey scores).

Construct Validity (Divergent validity)

- **Construct #1 =** Sleep **quality**
- **Construct #2 =** Sleep **quantity**
- **Note**: Quality vs. Quantity of Sleep
- Check convergent validity for sleep quality construct.
  - Let's imagine that we established a strong relationship between the two sets of scores from the two different measurement procedures under the two research methods for each construct.
- Did we include sleep quality within the same set of measures (questions in survey) used to measure sleep quality.

Construct Validity (Divergent validity)

- ➢ **Construct #1 =** Sleep **quality**
- ➢ **Construct #2 =** Sleep **quantity**
- ➢ **Note**: Quality vs. Quantity of Sleep
- ➢ Are the sleep quality and sleep quantity are part of the same construct or are two different constructs?
    - ○ if sleep quality and sleep quantity are two different constructs, but we measured them as if they were the same construct, we have introduced a confounding variable that will inevitably reduce the internal validity of our study

Construct Validity (Divergent validity)

- **Construct #1 =** Sleep **quality**
- **Construct #2 =** Sleep **quantity**
- **Note**: Quality vs. Quantity of Sleep
- To achieve this, we use the same research methods;
  - that is, we ask participants to complete a survey, as well as observing participants whilst sleeping.
- However, the survey contains (a) questions that measure sleep quality and (b) questions that measure sleep quantity.
- Similarly, when we observe participants, we record scores separately for (a) sleep quality and (b) sleep quantity.

Construct Validity (Divergent validity)

- **Construct #1 =** Sleep **quality**
- **Construct #2 =** Sleep **quantity**
- **Note**: Quality vs. Quantity of Sleep
- In order to assess whether the two constructs (i.e., sleep quality and sleep quantity) are different
- We first need to find that both constructs have convergent validity.
  - (1) Therefore, there should be a strong relationship between the survey scores and observational scores for (a) sleep quality and (b) sleep quantity.

Construct Validity (Divergent validity)

- **Construct #1 =** Sleep **quality**
- **Construct #2 =** Sleep **quantity**
- **Note**: Quality vs. Quantity of Sleep
- Next, we need to find that these two constructs are distinct;
  - that is, that we have divergent validity.
  - Therefore, there should be little or no relationship between (a) the survey scores for sleep quality and the survey scores for sleep quantity and (b) the observational scores for sleep quality and the observational scores for sleep quantity.

Construct Validity

- Construct validity can start to be established when:
  - both convergent and divergent validity are established
  - because construct validity is something that is built over time. No single study can establish construct validity.

Criterion Validity

- is used to predict future or current performance whether it correlates test results with another criterion of interest.
- We use a criterion - a well-established measurement procedure - to create a new measurement procedure to measure the construct you are interested in.

Example:

➢ If a physics program designed a measure to assess cumulative student learning throughout the major. The new measure could be correlated with a standardized measure of ability in this discipline, such as the GRE subject test. The higher the correlation between the established measure and new measure, the more faith stakeholders can have in the new assessment tool.

Criterion Validity (Predictive validity)

- Predictor variables, including number of new homes purchased, building permits awarded, interest rates on mortgages, and employment rate have high criterion validity in predicting the prices of houses.
- **Predictor Variable:** Building permits issued, interest rates on mortgages, employment rate
- **Criterion Variable:** Housing prices
- The housing market is a classic indicator of economic performance. The volume of sales each quarter are affected by numerous factors, including: the employment rate, interest rates, building supply, and consumer confidence, just to name a few.

It refers to how well a person's score on one variable predicts their score on a second variable.

Criterion Validity (Predictive validity)

- Predictor variables, including number of new homes purchased, building permits awarded, interest rates on mortgages, and employment rate have high criterion validity in predicting the prices of houses.
- **Predictor Variable:** Building permits issued, interest rates on mortgages, employment rate
- **Criterion Variable:** Housing prices
- Each one of those factors can be measured and correlated with the housing market. Some factors have strong criterion validity, while others may have moderate or low criterion validity. However, when economists put them all together, the ability to predict the housing market improves significantly.

Criterion Validity (Concurrent validity)

- Step counters that you wear on your watch apparently have high criterion validity. To test this, Adamakis (2021) got people to jog on a treadmill, counted their steps, then compared it to the results on the step counter. The step counters did pretty well!
- **Predictor Variable:** Steps recorded on a step counter
- **Criterion Variable:** Actual steps walked
- In this study, thirty adults wore two smartphones (one Android and one iOS), while running four apps: Runtastic Pedometer, Accupedo, Pacer, and Argus. They walked and jogged on a treadmill at three different speeds for 5 minutes. Two research assistants counted every step they took with a digital counter.
- Criterion validity of the apps was then assessed by comparing the data from the apps with the 100% accurate digital counters.

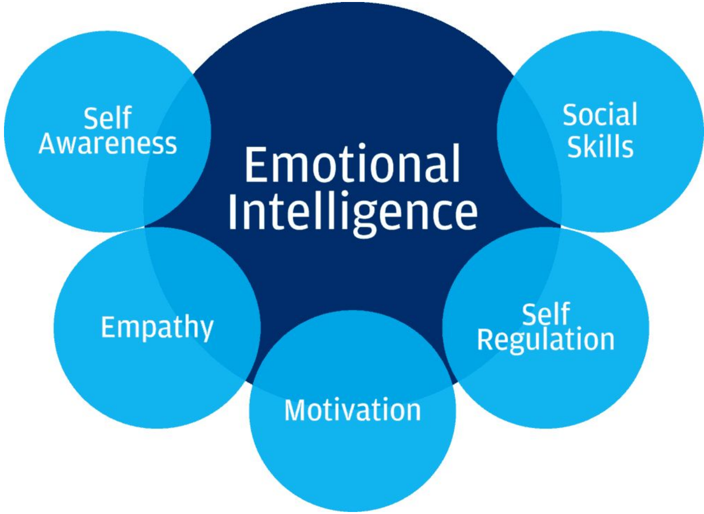
It offers a way of establishing a test's validity by comparing it to another similar test that is known to be valid. If the two tests correlate, then the new study is believed to also be valid.

Criterion Validity (Concurrent validity)

- Step counters that you wear on your watch apparently have high criterion validity. To test this, Adamakis (2021) got people to jog on a treadmill, counted their steps, then compared it to the results on the step counter. The step counters did pretty well!
- **Predictor Variable:** Steps recorded on a step counter
- **Criterion Variable:** Actual steps walked
- In this study, thirty adults wore two smartphones (one Android and one iOS), while running four apps: Runtastic Pedometer, Accupedo, Pacer, and Argus. They walked and jogged on a treadmill at three different speeds for 5 minutes. Two research assistants counted every step they took with a digital counter.
- Criterion validity of the apps was then assessed by comparing the data from the apps with the 100% accurate digital counters.

Content validity

- It refers to whether a test or scale is measuring all of the components of a given construct.
- is the degree to which a test or assessment instrument evaluates all aspects of the topic, construct, or behavior that it is designed to measure.
- **For example,** if there are five dimensions of emotional intelligence (EQ), then a scale that measures EQ should contain questions regarding each dimension
- Content validity assesses my test to see if it covers suitable material for that subject area at that level of expertise. In other words, does my test cover all pertinent facets of the content area? Is it missing concepts?



Goodness of measurement – Example 1

When a patient loses faith in the medicine his doctor prescribes, it loses much of its power to improve his health. He may skip doses, and in the end may decide doctors cannot help him and let treatment lapse all together. For similar reasons, when selecting a test one must consider how worthwhile it will appear to the participant who takes it and other non-expert who will see the results

Face validity

Goodness of measurement – Example 2

The student admission department wants to know how well do SAT scores predict college GPA in order to enroll the future top performers.

Criterion (Predictive) validity

Goodness of measurement – Example 3

A newly developed math test for the SAT will need to be validated before giving it to thousands of students. So, the new version of the test is administered to a sample of college math majors along with the old version of the test.

Criterion (Concurrent) validity

Goodness of measurement – Example 4

The math portion of the SAT contains questions that require skills in all types of math: arithmetic, algebra, geometry, calculus, and many others.

Content validity

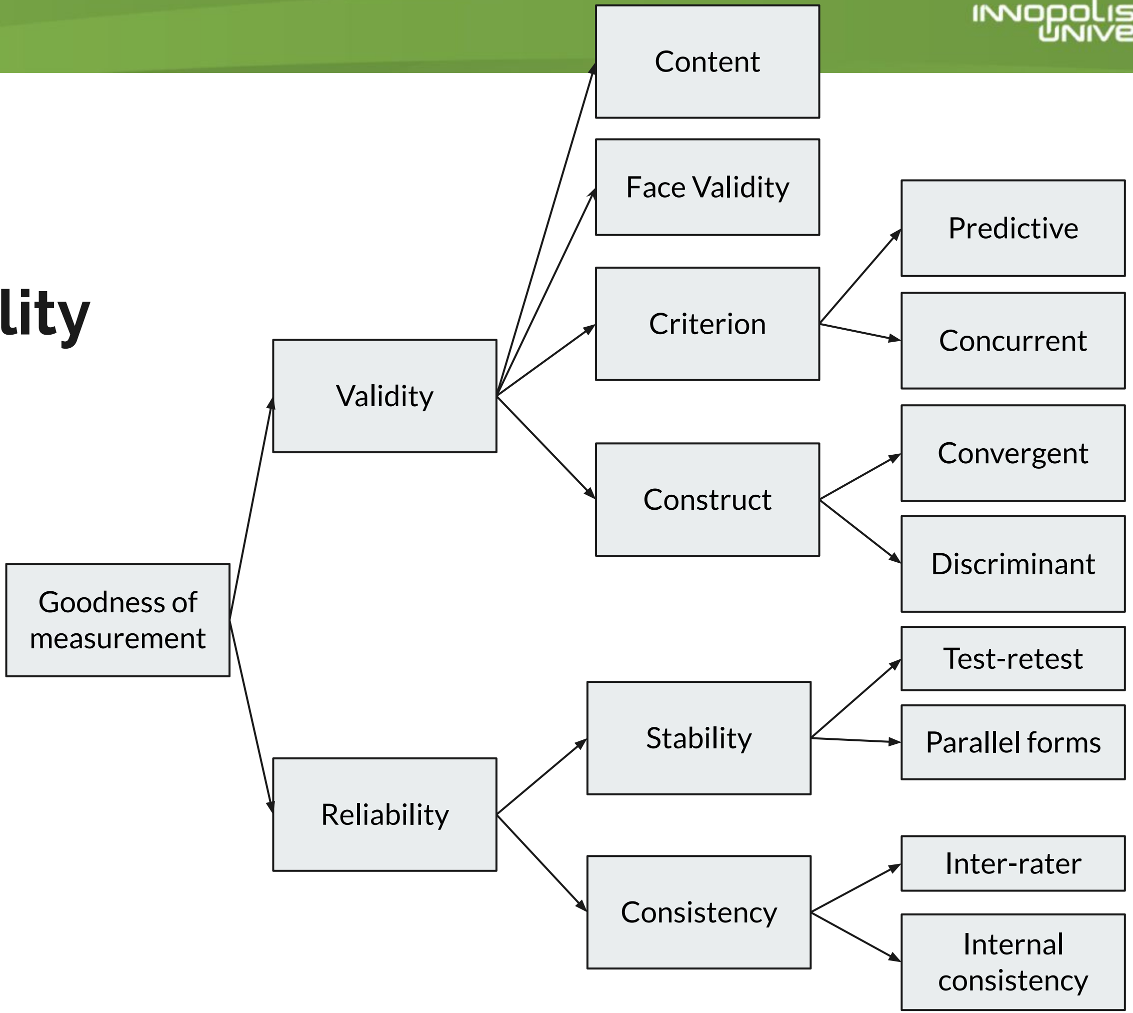
Relationship between reliability and validity

★ A valid test must be reliable.

★ A reliable test need not be valid.

Validity & Reliability

Practice: Suggest a research construct, measurement method and specify how to check the goodness of the measurement for one of the criteria shown on the right.

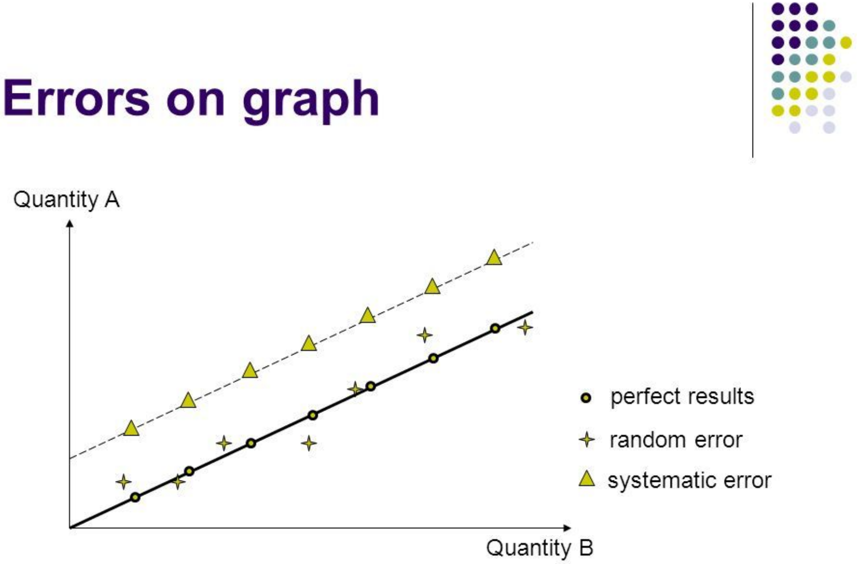


Measurement errors

- No matter how careful you are, there is always error in a measurement. Error is not a "mistake"—it's part of the measuring process.
- In science, measurement error is called experimental error or observational error. There are two broad classes of experimental errors: **random error** and **systematic error.**

Measurement errors

- **Random errors**
  - It varies unpredictably from one measurement to another,
  - are unavoidable, but cluster around the true value
- **Systematic errors**
  - It has the same value or proportion for every measurement.
  - can often be avoided by calibrating equipment, but if left uncorrected, can lead to measurements far from the true value



Errors on graph

Measurement errors – Examples

When weighing yourself on a scale, you position yourself slightly differently each time.

Random error

Measurement errors – Examples

Measured distance is different using a new cloth measuring tape versus an older, stretched one.

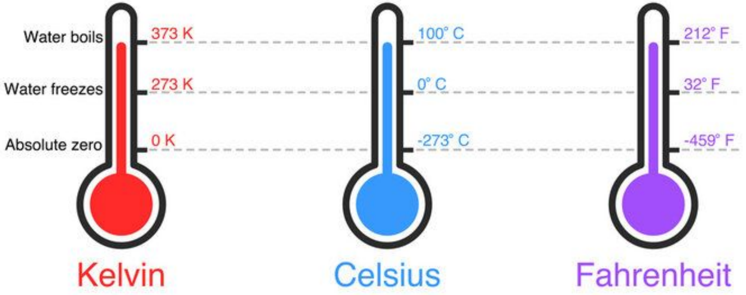
Systematic error

- Proportional errors of this type are called **scale factor errors**.

Measurement errors – Examples

An improperly calibrated thermometer may give accurate readings within a certain temperature range, but become inaccurate at higher or lower temperatures.

Systematic error



Measurement errors – Examples

Irregular changes in the heat loss rate from a solar collector due to changes in the wind.

Random error

Measurement errors – Examples

You want to measure the height of a tree using a measuring tape. The tree's true height is 10 feet. You take repeated measurements. The first measurement is 10.2 feet, the second is 9.9 feet, and the third is 10.1 feet.

Random error

- variations in the measuring tape
- the angle you look at the tape
- the sun in your eyes
- the wind blowing the tape

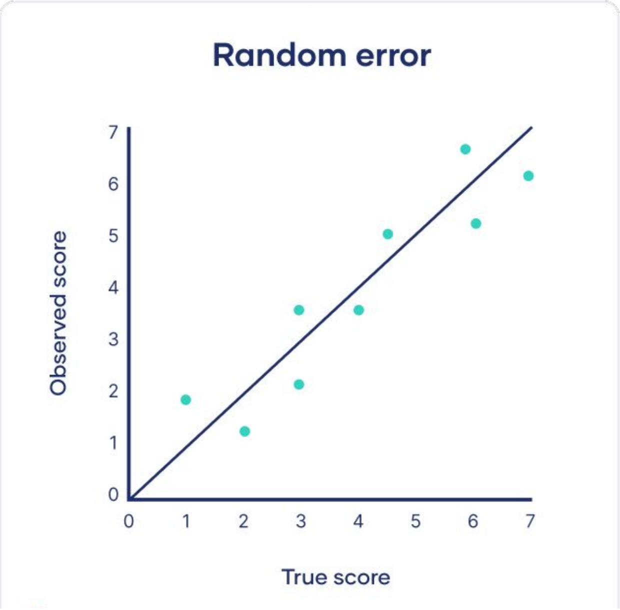
Measurement errors – Examples

Forgetting to tare or zero a balance produces mass measurements that are always “off” by the same amount.

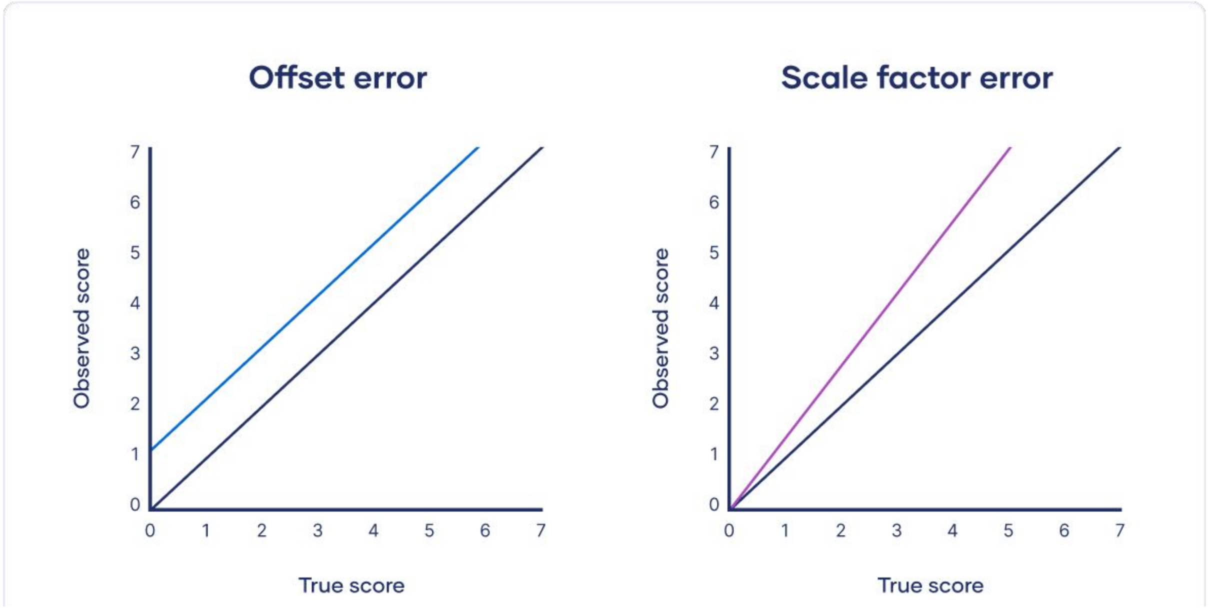
Systematic error

- Offset error

Measurement errors



Random error



Offset error

Scale factor error

Reducing Random errors

- **Take repeated measurements**
  - If you take multiple measurements of the same construct, you can **average** them together to get a more precise result
- **Increase your sample size**
  - The more data points you have, the less random error will affect your results.
  - That’s why larger sample sizes are generally better than smaller ones regarding precision and statistical power.
- **Increase the sensitivity of measuring instruments**
  - Use more sensitive instruments or calibrate them regularly.
- **Control other variables**
  - By controlling all relevant variables (controlled experiments), you can minimize sources of error and get more accurate results.

Reducing Systematic errors

- **Triangulation**
  - use multiple techniques to record observations so you’re not relying on only one instrument or method.
- **Regular calibration**
  - frequently comparing what the instrument records with the value of a known, standard quantity reduces the likelihood of systematic errors affecting your study.

Summary

- Goodness of measurement
  - Reliability
  - Validity
- Measurement errors
  - Random error
  - Systematic error

References

- https://explorable.com/types-of-validity
- https://www.researchgate.net/profile/Ram-Bajpai-3/publication/271186978_Goodness_of_Measurement_Reliability_and_Validity/links/5503164b0cf24cee39fd591b/Goodness-of-Measurement-Reliability-and-Validity.pdf
- https://chfasoa.uni.edu/reliabilityandvalidity.htm
- https://www.thoughtco.com/random-vs-systematic-error-4175358#:~:text=Random%20error%20causes%20one%20measurement,It%20is%20predictable.
- https://statisticsbyjim.com/basics/random-error-vs-systematic-error/
- https://datatab.net/tutorial/fleiss-kappa
- https://helpfulprofessor.com/
- https://gradcoach.com/research-constructs/
- https://dissertation.laerd.com/convergent-and-divergent-validity.php

Attendance

https://baam.duckdns.org

Questions?