DEPARTMENT OF MATHEMATICAL SCIENCES
(Faculty of Science)
BAUCHI STATE UNIVERSITY GADAU
2023/2024 Academic Semester

Course Code and Title: STAT 4313 Psychometric
Credit Unit: (2 Units)

---

1 Introduction to Psychometric

Psychometrics is a way to study how we measure things in psychology. It helps us make sure our tests are reliable (give the same results every time) and valid (test what they are supposed to). We use psychometrics to study things like intelligence and personality, and it helps us understand how our brain works. Psychometrics is important in many areas of psychology, like helping people who need it or figuring out who would be good at a job. Psychometrics is the scientific study of measurement in psychology. It is concerned with the development, validation, and application of psychological tests and assessments. The field of psychometrics has become increasingly important in recent years due to the growing need for reliable and valid measures of psychological constructs such as intelligence, personality, and mental health. Psychometrics is used in a variety of settings, including clinical psychology, educational psychology, and industrial-organizational psychology. One of the fundamental concepts in psychometrics is reliability. Reliability refers to the consistency of a test or assessment over time, across different raters or observers, and with different samples of people. A reliable test is one that produces consistent results when administered to the same person on different occasions. There are several methods for measuring reliability, including test-retest reliability, inter-rater reliability, and internal consistency. Test-retest reliability involves administering the same test to the same group of people at two different points in time and comparing the scores. Inter-rater reliability involves comparing the scores of two or more raters who have scored the same test or assessment. Internal consistency refers to the extent to which the items on a test or assessment are related to each other and measure the same construct. Another important concept in psychometrics is validity. Validity refers to the extent to which a test or assessment measures what it is intended to measure. There are several types of validity, including content validity, criterion-related validity, and construct validity. Content validity refers to the extent to which a test or assessment covers all aspects of the construct being measured. Criterion-related validity refers to the extent to which a test or assessment predicts performance on a specific criterion, such as job performance or academic achievement. Construct validity refers to the extent to which a test or assessment measures the underlying psychological construct it is intended to measure. Psychometrics is an important field of study in psychology that is concerned with the measurement of psychological constructs. It involves the development, validation, and application of psychological tests and assessments. Key concepts in psychometrics include reliability and validity, which are essential for ensuring that tests and assessments are accurate and useful. Psychometrics is used in a variety of settings, including clinical psychology, educational psychology, and industrial-organizational psychology, and is essential for understanding and evaluating psychological phenomena.

2 Examples

2.1 Example 1

In educational psychology, a teacher wants to measure the reading ability of their students. They develop a reading comprehension test for their class and administer it twice, two weeks apart. The teacher calculates the test-retest reliability by comparing the scores from both administrations of the test. They find that the test is highly reliable, meaning that students' scores are consistent over time, indicating that their reading ability has not changed significantly. This test can be used to accurately measure the reading ability of the students in this class.

2.2 Example 2

In clinical psychology, a therapist wants to assess their client's depression symptoms. They use a standardized depression assessment tool that has been validated through several studies. The therapist explains to the client that the tool has high construct validity, meaning that it is accurately measuring depression symptoms. The therapist can use this information to help diagnose and treat the client's depression symptoms.

2.3 Example 3

In industrial-organizational psychology, a company wants to assess the personality traits of potential job candidates. They use a personality assessment tool that has been validated and has high inter-rater reliability. The HR department can use this tool to accurately measure the personality traits of job candidates and determine if they are a good fit for the company culture and job requirements.

3 A Brief History of Psychometrics

Psychometrics is a branch of psychology that deals with the measurement of human abilities, traits, and characteristics. It involves the use of standardized tests and other assessment techniques to quantify and evaluate various psychological phenomena. It is a complex and dynamic field that has evolved over time, driven by advances in technology and changes in societal needs. One aspect of psychometrics that is particularly interesting is its history, which offers insights into how the field has developed and expanded over the years. The history of psychometrics can be traced back to the 19th century, when the French psychologist Alfred Binet developed the first standardized intelligence test. This test was designed to help identify children who were struggling in school and needed extra support. Binet's test was based on the concept of mental age, which refers to the level of intellectual functioning typically associated with a particular age group. This idea was later refined by the American psychologist Lewis Terman, who created the Stanford-Binet Intelligence Scale, which became one of the most widely used intelligence tests in the world. Over the years, psychometrics has grown to encompass a wide range of assessment techniques, including personality tests, aptitude tests, and achievement tests. These tests are used for a variety of purposes, such as selecting candidates for employment or admission to educational programs, diagnosing psychological disorders, and evaluating the effectiveness of interventions. Psychometricians today use sophisticated statistical methods to develop and validate tests, ensuring that they are reliable, valid, and fair to all individuals who take them. In conclusion, the history of psychometrics is a fascinating subject that offers a glimpse into the evolution of this important

field. From its humble beginnings with Alfred Binet's intelligence test to the sophisticated assessments used today, psychometrics has come a long way. As society's needs continue to change, psychometricians will undoubtedly continue to develop new and innovative ways to measure and evaluate human abilities, traits, and characteristics.

4 Sub-fields of Psychometrics

Psychometrics is the field of study that focuses on the measurement of psychological traits, such as intelligence, personality, and abilities. It is a branch of psychology that has gained significant importance over the years, as it provides a scientific approach to measuring psychological constructs. Psychometricians use statistical methods and theories to develop and validate tests and measurement tools for various purposes, such as educational and clinical assessments, personnel selection, and research. One of the fundamental aspects of psychometrics is the sub-fields that it comprises. These sub-fields are essential to understand the different aspects of psychological measurement, and they include classical test theory, item response theory, and factor analysis. Classical test theory is a sub-field of psychometrics that deals with the measurement of psychological constructs using observed scores. It involves the use of reliability and validity assessments to ensure that the tests used to measure a psychological construct are consistent and accurate. Item response theory is another sub-field of psychometrics that focuses on the individual items used to measure psychological constructs. It involves the analysis of the relationships between the responses to each item and the overall construct being measured. This sub-field has gained significant importance in recent years due to the increasing use of computerized adaptive testing, which uses item response theory to select the most appropriate items to measure a construct. Factor analysis is a statistical method used in psychometrics to identify the underlying dimensions of a set of variables. It involves identifying the common factors that contribute to the variation in the observed scores and grouping them into factors that represent the different dimensions of a construct. This sub-field is particularly useful in personality assessment, as it allows for the identification of different personality traits and their relationships to each other. In conclusion, psychometrics is an essential field of study that focuses on the measurement of psychological constructs. It involves the use of various sub-fields, such as classical test theory, item response theory, and factor analysis, to develop and validate tests and measurement tools. Understanding these sub-fields is crucial for anyone interested in the field of psychometrics, as they provide a comprehensive understanding of the different aspects of psychological measurement.

5 Concrete examples

5.1 Classical Test Theory

A psychologist wants to measure the IQ of a group of 10-year-old children in a school with a standardized test. He uses a reliability analysis to confirm that the test is consistent and produces similar results over time. He also uses a validity analysis to check that the test measures intelligence and not something else, such as reading ability.

5.2 Item Response Theory

A company wants to hire a new employee and uses a computerized adaptive test to measure their problem-solving skills. The test uses item response theory to select the most appropriate questions for the individual, based on their previous answers. The test adjusts the difficulty

level of the questions to match the ability of the person being tested, providing a more accurate assessment of their skills.

5.3 Factor Analysis

A researcher is interested in understanding the different dimensions of depression in a group of adults. She uses a factor analysis to identify the common factors that contribute to the variation in the observed scores. Based on the results of the factor analysis, she identifies three different depression dimensions such as "anhedonia," "hopelessness," and "insomnia" and their unique characteristics. These concrete examples demonstrate how psychometricians use statistical methods and theories to develop and validate tests and measurement tools for various purposes such as clinical assessments, personnel selection, and research.

6 Additional Examples

1. A school district is interested in assessing the intelligence of their students to ensure that they are receiving appropriate educational services. They hire a psychometrician to develop an intelligence test that is reliable and valid. The psychometrician uses classical test theory to ensure that the test is consistent and accurate, and item response theory to identify the most appropriate items to measure intelligence.

2. A hospital is looking to hire new nurses and wants to ensure that they are selecting the most qualified candidates. They hire a psychometrician to develop a personnel selection test that is reliable and valid. The psychometrician uses classical test theory to ensure that the test is consistent and accurate, factor analysis to identify the different dimensions of nursing skills and abilities, and item response theory to select the most appropriate items to measure these dimensions.

3. A researcher is interested in studying the relationship between personality traits and job performance. They hire a psychometrician to develop a personality assessment that is reliable and valid. The psychometrician uses classical test theory to ensure that the test is consistent and accurate and factor analysis to identify the different personality traits being measured and their relationships to each other.

7 Introduction to True and Error Scores

Psychometric testing involves the measurement of psychological constructs such as intelligence, personality traits, or attitudes. To effectively measure these constructs, we must understand the concepts of true and error scores.

7.1 True Score

The true score represents the hypothetical perfect measurement of an individual's true ability or trait. It is the score that would be obtained if our measurement instruments were perfectly reliable and free from any errors or biases.

7.2 Error Score

The error score represents the discrepancy between the true score and the observed score. It includes random errors, such as fluctuations in response due to mood or fatigue, as well as systematic errors, such as biases in the measurement instrument or scoring process.

7.3 Example

Let's consider an example of a student, John, taking an intelligence test. If John's true intelligence level is 120 (true score), but due to distractions during the test or ambiguous questions, he scores 115 (observed score), then the error score would be 5 (115 - 120).

7.4 Calculating True Scores

In psychometric testing, we aim to estimate the true scores of individuals based on their observed scores and the reliability of the measurement instrument. One of the classical methods for estimating true scores is through the use of the classical test theory model.

The observed score (X), therefore, is the sum of your true score and error score:

$$X = T + E$$

Important Points:

- Error scores are assumed to be random and independent of the true score.
- Errors can be positive or negative, meaning they can either inflate or deflate your observed score.
- The goal of psychometrics is to minimize the influence of error scores and obtain a more accurate estimate of the true score.

7.5 Understanding Reliability Coefficient

In psychology and other fields, when we talk about reliability, we're essentially asking: "How consistent is a measurement?" The reliability coefficient is a number that helps us understand just that.

Imagine you have a weighing scale. You step on it multiple times, and each time you get a slightly different reading. Now, if the scale is reliable, it means that even though the readings may vary a bit each time, they should still be pretty close to each other.

So, the reliability coefficient is like a score that tells us how much we can trust the consistency of our measurement tool. It's a number between 0 and 1. The closer it is to 1, the more consistent or reliable our measurement is. If it's closer to 0, it means our measurements are not very consistent.

For example, if you take a test and your score on different occasions is always very similar, then that test has a high reliability coefficient. But if your scores vary a lot each time you take the test, then the reliability coefficient would be lower.

In essence, the reliability coefficient helps us know if we can rely on our measurement tool to give us consistent results. The higher the reliability coefficient, the more confident we can be in the accuracy of our measurements.

Example:

Suppose we have a test with 50 items measuring mathematical aptitude. A student scores 40 out of 50 on the test. The reliability coefficient of the test is 0.80. We want to estimate the student's true mathematical aptitude score using the Classical Test Theory model.

Solution:

Given data:

- Observed Score (X): 40
- Reliability Coefficient (r): 0.80

We can use the formula for calculating true scores in the CTT model:

$$\text{True Score} = \frac{\text{Observed Score}}{\sqrt{\text{Reliability Coefficient}}}$$

Substituting the given values:

$$\text{True Score} = \frac{40}{\sqrt{0.80}} \approx \frac{40}{0.894} \approx 44.721$$

Therefore, the student's estimated true mathematical aptitude score is approximately 44.721.

This calculation demonstrates how the Classical Test Theory model can be used to estimate true scores based on observed scores and the reliability coefficient of a test.

8 Methods of Measuring Reliability

Reliability is a crucial aspect of psychometric testing, ensuring consistency and stability in measurements. Several methods are commonly used to assess reliability, including test-retest reliability, inter-rater reliability, and internal consistency.

8.1 Test-Retest Reliability

Test-retest reliability assesses the consistency of a measure over time by administering the same test to the same group of individuals on two separate occasions and correlating their scores. Let's illustrate this with an example:

Example:

Suppose a group of students takes a mathematics test on two occasions separated by two weeks. The correlation coefficient between their scores on the two tests is computed to determine test-retest reliability. Here are the scores:

- First Test Scores: $75, 80, 85, 70, 90$
- Second Test Scores: $72, 82, 83, 68, 88$

Using the Pearson correlation coefficient formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \times \sqrt{\sum(Y_i - \bar{Y})^2}}$$

where X_i and Y_i are the individual scores, and $\bar{X}$ and $\bar{Y}$ are the mean scores.

Substituting the values:

$$r = \frac{(75-80)(72-80) + (80-80)(82-80) + ... + (90-80)(88-80)}{\sqrt{(75-80)^2 + (80-80)^2 + ... + (90-80)^2} \times \sqrt{(72-80)^2 + (82-80)^2 + ... + (88-80)^2}}$$

After computation, suppose we find $r = 0.90$. This indicates a strong positive correlation between the scores on the two occasions, suggesting high test-retest reliability.

8.2 Inter-Rater Reliability

Inter-rater reliability measures the consistency of ratings or judgments made by different raters or observers. Let's demonstrate this with an example:

Example:

Two psychologists independently rate the anxiety levels of the same group of patients using a standardized anxiety scale. The ratings are then correlated to determine inter-rater reliability. Here are their ratings:

- Psychologist 1 Ratings: $3, 4, 2, 5, 3$
- Psychologist 2 Ratings: $4, 3, 2, 4, 3$

Again, using the Pearson correlation coefficient formula, suppose we find $r = 0.85$. This high correlation indicates strong agreement between the two raters, suggesting high inter-rater reliability.

8.3 Internal Consistency

Internal consistency assesses the extent to which items within a test measure the same underlying construct. One commonly used measure of internal consistency is Cronbach's alpha. Let's see an example:

Example:

Suppose we have a 20-item questionnaire designed to measure job satisfaction. After administering the questionnaire to a sample of employees, we calculate Cronbach's alpha to assess internal consistency. Here are the scores:

- Total Score Variance: 200
- Variance of Each Item's Scores: 10 (assuming equal variance for simplicity)
- Number of Items: 20

Using the formula for Cronbach's alpha:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \text{item variance}}{\text{total score variance}}\right)$$

Substituting the values:

$$\alpha = \frac{20}{20-1}\left(1 - \frac{20 \times 10}{200}\right)$$

After computation, suppose we find $\alpha = 0.80$. This indicates good internal consistency, suggesting that the items in the questionnaire measure the same underlying construct of job satisfaction reliably.

These examples illustrate how different methods of measuring reliability are applied and interpreted in practice.

Additional Example:

Consider a questionnaire designed to measure the perceived risk of cholera due to drinking contaminated water. The questionnaire consists of Likert scale questions rated on a scale from 1 to 5, with 1 indicating "Strongly Disagree" and 5 indicating "Strongly Agree." The questionnaire contains 5 items related to the perceived risk of cholera from drinking contaminated water. Participants are asked to rate each statement according to their level of agreement.

The Likert scale questions are as follows:

1. I am concerned about the risk of cholera from drinking contaminated water.
2. I believe that drinking contaminated water increases the likelihood of getting cholera.
3. I take precautions to avoid drinking water that may be contaminated.
4. I am aware of the importance of clean water for preventing cholera.
5. I have heard about cases of cholera in my community.

Suppose we collect responses from 5 participants, and each participant provides a rating for each of the 5 Likert scale questions. We compute the total score for each participant by summing up their responses to all 5 items.

Participant	Q1	Q2	Q3	Q4	Q5
1	4	5	3	4	2
2	3	4	2	5	3
3	5	4	3	4	5
4	2	3	1	4	2
5	4	5	4	3	4

We then calculate the mean score across all participants and the variance of the total scores to assess the internal consistency of the questionnaire using Cronbach's alpha.

For example, let's compute Cronbach's alpha for the sample data:

$$\alpha = \frac{k}{k-1}\left(\frac{S_y^2 - \sum S_y^2}{S_y^2}\right)$$

Now, let's compute Cronbach's alpha using the provided sample data:

$$\text{Variance of Total Scores } (S_y^2) = \frac{(18-17.6)^2 + (17-17.6)^2 + ... + (20-17.6)^2}{4}$$

$$\text{Variance of Total Scores } (S_y^2) \approx 12.3$$

$$\text{Variance of Items } (\sum S_y^2) = 1.3 + 0.7 + 1.3 + 0.5 + 1.7$$

$$\text{Variance of Items}(\sum S_y^2) = 5.5$$

$$\text{Cronbach's Alpha } (\alpha) = \frac{5}{5-1}\left(\frac{12.3 - 5.5}{12.3}\right)$$

$$\text{Cronbach's Alpha } (\alpha) \approx 0.69$$

The computed Cronbach's alpha for the sample data is approximately 0.69, indicating good internal consistency of the questionnaire.

8.4 Kuder-Richardson Formula 20 (KR-20)

Kuder-Richardson Formula 20 (KR-20) is a measure of internal consistency reliability, specifically designed for binary (dichotomous) items. It is commonly used in educational and psychological assessments where test items have two response options (e.g., true/false, yes/no).

The formula for computing KR-20 is as follows:

$$KR-20 = \frac{N}{N-1}\left(1 - \frac{\sum p_i(1-p_i)}{\sigma_X^2}\right)$$

where:

$$\begin{aligned} KR-20 &= \text{Kuder-Richardson Formula 20 coefficient} \\ N &= \text{number of test items} \\ p_i &= \text{proportion of examinees who answered item } i \text{ correctly} \\ \sigma_X^2 &= \text{variance of the total scores} \end{aligned}$$

In this formula, $p_i(1-p_i)$ represents the variance of each item, and σ_X^2 represents the variance of the total scores. The numerator N is the number of test items, and the denominator $(N-1)$ is used as a correction factor.

KR-20 ranges from 0 to 1, where a value closer to 1 indicates higher internal consistency reliability. Typically, a value above 0.7 is considered acceptable for research purposes.

Worked Example

Suppose we have a 10-item test assessing students' understanding of mathematical concepts, where each item is scored as either correct (1) or incorrect (0). The proportions of students who answered each item correctly are as follows:

Participant	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	1	1	1	1	1	0	1	1	1
2	1	1	1	1	1	1	0	0	1	0
3	1	1	1	0	1	0	0	0	1	1
4	0	0	1	0	1	0	0	0	1	0
5	0	0	0	0	1	0	1	1	0	1

We can compute the KR-20 coefficient for this test using the provided formula.

$$\begin{aligned} KR-20 &= \frac{10}{10-1}\left(1 - \frac{(0.6)(0.4) + (0.60)(0.40) + \ldots + (0.60)(0.40)}{\sigma_X^2}\right) \\ &= \frac{10}{9}\left(1 - \frac{0.24 + 0.24 + \ldots + 0.24}{6.8}\right) \\ &= \frac{10}{9}\left(1 - \frac{1.84}{6.8}\right) \\ &= \frac{10}{9}\left(1 - 0.27\right) \\ &= \frac{10}{9} \times (0.73) \\ &\approx 0.81 \end{aligned}$$

Thus, the computed KR-20 coefficient for this test is approximately 0.81.

8.5 Split-Half Reliability

Split-Half Reliability assesses the consistency of a test by dividing it into two halves and comparing the scores obtained on each half. There are various ways to split a test, such as splitting odd-even items or randomly dividing the items.

8.5.1 Explanation:

Suppose we have a questionnaire with 10 items measuring job satisfaction. We want to calculate the split-half reliability of this questionnaire using Spearman's rank correlation coefficient.

Step 1: Splitting the Questionnaire

We randomly split the questionnaire into two halves. Let's denote the first half as Form A and the second half as Form B.

Form A: Items 1, 3, 5, 7, 9
Form B: Items 2, 4, 6, 8, 10

Step 2: Scoring

We score each respondent's answers to the items in Form A and Form B separately. Let's assume we have scores for 10 respondents as follows:

$$\begin{aligned} &\text{Form A scores:} && 20, 22, 18, 21, 19, 23, 20, 24, 21, 22 \\ &\text{Form B scores:} && 18, 20, 17, 19, 18, 21, 19, 22, 20, 20 \end{aligned}$$

Step 3: Calculating Spearman's Rank Correlation

We calculate Spearman's rank correlation coefficient (ρ) between the scores of Form A and Form B.

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where d is the difference in ranks between corresponding pairs of scores, and n is the number of pairs.

For the half correlation:

$$\begin{aligned} n &= 10 \\ \rho &= 1 - \frac{6 \times 324}{10(10^2 - 1)} \\ &= 1 - \frac{1944}{990} \\ &= 1 - 1.964 \\ &= -0.964 \end{aligned}$$

Participant	Form A	Form B	Rank A	Rank B	Difference	D^2
1	20	18	7	3	4	16
2	22	20	3	7	-4	16
3	18	17	10	1	9	81
4	21	19	5	5	0	0
5	19	18	9	2	7	49
6	23	21	2	9	-7	49
7	20	19	7	3	4	16
8	24	22	1	10	-9	81
9	21	20	5	5	0	0
10	22	20	3	7	-4	16

The Spearman's rank correlation coefficient ($\rho = -0.964$) indicates a Strong negative correlation between the scores of the two halves of the questionnaire. Therefore, we can conclude that the questionnaire demonstrates strong split-half reliability.

The Spearman-Brown Formula is expressed as:

$$\rho_{Total} = \frac{2\rho_{half}}{1 + \rho_{half}(n-1)}$$

$$\rho_{Total} = \frac{2(-0.964)}{1 + (-0.964)(10-1)}$$

$$\rho_{Total} = 0.25$$

Example 2:

Consider a psychological assessment consisting of 30 items measuring anxiety. The test is split into two halves by randomly assigning half of the items to Form A and the other half to Form B. Participants complete both forms, and the scores on each form are correlated to determine split-half reliability.

8.6 Spearman-Brown Formula

The Spearman-Brown Formula estimates the reliability of a test after it has been lengthened or shortened. It predicts the reliability of a test based on the correlation between two halves of the test and the original test length or the new length.

8.6.1 Explanation:

The Spearman-Brown Formula is expressed as:

$$r_{new} = \frac{2r_{old}}{1 + r_{old}(n-1)}$$

where r_{old} is the correlation coefficient between the two halves of the original test, n is the original test length, and r_{new} is the estimated reliability for the new test length.

Example 1:

Suppose a 20-item vocabulary test has a split-half reliability coefficient of 0.80. An educational psychologist wants to estimate the reliability if the test is expanded to 40 items. Using the Spearman-Brown Formula:

$$r_{new} = \frac{2 \times 0.80}{1 + 0.80 \times (20-1)} = \frac{1.60}{1 + 0.80 \times 19} = \frac{1.60}{1 + 15.20} = \frac{1.60}{16.20} \approx 0.099$$

Example 2:

Consider a personality assessment with 50 items and a split-half reliability coefficient of 0.75. If the test is shortened to 25 items, we can use the Spearman-Brown Formula to estimate the new reliability:

$$r_{new} = \frac{2 \times 0.75}{1 + 0.75 \times (50 - 1)} = \frac{1.50}{1 + 0.75 \times 49} = \frac{1.50}{37.25} \approx 0.040$$

8.7 Parallel Form Reliability

Parallel form reliability, also known as equivalent form reliability, assesses the consistency of different versions of a test designed to measure the same construct. Let's illustrate this with an example:

Example:

Suppose a group of students takes two versions of a mathematics test, Form A and Form B, intended to measure the same mathematical knowledge. The scores on both forms are then correlated to determine parallel form reliability. Here are the scores:

- Form A Scores: $80, 85, 90, 75, 85$
- Form B Scores: $82, 87, 88, 73, 84$

Using the Pearson correlation coefficient formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \times \sqrt{\sum(Y_i - \bar{Y})^2}}$$

where X_i and Y_i are the individual scores, and $\bar{X}$ and $\bar{Y}$ are the mean scores.

Substituting the values:

$$r = \frac{(80 - 85)(82 - 85) + (85 - 85)(87 - 85) + \ldots + (85 - 85)(84 - 85)}{\sqrt{(80 - 85)^2 + \ldots + (85 - 85)^2} \times \sqrt{(82 - 85)^2 + \ldots + (84 - 85)^2}}$$

After computation, suppose we find $r = 0.97$. This high correlation indicates strong agreement between the scores on the two test forms, suggesting high parallel form reliability.