In this module we will cover test reliability, standard error of measurement, and confidence bands.
You will learn …

• Test reliability
• Standard error of measurement
• Confidence bands

The reliability of a test refers to the stability or consistency of test scores and is an important aspect of test quality.

The standard error of measurement of a test provides an estimate of the amount of error in a test.

Confidence bands are based on the standard error of measurement and provide additional information of what the student's true level of achievement probably is.
There are several definitions used to describe the reliability of a test: The degree to which a test yields a similar score for the same person upon retesting. Or, the stability of test scores over repeated administrations. Or, the consistency with which a test yields the same score for a person taking the test several times.
There are several ways to estimate the reliability of a test: test-retest, alternate forms, and internal consistency.
The test-retest method for estimating reliability is exactly what its name implies. The process for computing the test-retest reliability is: Give the test for the first time; then wait a predetermined time period such as 2 weeks or 4 weeks. Give the exact same test for the second time and finally, calculate the correlation between two test scores. The correlation coefficient is the test-retest reliability.
You should know about several issues regarding test-retest reliability.

There should be no instruction between two test administrations.

Memory or experience may impact the test-retest reliability. Longer time interval between two test administrations would reduce the impact of memory.

The test should also be administered under the same or equivalent conditions. For example, if the first administration of the test is given after lunch then the second one should also be given after lunch.
Alternate forms reliability is similar to test-retest reliability. To compute alternative forms reliability, two equivalent forms should be created and administered at two different times, then the correlation between these two sets of scores is calculated. Because two different forms are used, the time interval between two administrations can be shorter.
Alternate (Parallel) Forms method

Some issues for test-retest reliability

- Eliminates memory issue
- Very difficult to obtain equivalent forms
- Tests need to be administered under same conditions

Alternate forms reliability can eliminate the memory issue that occurs in the test-retest reliability. However, it is quite difficult to obtain equivalent forms. Two equivalent forms need to be administered under the same conditions to reduce the errors associated with the reliability.
Internal consistency reliability measures the degree to which items on a form have something in common. In other words, it measures whether the test is comprised of a single, fairly consistent concept.

There are multiple types of internal consistency reliability, such as split half, Kuder-Richardson formula 20, and Cronbach’s Alpha.
To compute split-half reliability, the tests are administered for a group of examinees.

After administration, the test items are randomly divided into two halves and the total score for each half is computed. Then the correlation between the two total scores is computed.

Since a longer test is more reliable, the split-half approach underestimates reliability for the full test. Therefore, the Spearman-Brown formula is used to adjust split-half reliability.

Since a single test is used to make two shorter alternative forms for split-half reliability, only one test administration is needed and no memory or practice can confound the results.
Kuder-Richardson formula 20 (KR20)

- Most accurate estimate
- Difficult to calculate
- Equal to the average of all possible split-half coefficients
- Only for binary data (right or wrong)

Kuder-Richardson formula 20, also called KR20, is the most accurate estimate of internal consistency reliability.

KR20 is difficult to compute because it requires the percentages of students passing each item on the test.

The resulting KR20 coefficient equals to the average of all possible split-half coefficients.

The disadvantage of KR20 is that it can only be used for binary data.
Cronbach’s Alpha is closely related to the Kuder-Richardson formula. It has a complex computation procedure. Cronbach’s alpha has the advantage of being applicable to the test with partial credits or multiple-level responses. For instance, test items may ask the subjects to rate their responses between one and five.
Internal consistency reliability are only appropriate for tests that consist of items measuring a single, fairly consistent concept. For example, a mathematics test is intended to measure only the mathematical ability.

However, some tests have rather diverse content and therefore the internal consistency will yield inflated estimates of reliability. For example, speededness can influence the test score.

If there are time constraints imposed during test administration, the test becomes diverse, measuring multiple concepts (such as the mathematical ability and speededness).

Unfortunately, most achievement tests are at least a little speeded.
Inter-rater reliability

Used for essays and other performance-based tasks.

Process
1. Give the test
2. Have judge 1 rate answers
3. Have judge 2 rate answers
4. Correlate judges’ ratings

Inter-rater reliability is an estimate used for essays and other performance-based tests.

To compute the inter-rater reliability, administer the test, and then have two judges independently rate the student responses.

The correlation coefficient between the two judges' ratings is the inter-rater reliability.
There are some principals related to the interpretation of reliability coefficients. They may be useful as guidelines in evaluating the reliability of scores from a test.

First, group variability affects test reliability. As the group variability increases, the test reliability increases. That is, higher coefficient results from heterogeneous groups than from homogeneous groups.

Second, scoring reliability affects test reliability. As scoring reliability decreases, the test reliability decreases. That is, if tests are scored inconsistently, error is introduced that will decrease the test reliability.

Third, test length affects test reliability. Under the situation of other factors being equal, as test length increases, the reliability increases.

Fourth, test difficulty affects test reliability. As test difficulty becomes too easy or too difficult, the reliability decreases.
All test scores are inaccurate because they contain some amount of error.

An obtained test score is comprised of both truth and error.

You can simply keep in mind the following equation: “Obtained score” equals “true score” plus/minus “error score.”
If we knew examinees’ true scores, we would expect to see:

– Obtained scores that are usually different from true scores, and

– Obtained scores that are sometimes higher, and sometimes lower than true scores due to error score.
True scores and error scores are always unknown but they can be estimated through the standard error of measurement (or SEM). Actually, the SEM is a standard deviation of test error scores.

Regarding error scores of the test, we always assume that they are “normally distributed” with a mean of “zero” and a standard deviation called the Standard Error of Measurement.

Given what we know about normal distributions, we can estimate the ways Error scores are impacting True scores on any given test.
There are several major sources of error scores. These sources include: within the test takers, within the test, in the test administration, and in the test scoring.

Test takers may have error based on their health, mental concentration, and the guessing factor.

Tests may provide error from poorly written items or items that are not on the reading level of the students.

The test administration may include disruptions, issues with the instructions, attitudes, and time limits.

Test scoring such as scoring keys, answer sheet marks, and hand scoring issues can be the source of error.
As mentioned earlier, the **standard error of measurement** provides an estimate of the amount of error in a test and is a standard deviation of test error scores. But we never know the error scores. How can we obtain SEM? Fortunately, SEM can be derived from the standard deviation of obtained scores and the test reliability.

The formula of SEM is expressed as follows:

SEM equals standard deviation times square root of (1 minus test reliability).

Based on the SEM formula, we can see that the SEM and test reliability are inversely related. That is, the *more* reliable the test is, the *smaller* standard error of measurement the test has. In contrast, the *less* reliable the test is, the *larger* standard error of measurement the test has.
How can we use the SEM to estimate an examinee’s true score?

When we have an estimate of the SEM, we can construct “confidence bands” around an examinee’s obtained score. The confidence band allows us to consider a range of scores where the examinee’s true score probably lies rather than using just a single value (the obtained test score), that always contains error.

Confidence bands are often used in score reports. The information from confidence bands is very useful because it provides a more accurate picture of what the student’s true level of achievement probably is.
Here is a practical example of confidence bands.

If a student’s obtained score is 40 and the standard error of measure is 3, the following conclusions can be drawn about the student’s true score.
We can be 68% (or 68.26%) sure that the true score of the student is between 37 and 43 which is within plus/minus 1 SEM of the obtained score.
Based on the normal distribution, 68% of the curve is between plus/minus one standard error of measurement (or SEM). Therefore, the true score of 68% confidence band is 40 plus/minus 1 times 3. which is between 37 and 43.
We can be approximately 95% sure that the true score is between 34 and 46, which is within plus/minus two SEM of the obtained score. As you can see in the normal distribution curve, approximately 95% of the curve is between plus/minus two standard error of measurement (or SEM). Therefore, the true score of 95% confidence band is 40 plus/minus 2 times 3, which is between 34 and 46.
We can be approximately 99% sure that the true score is between 31 and 49, which is within plus/minus three SEM of the obtained score. As you can see in the normal distribution curve, approximately 99% of the curve is between plus/minus three SEM. Therefore, the true score of 99% confidence band is 40 plus/minus 3 times 3, which is between 31 and 49.
Because of the inverse relationship between reliability and SEM, a more reliable test will have smaller bands.

Here is an example to illustrate this statement.

Let us compute 68% ranges on two tests with differing SEM. Both tests have an obtained score of 127. Test 1 has SEM of 2 and Test 2 has SEM of 10.

The 68% confidence band for test 1 is 127 plus and minus 1 times 2, which is between 125 and 129.

The 68% confidence band for test 2 is 127 plus and minus 1 times 10, which is between 117 and 137.
Let us do some practice about SEM and confidence bands.

First, calculate the SEM for a test with a standard deviation of 10 and a reliability of .84 and draw a normal curve showing the error score distribution in SEM units.

Recall the formula for SEM. Standard deviation of measurement equals standard deviation times square root of 1 minus reliability.

In this case, SEM equals 10 times square root of (1 minus .84). The SEM for the test with standard deviation of 10 and reliability of .84 is 4.
Second, use the SEM we computed previously to calculate the range within which we could be 68%, 95%, and 99% sure Annie’s true score lies. Annie received a score of 16 on this test.

Recall the error scores distribution that 68%, 95%, and 99% sure represent plus/minus one SEM, plus/minus two SEM, and plus/minus three SEM, respectively.

Therefore, we can be 68% sure Annie’s true score lies the obtained score plus/minus 1SEM, which is 16 plus/minus 1 times 2. The range of 68% sure for Annie’s true score is between 14 and 18.

95% band for Annie’s true score lies the obtained score plus/minus 2 SEM, which is 16 plus/minus 2 times 2. The 95% band for Annie’s true score is between 12 and 20.

99% band for Annie’s true score lies the obtained score plus/minus 3 SEM, which is 16 plus/minus 3 times 2. The 99% band for Annie’s true score is between 10 and 22.
Third, draw a normal curve showing the error score distribution in SEM units for the test Annie took.

The graph shows the normal distribution of error scores. For the test Annie took, $X$ represents 16. Since SEM equals 2, $+1 \text{ SEM}$ equals 18, $+2 \text{ SEM}$ equals 20, and $+3 \text{ SEM}$ equals 22. Negative 1 SEM equals 14, negative 2 SEM equals 12, and negative 3 SEM equals 10.
Two terms we’ve been considering recently sound very similar, but actually refer to distinctly different things.

These terms are standard deviation and standard error of measurement.

A standard deviation tells us about the variability, or spread in a distribution of test scores. It tells us how similar or different we might expect examinees’ scores to be to each other.

The standard error of measurement, on the other hand, is related to the accuracy of any obtained test score. We use it to come up with a range around a specific test score within which we have some confidence that the examinee’s true score probably lies. The standard error of measurement gives us a ‘confidence band’ range around any specific test score.

Keep these definitions in mind, and you’ll be better able to interpret test score data!
In this example, Gillian earns a score of 350 on a Scientific Reasoning test. The test has a mean of 300 and a standard deviation of 50.

The internal consistency reliability estimate of this test is .91. Thus the SEM is 15.
Interpreting example of SD and SEM

- Using the SD – her score is 1 SD above the mean (a z-score of 1); this is equivalent to a percentile rank score of 84%.
- Using the SEM – we can be 68% confident that her true score is between 335-365.
- Using the SEM – we can be 95% confident that her true score is between 320-380.

Now we will interpret Gillian’s score using the SD and SEM.

Using the standard deviation, her score is 1 standard deviation above the mean (which is a z-score of 1); this is equivalent to a percentile rank score of 84%.

Using the standard error of measurement, we can be 68% confident that her true score is between (350 plus/minus 1 times 15), which is between 335 and 365.

Furthermore, using the SEM, we can also be 95% confident that her true score is between 350 plus/minus 2 times 15, which is between 320 and 380.
The accuracy of an obtained test score can be represented on a student’s score report through the graphical use of the SEM.

• The score report can include a graphical bar showing both the student’s obtained score and “confidence bands” based on the SEM.
Using the score report table answer the following questions:

What is the student’s obtained score in Reading (the “point estimate”)?

What is the range of scores within which we are 68% sure that his true score lies?

What is the range of scores within which we are 95% sure that his true score lies?

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Score</th>
<th>68% range (X ± 1SD)</th>
<th>95% range (X ± 2SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>103</td>
<td>100-106</td>
<td>97-109</td>
</tr>
<tr>
<td>Listening</td>
<td>104</td>
<td>101-107</td>
<td>98-110</td>
</tr>
<tr>
<td>Writing</td>
<td>105</td>
<td>102-108</td>
<td>99-111</td>
</tr>
<tr>
<td>Social Studies</td>
<td>98</td>
<td>95-101</td>
<td>92-104</td>
</tr>
<tr>
<td>Science</td>
<td>100</td>
<td>97-103</td>
<td>91-106</td>
</tr>
<tr>
<td>Math</td>
<td>91</td>
<td>88-94</td>
<td>85-97</td>
</tr>
</tbody>
</table>

* M=100, SD=10
Example with single test score

- What is the student’s obtained score in Reading (the “point estimate”)?
- What is the range of scores within which we are 68% sure that his true score lies?
- What is the range of scores within which we are 95% sure that his true score lies?

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Score</th>
<th>68% range (X ± 1SD)</th>
<th>95% range (X ± 2SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>103</td>
<td>100-106</td>
<td>97-109</td>
</tr>
<tr>
<td>Listening</td>
<td>104</td>
<td>101-107</td>
<td>98-110</td>
</tr>
<tr>
<td>Writing</td>
<td>105</td>
<td>102-108</td>
<td>99-111</td>
</tr>
<tr>
<td>Social Studies</td>
<td>98</td>
<td>95-101</td>
<td>92-104</td>
</tr>
<tr>
<td>Science</td>
<td>100</td>
<td>97-103</td>
<td>91-106</td>
</tr>
<tr>
<td>Math</td>
<td>91</td>
<td>88-94</td>
<td>85-97</td>
</tr>
</tbody>
</table>

*M=100, SD=10

The student's point estimate is his actual obtained score. In this example, the student earned a score of 103 on the reading test and we can be 68% sure that his true score lies between 100 and 106. And we can also be 95% sure that his true score lies between 97 and 109.
Further, a battery of scores for a student can be compared using confidence bands to see if “real” differences exist in the student’s pattern of strengths and weaknesses or if the observed differences in test scores are just due to chance.
Let’s compare confidence bands around *multiple* test scores.
Using the 68% confidence bands, does there appear to be a real difference between the student’s performance in Science and Math?
To assess whether there are real differences in the scores, we will examine the confidence bands to see if there is overlap between the bands of interest.

Since there is no overlap between the bands of math and science, we can be 68% sure that there is a real performance difference in science and math for that student. That is, the student performed better on science than on math.
Now let's compare the confidence bands at the 95% accuracy level.

Using the 95% confidence bands, does there appear to be a real difference between the student's achievement in **Science** and **Math**?
Example at 95% accuracy level

- 95% sure that there was no real difference between the student’s achievement in science and math (i.e., there was overlap)
- 95% sure that there was real difference between the student’s achievement in writing and math (i.e., no overlap)

When using the 95% accurate level, there was not a real difference in the student’s achievement in science and math because there was overlap between the bands of science and math. At the 95% level, the student’s achievement difference in science and math can be attributed to chance.

As for the difference between writing and math, we can be 95% sure that there was real difference between writing and math for that student because there was no overlap between the bands of writing and math at the 95% level.
Use of “band interpretation” helps us avoid putting too much trust:
- in specific test score values, and
- in small differences in test scores from a battery of tests

Use of “band interpretation” helps us avoid putting too much trust “in specific test score values” and “in small differences in test scores” from a battery of tests so that our interpretations of test scores are appropriate and reasonable.
Confidence bands guidelines

- As a general guideline, use the:
  - 95% level as a signal to the school and parents
  - 68% level as a signal to you

When should we use confidence band at a 95% level or a 68% level? Here is a general guideline.

If we are going to make important decisions about a student or would like to find the real difference in two subjects for that student, the conservative 95% approach should be taken. The differences at the 95% level can be a signal to the school and to the parents.

On the other hand, if little is at risk about a wrong decision for the student, then less conservative 68% level can be used. The differences at the 68% level can be a signal to you.
As you can see in these two graphs, it is easy to find the differences between any pairs of two subjects using confidence bands at the 68% level for a particular student. The confidence band at the 68% level is less conservative.

In contrast, it is much harder to find the differences between any two subjects using confidence bands at the 95% level for a particular student. The confidence band at the 95% level is called a conservative approach.