**Presenter's Notes for PowerPoint:**
**Reliability**

| Slide #<br>Title | Presenter Notes |
|---|---|
| 1. Reliability for Teachers | Welcome to Module 2 Understanding Reliability for Teachers! This module provides teachers with basic information they need to understand what reliability is and why it is so important. Both reliability and validity are necessary for a teacher to make accurate inferences about student learning based on assessment results. For this reason, you need to study both this module and the Validity module to understand quality assessment of student learning. |
| 2. Essential Questions | This is module answers the essential questions:<br>• What is test reliability?<br>• What are the three types of reliability?<br>• What are the some of the issues related to reliability?<br>• How can teachers ensure that tests are reliable? |
| 3. Reliability is Essential | Reliability has mainly been a concern of developers of standardized and large-scale assessments. Claiming that a test is highly reliable helps publishers to sell tests and make money. It also allows state assessment contractors to justify the quality of their work. For researchers, a highly reliable test is a prerequisite to investigating the validity of the test, because test scores can be reliable but not valid, yet cannot be valid unless they are reliable. |
| 4. Reliability Represents Consistency | No notes for this slide. |
| 5. Tests – Reliable / Results - Valid | Because reliability and validity are interrelated it is important to understand the distinction between the two.<br><br>Validity refers to what inferences can be made about the test's results. But a test cannot be valid or invalid. It is the results of a test that are valid or invalid. So if you are talking about validity related to a test that is intended to measure students' computational abilities you would say "The results of this test are a valid representation of student's computational abilities."<br><br>But reliability does refer to the test. A test is or is not reliable.<br><br>So a test that isn't reliable cannot provide valid results. However, just because a test is reliable doesn't mean that test is valid. A test needs to be both reliable and valid to be useful. For example, you can't use the results from a reliable test that measures student's reading comprehension as a valid representation of student's computational abilities. |
| 6. What is Reliability? | Consistency, of course, implies some sort of comparison between at least two measurements. But before we look at multiple test scores, we need to understand what a single test score actually represents.<br><br>It would be easy to assume that if a test is simply consistent that it is reliable. However, educators use three different ways of determining reliability. A test that is reliable in one way may not be reliable in another way. Popham tells us that "An assessment literate teacher needs to understand it is a particular kind of reliability evidence that indicates whether a given test is consistent regarding the specific purpose for which the test needs to be consistent." |
| 7. Three Varieties of Reliability | The three varieties of reliability evidence are stability, alternative form, and internal consistency:<br>• Stability is the consistency of results between two time-separated testing occasions.<br>• Alternate form is the consistency of results between two different forms of a test. |

| | |
|---|---|
| | • Internal consistency is the consistency in the way a test's items function.<br>The three kinds of reliability are not interchangeable. But all three rely on statistical analyses that are correlational. |
| 8. Reliability Depends on Correlational Analysis | Before we can understand how to determine the three types of reliability we need to understand correlational analyses. These are correlation-based or score consistency and classification consistency. Test reliability depends upon score consistency and classification consistency. |
| 9. Correlation Coefficients | We will start with correlation or score consistency reliability. There are a variety of ways to compute and interpret correlation-based reliability but what is important to know is that correlational procedures take two sets of scores from the same group of test-takers. Then those scores are analyzed to see how closely they are related. The result of this analysis is called the correlation coefficient. The correlation coefficient is also sometimes referred to as the Pearson correlation coefficient after its originator Karl Pearson. The correlation coefficient is usually represented by the small letter r. This can represent a positive relationship, no relationship, or a negative relationship.<br>The closer to 1.0 the correlation coefficient is, the closer the relationship between the results of two sets of scores or the greater the score consistency. A correlation coefficient of zero signifies no relationship at all. If a correlation coefficient is below zero, the two scores show a negative relationship.<br>Here you see a Table from Mastering Assessment Booklet, Reliability: What Is It and Is It Necessary? that illustrates simple interpretations of correlation coefficients. |
| 10. No Pre-Determined Correlation Coefficient | It's important to understand that there is no predetermined reliability coefficient that tests must attain in order to show consistency of a test's scores. Test users have to decide if a test is reliable enough based up the score-consistency evidence available.<br>Score consistency reliability is an estimate of a test's consistency derived from the test taker's scores. Teachers can decide how much they trust a national, state, or district test based upon the score consistency correlations that come with those tests. |
| 11. Activity One | No notes for this slide. |
| 12. Classification Consistency Reliability | Classification consistency reliability is a representation of the proportion of students who are classified identically on two different test forms or two different administrations of the same test. Classification consistency has become more and more important to educators. This is due to recent requirements that schools' achieve a level of proficiency on tests to avoid consequences such as being placed "on improvement" because not enough students earn a level of proficient or above on state tests based upon pre-determined cut scores. |
| 13. Example of Classification Consistency (Good Reliability) | No notes for this slide. |
| 14. Example of Classification Consistency (Poor Reliability) | No notes for this slide. |
| 15. Issues Related to Classification Consistency | We need to be aware of two issues related to classification consistency. The first is inter-rater reliability. Inter-rater agreement is the degree of agreement in the ratings that two or more observers assign to the same behavior or performance. The more classifications there are the less chance there is of inter-rater agreement. |
| 16. Example form | In general, the key to improving inter-rater reliability is to have both a clearly defined |

| | |
|---|---|
| KSDE 6-TRAIT SCORING MANUAL | rubric and a set of student responses that illustrate the various characteristics of each score point of the rubric. Remember the anchor papers for the state's writing assessment? Here we see a pre-scored writing sample from the Kansas State Department of Education's 6-TRAIT SCORING MANUAL. |
| 17. Issues Related to Classification Consistency | Two types of errors are likely to occur when cut scores on tests are used to classify students. The first error is setting cut scores too high. The second error is setting cut scores too low. These errors of classification do not occur because someone made a mistake. The errors occur because no test can be perfectly reliable and because no method of setting cut scores is perfect.<br><br>The errors happen when cut scores are used to determine who will pass and who will fail. If the cut scores are set too high, students who really deserve to pass will fail. If the cut scores are set too low, students who really deserve to fail will pass. Moving the cut score up or down to reduce one type of error will necessarily increase the chances of making the other type of error. For example, it is possible to reduce the number of students who pass, but who really deserve to fail, by raising the cut score. The cost of doing so, however, is to increase the number of students who fail but who really deserve to pass. Good test development and good practices for setting cut scores can reduce the number of errors of classification, but no way exists to reduce the errors to zero.<br><br>The people involved in setting cut scores should consider both types of errors in making their judgments and decide which type of error they consider more harmful. The cut scores should reduce the more harmful type of error. |
| 18. For tests. . . | For tests that are used with cut scores, it is important to get answers to these questions. |
| 19. The reliability of classification is not perfect. | Students with similar scores on a test tend to be similar in what they know about the subject tested. Most tests cannot distinguish well between students with scores that are very close to one another. Whenever a cut score is used, however, students with scores just above the cut score and students with scores just below the cut score will be classified differently. What this means is that students who score near the cut score may pass or fail a test because of random fluctuations. |
| 20. Cut Scores & Classification Consistency | Even scorers who have been well trained will disagree occasionally about papers that are near the borderlines of score differences. For example, a response that one scorer gives a high 2, another scorer may legitimately see as a low 3. This discrepancy becomes a problem if the cut score requires a minimum of 3 to be classified as proficient. The discrepancy would not affect the reliability of classification, however, if scores of both 2 and 3 were considered basic. |
| 21. Reliable Tests = Classification Consistency | The more reliable a test is, however, the less likely it is that the scores will be affected by large random fluctuations. All other things being equal, longer tests will be more reliable than shorter tests; and objectively scored tests will be more reliable than subjectively scored tests. |
| 22. Activity Two | There are no notes with this slide. |
| 23. Stability Reliability | Stability reliability represents the consistency of a test's results when the test has been administered on two time-separated occasions. |
| 24. Determining Stability Reliability | Remember, to determine score consistency the most common way to compute reliability is to calculate a correlation coefficient between the two sets of students' scores. Large scale commercial and state assessments often have stability coefficients in the range of .80 to .90. For teacher–made tests the stability coefficient is usually smaller within the range of .60 to .70.<br><br>Most large-scale educational tests usually do stability reliability studies by using small samples of students randomly chosen to participate in a test-retest study. A short time (usually one or two weeks) later the same students are asked to take the same test once again. However, stability reliability isn't often collected because |

| | |
|---|---|
| | teachers don’t have time to do test-retest reliability studies and commercial tests developers don’t retest because it’s expensive. They would rather use the money-saving one test internal reliability analyses. |
| 25. Standard Error of Measurement | Standard error of measurement or SEM is an estimate of the consistency of a student’s score if the student had retaken the same test over and over again. Teachers often center their classroom instructional decisions on students as individuals, not on students as a group. When a student scores 75 on a 100 point test, we want to be confident that if the test is given again to the student, the score would be close to 75. The more consistently a test measures a student’s performance, the more confidence a teacher has about the student’s score – and what can be inferred from that score. To determine measurement consistency for one student at a time it is necessary to use the standard error of measurement. |
| 26. Standard Deviation | If we don’t know what standard deviation is, we can’t understand where we come up with the standard error of measurement. This is the short definition of standard deviation: standard deviation of test scores is a statistical indicator of how spread out a set of test scores is. This is also called test score distribution. If a group of test scores are close together and there isn’t much difference between them, the standard deviation will be small. If the test scores are spread out the standard deviation will be large. |
| 27. Formula for Computing Standard Error of Measurement | We need to know about standard deviation to understand standard error of measurement. If the standard error of measurement for a test is small this is a good thing. And the standard error of measurement is smaller when the standard deviation of the test scores is small and the reliability coefficient is large. Remember that tests publishers will provide you with the reliability coefficient. Here you can see the simple formula for Standard Error of Measurement.<br><br>The standard error of measurement is an important error estimation strategy that teachers use. Because educational tests are imprecise, a standard error of measurement provides insight into just how imprecise a test is in an understandable way.<br><br>One of the best things about standard error of measurement is that it reminds us that tests are NOT perfect! Even great teachers make mistakes in grading and so do educational test makers. But an assessment literate teacher will have an understanding of what acceptable and unacceptable levels of imprecision are and how reliability can be interpreted to find those levels for any test. |
| 28. Alternative-Form Reliability | Alternate-form reliability is the consistency of test results between two different – but equivalent – forms of a test. Alternate-form reliability is used when it is necessary to have two forms of the same tests. This may be necessary for test security when there are opportunities to retake a test by taking a different form of the exam. It would not be useful for one form of an exam to be more difficult or easier than the other if both exams are supposed to measure the same thing. |
| 29. Determining Alternative-Form Reliability | To determine alternate form reliability two forms of the same test are administered to students and students’ scores are correlated on the two test forms. The resulting coefficient is called the alternate-form coefficient of reliability. Alternative-form reliability is needed whenever two test forms are being used to measure the same thing. Ideally, the administration of the two forms should be done in a short time span. |
| 30. Internal Consistency Reliability | Internal reliability deals with the test items. This is different from stability and alternative form reliability which deal with the way tests takers perform. Internal consistency represents the degree to which items in a test are functioning in a similar way. An internal consistency estimate of reliability is computed using only a single administration of the test. And that is one of the reasons internal consistency is used more often.<br><br>So internal reliability demonstrates that the test items are functioning consistently. |

| | |
|---|---|
| | For example that what is intended to be measured is being measured consistently and isn’t impacted by what is not being measured. For example a math test’s results are not being impacted by reading comprehension. |
| 31. Formulae for Computing Internal Consistency (Terminology) | There are a number of different formulae used to compute a test’s internal consistency. These include:<br>• The Kuder-Richardson or K-R formula used for right/wrong answers such as multiple choice test items.<br>• The Cronbach Coefficient Alpha used for items in which students are given points such as essay questions.<br>• The Dichotomous formula that is used for right/wrong answers.<br>• And the Ploytomous formula used for test items that have multiple answers. |
| 32. The Most Common . . . | The most common is the Kuder-Richardson or K-R formula. You don’t need to know these formulae. What you need to know is that a K-R value of .95, like the correlation coefficient, shows a strong positive consistency. So that the closer the K-R value is to 1.00 the higher the internal reliability. |
| 33. Activity Three | There are no notes for this slide. |
| 34. Review of Reliability | There are no notes for this slide. |
| 35. Review of Reliability | There are no notes for this slide. |
| 36. Review of Reliability | There are no notes for this slide. |
| 37. Improving Classroom Tests’ . . . | There are no notes for this slide. |
| 38. Improving Classroom Tests’ . . . | There are no notes for this slide. |
| 39. Improving Classroom Tests’ . . . | There are no notes for this slide. |
| 40. Activity Four | There are no notes for this slide. |