Module 4B: Reliability and Validity

by Dr. Tim Brophy

Hello. I'm Tim Brophy, and welcome back to Passport to Great Teaching, Creative Assessment. This is module 4b, when we're going to dig a little deeper now into reliability and validity. But first, we're going to review these terms.

So let's review validity. Remember, validity is the extent to which the inferences from the results of an assessment match the assessment's intended purpose. Because you give an assessment or a test or an exam for a reason, and you're going to use those scores in a way that is going to help to inform that purpose and to give you the information that you need.

Now there are some kinds of validity evidence that you can gather to help you know and strengthen the validity of your work. The first is called content validity and this is the most basic question we ask on any test that we develop. Do the test items sufficiently cover the material that we're testing?

There's also something called predictive validity or criterion or empirical validity. This has to do with the extent to which whatever it is you're testing or the measure you're using predicts the outcome of another measure when the criterion is not in the future it becomes concurrent validity. So if you were looking at two tests that purportedly test the same thing-- say, part test version 1, test version 2-- and you want them to test the same thing, concurrent validity would have to do with the degree to which both of those tests are actually producing results that match your intended purpose of the test.

Then there's construct validity, a little bit more difficult to establish. Usually requires some review by colleagues in the field who understand the underlying constructs and the concepts that you're testing. So this has to do with the extent to which the measure yields results as predicted by a theory or a concept or as I said before, a construct.

Now let's review reliability. Remember, reliability has to do with consistency of measurement, and basically this is a numerical value or a statistic that we can calculate that expresses the degree to which a test consistently produces the same results across multiple administrations. So internal consistency is something we discuss in relationship to reliability because this is a measure that's based on correlations between different items on the same test. So a correlation is a statistical estimate of the strength and direction of a relationship between two continuous variables.

Now in this case, the two continuous variables would be scores on a test that perhaps one student has produced two scores on two tests and we want to correlate how well the two tests actually measure the same thing. So for every observed change in one variable, there's a related observed change in the other, and that's the essence of correlation. They vary together, but the degree to which they vary together and how we can interpret that

relationship is what a correlation statistic tells us. Now correlations can be positive or negative, and we'll talk a little bit later in another module about how to interpret the ranges of scores that we get in these correlation-- statistical analyses.

Now talking about test reliability, we can assess this in several different ways. We can estimate it. There is a parallel forms test reliability that as a Pearson product moment correlation coefficient, which we will not get into much detail here at all. There is also a test retest reliability when we can correlate two administrations of the same test, which is, I don't know how many of you do that, but it's a way to check the reliability of the test. Then there is a split-halves reliability called Flanagan's formula. Now in your resources for this module, you'll find a list of reliability studies that you can do, a list of reliability formulas as well, and Flanagan's is one of those.

Now there is a Kuder-Richardson formula as well, both 20 and 21. And these are tests that have dichotomous-- tests that have dichotomous items. Now dichotomous simply means there are two possible answers-- correct or incorrect. And again, that formula is on that Test Reliabilities page.

So when we're calculating reliability and the kinds of tests that we make as teachers and professors, one easy way, I think, is the use Flanagan's formula. We'll calculate what's called a split-half reliability. So here's how you do it.

You're going to divide your quiz or test into two equal parts, like maybe you have 100 questions, so you have the first 50 questions and the second 50 questions if you divide it in two. You're going to calculate the variance for each part. So you're going to apply then this formula.

So here it is. And when you add the variances of parts a and b, you're going to divide the sum by the variance of the total test. And subtract it from one and multiply that result by two. The result is the split-half reliability of your quiz.

Now I'm going to make this a lot simpler in just a moment because I'm going to show you how to do this. You can do it by hand, and you can also calculate it using a statistical software program as well if you have that. Now remember, good tests have reliability coefficients of 0.70 or higher. Remember I told you, reliability can be anywhere from 0 to 1.

So let's work an example. OK. We're going to calculate the split-half reliability of a history quiz. So I told you before about opening the formula for the Reliability Studies page in the resources folder. So you might want to get that out if you don't already have that open. There are a number of reliability formulas on the page, but we're going to-- and you can look at those on your own, but we're just going to look at the split-half reliability or the Flanagan's formula.

So in this calculation, you see that the teacher, you, would divide your test into two even parts and use the subscores on each of those parts as well as the total score to calculate the reliability. So here's how we do it. Easy calculation for teacher made tests. It's most practical that we can use.

First, you're going to split the questions, as I said, into part a and part b. You're going to calculate the subscores for each student on each part. Then you'll calculate the variances of the subscores on the two parts and then the variance of the total scores.

Now you enter these numbers into the formula and calculate it. So notice the key, S sub a squared is the variance of part a. S sub b squared is the variance of part b. S squared is the variance of the total score. And you see that all falls within that big bracket that we're going to multiply by two after we subtract it from one and that will give us the r or the reliability.

Now here's the example. So we have some history quiz results. These students have taken this test. There are 10 students on this, and the maximum score they can get is 10. So we've divided in quiz part a, the scores they got on that part, and quiz part b, the score they got on that part, and you see the total score. So now we have all the numbers that we need in order to calculate this split-half reliability.

Now to calculate the variance, I suggest you use the standard deviation online calculator that I've linked here because you'll enter the scores for part a and part b and the total one at a time and then you get the three variances. It happens in just a split second. You can cut and paste and put it right in there and it'll happen. So I would suggest you try pausing this video right now and doing this on your own before you continue. But if you want to continue, go right ahead because I want to show you the numbers and we'll calculate the reliability together.

Here are the results. The part a variance is 1.21. Part b variance is 0.9, and the total variance is 2.71. So now we're going to place the figures into the formula.

So remember Flanagan's formula? Here's our data plugged into it. It's r equals 2 times 1 minus 1.21 plus 0.9 over 2.71, which translates to 2 times 1 minus 2.11 divided by 2.71, and that's going to equal, when we subtract it from 0.78 times 2-- excuse me, we take 1 minus 0.78, we get 0.22, and we multiply that times 2 and we get a reliability of 0.44. Now let me go back to that and show you that again so you can see that a little bit longer. That's how it works when we do it by hand.

Now when we interpret this coefficient, well, a split-half reliability is 0.44. Remember, I told you before, reliability coefficients should be at 0.7 or higher for a test to be considered reliable. So this quiz is really not reliable by that statistical calculation. So this teacher really needs to examine the performance of each item and revise, and we'll talk more about interpreting item analysis a little later on in another module.

So next we're going to talk about how you can determine difficulty and discrimination. So let's talk pause to think here. A teacher has given a test in a calculus class. The teacher used a split-half reliability formula to calculate a reliability coefficient of 0.85. So what does this tell you?

All right. Now with that, that's the end of module 4b. Module 4c is all about difficulty in discrimination.