Reliability and validity matrix

TEST of Reliability | Application and APPROPRIATENESS| Strengths| Weaknesses| Internal Consistency| This measure of reliability is appropriate when trying to determine the difference in reliability from shortening or lengthening a test (Cohen & Swerdlik, 2010). Here I am specifically referring to the Spearman-Brown formula being used to determine internal consistency. A researcher could also use other measures of internal consistency meant for heterogeneous test items, such as Inter-item consistency. The reliability of a test increases with an increase in the number of test items. One of the strengths of the Spearman-Brown Formula is that is can determine how much more or less reliable a test is as a researcher lengthens or shortens the test. This measure can also work in reverse and tell a researcher how many items they need to add to reach a certain reliability coefficient. | The problem with the use of the Spearman-Brown formula to determine internal consistency is that it is only affective with homogenous test items, that is items that are the same difficulty and length.

Also, tests of reliability are higher for whole-test vs. half-test applications of the formula, which means that lengthier tests work better with this instrument. | Split-half| The split-half form of measuring reliability entails creating two halves in the same test that can be compared in the same manner as the parallel form of reliability testing uses. This type of measurement is appropriate when using odd-even reliability or random assignment splits, but is most applicable when designing mini-parallel forms of the same test.

In this instance, each half is, “…as nearly equal as humanly possible—in format, stylistic, statistical, and related aspects” (Cohen & Swerdlik, 2010, p.

145). | The strength of this kind of measure is that it is less time-consuming and less cumbersome for test-takers than the parallel form, but is also a good measure of internal consistency. This type of measurement also help keep in check intermediary variables that might introduce error variance into the analysis, since the both parallel portions of the test are taken at once. However, there are several intermediary variables that are enhanced by this form of measuring reliability: fatigue that is felt during the second part of the test but not the first and variance in the difficulty or content of the items in the first half vs. the second half. It is also not advised to simply split a test down the middle. The different halves should have the same content and difficulty of question for the measure of reliability to be accurate. Test/retest| This type of test is applicable when the construct being measured is relatively stable over time, but is inappropriate for constructs that are not stable over time (Cohen & Swerdlik, 2010). This is because test/retest reliability is based on taking the same test, with the same people, at two different times. If the construct being measured is purported to change over time, then the scores of the test would vary because of true variance, rather than error variance—which is the basis of reliability, the latter that is. An example of this principle might be an achievement test measuring grammatical skills.

If the test-taker undergoes a series of lessons on grammar between the first test and the second test, then the test will show variance, but not due to error but due to the intermediary variable of education. Test/retest reliability would be inappropriate in this situation. | The strength of this measurement of reliability are in tests that, “…employ outcome measures such as reaction

time or perceptual judgment" (Cohen & Swerdlik, 2010, p. 143). This is because these types of psychometric traits do not vary greatly over time and are not sensitive to many types of intervening variable. The weakness of test/retest reliability is, of course, that the underlying constructs being tested can change over time, and therefore lower the test/retest reliability due to true variance rather than error variance. In this case, the overall reliability of a test might be seen as lower even though the actual measurement of the construct is stable (it is just that the construct itself varies). | Parallel and alternate forms| Both parallel and alternative forms of test reliability utilize multiple instances of the same test items at two different times with the same participants (Cohen & Swerdlik, 2010).

These types of measures of reliability would be most appropriate with tests that measure traits that are stable over a long period of time and inappropriate when measuring finite emotional states or anxiety levels. | The strength of this measure of reliability is that it measures the core construct through several variances of the same test item. If equivalent scores are found on multiple forms of the same test item, then the reliability of the test will go up. Moreover, there are ways to perform this type of reliability analysis without having the test-taker undergo multiple examinations: internal onsistency estimate of reliability. This type of analysis would save time and money. | Designing these types of measures are time-consuming, expensive, and tiresome for the test-taker who has to take variations of the same test items over and over again. Also, these forms of testing reliability are not dependable for measuring constructs that change over time, such as anxiety levels. Another weakness is that if the tests are taken some time

apart, then intervening variables might have an effect on the scores, thereby increasing error variance. Test of Validity| Application and APPROPRIATENESS| Strengths| Weaknesses| Face validity| Face validity is a description of the subjective perception of the test-taker of the test's validity (Cohen & Swerdlik, 2010). This measure is not so much a quantification of the test's actual validity, but a measure of the test-taker's perception of the test's validity. Face validity is most appropriate when measuring the test-takers confidence that a test measures what it purports to measure. The strength of face validity is that if the test-taker has confidence in the validity of test, then they are more likely to take the test, and further the test user is more likely to administer the test. Without face validity, the test might be perfectly valid, but it is not administered or taken properly because the user/taker does not have confidence in the test. | The weakness of face validity is that it might not measure actual validity. A test can appear to be valid to the user/taker while also being completely invalid for the construct/time/place of the test.

A good example might be the inkblot test. Psychologists that adhere to the psychodynamic perspective of psychopathology would say that the test is perfectly valid for determining personality characteristics, but the test taker might not understanding how the test applies to personality development, thereby undermining the face validity of the test. | Content validity| Measures of content validity are most useful in situations a test designer is trying to create test items that match the content of the material being tested (Cohen & Swerdlik, 2010).

For instance, a final course exam should test the content area that the course covered. Further, this measure might not be applicable in situations where the skills that the test designer are looking for in the applicant are not currently part of the skill-set of the already employed, such as in cases of new positions. | One of the strengths of content validity is that it can used to work backwards from job responsibilities to job applicant requirements.

First, the test designer would examine veteran workers perform their job, and then design an application process that looks for these qualities in a potential employee. The items that are judge essential for the job are the ones that are most advantageous for the applicant to possess. | The downfall of content validity is that the perspective of the material being covered is culturally and chronologically subjective, meaning that the questions can have different answers in different areas of the world or at different times.

Therefore, the test items must be culturally and chronologically accurate for the test-takers for content validity to be used. | Criterion related| I know this is personal opinion, but I think that criterion-related validity is the most powerful of all of the methods of verifying validity—especially concurrent validity. This type of validity is used to verify that the criterion that the test score purports to represent is actually in the sample of individuals being tested (Cohen & Swerdlik, 2010).

For instance, a group of people who have already been diagnosed with schizophrenia could be tested using a new instruments and if they all score high on the test for schizophrenia, then the test can be said to have acceptable validity. | One of the strengths of criterion-related validity is that

it is a very powerful measure of the actual validity of a test score. This type of validity uses methods external to the test itself to verify that the test covers the subject matter and criterion that it purports to cover. This fact alone makes this measure the most objective and verifiable of the measures of validity. A weakness of content validity is that criterion contaminations can occur, which is when the same predictor measure and criterion measure are used. As an example, if the diagnosis of a mental disorder by a panel of diagnosticians is used both as the test criterion and the measure of test validity. | Construct| Construct validity is the umbrella under which all of the other sub-types of validity fall (Cohen & Swerdlik, 2010). Construct validity is appropriate to use in cases where a test is trying to measure some underlying construct, such as intelligence or anxiety.

I suppose this measure of validity might not be appropriate in situations where there is not one clear construct that is being measured, such as generalized achievement tests. | One of the main strengths of construct validity is that the procedures used to verify underlying constructs follow the edicts of the scientific method. A hypothesis is formulated, predicting that if someone possesses in great quantity the construct of intelligences—as verified through other measures—then they will score high on a test purporting to measure intelligence.

In this way, a predictions is made based on scientific facts and then the test is used to determine if the prediction holds true. If it does not, then the test items, predictions, or underlying construct might need to be revised. | The downfall of this measure of validity is that if there is not one clear construct or if the construct is vaguely defined, then the validity of the test score is not

measurable. So, the validity of the test rests on the underlying construct definition and specificity. |