4.0 SYNOPSIS

Topic 4 defines the basic principles of assessment (reliability, validity, practicality, washback, and authenticity) and the essential sub-categories within reliability and validity.

4.1 LEARNING OUTCOMES

By the end of this topic, you will be able to:

1. define the basic principles of assessment (reliability, validity, practicality, washback, and authenticity) and the essential sub-categories within reliability and validity;

2. explain the differences between validity and reliability;

3. distinguish the different types of validity and reliability in tests and other instruments in language assessment.

4.2 FRAMEWORK OF TOPICS
INTRODUCTION

Assessment is a complex, iterative process requiring skills, understanding, and knowledge-in the exercise of professionally judgment. In this process, there are five important criteria that the testers ought to look into for “testing a test”: reliability, validity, practicality, washback and authenticity. Since these five principles are context dependent, there is no priority order implied in the order of presentation.

4.4 RELIABILITY (consistency)

Reliability means the degree to which an assessment tool produces stable and consistent results. It is a concept, which is easily being misunderstood (Feldt & Brennan, 1989).

Reliability essentially denotes ‘consistency, stability, dependability, and accuracy of assessment results’ (McMillan, 2001a, p.65 in Brown, G. et al, 2008). Since there is tremendous variability from either teacher or tester to teacher/tester that affects student performance, thus reliability in planning, implementing, and scoring student performances gives rise to valid assessment.

Fundamentally, a reliable (trustworthy) test is consistent and dependable. If a tester administers the same test to the same test-taker or matched test-takers on two circumstances, the test should give the same results. In a validity chain, it is stated that test administrators need to be sure that the scoring performance has to be carried out properly. If scores used by the tester do not reflect accurately what the test-taker actually did, would not be rewarded by another marker, or would not be received on a similar assessment, then these scores lack reliability. Errors occur in scoring in any ways—for example, giving Level 2 when another rater would give Level 4, adding up marks wrongly, transcribing scores from test paper to database inaccurately, students performing really well on the first half of the assessment and poorly on the second half due to fatigue, and so on. Thus, lack of reliability in the scores students receive is a treat to validity.

According to Brown (2010), a reliable test can be described as follows:

- Consistent in its conditions across two or more administrations
- Gives clear directions for scoring / evaluation
- Has uniform rubrics for scoring / evaluation
- Lends itself to consistent application of those rubrics by the scorer
- Contains item / tasks that are unambiguous to the test-taker
4.4.1 Rater Reliability

When humans are involved in the measurement procedure, there is a tendency of error, biasness and subjectivity in determining the scores of similar test. There are two kinds of rater reliability namely inter-rater reliability and intra-rater reliability.

Inter-rater reliability refers to the degree of similarity between different tester or rater; can two or more testers/raters, without influencing one another, give the same marks to the same set of scripts (contrast with intra-rater reliability).

One way to test inter-rater reliability is to have each rater assign each test item a score. For example, each rater might score items on a scale from 1 to 10. Next, you would calculate the correlation (connection) between the two ratings to determine the level of inter-rater reliability. Another means of testing inter-rater reliability is to have raters determine which category each observation falls into and then calculate the percentage of agreement between the raters. So, if the raters agree 8 out of 10 times, the test has an 80% inter-rater reliability rate. Rater reliability is assessed by having two or more independent judges score the test. The scores are then compared to determine the consistency of the raters’ estimates.

Intra-rater reliability is an internal factor. In intra-rater reliability, its main aim is consistency within the rater. For example, if a rater (teacher) has many examination papers to mark and does not have enough time to mark them, s/he might take much more care with the first, say, ten papers, than the rest. This inconsistency will affect the students’ scores; the first ten might get higher scores. In other words, while inter-rater reliability involves two or more raters, intra-rater reliability is the consistency of grading by a single rater. Scores on a test are rated by a single rater/judge at different times. When we grade tests at different times, we may become inconsistent in our grading for various reasons. Some papers that are graded during the day may get our full and careful attention, while others that are graded towards the end of the day are very quickly glossed over. As such, intra-rater reliability determines the consistency of our grading.

Both inter-and intra-rater reliability deserve close attention in that test scores are likely to vary from rater to rater or even from the same rater (Clark, 1979).
4.4.2 Test Administration Reliability

There are a number of reasons which influences test administration reliability. Unreliability occurs due to outside interference like noise, variations in photocopying, temperature variations, the amount of light in various parts of the room, and even the condition of desk and chairs. Brown (2010) stated that he once witnessed the administration of a test of aural comprehension in which an audio player was used to deliver items for comprehension, but due to street noise outside the building, test-taker sitting next to open windows could not hear the stimuli clearly. According to him, that was a clear case of unreliability caused by the conditions of the test administration.

4.4.3 Factors influencing Reliability

![Factors that can affect the reliability of a test](image)

Figure 4.4.3 Factors that affect the reliability of a test

The outcome of a test is influenced by many factors. Assuming that the factors are constant and not subject to change, a test is considered to be reliable if the scores are consistent and not different from other equivalent and reliable test scores. However, tests are not free from errors. Factors that affect the reliability of a test include test length factors, teacher and student factors, environment factors, test administration factors, and marking factors.

a. Test length factors

In general, longer tests produce higher reliabilities. Due to the dependency on coincidence and guessing, the scores will be more accurate if the duration of the test is longer. An objective test has higher consistency
because it is not exposed to a variety of interpretations. A valid test is said to be reliable but a reliable test need not be valid. A consistent score does not necessary measure what is intended to measure. In addition, the test items that are the samples of the subject being tested and variation in the samples may be found in two equivalent tests and there can be one of the causes test outcomes are unreliable.

b. **Teacher-Student factors**

In most tests, it is normally for teachers to construct and administer tests for students. Thus, any good teacher-student relationship would help increase the consistency of the results. Other factors that contribute to positive effects to the reliability of a test include teacher’s encouragement, positive mental and physical condition, familiarity to the test formats, and perseverance (determination) and motivation.

c. **Environment factors**

An examination environment certainly influences test-takers and their scores. Any favourable environment with comfortable chairs and desks, good ventilation, sufficient light and space will improve the reliability of the test. On the contrary, a non-conducive environment will affect test-takers’ performance and test reliability.

d. **Test administration factors**

Because students’ grades are dependent on the way tests are being administered, test administrators should strive to provide clear and accurate instructions, sufficient time and careful monitoring of tests to improve the reliability of their tests. A test-re-test technique can be used to determine test reliability.

e. **Marking factors**

Unfortunately, we human judges have many opportunities to introduce error in our scoring of essays (Linn & Gronlund, 2000; Weigle, 2002). It is possible that our scoring invalidates many of the interpretations we would like to make based on this type of assessment. Brennan (1996) has reported that in large-scale, high-stakes marking panels that are tightly trained and monitored marker effects are small. Hence, it can be concluded
that in low-stakes, small-scale marking, there is potentially a large error introduced by individual markers. It is also common that different markers award different marks for the same answer even with a prepared mark scheme. A marker’s assessment may vary from time to time and with different situations. Conversely, it does not happen to the objective type of tests since the responses are fixed. Thus, objectivity is a condition for reliability.

4.5 VALIDITY

Validity refers to the evidence base that can be provided about appropriateness of the inferences, uses, and consequences that come from assessment (McMillan, 2001a). Appropriateness has to do with the soundness (accuracy), trustworthiness, or legitimacy of the claims or inferences (conclusion that testers would like to make on the basis of obtained scores. Clearly, we have to evaluate the whole assessment process and its constituent (component) parts by how soundly (thoroughly) we can defend the consequences that arise from the inferences and decisions we make. Validity, in other words, is not a characteristic of a test or assessment; but a judgment, which can have varying degrees of strength.

So, the second characteristic of good tests is validity, which refers to whether the test is actually measuring what it claims to measure. This is important for us as we do not want to make claims concerning what a student can or cannot do based on a test when the test is actually measuring something else. Validity is usually determined logically although several types of validity may use correlation coefficients.

According to Brown (2010), a valid test of reading ability actually measures reading ability and not 20/20 vision, or previous knowledge of a subject, or some other variables of questionable relevance. To measure writing ability, one might ask students to write as many words as they can in 15 minutes, then simply count the words for the final score. Such a test is practical (easy to administer) and the scoring quite dependable (reliable). However, it would not constitute (represent) a valid test of writing ability without taking into account its comprehensibility (clarity), rhetorical discourse elements, and the organisation of ideas.

The following are the different types of validity:

- **Face validity**: Do the assessment items appear to be appropriate?
• **Content validity:** Does the assessment content cover what you want to assess? Have satisfactory samples of language and language skills been selected for testing?

• **Construct validity:** Are you measuring what you think you're measuring? Is the test based on the best available theory of language and language use?

• **Concurrent (parallel) validity:** Can you use the current test score to estimate scores of other criteria? Does the test correlate with other existing measures?

• **Predictive validity:** Is it accurate for you to use your existing students' scores to predict future students' scores? Does the test successfully predict future outcomes?

It is fairly obvious that a valid assessment should have a good coverage of the criteria (concepts, skills and knowledge) relevant to the purpose of the examination. The important notion here is the purpose.

![Figure 4.5: Types of Validity](image)
4.5.1 Face validity

Face validity is validity which is “determined impressionistically; for example by asking students whether the examination was appropriate to the expectations” (Henning, 1987). Mousavi (2009) refers face validity as the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgement of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers.

It is pertinent (important) that a test looks like a test even at first impression. If students taking a test do not feel that the questions given to them are not a test or part of a test, then the test may not be valid as the students may not take it seriously to attempt the questions. The test, hence, will not be able to measure what it claims to measure.

4.5.2 Content validity

Content validity is concerned with whether or not the content of the test is sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure” (Henning, 1987). The most important step in making sure of content validity is to make sure all content domains are presented in the test. Another method to verify validity is through the use of Table of Test Specification that can give detailed information on each content, level of skills, status of difficulty, number of items, and item representation for rating in each content or skill or topic.

We can quite easily imagine taking a test after going through an entire language course. How would you feel if at the end of the course, your final examination consists of only one question that covers one element of language from the many that were introduced in the course? If the language course was a conversational course focusing on the different social situations that one may encounter, how valid is a final examination that requires you to demonstrate your ability to place an order at a posh restaurant in a five-star hotel?
4.5.3 Construct validity

Construct is a psychological concept used in measurement. Construct validity is the most obvious reflection of whether a test measures what it is supposed to measure as it directly addresses the issue of what it is that is being measured. In other words, construct validity refers to whether the underlying theoretical constructs that the test measures are themselves valid. Proficiency, communicative competence, and fluency are examples of linguistic constructs; self-esteem and motivation are psychological constructs.

Fundamentally every issue in language learning and teaching involves theoretical constructs. When you are assessing a student’s oral proficiency for instance. To possess construct validity, the test should consist of various components of fluency: speed, rhythm, juncture, (lack of) hesitations, and other elements within the construct of fluency. Tests are, in a manner of speaking, operational definitions of constructs in that their test tasks are the building blocks of the entity that is being measured (see Davidson, Hudson, & Lynch, 1985; T. McNamara, 2000).

4.5.4 Concurrent validity

Concurrent validity is the use of another more reputable and recognised test to validate one’s own test. For example, suppose you come up with your own new test and would like to determine the validity of your test. If you choose to use concurrent validity, you would look for a reputable test and compare your students’ performance on your test with their performance on the reputable and acknowledged test. In concurrent validity, a correlation coefficient is obtained and used to generate an actual numerical value. A high positive correlation of 0.7 to 1 indicates that the learners’ score is relatively similar for the two tests or measures.

For example, in a course unit whose objective is for students to be able to orally produce voiced and unvoiced stops in all possible phonetics environments, the results of one teacher’s unit test might be compared with an independent assessment such as a commercially produced test of similar phonemic proficiency. Since criterion-related evidence usually falls into one of two categories of concurrent and predictive validity, a classroom test designed to assess mastery of a point of grammar in a communicative use will have
criterion validity if test scores are verified either by observed subsequent behaviour or by other communicative measures of grammar point in question.

4.5.5 Predictive validity

Predictive validity is closely related to concurrent validity in that it too generates a numerical value. For example, the predictive validity of a university language placement test can be determined several semesters later by correlating the scores on the test to the GPA of the students who took the test. Therefore, a test with high predictive validity is a test that would yield predictable results in a latter measure. A simple example of tests that may be concerned with predictive validity is the trial national examinations conducted at schools in Malaysia as it is intended to predict the students’ performance on the actual SPM national examinations. (Norleha Ibrahim, 2009)

As mentioned earlier validity is a complex concept, yet it is crucial to the teacher’s understanding of what makes a good test. It is good to heed Messick’s (1989, p. 36) caution that validity is not an all-or-none proposition and that various forms of validity may need to be applied to a test in order to be satisfied worth its overall effectiveness.

What are reliability and validity? What determines the reliability of a test?

What are the different types of validity? Describe any three types and cite examples.


4.5.6 Practicality

Although practicality is an important characteristic of tests, it is by far a limiting factor in testing. There will be situations in which after we have already determined what we consider to be the most valid test, we need to reconsider the format purely because of practicality issues. A valid test of spoken interaction, for example, would require that the examinees be relaxed, interact with peers and speak on topics that they are familiar and comfortable with. This sounds like the kind of conversations that people have with their friends while sipping afternoon tea by the roadside stalls. Of course such a situation
would be a highly valid measure of spoken interaction – if we can set it up. Imagine if we even try to do so. It would require hidden cameras as well as a lot of telephone calls and money.

Therefore, a more practical form of the test especially if it is to be administered at the national level as a standardised test, is to have a short interview session of about fifteen minutes using perhaps a picture or reading stimulus that the examinees would describe or discuss. Therefore, practicality issues, although limiting in a sense, cannot be dismissed if we are to come up with a useful assessment of language ability. Practicality issues can involve economics or costs, administration considerations such as time and scoring procedures, as well as the ease of interpretation. Tests are only as good as how well they are interpreted. Therefore tests that cannot be easily interpreted will definitely cause many problems.

4.5.7 Objectivity

The objectivity of a test refers to the ability of teachers/examiners who mark the answer scripts. Objectivity refers to the extent, in which an examiner examines and awards scores to the same answer script. The test is said to have high objectivity when the examiner is able to give the same score to the similar answers guided by the mark scheme. An objective test is a test that has the highest level of objectivity due to the scoring that is not influenced by the examiner’s skills and emotions. Meanwhile, subjective test is said to have the lowest objectivity. Based on various researches, different examiners tend to award different scores to an essay test. It is also possible that the same examiner would give different scores to the same essay if s/he is to re-check at different times.

4.5.8 Washback effect

The term 'washback' or backwash (Hughes, 2003, p.1) refers to the impact that tests have on teaching and learning. Such impact is usually seen as being negative: tests are said to force teachers to do things they do not necessarily wish to do. However, some have argued that tests are potentially also 'levers for change' in language education: the argument being that if a bad test has negative impact, a good test should or could have positive washback (Alderson, 1986b; Pearson, 1988).
Cheng, Watanabe, and Curtis (2004) offered an entire anthology to the issue of washback while Spratt (2005) challenged teachers to become agents of beneficial washback in their language classrooms. Brown (2010) discusses the factors that provide beneficial washback in a test. He mentions that such a test can positively influence what and how teachers teach, students learn; offer learners a chance to adequately prepare, give learners feedback that enhance their language development, is more formative in nature than summative, and provide conditions for peak performance by the learners.

In large-scale assessment, washback often refers to the effects that tests have on instruction in terms of how students prepare for the test. In classroom-based assessment, washback can have a number of positive manifestations, ranging from the benefit of preparing and reviewing for a test to the learning that accrues from feedback on one’s performance. Teachers can provide information that “washes back” to students in the form of useful diagnoses of strengths and weaknesses.

The challenge to teachers is to create classroom tests that serve as learning devices through which washback is achieved. Students’ incorrect responses can become a platform for further improvements. On the other hand, their correct responses need to be complimented, especially when they represent accomplishments in a student’s developing competence. Teachers can have various strategies in providing guidance or coaching. Washback enhances a number of basic principles of language acquisition namely intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment, among others.

Washback is generally said to be either positive or negative. Unfortunately, students and teachers tend to think of the negative effects of testing such as “test-driven” curricula and only studying and learning “what they need to know for the test”. Positive washback, or what we prefer to call “guided washback” can benefit teachers, students, and administrators. Positive washback assumes that testing and curriculum design are both based on clear course outcomes, which are known to both students and teachers/testers. If students perceive that tests are markers of their progress towards achieving these outcomes, they have a sense of accomplishment. In short, tests must be part of learning experiences for all involved. Positive
washback occurs when a test encourages good teaching practice.

Washback is particularly obvious when the tests or examinations in question are regarded as being very vital and having a definite impact on the student’s or test-taker’s future. We would expect, for example, that national standardised examinations would have strong washback effects compared to a school-based or classroom-based test.

4.5.9 Authenticity

Another major principle of language testing is authenticity. It is a concept that is difficult to define, particularly within the art and science of evaluating and designing test. Citing Bachman and Palmer (1996) in Brown (2010) authenticity is “the degree of correspondence of the characteristics of a given language test task to the features of a target language task” (p.23) and then suggested an agenda for identifying those target language tasks and for transforming them into valid test items.

Language learners are motivated to perform when they are faced with tasks that reflect real world situations and contexts. Good testing or assessment strives to use formats and tasks that reflect the types of situation in which students would authentically use the target language. Whenever possible, teachers should attempt to use authentic materials in testing language skills.

4.6.0 Interpretability

Test interpretation encompasses all the ways that meaning is assigned to the scores. Proper interpretation requires knowledge about the test, which can be obtained by studying its manual and other materials along with current research literature with respect to its use; no one should undertake the interpretation of scores on any test without such study. In any test interpretation, the following considerations should be taken into account.
A. Consider Reliability:
Reliability is important because it is a prerequisite to validity and because the degree to which a score may vary due to measurement error is an important factor in its interpretation.

B. Consider Validity:
Proper test interpretation requires knowledge of the validity evidence available for the intended use of the test. Its validity for other uses is not relevant. Indeed, use of a measurement for a purpose for which it was not designed may constitute misuse. The nature of the validity evidence required for a test depends upon its use.

C. Scores, Norms, and Related technical Features:
The result of scoring a test or subtest is usually a number called a raw score, which by itself is not interpretable. Additional steps are needed to translate the number directly into either a verbal description (e.g., pass or fail) or into a derived score (e.g., a standard score). Less than full understanding of these procedures is likely to produce errors in interpretation and ultimately in counseling or other uses.

D. Administration and Scoring Variation:
Stated criteria for score interpretation assume standard procedures for administering and scoring the test. Departures from standard conditions and procedures modify and often invalidate these criteria.

Study some of commercially produced tests and evaluate the authenticity of these tests/test items.

Discuss the importance of authenticity in testing.

Based on samples of formative and summative assessments, discuss aspects of reliability/validity that must be considered in these assessments.

Discuss measures that a teacher can take to ensure high validity of language assessment for the primary classroom.