Tikrit University

College of Education for Humanities

English Department

PhD (Second Course) 2025-2026

An Advanced Course in EFL Communicative Testing, Measurement and Evaluation

Principles of Language Assessment

(3)

Prof. Dunia Tahir (PhD)

Brown, H. D., & Abeywickrama, P. (2018). ***Language Assessment Principles and Classroom Practices*** **(3rd ed.). Pearson Education.**

2026.A.D. **1447.A.H.**

Chapter 2 Principles of Language Assessment

Principles of Language Assessment

How do you know whether a test is effective, appropriate, useful, or, in down-to-earth terms, a "good" test? For the most part, that question can be answered by responding to such questions as:

1. Can it be given within appropriate administrative constraints?
2. Is it dependable?
3. Does it accurately measure what you want it to measure?
4. Does the language in the test represent real-world language use?
5. Does the test provide information that is useful for the learner?

These questions help to identify five cardinal criteria for "testing a test": practicality, reliability, validity, authenticity, and washback. We will look at each one here.

Q1:What are the five major principles of language assessment?

Q2:What is the main purpose of language assessment principles?

Q3:Wht are the principles of language assessment considered context-dependent?

1. Practicality

Practicality refers to the logistical, down-to-earth, administrative issues involved in making, giving, and scoring an assessment instrument. These include "costs, the amount of time it takes to construct and to administer, ease of scoring, and ease of interpreting/reporting the results" (Mousavi, 2009, p. 516). A test that fails to meet such criteria is impractical. Consider the following attributes of practicality:

A practical test….

1. stays within budgetary limits

2. can be completed by the test-taker within appropriate time constraints
3. has clear directions for administration
4. appropriately utilizes available human resources
5. does not exceed available material resources
6. considers the time and effort involved to both design and score

A test of language proficiency that takes a student 5 hours to complete is impractical—it consumes more time than is available to accomplish its objective. A test that requires individual one-on-one proctoring is impractical for a group of several hundred test-takers and only a handful of examiners. A test that requires a few minutes for a student to take and several hours for an examiner to evaluate is impractical for most classroom situations.

Q1:What does practicality mean in language assessment?

Q2:What makes a test impractical?

Q3:Mention three characteristics of a practical test.

Q4:Why is a five-hour language test considered impractical?

Q5:Why is heavy reliance on subjective scoring sometimes impractical?

Q6:How does technology affect practicality?

Q7:What problem occurred during the English placement test administration?

Q8:How did the administrator respond to the problem?

Q9:Why did scoring the dictation cause difficulty?

Q10:What practicality issue affected the scoring process?

Q11:What lesson does this incident teach about practicality?

2. Reliability

A reliable test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. We might capsulate the principle of reliability in the following:

Reliable test…

1. Has consistent conditions across two or more administrations
2. gives clear directions for scoring/evaluation
3. has uniform rubrics for scoring/evaluation
4. lends itself to consistent application of rubrics by the scorer
5. contains items/tasks that are unambiguous to the test-taker

The issue of the reliability of tests can be better understood by considering a number of factors that can contribute to their unreliability. We examine four possible sources of fluctuations in (1) the student, (2) the scoring, (3) the test administration, and (4) the test itself.

Q1:What is reliability in language testing?

Q2:When is a test considered reliable?

Q3:List three characteristics of a reliable test.

Q4:Why are clear scoring rubrics important for reliability?

Q5:What are the four main sources of unreliability in tests?

2.1 Student-Related Reliability

The most common learner-related issue in reliability is caused by temporary illness, fatigue, a “bad day,” anxiety, and other physical or psychological factors, which may make an observed score deviate from one’s “true” score. Also included in this category are such factors as a test-taker’s **test-wiseness**, or strategies for efficient test-taking.

Q1:What factors can affect student- related reliability?

Q2:Can teachers reduce student-related unreliability?

2.2 Rater Reliability

Human error, subjectivity, and bias may enter into the scoring process. **Inter-rater reliability** occurs when two or more scorers yield consistent scores of the same test. Failure to achieve inter-rater reliability could stem from lack of adherence to scoring criteria, inexperience, inattention, or even preconceived biases.

Rater-reliability issues are not limited to contexts in which two or more scorers are involved. **Intra-rater reliability** is an internal factor, a common occurrence for classroom teachers. Such reliability can be violated in cases of unclear scoring criteria, fatigue, bias toward particular "good" and "bad" students, or simple carelessness. If faced with grading up to 40 essay tests (for which there is no absolute right or wrong set of answers) within only a week, you might recognize that the standards applied to the first few tests will differ from those applied to the last few.

Q1:What is inter-rater reliability?

Q2:What causes low inter-rater reliability? What factors can reduce inter-rater reliability?

Q3:Why is rater training important?

Q4:Why is rater reliability difficult to achieve in writing tests?

2.3 Test Administration Reliability

Unreliability may also result from the conditions in which the test is administered. We once witnessed the administration of a test of aural comprehension in which an audio player was used to deliver items for comprehension, but because of street noise outside the building, students sitting next to open windows could not hear the stimuli accurately. This was a clear case of unreliability caused by the conditions of the test administration. Other sources

of unreliability are found in photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the condition of desks and chairs.

Q1:What is test administration reliability?

Q2: Give examples of factors that reduce test administration reliability.

2.4 Test Reliability

Sometimes the nature of the test itself can cause measurement errors. Tests with multiple-choice items must be carefully designed to include a number of characteristics that guard against unreliability. For example, the items need to be evenly difficult, distractors need to be well designed, and items need to be well distributed to make the test reliable.

In classroom-based assessment, test unreliability can be caused by many factors, including rater bias. This typically occurs with **subjective tests** with open-ended responses (e.g., essay responses) that require a judgment on the part of the teacher to determine correct and incorrect answers. **Objective tests**, in contrast, have predetermined fixed responses, a format that of course increases their test reliability.

Q1:How can the nature of a test affect reliability?

Q2:Why are objective tests usually more reliable than subjective tests?

Q3:What happens when a test has too many items?

Q4:How can test characteristics interact with student-related factors?

3. Validity

By far the most complex criterion of an effective test—and arguably the most important principle—is validity, "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Gronlund, 1998, p. 226). In somewhat more technical terms, Samuel Messick (1989), who is widely recognized as an expert on validity, defined validity as "an integrated evaluative judgment

of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 11). We might infer from these definitions the following attributes of validity:

A valid test...

1. measures exactly what it proposes to measure
2. does not measure irrelevant or "contaminating" variables
3. relies as much as possible on empirical evidence (performance)
4. involves performance that samples the test's criterion (objective)
5. offers useful, meaningful information about a test-taker's ability
6. is supported by a theoretical rationale or argument

A valid test of reading ability actually measures reading ability—not 20/20 vision, or previous knowledge of a subject, or some other variable 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 would be easy to administer (practical), and the scoring quite dependable (reliable), but it would not constitute a valid test of writing ability without some consideration of comprehensibility, rhetorical discourse elements, and the organization of ideas, among other factors.

Q1:What is validity in language assessment?

Q2:How did Samuel Messick define validity?

Q3:Why is validity considered the most important principle of testing?

Q4:Why is counting the number of words written not a valid measure of writing ability?

Q5:How can the validity of a test be established?

3.1 Content-Related Evidence

If a test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behavior measured, it can claim content-related evidence of validity, often popularly referred to as **content-related validity**. You can usually identify content-related evidence observationally if you can clearly define the achievement you are measuring. A test of tennis competency that asks someone to run a 100-yard dash obviously lacks content validity.

Another way of understanding content validity is to consider the difference between **direct** and **indirect testing**. Direct testing involves the test-taker in actually performing the target task. In an indirect test, learners do not perform the task itself but rather a task that is related in some way. For example, if you intend to test learners' oral production of syllable stress and your test task is to have learners mark (with written accent marks) stressed syllables in a list of written words, you could, with a stretch of logic, argue you are **indirectly testing their oral production**. A direct test of syllable production would require that students actually produce target words orally.

Q1:What is content-related evidence of validity?

Q2:Why do multiple-choice grammar questions lack content validity for assessing speaking ability?

Q3:What is one way to identify content-related evidence?

Q4:Why does the article quiz have some content validity?

Q5:Why may some standardized language proficiency tests lack content validity?

Q6:What is the difference between direct and indirect testing?

Example of indirect testing: Asking students to mark stressed syllables in written words to rest oral syllable stress.

Example of direct testing: Requiring students to orally produce stressed syllables or target words.

Q7:What is a practical rule for achieving content validity in classroom assessment?

Q8:What should a test in a listening / speaking class include?

Q9:Why is content-related evidence especially important in classroom testing?

3.2 Criterion-Related Evidence

A second form of evidence of the validity of a test may be found in what is called criterion-related evidence, also referred to as **criterion-related validity**, or the extent to which the "criterion" of the test has actually been reached. In the case of teacher-made classroom assessments, criterion-related evidence is best demonstrated through a comparison of results of an assessment with results of some other measure of the same criterion. For example, in a course unit whose objective is for students to orally produce voiced and voiceless stops in all possible phonetic environments, the results of one teacher's unit test might be compared with an independent assessment—possibly a commercially produced test in a textbook—of the same phonemic proficiency.

Criterion-related evidence usually falls into one of two categories: (1) concurrent and (2) predictive validity. A test has **concurrent validity** if its results are supported by other concurrent performance beyond the assessment itself. For example, the validity of a high score on the final exam of a foreign-language course will be substantiated by actual proficiency in the language. The **predictive validity** of an assessment becomes important in the case of placement tests, admissions assessment batteries, and achievement tests designed to determine students' readiness to "move on" to another unit. The assessment criterion in such cases is not to measure concurrent ability but to assess (and predict) a test-taker's likelihood of future success.

Q1:What is criterion-related evidence of validity?

Q2:How is criterion-related evidence demonstrated in classroom tests?

Example of criterion-related evidence: Comparing a teacher's pronunciation test with a commercially produced phonetic test measuring the same ability.

Q3:What are the two types of criterion-related validity?

Concurrent validity: It exists when test results are supported by other performance measures taken at the same time.

Predictive validity: It is the ability of a test to predict a test-taker's future performance or success.

Q4:Where is predictive validity especially important?

3.3 Construct-Related Evidence

A third kind of evidence that can support validity, but one that does not play as large a role for classroom teachers, is construct-related validity, commonly referred to as construct validity. **A construct is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions.** Constructs may or may not be directly or empirically measured—their verification often requires inferential data. Proficiency and communicative competence are examples of linguistic constructs; self-esteem and motivation are psychological constructs. Virtually every issue in language learning and teaching involves theoretical constructs. Tests are, in a manner of speaking, operational definitions of constructs in that their test tasks are the building blocks of the entity measured.

Imagine, **for example**, that you have been given a procedure for conducting an oral interview. The scoring analysis for the interview includes several factors in the final score:

1. pronunciation
2. fluency
3. grammatical accuracy
4. vocabulary use
5. sociolinguistic appropriateness

The justification for these five factors lies in a theoretical construct that claims they are major components of oral proficiency. So, if you were asked to conduct an oral proficiency interview that evaluated only pronunciation and grammar, you could be justifiably suspicious about the construct validity of that test.

Q1:What is a construct-related validity?

Q2:What is a construct?

Example of linguistic construct: Proficiency, communicative competence.

Example of psychological constructs: self-esteem, motivation.

Q3:Why is construct validity important in language assessment?

Q4: Is formal construct validation always required for classroom teachers?

Q5:What factors may be used to score an oral interview?

Q6:Why are these factors included in an oral proficiency test?

3.4 Consequential Validity (Impact)

Two other categories—in addition to the three widely accepted forms of evidence— may be of some interest and utility in your own quest to validate classroom tests. Consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its effect on the preparation of test-takers, and the (intended and unintended) social consequences of a test's interpretation and use.

Bachman et al., (2010), use the term **impact** to refer to consequential validity, perhaps more broadly encompassing the many consequences of assessment, before and after a test administration. The impact of test-taking and the use of test scores can, according to Bachman and Palmer (2010, p. 30), be seen at both **a macro level** (the effect on society and educational systems) and a micro level (the effect on individual test-takers).

Q1:What is consequential validity?

Q2:What does consequential validity include besides test accuracy?

Q3:At what level can the impact of a test be observed?

Q4:What is washback in language assessment?

Q5:At what level is washback especially important?

Q6:Why should teachers consider washback when designing assessments?

3.5 Face Validity

An offshoot of consequential validity is the extent to which "students view the assessment as fair, relevant, and useful for improving learning". Face validity refers to "the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers" (Mousavi, 2009, p. 247).

Teachers can increase a student's perception of fair tests by using:

1. formats that are expected and well-constructed with familiar tasks
2. tasks that can be accomplished within an allotted time limit
3. items that are clear and uncomplicated
4. directions that are crystal clear
5. tasks that have been rehearsed in their previous course work
6. tasks that relate to their course work (content validity)
7. level of difficulty that presents a reasonable challenge

Q1:What is face validity?

Q2:Whose judgment is face validity based on ?

Q3:Is face validity an empirically measurable type of validity?

Q4:Why is face validity still important in classroom assessment?

Q5:How can teachers increase students' perception of face validity?

4. Authenticity

A fourth major principle of language testing is authenticity, a concept that is difficult to define, especially within the art and science of evaluating and designing tests. Bachman and Palmer (1996) defined authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task" (p. 23). Essentially, when you make a claim for authenticity in a test task, you are Saying that this task is likely to be enacted in the real world.

An authentic test...

1. contains language that is as natural as possible
2. has items that are contextualized rather than isolated
3. includes meaningful, relevant, interesting topics
4. provides some thematic organization to items, such as through a
5. story line or episode
6. offers tasks that replicate real-world tasks

Q1:What is authenticity in language testing?

Q2:Why is authenticity difficult to define and measure?

Q3:Why is authenticity important in language assessment?

Q4:Why do many test items lack authenticity?

Q5:What are the characteristics of an authentic test?

Q6:How has authenticity in testing changed in recent years?

5. Washback

A facet of consequential validity is “the effect of testing on teaching and learning” (Hughes, 2003, p. 1), otherwise known in the language assessment field as washback. Messick (1996, p. 241) reminded us that the washback effect may refer to both the promotion and the inhibition of learning, thus emphasizing what may be referred to as beneficial versus harmful (or negative) washback.

Test That Provides Beneficial Washback ...

1. positively influences what and how teachers teach
2. positively influences what and how learners learn
3. offers learners a chance to adequately prepare
4. gives learners feedback that enhances their language development
5. is more formative in nature than summative
6. provides conditions for peak performance by the learner

Q1:What is washback in language assessment?

Q2:What are the two types of washback?

Q3:Why is washback important in language education?

Q4:What role do teachers play in washback?

Q5:What is beneficial washback? What are the main features of a test that provides beneficial washback?

Q6:How does classroom-based assessment promote positive washback?

Q7:Why may formal tests fail to provide beneficial washback?

Q8:What challenge do teachers face in achieving washback?

Q9:How do formative tests support washback?

Q10:Why should summative tests also include feedback?

Q11:Why can low scores on test subsections be useful?

Q12:Why is cooperative classroom environment important for washback?

Applying Principles to Classroom Testing

Q1:What are the five principles of language assessment? Why are these five principles important?

Q2:Which principle is considered the most important and why?