بسم الله الرحمن الرحيم

Principles of Language Testing

(Characteristics of a good language test)

-Practicality, Reliability, Validity, Authenticity, Washback-

-5th Lecture-

Ms. Zainab Jaafar

MA in Applied Linguistics

Practicality

Practicality

- It is the logistical, down-to-earth, administrative issues involved in making, giving, and scoring an assessment instrument.

Administrative issues

- Costs
- Time of construction & administration
- Ease of scoring
- Ease of interpreting/ reporting the results

* A test that doesn’t meet these criteria is *impractical test*

Characteristics of a practical test

- Stays within budgetary limits
- Can be completed by the test-taker within appropriate time constraints
- Has clear directions for administration
- Appropriately utilizes available human resources
- Does not exceed available material resources
- Considers the time and effort involved to both design and scoring

Examples of impractical tests

- A 5 hours proficiency test
- An individual one-on-one proctoring test
- A few minutes test for a student to take and several hours for an examiner to evaluate
- A test scored only by a computer
- A test that relies too heavily on the subjectivity of the scorer

Reliability

Reliability

- A reliable test is consistent and dependable (i.e., the test should yield similar results if it is given to the same student or matched students on two different occasions.

Factors that affect the reliability of a test

1. the student
2. the scoring
3. the administration of a test, &
4. the test itself

Characteristics of a reliable test

- Has consistent conditions across two or more administrations
- Gives clear directions for scoring/evaluation
- Has uniform rubrics for scoring/evaluation
- Lends itself to consistent application of rubrics by the scorer
- Contains items/tasks that are unambiguous to the test-taker

1. Student-Related Reliability

Factors affect the student-related reliability

- illness, fatigue, a “bad day," anxiety, and other physical or psychological factors
- a test-taker’s test-wiseness, or
- strategies for efficient test-taking

2. Rater Reliability

Rater Reliability

- Human error, subjectivity, and bias in scoring

1. Inter-rater Reliability

   Occurs when two or more scorers yield consistent scores of the same test

Failure to achieve interrater reliability is due to:

- lack of adherence to scoring criteria, inexperience, inattention, or even preconceived biases

Rater Reliability

2. Intra-rater Reliability

- Is an internal factor, including unclear scoring criteria, fatigue, bias toward particular “good" and “bad” students, or simple carelessness.

Example of intra-rater unreliability

- Scoring 40 essays test within a week. Scoring the first essays will differ from the last ones as the scorer might get tired. Thus, the result may be an inconsistent evaluation across all tests.

Solution

- Read through about half of the tests before rendering any final, then cycle back through the whole set of tests to ensure even-handed judgment.
- Use an analytical scoring instrument to increase both inter- and intra-rater reliability

3. Test Administration Reliability

Test Administration Reliability

- The conditions in which the test is administered.

Example of test administration unreliability

- Street noise that prevents students from hearing an audio player of an aural comprehension test
- Photocopying variations
- The amount of light in different parts of the room
- Variations in temperature
- Condition of desks and chairs

4. Test Reliability

Test Reliability

- The nature of the test itself that may cause measurement errors

Example of test reliability

- A well-designed test of multiple-choice items in which the items are evenly difficult and well distributed, and distractors are well designed.
- Test reliability is increased through objective tests which have predetermined fixed responses.

4. Test Reliability

Examples of test unreliability

- Subjective tests with open-ended responses (e.g., essay responses) in which the teacher determines correct and incorrect answers that leads to rater bias.
- Poorly written test items (items that are ambiguous or have more than one correct answer).
- A test with too many items where the test-takers becomes fatigued by the time they reach the later items and hastily respond incorrectly.
- Timed tests that affect students who do not perform well on a test with a time limit.

Validity

Validity

- The most complex and the most important criterion of an effective test
- It is a matter of degree, not all or none

Types of validity

1. Content Validity
2. Criterion Validity
3. Construct Validity
4. Consequential Validity
5. Face Validity

Characteristics of a valid test

- Measures exactly what it proposes to measure
- Does not measure irrelevant or “contaminating” variables
- Relies as much as possible on empirical evidence (performance)
- Involves performance that samples the test’s criterion (objective)
- Offers useful, meaningful information about a test-taker’s ability
- Is supported by a theoretical rationale or argument

1. Content Validity

Content Validity

- A test should actually samples the subject matter about which conclusions are to be drawn, and requires the test-taker to perform the behavior measured.

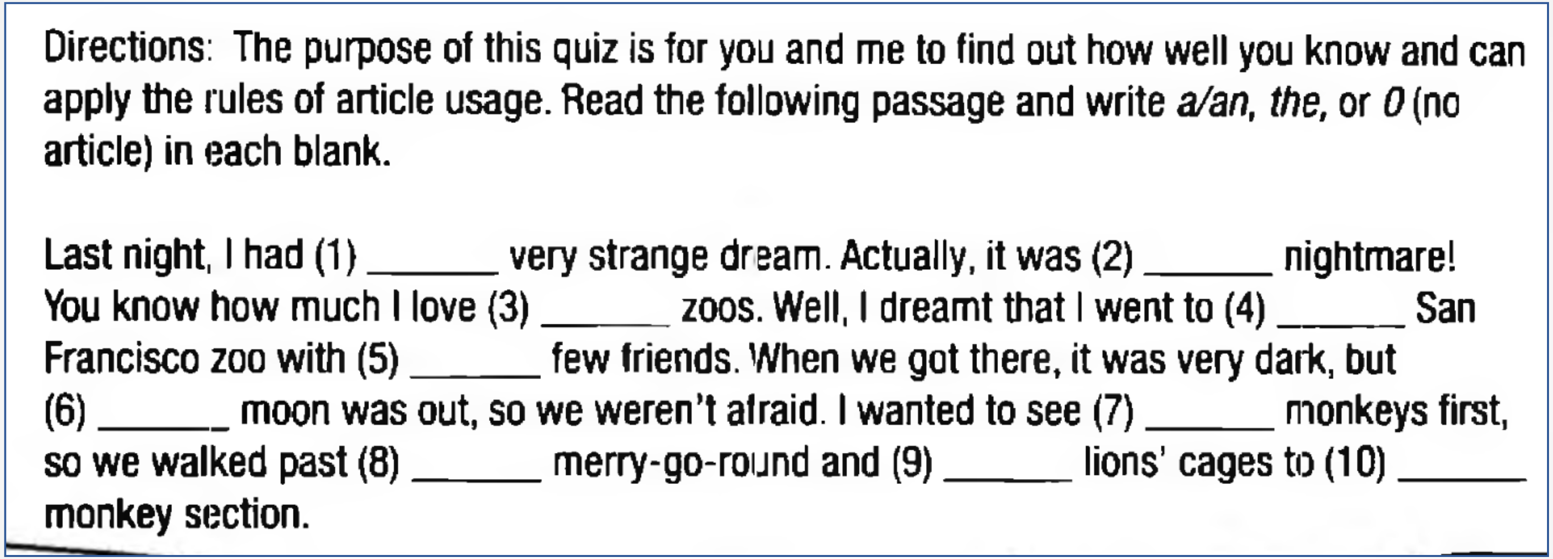
Example of a content-valid test

- A speaking test that requires the learner to actually speak within some sort of authentic context.

Examples of tests lacking content validity

- A speaking test that requires the learner to answer paper-and-pencil multiple choice questions.
- A test that covers only 2 objectives of a course with 10 objectives.

Conversation Test with Low Content Validity



Directions: The purpose of this quiz is for you and me to find out how well you know and can apply the rules of article usage. Read the following passage and write *a/an*, *the*, or *0* (no article) in each blank.

Last night, I had (1) ______ very strange dream. Actually, it was (2) ______ nightmare! You know how much I love (3) ______ zoos. Well, I dreamt that I went to (4) ______ San Francisco zoo with (5) ______ few friends. When we got there, it was very dark, but (6) ______ moon was out, so we weren't afraid. I wanted to see (7) ______ monkeys first, so we walked past (8) ______ merry-go-round and (9) ______ lions' cages to (10) ______ monkey section.

Teaching: **The use of articles in conversation**

Test: **Paper- and- pencil test on the use of articles**

2. Criterion Validity

Criterion Validity

- The extent to which the "criterion" of the test has actually been reached.
- A comparison of results of an assessment with results of some other assessment measure of the same criterion

Example of a criterion valid test

- The results of a classroom oral test of voiced and voiceless stops done by a teacher, might be compared with an independent assessment of the same phonemic proficiency.

Types of Criterion Validity

1. **Concurrent validity:** a test results are supported by other concurrent performance beyond the assessment itself (ex. a classroom test results of some skill are compared to a commercially produced test results of the same skill).
2. **Predictive validity:** The assessment criterion is to assess (and predict) a test-taker's likelihood of future success (ex. placement tests, admission tests, and achievement tests that enable students to move on to another unit".

3. Construct Validity

Construct Validity

- does not play at large a role for classroom teachers.
- is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions.
- Ex. of linguistic constructs: Proficiency, communicative competence, & fluency.
- Ex. of psychological constructs: Self-esteem & motivation
- In language learning, every issue involves theoretical construct

Examples of tests with construct validity

- Oral fluency test that includes the components of fluency “speed, rhythm, juncture, hesitation, etc”.
- Oral interview assessment that includes pronunciation, fluency, grammar, vocabulary use, & socio-linguistic appropriateness.

Examples of tests lacking construct validity

- A speaking test that requires the learner to answer paper-and-pencil multiple choice questions.
- A test that covers only 2 objectives of a course with 10 objectives.
- Written vocabulary quiz which asks students to define a set of words they studied in a unit which lexical objective was the communicative use of vocabulary.

4. Consequential Validity (Impact)

- The consequences of a test/ assessment
- Includes a test's 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.
- Consequential validity referred to as the impact "many consequences of assessment, before and after test administration". This impact involves both a macro level "the effect on society and educational system" and a micro level "the effect on the individual test-takers" & "washback".
- Ex. the use of standardized tests for such gatekeeping purposes as college admission "deprive[s] students of crucial opportunities to learn and acquire productive language skills,' causing test consumers to be "increasingly disillusioned with EFL testing. (p.34)

5. Face Validity

Face Validity

- the extent to which students view the assessment as fair, relevant, and useful for improving learning. (student perception of the fairness of a test).
- the degree to which a test 'looks' right and 'appears' to measure the knowledge or abilities it claims to measure
- is based on the subjective judgment of the examinees, the administrative personnel, and other psychometrically unsophisticated observers.
- cannot be empirically measured or theoretically justified under the category of validity
- is viewed by some educators as a superficial factor that is too dependent on the whim of the perceiver.
- But it is significant since it affects the student's performance which lead to student unreliability.

5. Face Validity

How to increase students' perception of fair test (Face Validity)?

Use:

- a well-constructed, expected format with familiar tasks
- tasks that can be accomplished within an allotted time limit
- items that are clear and uncomplicated
- directions that are crystal clear
- tasks that have been rehearsed in their previous course work
- tasks that relate to their course work (content validity)
- a difficulty level that presents a reasonable challenge

Authenticity

Authenticity

- The degree to which a test task simulate real-world tasks.

Examples of authentic tasks

- Reading passages selected from real-world sources.
- Listening comprehension sections feature natural language with hesitations, white noise, and interruptions.
- Sequenced episodes of meaningful units, paragraphs, or stories.

Characteristics of an authentic test

- Contains language that is as natural as possible.
- Has items that are contextualized rather than isolated .
- Includes meaningful, relevant, interesting topics.
- Provides some thematic organization to items, such as through a story line or episode.
- Offers tasks that replicate real-world tasks.

Washback

Washback

- The effect of testing on teaching and Learning
- The promotion and the inhibition of learning

Assessments with washback effect

1. **Informal performance assessment** (The teacher usually provides interactive feedback)
2. **Formal tests** (no beneficial washback if the students receive a simple letter grade or a single overall numerical score).

Characteristics of a test with a washback effect

- Positively influences what and how teachers teach
- Positively influences what and how learners learn
- Offers learners a chance to adequately prepare
- Gives learners feedback that enhances their language development
- Is more formative in nature than summative
- Provides conditions for peak performance by the learner

Washback

A test with washback effect..

- serves as a learning device
- students’ incorrect responses are windows of insight into further work
- students’ correct responses are praised

Benefits of washback

- Enhances basic principles of language acquisition like intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment, & others

How to enhance a test washback?

- Comment generously and specifically on test performance. (letter grades and numerical scores without feedback fosters competitive, not cooperative, learning).
- Give praise for strength as well as constructive criticism of weaknesses.
- Give strategic hints on how a student might improve certain elements of performance

*Washback can promote an atmosphere of dialogue between students and teachers regarding evaluative judgments.

Reference

- Brown, H. D. & Abeywickrama, P. (2012). Language Assessment: Principles & Classroom Practices. 2nd Ed. Pearson Education, USA.