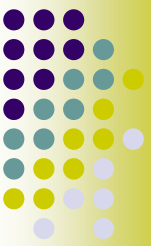
Unit 3
Test essential requirements: the issues of validity, reliability and practicality.



PRACTICALITY

1) Tests need to be practical, i.e. they need to be…

- **…within the means of financial limitations;**
- **…within time constraints, and**
- **…easy to administer, score, and interpret.**

Thus…

PRACTICALITY

- ...a test that is prohibitively expensive is impractical;
- ...a test of lg proficiency that would take students 10 hours to complete would be impractical;
- ...a test that requires individual one-to-one proctoring is impractical for a group of 500 test-takers and only a handful of examiners;

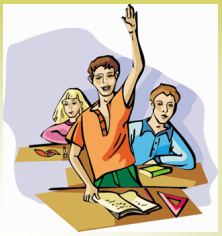
PRACTICALITY

- ...a test that takes students a few minutes to complete and several hours for the examiner to prepare and/or correct is impractical for a large number of testees and one examiner if results need to come out within a short time.

2) Tests need to have "instructional value", i.e. it should be possible to use the test to enhance the delivery of instruction in student populations.

PRACTICALITY

?



What does it mean that tests should have instructional value?

Since teaching & testing are interrelated, teachers should make clear & useful interpretations of test data (e.g. results) in order to understand their students & their lg learning processes better.

PRACTICALITY

The extent to which a test is practical may depend on whether it is designed to be norm-referenced or criterion-referenced.

NORM-REFERENCED TESTING

☛ **In norm-referenced tests, each test-taker's score is interpreted in relation to a mean, standard deviation, and/or percentile rank.**

PRACTICALITY

- The purpose in n-r tests is to place test-takers along a mathematical continuum in rank order.

- Typical of n-r tests are standardized tests intended to be administered to large audiences, with results quickly disseminated to test-takers (i.e. TOEFL).

PRACTICALITY

- Such tests must have fixed, predetermined responses in a format that can be electronically scanned. Practicality in these tests is of paramount importance.

CRITERION-REFERENCED TESTING

- Criterion-referenced tests, on the other hand, are designed to give test-takers feedback on specific course or lesson objectives (i.e. the "criteria").

PRACTICALITY

- Classroom tests involving smaller numbers & connected to a curriculum are typical of c-r testing.
- More time and effort on the part of the teacher/test-giver are usually required to deliver the feedback.
- It could be said that c-r testing considers practicality as a secondary issue in the design of the test; teachers sacrifice time & effort in order to offer students appropriate & useful feedback. That is, the instructional value of c-r testing is high.

RELIABILITY

- A reliable test is consistent & dependable.
- Sources of unreliability may be found in the test itself (test reliability) or in the scoring of the test (rater/scorer reliability).
- If you give the same test to the same subject on two different occasions, the test should yield similar results; it should have test reliability.

RELIABILITY

- Imagine a test with a listening comprehension component, which is delivered in a room with street noise coming in &, thus, impeding students to hear the text accurately. This is a clear case of test unreliability.
- Sometimes, tests can yield unreliable results for factors beyond the control of the test writer, such as illness on the part of the test-taker, a "bad day", no sleep on the night before the test, etc.

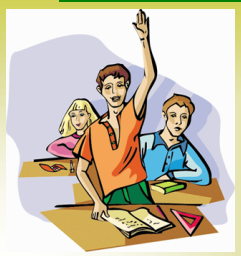
RELIABILITY

☛ Scorer reliability is the consistency of scoring by two or more scorers. The more subjective the scoring technique, the lower the scorer reliability.

In other words, if scoring directions are clear and specific as to the exact details the judge (test-rater or scorer) should attend to, then scoring can become reasonably consistent & dependable.

RELIABILITY

?



Is scorer reliability difficult or easy to achieve in tests of writing ability? Why?

It usually is difficult to achieve since writing proficiency involves numerous traits that are difficult to define. However, the careful specification of an analytical scoring instrument can increase scorer reliability.

RELIABILITY

How can test designers make tests more reliable?

- **Take enough samples of behavior.**
- **Exclude items which do not discriminate well between weaker & stronger students.**
- **Do not allow candidates too much freedom.**
- **Write unambiguous items.**
- **Provide clear and explicit instructions.**
- **Ensure that tests are well laid out and perfectly legible.**
- **Make candidates familiar with format and testing techniques.**

RELIABILITY

- Provide uniform & non-distracting conditions of administration.
- Use items that permit objective scoring.
- Make comparisons bt candidates as direct as possible.
- Provide a detailed scoring key.
- Train scorers.
- Agree acceptable responses & appropriate scores at outset of scoring process.
- Identify candidates by number, not name.
- Employ multiple, independent scoring.

VALIDITY

☛ What is VALIDITY?

✓ Validity can be defined as the degree to which the test actually measures what it is intended to measure.

Thus, a valid test of reading ability is one that actually measures reading ability and not, say, previous knowledge of a subject, or some other variable (e.g. writing ability).

VALIDITY

☛ How do we determine the VALIDITY of a test?

✓ By observation and theoretical justification, since there's no final, absolute, and objective measure of validity. We have to ask questions that yield convincing evidence that a test accurately and sufficiently measures the testee for the particular purpose of the test.
✓ Validity is by far the most complex criterion of a good test.

VALIDITY

E.g. 1

To measure writing ability, one might ask students to write as many words as they can in 15 min. Then count the words for the final score.
The test would be practical and reliable – easy to administer, and the scoring quite dependable.
However, it would hardly constitute a valid test of writing ability unless some consideration were given to communication & organization of ideas, among other factors.

VALIDITY

E.g. 2

A valid test of driving ability would include a sample of a person's behind-the-wheel behavior, wouldn't it?

However, in many places you only need to take a pencil-and-paper test to get your license renewed. Is this test valid?
It's at least doubtful that the written test actually predicts the quality of driving ability. This kind of test probably has little validity for predicting good driving; what it does measure is knowledge of regulations, a small part of total driving ability.

VALIDITY

In language tests, validity is supported by subsequent personal observation of teachers and peers.

The validity of a high score on the final exam of a FL will be substantiated by "actual" proficiency in the lg (assuming that it is true that a high score is indicative of high proficiency).

(Remember what was said on slide 16 about justification of validity)

VALIDITY

E.g.:

A classroom test designed to assess mastery of a point of grammar in communicative use will have validity if test scores correlate either with observed subsequent behavior or with other communicative measures of the grammar point in question.

VALIDITY

☛ **How can teachers be sure that a given test (whether standardized or constructed for classroom use) is valid?**

There are two major types of validation for classroom teachers:

- **content validity and**
- **construct validity.**

VALIDITY

✓ **A test that has content validity is one that actually involves the test-taker in a sample of the behavior that is being measured.**

E.g.: If you want to assess a person's ability to speak a SL in a conversational setting, a test which asks the learner to answer paper-and-pencil multiple-choice questions requiring grammatical judgements does not achieve content validity. But a test that requires the learner actually to speak within some sort of authentic context does.

VALIDITY

A concept that is very much related to content validity is face validity, which asks the question:

does the test, on the "face" of it, appear to test what it is designed to test?

VALIDITY

☛ **What is FACE VALIDITY?**

✓ **A test is said to have face validity when it looks as if it measures what it is supposed to measure.**

Thus, a test that pretended to measure pronunciation ability but did not required the test-taker to speak might be thought to lack face validity.

✓ **Face validity is important bc tests which lack it may not be accepted by candidates, teachers, education authorities, or employers. One such test may simply not be used.**

VALIDITY

And, if used, the candidates' reaction to it may mean that they do not perform on it in a way that truly reflects their ability.

In other words, to achieve peak performance on a test, a test-taker needs to be convinced that the test is indeed testing what it claims to test.

✓ **Face validity is almost always perceived in terms of content: if the test samples the actual content of what the learner has achieved or expects to achieve, then f.v. will be perceived.**

VALIDITY

☛ What is CONSTRUCT VALIDITY?

✓ One way to look at c.v. is to ask: does this test actually tap into the theoretical construct as it has been defined?

✓ Proficiency = construct

✓ Communicative competence = construct

etc.

VALIDITY

✓ Tests are, in a manner of speaking, operational definitions of such constructs in that they operationalize the entity which is being measured.

✓ Teachers, therefore, need to be satisfied that a particular test is an adequate definition of a construct.

✓ A general proficiency test that consists of, say, grammatical judgement items, reading comprehension items, and listening comprehension items is defining "proficiency" as consisting of those three modes of performance.

VALIDITY

- ✓ **Most tests constructed for classroom purposes (not standardized) can be validated adequately thru content –i.e. if the test samples the outcome behavior, then validity has been achieved.**
- ✓ **However, when content validity is low or questionable, construct validity becomes essential.**
- ✓ **Standardized tests built to be given to large numbers of students tend to suffer from poor content validity but are redeemed thru their construct validity.**

VALIDITY

- ✓ **Take the TOEFL. This test does not sample oral production, although there's no doubt that oral production is an important part of succeeding academically in a university course of study.**
- ✓ **The TOEFL's absence of oral production content is justified by research that has shown positive correlations bt oral production and the behaviors actually sampled on the TOEFL (listening, reading, grammaticality detection, and writing).**

VALIDITY

Because of the crucial need to offer a financially affordable proficiency test & the high cost of administering and scoring oral production tests, the omission of oral content from the TOEFL has been accepted as a necessity in the area of language learning testing.