Assessment and Evaluation Fourth Year

Dr. Dhea Mizhir

College of Education Ibn Rushd/ English Language Department

Types of Tests

- **Why do we need tests?**

  - Test scores → educational decisions

    inference

  - Test scores ← performance/ true ability

Backwash effect

- Impact of tests on teaching and learning
- Beneficial/harmful backwash effect
- Ex.

Testing and Assessment

- Assessment: tests, projects, observation of performance, portfolios, etc.
- Tests are one form of assessment

Formative vs. summative assessment

- Formative assessment:

  check progress of learning

- Summative assessment:

  end of program check

Types of tests (purposes)

- Proficiency tests
- Diagnostic tests
- Placement tests
- Achievement tests
- Aptitude tests
- Admission tests
- Progress tests
- Language dominance tests

Proficiency tests

- Measure general ability in a language
- Regardless of previous training

Diagnostic tests

- Identify students' strengths and weaknesses
- To benefit future instruction
- Difficult to construct. Lack of good ones.

Placement tests

- To assign students to classes/programs appropriate to their level of proficiency
- Define characteristics of each level of proficiency

Achievement tests

- Measure how successful students are in achieving objectives of a lesson/course/curriculum
- Closely related to the content of a particular lesson/course/ curriculum
- Syllabus content approach OR course objectives approach?
- Final achievement tests / progress achievement tests (formative assessment)
- Frequency?

Aptitude tests

- To predict a person's future success in learning a (any) foreign language
- Taken before actual learning

Admission tests

- to provide information about whether a candidate is likely to succeed

Progress tests

- tests—to assess students' mastery of the course material (during the course)

Language dominance tests

- to assess bilingual learners' relative strength of the 2 languages

Direct vs. indirect testing

- **Direct testing**:

  -Requires Ss to perform the skill to be measured

- **Indirect testing:**

  -Measures the abilities underlying the skills to be measured

  -Ex. A writing test that requires Ss to identify grammatical errors in sentences

- **Semi-direct testing:**

  -tape recorded speaking test

Problems

- Direct testing:

  -practicality (limited resources)

  -small sample of tasks

- Indirect testing:

  -nature of the trait to be measured

  -relationship b/w test performance and skills tested

Discrete point vs. integrative tests

- Discrete point tests:

  -Focus on one linguistic element at a time

  -Assumption: language can be broken down into separate element

  -tend to be indirect

- Integrative tests:

  -Requires to students to combine many linguistic elements

  -Unitary trait/competence hypothesis (Oller)

  -tend to be direct

  -Ex. Composition, dictation, cloze tests, note-taking

Norm v.s. Criterion-referenced tests

Test type	Criterion-Referenced Tests	Norm-Referenced Tests
Purpose	To classify students according to whether they have met the established standards	To show how a student's performance compares to that of other test-takers
Result	Percentage; descriptive	Percentile, grade equivalence
Features	Comparison with a set criterion. Direct info on what the Ss can do. More motivating. Cut-off score. Not affected by other test-takers' performance.	Comparison with other test-takers. Will be affected by others' performance.
Example		

Objective vs. subjective tests

- Scoring of tests
- Objective tests:

  -Requires no judgment from the scorer

  -Ex. Multiple choice, T/F tests
- Subjective tests:

  -Requires judgment from the scorer

  -Ex. Essay questions, composition
- Different degrees of subjectivity

History of language testing

- Prescientific period (b/f 1950s)

  GTM, reading-oriented methods
- Psychometric-structuralist period (1950s-1960s)

  structural linguistics, behavioral psychology, discrete point tests
- Integrative-sociolinguistic period (a/f 1960s)

  communicative language ability

Communicative competence

- Grammatical competence
- Discourse competence
- Sociolinguistic competence
- Strategic competence

Communicative language testing

- Communicative nature of tasks
- Authenticity of tasks

Computer Adaptive Testing (CAT)

- Saves time and effort
- Start with average level of difficulty, lower/increase levels of difficulty according to test taker's performance
- Needs a bank of items graded by difficulty

Characteristics of Language Assessment

Overview

- What are the characteristics of language testing?
- How can we define them?
- What factors can influence them?
- How can we measure them?
- How do they interrelate?

Reliability

Related to accuracy, dependability and consistency

According to Henning [1987], reliability is

- a measure of accuracy, consistency, dependability, or fairness of scores resulting from the administration of a particular examination e.g. 75% on a test today, 83% tomorrow – problem with reliability.

Validity: internal & external

Construct validity [internal]

- the extent to which evidence can be found to support the underlying theoretical construct on which the test is based

Content validity [internal]

- the extent to which the content of a test can be said to be sufficiently representative and comprehensive of the purpose for which it has been designed

Validity [2]

Response validity [internal]

- the extent to which test takers respond in the way expected by the test developers

Concurrent validity [external]

- the extent to which test takers' scores on one test relate to those on another externally recognised test or measure

Validity [3]

Predictive validity [external]

- the extent to which scores on test Y predict test takers' ability to do X e.g. IELTS + success in academic studies at university

Face validity [internal/external]

- the extent to which the test is perceived to reflect the stated purpose e.g. writing in a listening test – is this appropriate? depends on the target language situation i.e. academic environment

Validity [4]

- 'Validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put.'

Alderson [2002: 5]

Practicality

The ease with which the test:

- items can be replicated in terms of resources needed e.g. time, materials, people
- can be administered
- can be graded
- results can be interpreted

Factors which can influence reliability, validity and practicality…

Test

- time allowed
- clarity of instructions
- use of the test
- selection of content
- sampling of content
- invalid constructs

Test taker

- familiarity with test method
- attitude towards the test i.e. interest, motivation, emotional/mental state
- degree of guessing employed
- level of ability

Test administration

- consistency of administration procedure
- degree of interaction between invigilators and test takers
- time of day the test is administered
- clarity of instructions
- test environment – light / heat / noise / space / layout of room
- quality of equipment used e.g. for listening tests

Scoring

- accuracy of the key e.g. does it include all possible alternatives?
- inter-rater reliability e.g. in writing, speaking
- intra-rater reliability e.g. in writing, speaking
- machine vs. human

How can we measure reliability?

Test-retest

- same test administered to the same test takers following an interval of no more than 2 weeks

Inter-rater reliability

- two or more independent estimates on a test e.g. written scripts marked by two raters independently and results compared

Measuring reliability [2]

Internal consistency reliability estimates

e.g.

- Split half reliability
- Cronbach’s alpha / Kuder Richardson 20 [KR20]

Split half reliability

- test to be administered to a group of test takers is divided into halves, scores on each half correlated with the other half
- the resulting coefficient is then adjusted by Spearman-Brown Prophecy Formula to allow for the fact that the total score is based on an instrument that is twice as long as its halves

Reliability is influenced by …..

- the longer the test, the more reliable it is likely to be [though there is a point of no extra return]
- items which discriminate will add to reliability, therefore, if the items are too easy / too difficult, reliability is likely to be lower
- if there is a wide range of abilities amongst the test takers, test is likely to have higher reliability
- the more homogeneous the items are, the higher the reliability is likely to be

Construct validity

- evidence is usually obtained through such statistical analyses as factor analysis [looks for items which group together], discrimination; also through retrospection procedures

Content validity

- this type of validity cannot be measured statistically; need to involve experts in an analysis of the test; detailed specifications should be drawn up to ensure the content is both representative and comprehensive

Response validity

- can be ascertained by means of interviewing test takers [Henning]; asking them to take part in introspection / retrospection procedures [Alderson]

Concurrent validity

- determined by correlating the results on the test with another externally recognised measure. Care needs to be taken that the two measures are measuring similar skills and using similar test methods

Predictive validity

- can be determined by investigating the relationship between a test taker's score e.g. on IELTS/TOEFL and his/her success in the academic program chosen
- problem - other factors may influence success e.g. life abroad, ability in chosen field, peers, tutors, personal issues, etc.; also time factor element

Reliability vs. validity?

- 'an observation can be reliable without being valid, but cannot be valid without first being reliable. In other words, reliability is a necessary, but not sufficient, condition for validity.'

[Hubley & Zumbo 1996]

- ‘Of all the concepts in testing and measurement, it may be argued, validity is the most basic and far-reaching, for without validity, a test, measure or observation and any inferences made from it are meaningless’

[Hubley & Zumbo 1996, 207]

Reliability vs. validity [2]

- even an ideal test which is perfectly reliable and possessing perfect criterion-related validity will be invalid for some purposes

[Henning 1987]

Practicality

Designing and developing good test items requires

- working with other colleagues
- materials i.e. paper, computer, printer etc.
- time

Some items look very attractive but this attraction has to be weighed against these factors.

References

- Alderson, J. C 2002 *Conceptions of validity and validation.* Paper presented at a conference in Bucharest, June 2002.

- Angoff, 1988 Validity: An evolving concept. In H. Wainer & H. Braun [Eds.] *Test validity* [pp. 19-32], Hillsdale, NJ: Erlbaum.

- Bachman, L. F. 1990 *Fundamental considerations in language testing*. Oxford: O.U.P.

- Cumming A. & Berwick R. [Eds.] *Validation in Language Testing* Multilingual Matters 1996

- Hatch, E. & Lazaraton, A. 1991 *The Research Manual - Design & Statistics for Applied Linguistics* Newbury House