MEETING I

TESTING, ASSESSING AND TEACHING

WHAT IS A TEST?

- Method of measuring a person’s ability, knowledge, or performance in a given domain.

component of a TEST

Method → instrument (a set of technique

Measure → data

individual ability (trait) → competence

Test and Data

- Test: An instrument or activity used to accumulate data on a person’s ability/competence to perform a specified task. In kinesiology the content of these tests are usually either cognitive, skill, or fitness.
- Data: The translation of behavior into a numerical or verbal descriptor which is then recorded in written form.

Why Administer Tests?

- To measure individual differences on a specific trait (behavior).

Discussion:

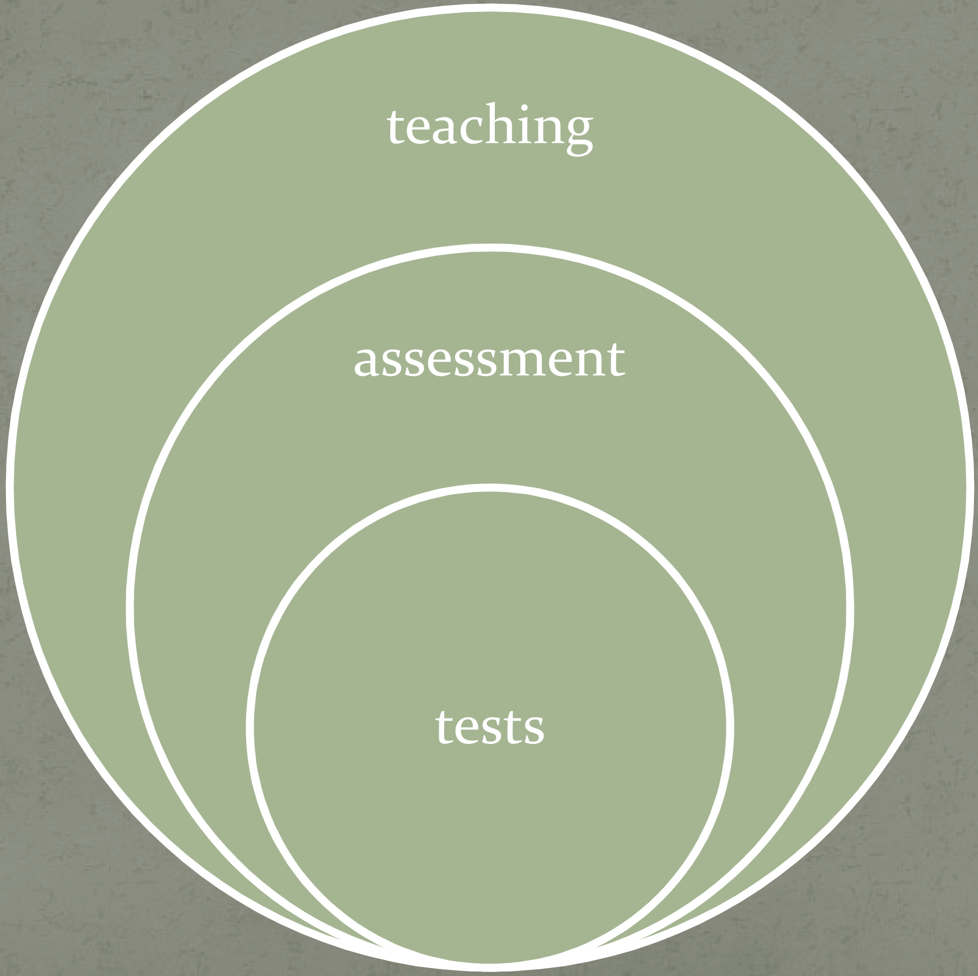
- Is a test "good" if everyone/anyone scores 100%? Or, is a test "good" if everyone/anyone scores 0%?

- To test: To measure **individual differences** on a specific trait (behavior).

Use of Tests

- Reflection
- Evaluation
- Motivation
- Achievement
- Improvement
- Diagnosis
- Prescription
- Grading
- Classification
- Prediction

ASSESSMENT AND TEACHING



INFORMAL AND FORMAL ASSESSMENT

Informal Assessment: no recording of the result

- Unplanned comments and responses
- Comments on students' writing

Formal assessment: record of the result

- Tests
- Journal
- Portfolio

Discussion

- One day Mr. Jono came to the class to give the students a writing assignment. First, he asked them to brainstorm their ideas. At the end of this first process, he wrote come notes on the result of their brainstorming. Some of the students got “Good Job”, but some got “Some more ideas please”. In this first process, he did not give any score to their writing.
- Next, he asked the students to make the outline for their compositions. He gave them 60 minutes to complete their outline. All the students would have to finish their outline in this process. After 60 minutes, he collected their outlines and scored them. Some students got 98 but some still got 60.

FORMATIVE AND SUMMATIVE ASSESSMENT

Formative Assessments

- Conducted during the process of forming students' competencies: What is it for?
- To motivate, improve students' competencies
- Examples:?

Summative Assessments

- Conducted (typically) at the end of a course.
- To measure, summarize students' competencies after a certain period of learning
- Examples?

NORM-REFERENCED AND CRITERION REFERENCED TESTS

NORM-REFERENCED TESTS

- Students' score is interpreted in relation to the mean (compared to the central tendencies)
- Example:?

Criterion-referenced tests

- Compared to the scores in the class/linked to the agreed criteria that if met show that a learner is ready to proceed to the next learning activity
- Example:?

APPROACHES TO LANGUAGE TESTING

- DISCRETE-POINT AND INTEGRATIVE TESTING

Discrete-point

- Language can be broken down into its components parts (skills, phonology, syntax, etc) and can be tested separatedly.

Integrative testing

- Indivisible view of language proficiency, that language skills and the vocabulary, grammar, etc. are integrated.

COMMUNICATIVE LANGUAGE TESTING

- What is it?
- That language tests should correspond to the demonstrable ways to language use in non-test situation.

PERFORMANCE-BASED ASSESSMENT

- Rely more on students' performance in doing the tasks than on formally structured tests.

CURRENT ISSUES IN CLASSROOM TESTING

NEW VIEWS OF INTELLIGENCE

- IQ ------ → 5 frames or even multiple intelligence
  - Spatial intelligence
  - Musical intelligence
  - Bodily-kinesthetic intelligence
  - Interpersonal intelligence
  - Intrapersonal intelligence

  EQ------ESQ------ what other .....Q’s?

TRADITIONAL AND ALTERNATIVE ASSESSMENT

Traditional assessment	Alternative assessment
1. One-shot, standardized exams 2. Timed, multiple-choice format 3. Decontextualized test items 4. Scores suffice for feedback 5. Norm-referenced scores 6. Focus on the ”right” answers 7. Summative 8. Oriented to product 9. Non-interactive performance 10. Fosters extrinsic motivation	1. Continuous long-term assessment 2. Untimed, free-response format 3. Contextualized communicative tasks 4. Individualized feedback and washback 5. Criterion-referenced scores 6. Open-ended, creative answers 7. Formative 8. Oriented to process 9. Interactive performance 10. Fosters intrinsic motivation

COMPUTER BASED TESTING

Important points

- Periodic assessments, both formal and informal can increase motivation by serving as milestones of students progress.
- Appropriate assessments aid in the reinforcement and retention of information.
- Assessments can confirm areas of strength and pinpoint areas needing further work.
- Assessments can provide a sense of periodic closure to modules within a curriculum.
- Assessments can promote student autonomy by encouraging students self evaluation of their progress.
- Assessments can spur learners to set goal for themselves.
- Assessments can aid in evaluating teaching effectiveness.

Principles of language testing

Overview

- What are the principles 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 e.g. 20°C here today, 20°C in North Italy – are they the same?

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]

AUTHETICITY

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 [1]

- quality of items
- number of items
- difficulty level of items
- level of item discrimination
- type of test methods
- number of test methods

Test [2]

- 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

Cronbach's Alpha [KR 20]

- this approach looks at how test takers perform on each individual item and then compares that performance against their performance on the test as a whole

- measured on a -1 to +1 scale like discrimination

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

How can we measure validity?

According to Henning [1987]

- non-empirically, involving inspection, intuition and common sense
- empirically, involving the collection and analysis of qualitative and quantitative data

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

References [2]

- Henning, G. 1987 *A guide to language testing: Development, evaluation and research* Cambridge, Mass: Newbury House
- Hubley, A. M. & Zumbo, B. D. A dialectic on validity: where we have been and where we are going. *The Journal of General Psychology 1996. 123[3] 207-215*
- Messick, S. 1988 The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun [Eds.] *Test validity* [pp. 33-45], Hillsdale, NJ: Erlbaum.
- Messick, S. 1989 Validity. In R. L. Linn [Ed.] *Educational measurement.* [3rd ed., pp 13-103]. New York: Macmillan.

Item-total Statistics

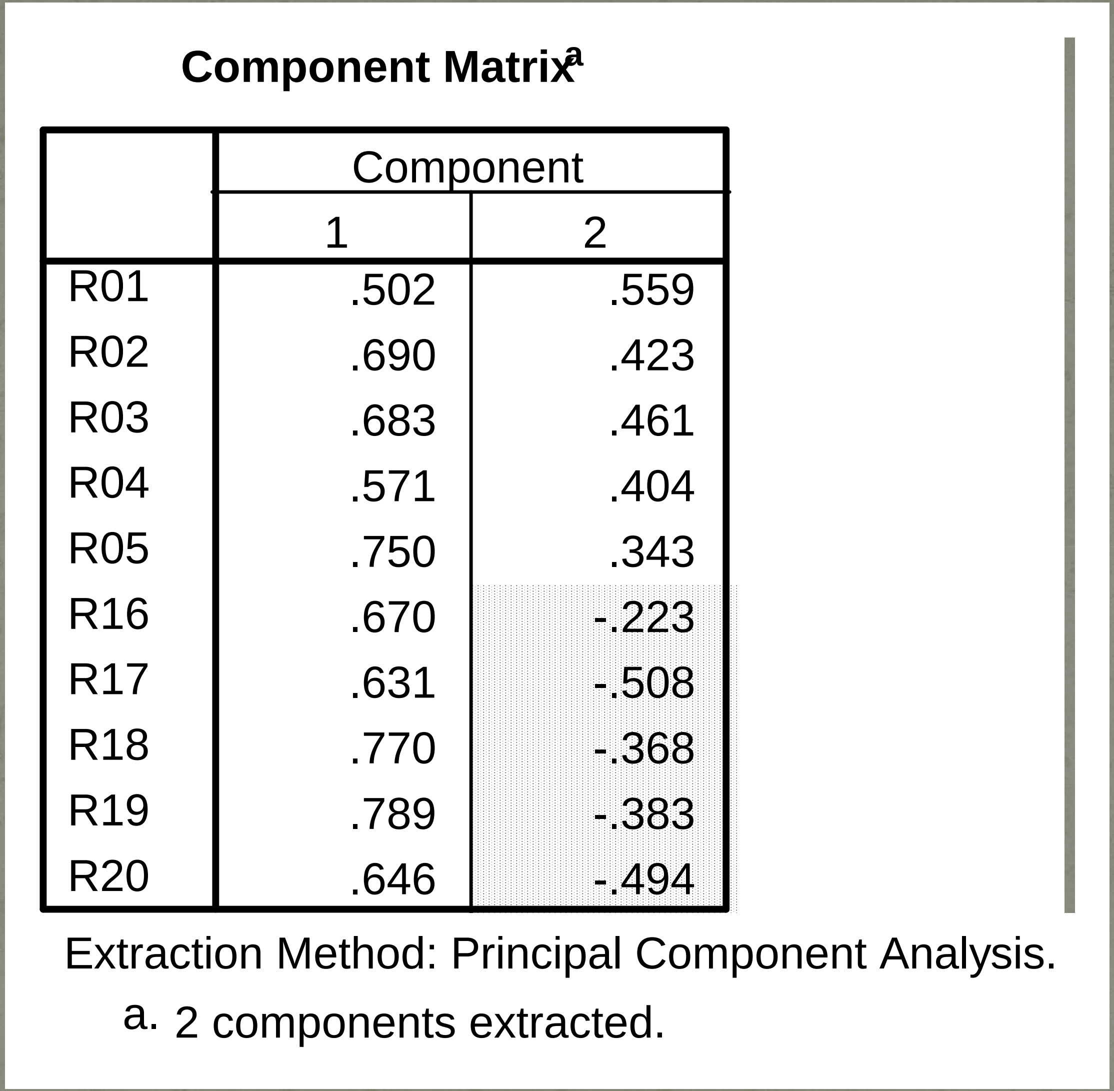
	Corrected Item-Total Correlation	Alpha if Item Deleted
R01	.5259	.7964
R02	.6804	.7594
R03	.6683	.7623
R04	.5516	.7940
R05	.7173	.7489
R16	.3946	.8288

N of Cases = 194.0 N of Items = 6 Alpha = .8121

Item-total Statistics

	Corrected Item Total Correlation	Alpha if Item Deleted
R16	.5773	.7909
R17	.5995	.7863
R18	.7351	.7553
R19	.7920	.7419
R20	.6490	.7753
R01	.1939	.8663

N of Cases = 194.0 N of Items = 6 Alpha = .8185



Component Matrix[a]

	Component	
	1	2
R01	.502	.559
R02	.690	.423
R03	.683	.461
R04	.571	.404
R05	.750	.343
R16	.670	-.223
R17	.631	-.508
R18	.770	-.368
R19	.789	-.383
R20	.646	-.494

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

DESIGNING CLASSROOM LANGUAGE TESTS

3RD MEETING

SOME CRITICAL QUESTIONS

- 1. What is the purpose of the test?
  - Placement or achievement, or…….
- 2. What are the objectives of the test?
  - What skills/competencies to measure
- 3. How will the test specifications reflect both the purpose and the objective
- 4. How will the test tasks be selected and the separate items arranged?
- 5. What kind of scoring, grading, and/or feedback is expected?

Test types

- Language aptitude tests
    - Measure capacity or general ability to learn a foreign language and **ultimate success** in that undertaking.
- Proficiency tests
    - Measuring global competence in language (all the skills)
- Placement tests
    - To place a student into a particular level or section of language curriculum or school.
- Diagnostic tests
    - To diagnose specified aspects of a language, e.g. to diagnose difficulties found by students in a certain skill.
- Achievement tests
    - Related directly to classroom lessons, units, or even a total curriculum.

SOME PRACTICAL STEPS TO TEST CONSTRUCTION

1. Assessing clear, unambiguous objectives
   - A teacher said that he wanted to test his students listening skills. Is it a clear objective? So, what is a good test objective?

- Task 1
- Read the SK/KD in SMP/SMA. Select some objectives for integrated-skills course. You can find the example on page 50.

2. Drawing up test specification
   - Test specifications: drawn from the selected objectives
   - Example: page 52

Task 2:

Write your test specification. Refer to the objectives you have chosen before.

3. Devising Test Tasks

Drafting the questions that you are going to use in a test.

Things to consider:

a. clear directions to each section

b. examples: need or not

c. Each item measures a certain objective.

d. Each item is stated in clear, simple language.

e. Each item in the multiple choice section should have good distractors.

f. Each item should have the right level of difficulty.

g. Each item should have sufficient authentic language.

h. The sum of all the items should reflect the learning objectives.

DESIGNING MULTIPLE-CHOICE QUESTIONS

- Strengths?
- Weaknesses?
- Characteristics:
  - All receptive or selective response item
  - Each item consists of a stem and some alternatives.
  - For each item, there is only one key, while the rest of the alternatives are distractors.

Steps to write a multiple-choice test

- Design each item to measure a specific objective.
- State both stem and options as simply and directly as possible.
- Make sure that the intended answer is clearly the only correct option.
- Use item indices to accept, discard, or revise items.

Writing a multiple-choice test

Develop your own 10 item-English multiple choice test by following the steps to design a multiple choice item. Here is an example.

Item	Skill & Objective	Stem	Alternatives
1	*Listening* Comprehension of WH-questions	When did the students have the final exam?	a. Yes, they did. b. Yesterday c. In the English class d. With Mr. Jono