Introduction to Testing and Assessment

Instructor: B. Dendrinos

Is testing a 'good' or a 'bad' thing?

- Language educators (and not only) are divided into two camps: the **teachers** and the **testers**.
- Teachers often say things like:
  - *Let's learn to teach before we learn to test*
  - *We deal with people, testers deal with statistics.*
- Testers think that teachers
  - tend to be unspecific about their aims and objectives
  - Are disinterested in finding out whether goals and objectives have been met

Can we do without teaching or testing?

◆ Probably yes because

  ▪ learning can occur in spite of teaching and/or **testing**, despite any kind of **formal evaluation**

  ▪ the outcomes of teaching can be **assessed** without any form of testing

  ▪ testing may be used **to measure** what people already know

Is testing synonymous with the terms below?

- ◆ **Evaluation**
  - ▪ Evaluation may focus on the effectiveness or impact of a program of instruction, examination or project. Students are usually not asked to evaluate while teachers carry out or take part in evaluation only in some contexts. 'Experts' or the authorities are most commonly legitimized to carry out formal evaluation
- ◆ **Measurement**
  - ▪ Measurement is the process of determining the amount or length of something when compared with a fixed unit (e.g. using a ruler to measure length). In language teaching measurement constitutes the quantification of language proficiency. Aspects of language knowledge, specific abilities and skills are *measurable* when there are transparent criteria and precise analysis of data.

Is assessment synonymous with testing?

◆ No, it is not.

◆ Assessment is a more encompassing term than testing.

◆ It is the process of gathering, interpreting, and sometimes recording and using information about students' responses to an educational task in order to provide the next learning step.

◆ Assessment is primarily concerned with providing teachers and/or students with feedback information.

◆ In language teaching, it is a local or global procedure though which one can appraise one or more aspects of language proficiency.

◆ Assessment is transparent when clear assessment criteria have been predetermined.

Is there one form of assessment?

◆ There are different forms of assessment, including:

■ Formative assessment

■ Summative assessment

■ Self-assessment

■ Peer assessment

Which are the most common forms of assessment?

◆ **Continuous assessment** refers to the activities required by students during the conduct of a course. It takes place within the normal teaching period and contributes to the final assessment.

◆ **Formative assessment** refers to observations which allow one to determine the degree to which students know or are able to perform a given task. It involves all those activities (assigned by teachers and performed by students) which provide information used as feedback so that teaching may meet students' needs. It can also include teacher assessment, feedback and feed-forward.

◆ **Summative assessment** is usually carried out at the conclusion of a unit or units of instruction, activity or plan, in order to assess acquired knowledge and skills at that particular point in time. It usually serves the purpose of giving a grade or making a judgment about the students' achievements in the course.

Are there other forms of assessment?

◆ Less frequent but increasingly important forms are:
  - **Self-assessment** occurs when an appraisal instrument is self-administered for the specific purpose of providing performance feedback, diagnosis and prescription recommendations rather than a pass/fail decision. Students engage in a systematic review of their progress and achievement, usually for the purpose of improvement. It may involve comparison with an exemplar, success criteria, or other criteria. It may also involve critiquing one's own work or a description of the achievement obtained.
  - **Peer assessment** occurs when students judge one another's work on the basis of reference criteria. This can occur using a range of strategies. The peer assessment process needs to be taught and students need to be supported by opportunities to practice it regularly in a supportive and safe (classroom) environment.

Does assessment include testing?

- Yes, it does.

- Testing is a particular kind of assessment which focuses on eliciting a specific sample of performance. The implication of this is that in designing a test we construct specific tasks that will elicit performance from which we can make the inferences we want to make about the characteristics of students, groups or individuals.

How do we test?

◆ There are different sorts of testing, including:

- ■ Achievement testing
- ■ Communicative testing
- ■ Competence testing
- ■ Diagnostic testing
- ■ Integrative testing
- ■ Performance testing
- ■ Progress testing
- ■ Proficiency testing
- ■ Psychometric testing

Which kind of testing is the most common?

◆ **Achievement testing**. It is used to determine whether or not students have mastered the course content and how they should proceed. The content of achievement tests, which are commonly given at the end of the course, is generally based on the course syllabus or the course textbook.

◆ **Progress testing**. It is used at various stages throughout a language course to determine learners' progress up to that point and to see what they have learnt.

◆ **Proficiency testing**. It is used to measure learners' general linguistic knowledge, abilities or skills without reference to any specific course.

- Some proficiency tests are intended to show whether students or people outside the formal educational system have reached a given level of general language ability.
- Others are designed to show whether candidates have sufficient ability to be able to use a language in some specific area such as medicine, tourism etc. Such tests are often called *Specific Purposes tests*.

Which kind of testing is the least common?

- **Diagnostic testing**, which seeks to identify those areas in which a student needs further help. These tests can be fairly general, and show, for example, whether a student needs particular help with one of the four language skills; or they can be more specific, seeking to identify weaknesses in a student's use of grammar.

- **Psychometric testing**, which is aimed at measuring psychological traits such as personality, intelligence, aptitude, ability, knowledge, skills which makes specific assumptions about the nature of the ability tested (e.g. that it is unidimensional and normally distributed). It includes a lot of discrete point items.

What do tests do?

◆ What a test will appraise or measure depends on what testers wish to know and what the testers believe a test to be. There is indeed a difference between:

◆ **Competence testing**, which is used to measure candidates' acquired capability to understand and produce a certain level of foreign language, defined by phonological, lexical grammatical, sociolinguistic and discourse constituents. In order to make test-takers' competence measurable and visible, testers turn of necessity to their actual performance which may indicate their competence.

◆ **Performance testing**, which includes direct, systematic observation of an actual student performance or examples of student performances and rating of that performance according to pre-established performance criteria. Students are assessed on the result as well as the process engaged in a complex task or creation of a product. / A performance test measures performance on tasks requiring the application of learning in an actual or simulated setting. Either the test stimulus, the desired response, or both are intended to lend a high degree of realism to the test situation.

Do all language tests aim at measuring communicative competence?

- No. They may test aspects of language knowledge and skills which are considered to be indicators of communicative competence.
- So, are all types of tests 'communicative'?
- Tests identified as 'communicative' are those which are interaction-based, open-ended (that is, responses cannot be predicted as in natural communicative environments), authentic, behavior-based and so on.
- Communicative tests are supposed to measure communicative competence which includes:
  - linguistic competence
  - sociolinguistic competence
  - strategic competence

Do tests test one or many things at a time?

◆ There are two different types of tests:

- **Integrative tests**, which include activities that assess skills and knowledge in an integrated manner (e.g., reading and writing, listening and speaking). Less attention is paid to specific lexicogrammatical points.

- **Discrete point tests**, which contain items that ideally reveal the candidate's ability to handle one level of language and one element of receptive or productive skills.

For whom are tests important?

◆ For almost all the people involved in the education process:

- the **learner** who wants to know how well s/he is doing, and also wants the 'piece of paper for professional and education purposes
- the **teacher** wants to know how the learner is progressing and whether and how well s/he herself is succeeding in his job
- the **parents**, who want to make sure that they're getting their money's worth
- **educational authorities** and others who have some interest in the learner's progress or his/her proficiency level
- the **potential employer** who relies heavily on what tests tell him/her about learner proficiency levels

Why else is testing important?

◆ Because of its backwash effect

◆ <u>What does this mean</u>? It is the effect that testing has on teaching. For better or worse, tests and exams exert control over what goes on in classrooms. This is because very many language classes are geared more or less directly to the tests or examinations the learners will end up taking. Teachers must often 'teach to' a test.

◆ <u>Is the quality of tests important for teaching</u>? Yes.
   ■ If the test is a bad one (or the teacher is too narrow in his/her interpretation of it), the result may be negative washback, where we can say that teaching suffers because of the test coming at the end of the course.
   ■ If the test is a good one, and its nature well understood by the teacher, the effect on the teaching may be very positive. There will be positive backwash.

Considerations when constructing a test

◆ There are two basic considerations when constructing a test. It must be **valid** and **reliable**. The first concept first:

◆ **Validity** is commonly defined as 'the extent to which [a test] measures what it is supposed to measure *and nothing else.* If a test is valid, the outsider who looks at an individual's score knows that it is a true reflection of the individual's skill in the area the test claims to have covered.

Kinds of validity

◆ **Content validity**. A test is said to have content validity if the items or tasks of which it is made up constitute a representative sample of items or tasks for the area of knowledge or ability to be tested (often related to a syllabus or a course).

◆ **Construct validity**. A test is said to have construct validity if the scores that a candidate gets on this relate in the same way to another test or form of assessment for the same aspect of knowledge.

◆ **Empirical validity**. A measure of the validity of a test arrived at by comparing the test with one or more criterion measures.

◆ **Face validity**. The extent to which a test appeals to candidates or to those choosing it on behalf of the candidates because it is considered to be an acceptable measure of the ability they wish to measure. It is sometimes referred to as 'test appeal'.

◆ **Predictive validity**. A type of validity based on the degree to which a test accurately predicts future performance. A language aptitude test for example, should have predictive validity because the results of the test should predict the ability to learn a foreign language.

◆

Important consideration in testing

◆ Reliability is another very important consideration when testing.

◆ **Relibility** refers to the consistency of a test. That is, if every time the test is administered it will have the same outcome. But reliability does not have to do with the content of the test alone; it has to do with marking in two ways:

  ■ ensuring that different raters give comparable marks to the same script

  ■ the same raters give the same marks on two different occasions to the same script

Kinds of reliability

Reliability is most often estimated with regard to:

- The **internal consistency** in a test; that is, if there is correlation among the variables comprising the test

- The results when **testing and re-testing**; that is, if there is correlation between two (or more) administrations of the same item, scale, or instrument for different times, locations, or populations, when the two administrations do not differ in other relevant variables

- **Inter-rater reliability**, which refers to the level of agreement between two or more evaluators/ judges/ raters on a particular instrument at a particular time. They are to apply their marks in a manner that is predictable and replicable. Therefore, note that inter-rater reliability is a property of the testing situation, and not of the instrument itself.

More about tests and testing

◆ How does one define what will be tested?

◆ How can tests and feedback provided be a positive asset in the educational process?

◆ What types of feedback can teachers provide to test-takers and how?

◆ Are tests the best tools for evaluation and assessment?

◆ What are the most important things to remember about raters and marking (closed and open-ended items) in integrative and discrete point tests?

Statements about feedback: True or false?

- The fact that the teacher gives feedback on student performance implies a power hierarchy: the teacher above, the student below.

- Assessment is potentially humiliating to the assessed person.

- Teachers should give their students only positive feedback, in order to encourage, raise confidence and promote feelings of success; negative feedback demoralizes.

- Giving plenty of praise and encouragement is important for the fostering of good teacher-student relationships.

- Very frequent approval and praise lose their encouraging effect; and lack of praise may then be interpreted as negative feedback.

- Teachers should not let students correct each other's work, as this is harmful to their relationships.