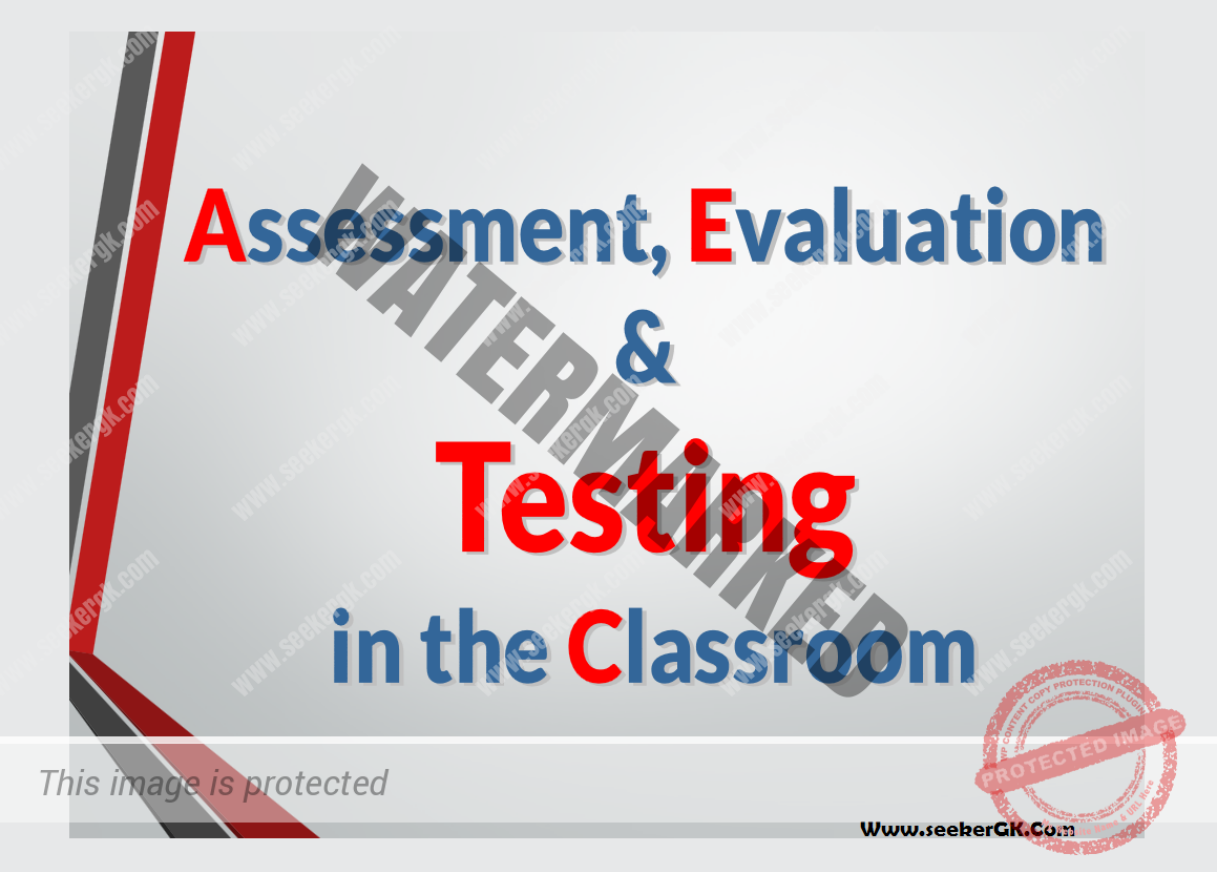

What are the difference among *Evaluation, Assessment and Testing.*

The concepts of evaluation, assessment and testing seem to differ for different authors.
For the purpose of this unit, we will use the terms as defined by **Harris and Mc. Cann(1994)**

Evaluation:

This concept involves looking at *all the factors* that influence the learning process, ex: syllabus objectives, course design, materials, methodology, teacher performance and assessment.

Assessment:

It involves measuring the performance of our students and the progress that they are making. It helps us to be able to diagnose the problems they have and to provide them with useful feedback.

This assessment can be of three kinds:
1) Informal assessment
2) Formal assessment (testing)
3) Self-assessment

1. **Informal assessment:**

   It is the observation of everyday performance. It is a way of collecting information about our students' performance in normal classroom conditions. It is done without establishing test conditions such as in the case of formal assessment. We intuitively assess them when speaking, writing, reading or listening. We can see which students are doing well and which students are having difficulties. We are also aware of their attitudes and effort.

2. **Formal Assessment:**

   This is synonymous of "testing". And there are two possible interpretations of it.
   1) It refers to what are often called examinations.
   This examination are often external (KET, PET, etc). They are administered to many students under standardized conditions. They assess a broad range of language. They are marked objectively or under standardized subjective marking schemes and are likely to be administered at the end of a course.

Other authors include all types of language tests under this term. This tests include the kind of tests commonly administered in class by the teacher, in order to assess learning. These tests are not so formal as the examinations of external bodies and their scope of action is limited to the context in hand. These tests are often administered to one class, for purposes internal to the class; they focus on a narrow range of language; they are assessed either objectively or subjectively; they are done to assist teaching and are often backward looking.

- **3. Self Assessment:**

  It refers when the students themselves assess their own progress.
  Dickinson(1997) says it is particularly appropriate:
  a) as a complement to self-instruction.
  b) to build autonomous and self-directed language learners.

c) It could give learners an opportunity to reflect on his/her learning in order to improve it.

Do you see any disadvantages on it? L

Both Dickinson and Mac Cann point the problems associated with the use of self-assessment in classrooms:
a) It cannot work in situations where marks have great intrinsic value and there is competition.
b) the time required to train students to use self-assessment can be significant.
c) Reliability problems. Can students make adequate, fair assessment of their performance? Will many students be tempted to give themselves unfairly high assessments of their performance?

In formal assessment, we also have the terms summative and formative introduced by Scriven (1967:43)
A) *Formative*: this refers to forms of assessment which aim to evaluate the effectiveness of learning at a time during the course (quizzes), in order to make future learning more effective.

b) *Summative*: the administration of this test may result in some judgement on the learner, such as 'pass' or 'fail'. The amount of contents assessed are usually several.

Formal Assessment can also refer to test types according to purpose.
The main types are listed below:
1) Aptitude tests
2) Placement tests
3) Diagnostic tests
4) Progress tests
5) Achievement tests
6) Proficiency tests

1. **Aptitude Tests:**
   These are designed to predict who will be a successful language learner and are based on the factors which are thought to determine an individual's ability to acquire a second or foreign language.
   They are usually large scale tests taking a long time to administer and with a number of components, each testing a different facet of language. They are also forward-looking tests, concerned with future language learning.
2. **Placement tests:**
   These tests are used to make decisions regarding the students' placement into

appropriate groups. They tend to be quick to administer and to mark. They are usually administered at the start of a new phase or language course. As a result, students are often put into homogenous groups for language study according to their present language ability.

3. **Diagnostic Tests:**
   These tests are usually syllabus based and they aim to determine the students' areas of strength and weaknesses in relation to the contents to be covered in the course.
4. **Progress Tests:**
   These tests are usually written and administered by a class teacher, and look back over recent work, perhaps the work of the last lesson or week. They usually therefore test a small range of language. (pop quizzes)
5. **Achievement Tests:**
   This tests come at the end of a relatively long period of learning, and whose content derives from the syllabus that has been taught over the period of time. They are usually large scale tests, covering a wide range of language and skills. These tests can be used for a variety of purposes, including promotion to a more advanced course, certification, or as an entry qualification to a job.
6. **Proficiency Tests:**
   These tests are based on a theory of language proficiency and the specific language abilities to constitute language proficiency. They are often related to specific academic or professional situations where English is needed. (PET, FCE, CAE, IELTS, TOEFL, etc)
7.

Three phases to categorize formal tests and compare them:

1) First generation tests.

2) Second generation tests.

3) Third generation tests.

1. **First generation tests:**

   These are broadly associated with the grammar translation approach to language learning. Candidates are asked to complete various questions such as compositions, translations, or simple questions and answer activities devoid of context.
   Ex: Write about a holiday you enjoyed. (200 words). These tests evaluate grammar,

vocabulary, punctuation, spelling and discourse structure. They lead to subjective scoring so this can lead to problems of reliability in marking.

The degree of agreement between 2 examiners about a mark for the same language sample is known as inter-rater reliability. The degree of agreement between one single examiner marking the same sample on 2 separate occasions is known as intra-rater reliability. *Both inter- and intra- rater reliability is low in first generation tests*.

- **Second generation tests:**

  Where first generation testing techniques had been marked subjectively, with the associated problems in standardizing marking to ensure fairness, language items could be assessed objectively through multiple choice testing of discrete language items. The text could be marked by a non-expert, by different people, or by the same person more than once, and the result would always be the same.

Questions in second generation testing normally measure one item of language, known as **discrete point**. Since each question tests one tiny aspect of language (ex: verb form, prepositions, etc), tests are often very long, so these tests are criticized because they do not sample **integrative language** as first generation tests.

- **Third generation tests:**

  The testing of integrative language, with the use of both objective and subjective testing formats, has come together in third generation tests. These are those tests which have come along the back of developments in communicative language teachings. Thus, communicative tests aim to emulate real life language use. Recent models of communicative language ability propose that it consists of both knowledge of language and the capacity for implementing that knowledge in communicative language use.

Examples of these tests could be authentic reading with some transfer of information such as correcting some notes taken from it, or writing a note with instructions about some aspect of household organization, or listening to an airport announcement to find the arrival time of a plane, or giving someone spoken instructions for how to get to a certain place. Third generation techniques are contextualized by their very nature as authentic. Candidates are asked to do tasks which have clear reference in reality. These tests assess integrative language, so they have to be assessed subjectively.

West (1990) gives a good summary of these principles of testing. The principles can be described in pairs:

1) Competence v/s Performance
2) Usage v/s Use
3) Direct v/s Indirect Assessment
4) Discreet Point v/s Integrative Assessment
5) Objective v/s Subjective Assessment
6) Receptive v/s Productive Skills
7) Backward and Forward-looking Assessment
8) Contextualized v/s Disembodied Language
9) Criterion Referenced and Norm-Referenced Assess.
10) Reliability v/s Validity

The opposition between members of a pair indicates some sort of tension that exists in language testing in general; generally the more that one test confirms to one of the pair, the less likely it is to exhibit characteristics of the other part of the pair. Thus, the more reliable a tets (multiple choice), the less valid it is likely to be (it tests only discrete items). This opposition corresponds with the differences between second & third generation testing.

1. **Competence v/s Performance**:
   Chomsky drew this distinction between the ideal knowledge all mature speakers hold in their minds (competence) and the flawed realization of it that comes out in language use (performance).
   Third generation testing is often called “performance testing”

2. **Usage v/s Use**:
Widdowson distinguished between language use and language usage.
For example, learners whose instruction has consisted of grammatical rules, will be required to produce sentences to illustrate the rules. These sentences are for Widdowson, examples of usage. Examples of usage can show the learner’s current state of competence, but will not necessarily indicate anything about the learner’s possible performance. He argues that performance teaching and testing require examples of language use, not usage.

3. **Direct v/s Indirect Assessment**:
Testing that assesses competence without eliciting peformance is known as indirect testing. Multiple choice testing fits this decription, since language is assessed without any production of language use form the learner. Conversely, direct tests use examples of performance as an indicator of communicative competence. These tests use testing tasks of the same type as language tasks in the real world.

4. **Discrete Point v/s Integrative Assessment**:
Indirect assessment is usually carried out through a battery of many items, each one of which only tests one small part of the language. Each item is known as a discrete-point item. The theory is that if there are enough of them, they give a good indication of the learner's underlying competence. Thus, testers require items which test the ability to combine knowledge of different parts of the language, these items are known as integrative or global. Ex: answering a letter, filling in a form, etc.

5. **Objective v/s Subjective Assessment**:
Objective assessment refers to test items that can be marked clearly as right or wrong, as in a multiple choice item. Subjective assessment requires that an assessor makes a judgement according to some criteria and experience. Most integrative test elements require subjective assessment. The difficulty in subjective assessment arises in trying to achieve some agreement over marks, both between different markers and with the same marker at different times.

6. **Receptive v/s Productive Skills**:
The receptive skills (reading and listening) tend themselves to objective marking. The productive skills (speaking and writing) are generally resistant to objective marking. So third generation testers are placing great emphasis on achieving a high degree of standardisation between assessors through training in the application of band descriptions or rubrics.

7. **Backward and Forward-looking Assessment**:
Competence based tests look backwards at a usage-based syllabus to see what degree has been assimilated by the learner. Third generation tests are better linked to the future use of language (looking forward), and their assessments of real language use also show mastery of a performance based syllabus.

8. **Contextualised v/s Disembodied Language**:
Disembodied language has little or no context. This is more evident in items of multiple choice, based on language usage. The items bear little relevance to each other and act as examples of disembodied language with no purpose other as part of a test.
Integrative items need a full context in order to function. The closer the items in an integrative test are to simulating real world language tasks, the fuller the context must be.

9. **Criterion referenced and Norm Referenced Assessment**:
Norm-referenced tests compare students with an average mark or a passing score, in order to make some type of pass/fail judgement of them. The problem with this type of testing is that it is not clear what the norm refers to. To know that a learner is a 4,0 in English and that is a

pass, tells us nothing of what he/she can actually do with the language. The fact that a 3,9 student is part of the “fail” ones and that he knows the same or probably more than the 4,0 one is not taken into account.

Criterion-referenced assessment compares students not against each other, but with success in performing a task. The results of a criterion-referenced test can be expressed by continuing the sentence “he/she is able to…..” where the ability may refer to some small or larger integrative language task. Often these tests lead to a profile of language ability, where the learner is seen as capable of completing certain tasks to the given standards, but not others.

10. **Reliability v/s Validity:**
Reliability refers to the consistency of the scoring of the test, both between different raters, and bewteen the same rater on different occasions.Objective testing should give perfect realiability. However, faulty tests (ambiguous multiple choice, wrong answers on an answer sheet, etc) can reduce the realiability of even an objective test.

The subjective testing inevitably associated with testing the productive skills reduces reliability but they are more valid because they test integrative knowledge of the language so they give the teacher the opportunity to see how students really use the language. Teacher can have a better view of their students’ competence of language.
So the higher relaibilty, the less valid or all the opposite.

Desirable characteristics for tests:
Apart from validity and reliability, we have three extra characteristics to pay attention to:
1) UTILITY
2) DISCRIMINATION
3) PRACTICALITY

Utility:a test which provides a lot of feedback to assist in the planning of the rest of a course or future courses.
Discrimination: the ability of a test to discriminate between stronger and weaker students.
Practicality: the efficiency of the test in physical terms. (Does it require a lot of equipment? Does it take a lot of time to set, administer or mark?)

Students’ Assignment:
Make an assessment of the test given by your teacher (Appendix 3.1) by answering the following questions.
1) Does it test performance or competence?
2) Does it ask for language use or usage?

3) Is it direct or indirect testing?
4) Is it a discrete point or integrative testing?
5) Is it objectively or subjectively marked?
6) What skills does it test?
7) Is it backwards or forward-looking?

8) Is language contextualised or disembodied?
9) Is it criterion-referenced or norm-referenced?
10) Would it have low/high reliability?
11) Comment on its validity.
12) Comment on its utility, discrimination and practicality.