Lifelong Learning Programme

This project has been funded with support from the European Commission.

VET Integrated Language Learning

VET INTegrated LANGuage lEarning/VINTAGE

PROJECT NUMBER: 2013 – 1- ES1-LEO05-68044

Document Title	
Date of Issue	March 2015
Author(s)	Sirpa Junge , Lorenzo Rocca
Contributors	
Organisation	Aul Hamburg - CVCL
Approval Status	
Number of Pages	
Recipients	
Method of Distribution	
Confidentiality Status	

Guidelines for Assessment of Language Learning

1. Introduction

The guidelines are designed to be of use to teachers, test developers and testing program administrators, as they work to ensure that assessments are fair and valid. These guidelines focus on content area assessment within the CEFR levels A1 – B1.

The main purpose of these guidelines and the tools, which can be found on the VINTAGE-project research centre, is to provide testing practitioners as well as educators with a framework to assist in making appropriate decisions regarding different assessment scenarios, including the innovative approach of language learning assessment **closely combined with the contents of the vocational education and training**.

The Common European Framework of Reference for Languages CEFR uses the term "**assessment**" to refer to the implementation of language competence, thereby focusing on learner performance and its analysis. This focus contrasts with the more global term, "evaluation". Assessment refers only to analyses about the level of learners' proficiency evident in their performance, whereas evaluation can also refer, for instance, to the quality of a course, the effectiveness of teaching, or the appropriateness of pedagogical materials.

All types of language tests are a form of assessment, but tests are not the only possible means for assessment. Assessing also implies informal checking or verification, which can be done in various ways, one of which is testing. All assessment involve collecting data for the purpose of making effective decisions, ranging from tests to checklists in continuous assessment as well as informal observations by a teacher, only to mention some examples.

One of the aims of the CEFR is to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications. For this purpose the Descriptive Scheme and the Common Reference Levels have been developed. Between them they provide a conceptual grid which users can exploit to describe their system. *Council of Europe 2001a: 21.*

Assessment has profited from the advances that have taken place in statistical analysis over the last century and it is now considered important for assessors to check their tests empirically as well as rationally. In addition, the upsurge in interest in the results of tests and examinations by governments, authorities and candidates has led to a testing industry and contemporaneously there is a huge increase in the numbers of scholars researching different aspects of testing and assessment. Additionally the amount of Language learners is continuously increasing in the EU states through Migration within the EU, through EU-Mobility projects for young people and through Migration from non-EU-states and Refugees

During the past twenty years there have been a number of focus changes in the assessment of learning and teaching. Learning assessment is no longer only summative, that means the exam at the end of the course. New methods have been developed and are used. The assessment of learning has shifted:

from written exams to course work
from teacher centered assessment to student-led assessment (from external to self-assessment)
from quantitative to qualitative assessment
from implicit to explicit criteria
from competition to cooperation
from assessment of the production to assessment of learning procedure
from assessment of contents to the assessment of skills
form assessment of the course to the assessment of modules
from indirect assessment to direct (natural) assessment
from artificial assessment to authentic assessment
from authority assessment to agreement assessment (*Brown etc. 1994*)

Although trends in testing as in other fields change over time, some principles of assessment are permanent and are not overly affected by current fashions. (s. page ____ The Concepts of Assessment). The most European countries are introducing or formalizing linguistic requirements for the purposes of migration, residency and citizenship (s. also *VINTAGE National Chapters*). National authorities often require language tests or other formal assessment procedures to be used. Test fairness is a particularly important quality when tests are related to migration, residency or citizenship.

There are two overarching types of assessment in educational settings**: informal** and **formal** assessments. Both types are useful when used in appropriate situations. Informal assessments are those assessments that result from teacher’s spontaneous day-to-day observations of how students behave and perform in class or at workplace. When teachers/instructors conduct informal assessments, they don’t necessarily have a specific agenda in mind, but are more likely to learn different things about students as they proceed through the day naturally. These types of assessments offer important insight into a student’s misconceptions and abilities that might not be represented accurately through other formal assessments.

Formal assessment on the other hand, is preplanned, systematic attempts by the teacher to ascertain what students have learned. The majority of assessments in educational settings are formal. Typically, formal assessments are used in combination with goals and objectives set forth at the beginning of a lesson or the language course. Formal assessments are also different from informal in that students can prepare ahead of time for them.

The European Language Portfolio ELP of the Council of Europe points out the importance of independent language learning in order to support individuals to achieve a fuller awareness of themselves as language learners and to develop language skills that they can deploy to meet their needs*. (DGIV/EDU/LANG(200) 33rev. rev. June 2004; January 2011).* The ELP is designed to take account of the entire learner's language and intercultural learning, whether it takes place inside or outside formal educational contexts. This is a very important step towards the entirety of language learning, which is the reality of migrants and mobility participants – the main VINTAGE-target group. Accordingly the assessment of all language learning has to consist of different assessment forms. The common reference levels of CEFR are summarized in the self-assessment grid and should be included in all ELP models.

2. Concepts and Types of Assessment (CEFR)

2.1. Concepts of Assessment

The Common European Framework provides a coherent and transparent set of definitions and descriptions which can be applied to establishing quality assurance procedures and quality control systems. The chapter 9 of the CEFR describes in detail the different options which need to be taken into account in the development of language learning / teaching activities. It distinguishes between norm- and criterion-referenced assessment, between achievement and proficiency testing, between formative and summative assessment and it presents the fundamental three concepts, which occur in any discussion of assessment and still are the basic considerations in spite of all changing trends. These three Validity, Reliability and Feasibility should be complemented by some more important concepts which are used in the field of assessment planning.

Validity

Validity is a term which is related to questions about what the test is actually assessing. Is the test telling you what you want to know? Does it measure what it is intended to measure? A test which is intended to measure a candidate's level of reading comprehension in a foreign language but instead tests intelligence or background knowledge is not valid.

There are several aspects of validity depending of the purpose for which the test is meant to be used. Of course the importance of the assessment is also an important factor. A weekly testing of vocabulary in the classroom has another importance than a test on its basis the authorities decide about a residence permit and needs of course another investigation of validity. This investigation should be carried out regarding to its use, not the test itself. The same test of oral German could be o.k. for young learners at school but not for migrants who need German for daily life or for technical studies.

One of these aspects is the *construct validity* which is often used as a synonym for validity. This term covers the whole construct being measured. Hereby can be used a combination of internal and external quantitative and qualitative methods. The next aspect is the *content validity.* Depending on the importance of the test the test designers should take care whether the items of the test are actually testing the content they are supposed to test. *The Face validity* means the understanding of non-professionals such as learners or parents. Does the test give them the impression of being appropriate? If the test lacks face validity, it may not work as it is meant to work. To assess *the criterion-related validity*, the learner's test scores may be correlated with other measures of the learner's language ability such as teachers "rankings" of the students, or with a similar test. Such measures assess the concurrent validity of the measure.

Reliability

The *reliability* of a test is an estimate of the consistency of its marks; a reliable test is one where, a learner will get the same mark if he or she takes the test twice on different days and with different examiners; it refers to consistency across situations, opportunities and people. Important within this concept is the precision of that classification with reference to a norm. It is important to specify the criteria chosen ("master"/"non-master"; rating on a scale B1/B1+" etc.) and the procedures implemented to reach a certain judgment. The high reliability does not necessarily imply that a test is good or interpretations of the results are valid, but for the valid interpretation of test results, scores must have acceptable reliability, because without it the results can never be dependable or meaningful.

Feasibility

The third key concept is *feasibility.* That means it is possible to cope with the practical terms and conditions. Time constraints and too many categories and criteria at the same time can lead to infeasible assessment settings.

All these key concepts are inseparably linked. The assessors have to consider both pedagogical reasons and external conditions in order to take decisions. Certain tensions can occur between these key concepts. For instance using external assessors can be helpful in targeting the object of assessment and in improving consistency in some situations but may be unfeasible in other situations.

Another important feature is **Accuracy of decisions**. Made in relation to a standard it can be in fact more important than reliability. If the assessment reports results as pass/fail or levels A2 + B1/B1+, how accurate are these decisions? The accuracy of the decisions will depend on the validity of the particular standard (e.g. Level B1) for the context. It will also depend on the validity of the criteria used to reach the decision and the validity of the procedures with which those criteria were developed.

Especially for the assessment of Language for Special Purposes (e.g. migrants at the working place) is **Authenticity** a very important aspect. Authenticity of task means that the test tasks (e.g. LSP test) should share critical features of tasks in the target language use situation of interest to the test takers. The intent of linking the test tasks to non-test tasks in this way is to increase the likelihood that the test taker will carry out the test task in the same way as the task would be carried in the actual target situation. (*Douglas, 2000*) The situational authenticity refers to the accuracy with which tasks and items represent language activities from real life. Interactional authenticity refers to the naturalness of the interaction between test taker and task and the mental processes which accompany it. The balance of different aspects in language tests requests often adaptation of the materials to learners' current level of language proficiency.

An objective for all assessment providers is to make their assessment as **fair** as possible. (s. e.g. *Standards for Educational and Psychological Testing AERA et al 1999).* These standards acknowledge three aspects of fairness: *fairness as lack of bias, fairness as equitable treatment in the testing process, and fairness as equality in outcomes of testing.* Several bodies have produces Codes of Fairness to assist test providers in the practical aspects of ensuring tests are fair (e.g. *ILTA Guidelines for Practice 2007*). The **fairness** of assessment enables an equal participation for everybody.

Any piece of assessment should also have positive **washback**; the effect of the test on the teaching must be beneficial. That means that the construct of the assessment shouldn't lead for example to candidates learning content by heart or achieving high marks by simply applying test-taking skills rather than genuine language skills. A good assessment should also be **clear;** the layout, the images, the format, Design and visual cues should be easily understood and legible.

2.2 Types of assessment

Assessment involves many factors related to contexts, cultures and assessment traditions. Choosing among different types of assessment requires carefully selecting procedures consistent with the assessment goal in its appropriate context. The CEFR makes a number of important distinctions in relation to assessment. The table 7 (p. 183) of the chapter 9 shows 26 types of assessment. These parameters have been organized by classifying them into 13 pairs which are at once distinct and complementary.

1.	*Achievment assessment* Assessment of what has been taught, is oriented to the course – internal perspective	*Proficiency assessment* Assessment of what someone can do/knows in relation to the application of the subject in the real world – external perspective
2.	*Norm-referencing (*NR) Placement of learners in rank order	*Criterion-referencing* Reaction against norm-referencing – the learner is assessed purely in terms of his ability in the subject
3.	*Mastery learning CR* A single „minimum competence standard“ is set to divide learners into „masters“ or „non-masters"	*Continuum CR* An individual ability is referenced to a defined continuum of all relevant degrees of ability
4.	*Continuous assessment* Assessment by the teacher and possibly by the learner throughout the course	*Fixed assessment points* Grades are awarded and decisions made on the basis of an examination – usually at the end or before the beginning of the course
5.	*Formative assessment*	*Summative assessment*

	An ongoing process of gathering information on the extent of learning, on strengths and weaknesses, which the teacher can feed back into his course planning	Sums up attainment at the end of the course with a grade
6.	*Direct assessment* What the candidate is actually doing	*Indirect assessment* Uses a test, usually in paper, which often assesses enabling skills
7.	*Performance assessment* The learner has to provide a sample of language in speech or writing in a direct test	*Knowledge assessment* The learner has to answer questions which can be of a range of different item types
8.	*Subjective assessment* A judgement by an assessor (judgement of the quality of performance)	*Objective assessment* Assessment in which the subjectivity is removed = indirect test in which the items have only one right answer e.g. multiple choice
9.	*Checklist rating* Judging a person in relation to a list of points deemed to be relevant for a particular level	*Performance rating* Judging that a person is at a particular level or band on a scale (vertical)
10.	*Impression* Fully subjective judgment made on the basis of experience of the learners performance in the class	*Guided judgment* Judgment in which individual assessor‘s subjectivity is reduced by complementing impression with conscious assessment
11.	*Holistic assessment* Making a global synthetic judgment	*Analytic assessment* Looking at different aspects separately
12.	*Series assessment* Involves a single assessment task in which performance is judged in relation to the categories in an assessment grid	*Category assessment* involves a series of isolated assessment tasks, which are rated with a simple holistic grade
13.	*Assessment by others* Judgment by the teacher or examiner	*Self-assessment* Judgments about your own proficiency

The arrangement of the types doesn’t represent a judgment of value; it rather aims at fostering awareness of the advantages and disadvantages of the different types. The arrangement in two columns can be seen as a way of establishing the two ends of a continuum.

The most relevant types for the VINTAGE target group will be discussed in the next Chapter VINTAGE – target groups, special needs and an innovative approach.

3. VINTAGE – target groups, special needs and an innovative approach

The mastery of local languages is a key asset for the aim group of VINTAGE: **mobile workers**, **students** enrolled in universities or engaged in an internship abroad and **migrants**, in order to improve their skills and competences.

Several projects and scientific articles promoted during the last 5 to 10 years coped with the scarce attention given to work-based and sectorial specific approaches to language learning (e.g. LSP), creating a first basis of methods and didactical resources. Integration between language learning and VET contents still represents, however, an open challenge. The project, addressing at least 3 different educational sectors (IVET level, Higher VET and Continuous Vocational Training, including Adult education), moves from the conviction that linguistic skills should and could be improved through learning strategies based on action, proximity to daily life and workplace settings, integration between vocational education and training contents and development of communication capabilities. That means using typical VET related settings – implying understanding instructions and lessons, questioning and answering, naming tools and procedures, describing and explaining work tasks and processes, etc. – in order to improve and assess language mastery.

More and more migrants, mobile workers or students learn to speak local languages following different routes, sometimes integrating **formal** or **non-formal** training (courses, e-learning) and **informal** learning, sometimes exploiting more or less only informal opportunities. Learning through "exposure" to a diverse linguistic context always plays a fundamental role. Informal learning processes occur at the workplace, listening and reading, exploiting the media, communicating with colleagues and friends. Web based learning wins recognition especially among young people, to its mobile and cheap fruition. In recent years, the validation of informal and non-formal learning and education has attracted considerable interest also among education promoters. The policy initiative of the Council of the European Union's *Recommendation on the validation of non-formal and informal learning (2012/C 398/01)* asks member states to introduce by 2018 "arrangements for the validation on non-formal and informal learning"…which will enable individuals to "obtain a full qualification, or, where applicable, part qualification, on the basis of validates non-formal and informal learning experiences…".

Specifically, the *Recommendation* suggests that the following elements might be included in such arrangements

(a) IDENTIFICATION of an individual's learning outcomes acquired through non-formal and informal learning;

(b) DOCUMENTATION of an individual's learning outcomes acquired through non-formal and informal learning;

(c) ASSESSMENT of an individual's learning outcomes acquired through non-formal and informal learning;

(d) CERTIFICATION of the results of the assessment of an individual's learning outcomes acquired through non-formal and informal learning in the form of a qualification, or credits leading to a qualification, or in another form, as appropriate.(

While in formal education systems, assessment generally implies some kind of examination, **assessment of informal and non-formal learning** often requires alternative approaches. In addition to portfolios, structured interviews, practical presentations, observations, mapping of learning needs, planning of individual learning, reflection and self-assessment have been suggested as possible methodologies.

The recognition of prior learning (RPL) is giving credit to what learners already know and can do regardless of whether this learning was achieved formally, informally or non-formally. The recognition of prior learning is particularly important for mature age migrants who will have a wealth of industry and life knowledge, skills and experience which could provide them with credit towards units of competency or a full qualification and support the learning of the host language. The learner will save time. There is no fundamental difference in the assessment of previously acquired skills and knowledge and the assessment of skills and knowledge achieved through a current learning programme.

The key questions regarding to assessment of non-formal and informal learning for which also the VINTAGE team will be trying to find solutions are:

- How can informal and non-formal learning be presented and documented?
- What methods and tools can be used to assess expertise gained through informal and non-formal contexts?
- Can and/or should informal non-formal learning be validated and certified?
- If informal and non-formal learning be validated, will such learning be recognised (even with certification)?

Integrating language learning and educational / vocational pathways from Initial Vocational Education and Training to continuous training is the opportunity and the challenge of holistic language learning. Interlinking formal and informal learning and taking in account workplace settings and new possibilities of learning, the VINTAG project will support teachers and trainers to develop their competences in managing these processes, providing them with a website including a Resource Centre with hints, suggestions, tools for designing, planning and managing personalized language learning pathways.

The project VET integrated language learning directly aims at promoting linguistic learning according to learners' **needs and expectations**. To consider the aims and operational objectives concerning the development, testing and valorization of an innovative learning model frames the heart of the project approach. Through integration of language learning in qualification pathways can the learner’s needs be combined with the external needs of industry and authorities.

The project and its concept are based on the comprehensive methodological framework “fide”, promoted and financed by the Swiss Federal Office for Migration. The aim of “fide” is the development of a comprehensive innovative methodology of language learning (French, German and Italian languages), based on action, proximity to learners’ daily life and needs (learning “scenarios”), valuation of informal and non-formal learning settings. In this project the training design aims at making language learning as close as possible to needs, expectations and learning possibilities of participants. The concept implies a diversified and tailor-made training design, adapted to the target audience, offering flexible planning of training times, and decentralized learning spaces close at the workplace. The branch specific nature of instructional design, streamlines the management of communicative situations at the workplace facilitating the integration between job-specific training opportunities and language learning.

This approach is fully coherent with the philosophy of CEFR. Levels descriptors of CEFR turned in recent years to a more pragmatic measurement of language proficiency. Language is conceived on the basic and advanced stage (at least until B1) primarily as a means of communication; this is also the basic idea underpinning the “fide” approach. Second language teaching mainly aims at enabling participants to communicate as successful as possible at the end of the course at a certain level (competence in action). A communication or interaction works successful if the speaker can "naturally" express him/herself and the contents of the communication is interpreted correctly by him/her and by the interlocutors. Intelligibility of communication is in turn hardly depending from contexts and situations, as well as oral communication is strongly supported by non- and para-verbal means. As a result, one important “piece in the puzzle” will be intercultural communication contents which will enable the learners to act and to interpret the acting of others in an appropriate way.

Moving from results already achieved in order to integrate **work-based contents** in language learning, aims at developing an **innovative approach** addressing difficulties and challenges of teachers and trainers in overcoming constraints of a specialized approach to language learning (sectorial didactics, standardized curricula and scholastic learning settings, etc.), enabling them to design personalized didactical activities valuing VET learning settings, and respecting at the same time individual needs, autonomy and responsibility of the learners. This innovative approach of VET integrated language learning request on the other hand an **innovative assessment** of learning outcomes, methods and tools, valuing self-evaluation and assessing communication skills embedded in qualification achievements.
Parallel to the CEFR the Council of Europe has been developing the **European Language Portfolio ELP**. The ELP has complementary pedagogical and reporting functions the make it appropriate for use in language programs for adult migrants, and will be an important method also for the assessment of VINTAGE target group. Because language learning is a lifelong process **autonomous learning** is a very important issue within all European

innovative projects and the ELP is designed to support the "autonomisation" of language learning and language learners. Correspondently from the Swiss "Fide" Project, drawing on principles of the *Common European Framework of Reference for Languages* (CEFR), and from the Irish pilot program IILT, providing evidence of new possible language learning pathways, following 3 key principles will be taken into account in the VINTAGE- Model:

1. All worthwhile language learning is firmly embedded in **language use**; that means enable the learners, learn vocabulary, familiarize themselves with grammatical patterns, practice pronunciation, actually using the target language, at work and in their daily lives
2. Learners' motivation depends on the extent to which their **needs** are being met, considering objective and social needs but also subjective and individual needs (which can be understood and responded to only in the process of teaching and learning)
3. In order to elicit and respond to learners' subjective/ individual needs we must acknowledge and respond to their **autonomy**. That means taking in account in learning design the identity of the learners, their responsibilities that help to define their learning goals, their personal priorities and preferences. We respond to the autonomy of our learners by giving them autonomy in the language classroom, making them responsible for (among other things) identifying learning targets and **co-evaluating learning outcomes.**

The Vintage-Project aims at developing and testing a **comprehensive strategy** (based on methodologies, guidelines, tools and resources, avoiding the definition of standardized curricula and units of learning) addressing the specific needs of teachers and trainers working with migrant and mobile learners, aiming at developing their linguistic skills in the hosting country, in combination with workplace experience and vocational education and training. Methodologies enhancing work and home-based learning will be integrated in pathways leading to certifications based on the CEFR (from A1 to B1 levels). The nature of the target group - migrant and mobile workers / students learning local languages as a second language – and the interlinks between learning and daily life settings will in addition ensure real interaction and learning opportunities with native speakers of the target language. The designing and testing of **assessment methodologies and tools** (summative and formative) will enable the recognition, validation and accreditation (when possible also certification) of linguistic and professional competences gained by the learners in informal and non-formal learning settings (verifying entry level and learning outcomes)

4. Factors which influence assessment of language learning

There are different issues to be considered when developing assessments and making decisions. The factors provide useful context for the assessment guidelines. Of course political and legal restraints (e.g. residency restraints) are a relevant factor but these have been discussed in other chapters of the project guidelines.

4.1. The structure elements of assessment

A factor for describing assessment can be the differentiation of the structure element terms of assessment. The question then would be:

From which obligatory elements the assessment situation is composed? We have at least

1. The assessor (who can be the learner himself, another person of the group, an external assessor)

2. The assessed learner (who is acting in a situation which is assessed)

3. The assessment objective (which can be the product or the process of the learning)

4. The interest of the assessment (which tells us why the assessment should be made and why exactly in this way)

5. The assessment tools (all the measures, regulations and tools, which are used to define the accuracy of the assessment action) *(Brown, et al. 1997*)

These structure elements as a whole issue can also be called “the assessment design and implementation”. The assessor can be a practitioner-assessor, i.e. the learning facilitator 8techer, lecturer and trainer), who has traditionally administered assessment in addition to facilitating learning. The assessor can also be a workplace supervisor, manager or team leader, provided that they are skilled in the process of assessment. In the self-assessment the learner him/herself plays also the role of the assessor. The assessment should be a common task. The other learners should be used as each other’s evaluators. The peer evaluation can be very beneficiary for promoting students’ co-operative working culture and community work skills.

For the assessed learner it is important to understand from the outset what their roles and responsibilities are in terms of their assessments. Also, learners must understand the process of assessment and why it is done in a particular way. They will also need to know what they can expect from the assessor and what assessor expects from them. The candidates should be treated with courtesy, respect and impartiality, regardless of their age, disability, ethnicity, gender, national origin, religion, sexual orientation or other personal characteristics. Assessments are meant to be as clear and transparent as possible; the test takers should receive a brief oral or written explanation prior to testing about the purposes for testing and given the manner in which the results will be used. (see e.g. I*LTA-guidelines for Practice).*

The objective doesn't require any additional specifications. The interest and the tools of the assessment should be specified. The interest consists of the objective and of the philosophical and practical considerations. This element can also be defined as a "business idea" of the assessment then the assessment always has some conceptual and functional grounds. Assessment is done because someone wants to classify the learners, to guide the learner's choices and give feedback to them. The interest conducts the assessment. The assessment tools, in a wide definition, contain the social and cultural context with the social and technical rules and operational models.

4.2. Linguistic, educational and cultural background factors

When developing assessments it is important to keep in mind the different linguistic background of learners. The cross-lingual interference can take place at all levels of the linguistic system, i.e. in phonology, morphology, syntax, semantics, pragmatics, and the lexicon. This factor can be taken in account especially within the formative assessment in multinational courses as well as in teaching and training for groups of a linguistically homogeneous background. Varying levels of proficiency in native language can also influence the teaching and the assessment results.

Also the writing skills in the native language have to be taken into consideration when designing an assessment. There has been a lively discussion about alphabetization during the last decades. If the learner's native language doesn't use Latin script (e.g. Bulgaria) or if the learner has for lack of attendance at school only few writing skills how the assessment still can be fair and what kind of assessment should be used. This is for the VINTAGE aim group a very important question.

As mentioned previously the second/foreign language teaching and coherently the assessment vary widely in the level of formal schooling the learners have in their native languages, because this affects not only native language proficiency – specially literacy in the native language – but also the level of content area skills and knowledge. The primery challenge for these learners is to transfer their existing content knowledge into host country language.

Cultural factors can also be potential sources of construct-irrelevant variance that add to the complexity of appropriately assessing language learning. Lack of familiarity with mainstream host country culture, for example, can potentially have an impact on test scores. There can be different assumptions about the testing situation or the educational environment in general, have different background knowledge and experience, or possess different sets of cultural values and beliefs, and therefore act differently.

4.3. The basic forms of assessment

The basic forms of assessment are internal, participatory and observing assessment. All these forms can occur alone or be combined with each other:

Internal assessment refers to assessment that logically and necessarily always is a part of the assessment. It is thus a self-assessment concept. The assessor and the assessed are one and the same person, the target is the continuing process of student's own learning operations, the interest is the development of the proficiency and the assessment tools are individual.

One of the most important tools of internal assessment is the European Language Portfolio ELP. Such educational portfolios allow learners to collect together certificates, attestations and good pieces of work to document and inform others about their learning achievements. It is designed to help learners to achieve a fuller awareness of their developing linguistic and cultural identity. Learners can also use a portfolio to describe, reflect on and plan their **learning process**, and to improve their learning strategies. The European Language Portfolio is based on the CEFR, which was established by the Council of Europe. There are Language Portfolios in many European countries, in many languages and for different age groups, from small children to adults. (see more in Chapter 5).

Participatory assessment is assessment carried out by two or more persons in cooperation. The assessors have a common objective and they assess the cooperation with and the acting of each other in order to solve the problem. Like the internal assessment, participatory assessment is also an integral part of rational action. In the Participatory assessment the assessor is a co-operation partner, the assessed operator the other part of the co-operation and the objective is the cooperation process itself, as well as the operational content. Assessment of the interest is monitoring and development of the joint action (performance) Assessment tools are interactive, i.e. operators have mutual consent or shared knowledge about the assessment tools.

Observing assessment refers to an activity form outside; the following, observing, controlling or determining assessment. In the monitoring assessment the assessor is not part of primary operation, but he is engaged as an independent "higher-level" operator. The observing assessment is external and normative monitoring and classification, directed to the operation. Normativity means that the assessor compares the performance of one student to a norm, which provides a standard for performance acceptability. Standard may be the content description or numerical standard. In the observing assessment the assessor can be an external evaluation expert, the assessed student may be an individual, a group or community, and the objective is the operation process or the products of the operation. The assessment interest is critical or questioning and assessment tools are professional, specialized or requiring special expertise.

4.4. Assessment modalities

These factors consider the question at what time of the learning the assessment will be applied. It can be diagnostic, formative or summative. In order to foster learning as versatile as possible it would be positive to combine these assessment modalities In the following a short description of all these applications:

Diagnostic assessment is situated at the beginning of the learning, as a pretest in order to find out learner's language skills and competences. **Pretesting** is normally used as an entry-level test in order to place the learners into the appropriate level course. For language learning a diagnostic test can also be used for preparing an appropriate curriculum for the tested group.

Formative Assessments are on-going assessments, reviews, and observations in a classroom gathering information about learner's learning during the progression of a course or program and usually repeatedly-to improve the learning of those students. Teachers use formative assessment to improve instructional methods and student feedback through teaching and learning processes (William & Black, 1996). One way of conceptualizing formative assessments are as "practice." The formative assessments has positive influence to learner's motivation and the formative methods are often used for activating the learning process.
The step from formative assessment to self-assessment is short. From the perspective of effectiveness, self-assessment plays a considerable role. To do self-assessment, learners need to have suitable tools at their disposal. The assumption that rating on a scale and rating on a checklist are complementary is fully justified and shown e.g. in the ELP.

Summative Assessments are given to assess what students do and do not know about a particular learning topic. They measure the level of success or proficiency that has been obtained at the end of an instructional unit. Summative Assessment is done at the conclusion of a course or some larger instructional period (e.g., at the end of the program). The purpose is to determine success or to what extent the program/project/course met its goals. Examples include final exams, state assessments, end-of-unit chapter tests, and benchmark assessments.

Course-Embedded Assessments refers to techniques that can be utilized within the context of a classroom (one class period, several or over the duration of the course) to assess students' learning, as individuals and in groups. Course-embedded assessments can be formative or summative. When used in conjunction with other assessment tools, course-embedded assessment can provide valuable information at specific points of a program. For example, faculty members teaching multiple sections of an introductory course might include a common pre-test to determine student knowledge, skills and dispositions in a particular field at program admission. There are literally hundreds of classroom assessment techniques, limited only by the instructor's imagination. Course-embedded assessment can have a variety of advantages, including purposeful reexamination of course objectives, sequencing, and content and feedback. They can be used to evaluate the development of student skills and provide feedback (formative) and they can be summative as well (evaluating final student product (MSU.edu, 2012).

5. Developing Test Items and Scoring Criteria

5.1Test development process

This part attempts to characterize the procedures involved in the development of a new test.

The first step is the identification of the requirements for a new test. It may be that a new test is needed in order to replace an existing one. If this is the case, the test developers will already have some clear ideas on the ways in which the new test should be an improvement on the old one. They may be taking account of a change in the candidate population, or of developments in testing theory. Whether the reasons for change are theoretical or practical, they are present from the start, contributing to the definition of the test.

5.1.1The planning phase

A situational analysis is carried out, identifying and describing the following main points.

a) The stakeholders. These are the people involved in the testing process, those who will design and develop the test, those who will administer it, those who will take it and those who will use the results. They include students, teachers, parents, school managers, government agencies and commercial enterprises. The test has to be generally acceptable to all these people in the sense that they need to understand and accept why the test is the way it is.

b) The purpose of the test. It is essential to have a clear view of why the new test is needed, and why it takes the precise form it does. Those involved in developing it must be able to account for its features. Its level of difficulty should be established. How it fits into the current system in terms of the objectives of the curriculum and current practices in teaching needs to be determined, and what future developments are planned have to be identified.

c) External influences. Expectations of how the ability or competence should be tested (for instance: speaking) in the context in which the test will be used have to be considered, with reference to commercially available tests, and the demands of educational policy and local conditions.

d) Internal factors. The new test may be being developed in the context of a school, university, examinations board, etc. Whichever kind of organization is involved, the test has to fit in with the existing working practices of that organization, and the level of knowledge of the theoretical background to the testing which exists there. The resources available in terms of the staff, technology, time and money which can be put into test development, administration, reporting results, replication and validation of the test also have to be assessed. A project plan is needed, so that objectives are stated, necessary resources identified and a time scale established.

5.1.2 The design phase

The design phase involves writing initial test specifications. This means focusing on the practical and professional considerations and constraints, based on the situational analysis, which affect test design and administration. Decisions on test design and content have to be made, which means making comparisons with other existing tests, and - if the new test is being added to an established examination as a new component - making sure that the new element falls into line with those which already exist.

In designing the test, the following factors, all of which interact with one another, have to be considered. It is important to attempt to account for and describe these interactions in order to allow for test validation to take place. For instance in designing a speaking test, it should be taken into account that no decisions can be made in isolation from each other; for example, the type and complexity of rating scales used will depend on the type of tasks set, and also on the examiners available. If plenty of native speaker examiners can be recruited, and there is the time and money available to give them thorough training, it is reasonable to expect them to put into practice a complex and demanding rating system. If, on the other hand, time and money are in short supply, and it is difficult to find suitably qualified examiners, a system of rating which demands less of them will have to be devised.

Throughout the entire process of test development, it is necessary to regard decisions as provisional, and to be ready to adapt to change and to return to an earlier stage in the process. Even when the test becomes live, it cannot be regarded as somehow permanently 'fixed', and areas needing change and improvement will emerge.

a) Candidates. Demographic features of the candidate population need to be considered, and a candidate profile produced. Their background, age and nationality will affect the choice of materials to use in testing them. The types of linguistic behavior they need to be able to produce, and the areas of language use in which their competence needs to be tested have to be defined, as does the ability level expected of them.

b) Examiners. In the case of subjective tests**,** decisions have to be made on what the desired qualifications are for examiners for the test, what training they will be given and what limits are placed on their behavior during testing. Examiners will have to be trained in using whatever method of rating and type of rating scales are devised for the test.

c) Tasks. Decisions have to be made over the format of the tasks used, and the types of prompts which will be used as a stimulus to elicit responses from candidates. A variety of task types may be available for the writers of test tasks to choose from, or they may have to adhere to the same sequence of task types in each version of the test.

d) Ratings. A Rating scheme should be defined both for objective and subjective tests. For the latter, each candidate will have to be rated on a scale or scales and each possible score being defined by descriptors. Criteria should be used as well such as: grammatical accuracy, range of vocabulary, etc. The more complex the rating to be made, the more carefully selected and highly trained the examiners will need to be.

Professional and practical considerations

At the design stage of the test, many considerations should be carefully addressed. These can be divided into professional and practical concerns.

a) Professional considerations

Professional considerations refer to what exactly has to be tested, and to the theoretical model of language ability on which the test developers choose to base their work. Test design has to take account of the real life situations in which candidates will need to use the language, and the level of competence in speaking which they will need. Choices have to be made concerning the types of real-life language events to be re-created in the test, in terms of such features as topics, the amount of language required in the candidate's response. Decisions also have to be made about the information on their performance which will be given to candidates. For example, they may be given scores, grades, a simple 'pass/fail' result or a detailed profile related to performance on each task.

b) Practical considerations

All aspects of the test have to be achievable from the practical point of view, given the resources of time, money, personnel etc. likely to be available, as revealed by the situational analysis. In any testing situation there are likely to be limits on these, and the test has to fit in with circumstances as they are, rather than how they might ideally be. Practical considerations affect test administration, candidates, examiners and ratings, tasks and materials, assessment and quality control procedures.

i) Administration. Considerations include: the number of staff available both to write tasks and help conduct tests, rooms available, the period of time over which testing must take place, the length of each testing session and how soon after testing the results have to be issued.

ii) Candidates considerations include: the number of candidates for assessment, the length of each assessment, etc.

iii) Examiners and ratings (particularly in the case of subjective tests), considerations include: how many examiners are available, whether they should all be native speakers, how they will be trained and how much time is available for training, etc,

iv) Tasks and materials considerations include: the number of phases or stages testing different aspects of competence, the task types to be used in each phase, what sorts of prompts will be used, etc.

v) Assessment considerations include: the way in which scores will be reported to candidates, in the case of subjective tests whether assessment will be made on a discrete point or holistic basis (and, depending on the answer to that, how many discrete points or how many scales will be used), what measures can be taken to ensure that scoring is reliable.

vi) Quality control procedures considerations include: what quality control methods can be used, how data will be collected and stored for analysis and validation and who should carry out analysis and validation.

5.1.3 The development phase

By the end of the design phase some sample materials for the test will have been written. The process of writing materials may reveal weak points in the initial specifications, which then have to be revised and amended. When this has been done, the process moves forward to the development phase.

At this stage the sample materials and prototypes of rating scales for subjective tests are trialed. This means that a group of students who are at the same level of competence as those who the test will aim to assess will be needed to act as volunteers in simulations of the testing situation.

Trialing yields a great deal of information about the test, as those who take part as candidates and as examiners can provide detailed feedback on many aspects of the test from their two points of view. Each can comment on their reactions to prompt materials, topics, task types and level of difficulty, test length, etc. Candidates can comment on the adequacy of the instructions they receive and the task format and how comfortable or otherwise they felt in the physical environment provided for testing. Examiners can give their reactions to the rating system they were asked to use. This feedback can be collected by means of reports or questionnaires.

When questionnaires, and any other data have been analyzed, an evaluation can be made of what they show. The progress of the test's development up to this point can be reviewed. As a result of this, changes may be made, to the specifications, task types and rating scales. It is possible that new sample materials will have to be written, and the process of trialing and analysis repeated, bearing in mind that this must be done within the time constraints imposed. It may happen that the process of trialing, analysis of results and evaluation will have to be repeated several times before the test is considered ready to administer on a regular basis, and enters its operational phase.

5.1.4 The operational/monitoring phase

Before the test becomes live, it is necessary, in the case of subjective tests, to have recruited and trained sufficient examiners to deal with likely numbers of candidates. It may also be necessary, if the test has more than extremely limited use (for instance, within one college) to train a team of writers who can produce test tasks, so that a bank of materials can be set up and a version of the test constructed when one is needed.

As the test goes through successive administrations, it will be monitored to ensure that a constant level of difficulty is maintained, and the data provided by results may be used for research purposes. Procedures should be put in place to validate it as a true measure of the skills it seeks to measure.

After it has been in use for some time it may become clear that the nature of the candidate population is changing, or that the tasks look outdated or out of step with current thought on language testing, and the need for revision may necessitate returning to the beginning of the cycle and developing a new test, or going back to the design phase, revising the specifications and going through the process of developing and trialing new materials.

5.2 Scoring criteria

In any kind of test, one essential task is to give a fair result to the candidate. The aim is always to achieve insistence on accuracy and consistency and to keep scoring error to a minimum.

One aspect strictly related to consistency is reliability. Reliability is an issue of particular importance in language testing and there are quite a number of issues related to reliability.

In terms of methods of marking and scoring and the particular problems or issues to be considered, the main classification is the following:

- objective marking
- subjective judgment

5.2.1 Reliability in objective tests

Reliability in testing is often defined as consistency of measurement: i.e. a candidate taking two versions of the same test on two occasions close in time to one another would be expected to get approximately the same score, and not find that one version was more difficult than the other.

Both internal and external factors can affect reliability in objective tests.

i) **Internal factors**

These concern features such as the number and quality of the items used. There are various statistical methods for checking test reliability in terms of internal consistency. These generate statistics in the form of coefficients; the most commonly used are Cronbach's alpha (coefficient alpha) and Kuder- Richardson 20 (K-R20).

ii) **External factors**

External factors which can affect reliability concern:

•external conditions: the score achieved could be affected by the quality of the room in which the test is taken (lighting, comfort, space, noise levels from outside). The person administering the test may be unsympathetic and unsupportive. Such factors are avoidable and should be dealt with in training and preparation procedures for the administration of the test.

•attitude and behavior of candidates: the score achieved could be affected by the extent to which candidates feel confident or nervous, healthy or off-color as well as by candidates who are uncooperative and unsympathetic to the idea of being tested – a reluctance to expand on answers in a speaking test, for example. The reliability of the result will also be affected by candidates who guess the answers in multiple choice tests and by candidates who have had intensive practice or no practice at all with the format and item types used in the test.

5.2.2 Reliability in subjective tests

The question of reliability in subjective tests concerns the quality of the judgments made by markers and examiners. What is aimed at is:

- inter-rater reliability: consistency of judgment between different markers of the same test
- intra-rater reliability: consistency of judgment by the same person on different occasions and under differing circumstances

Clearly this type of marking is far more influenced by human changeability and error than computerized or clerical marking. It is, therefore, unrealistic to expect total reliability. There are, however, many measures, the rigorous training of examiners, for example, which can be put in place to make subjective tests as reliable as possible.

This needs to be addressed in training. Markers also need to be monitored through making ransom checks on samples of each marker's work.

5.2.3 Some issues in marking and scoring subjective tests: marking the candidate's response

Tests of productive skills (speaking and writing) tend not to be item based, These tests make different demands on markers, and the process of examining these skills is generally regarded as more subjective. The questions: How can the assessment of subjective tests be fair?

Ensuring fair assessment in performance tests will depend on the following aspects of the process:

•**recruitment and induction of markers and examiners.** Skilled markers and examiners are needed for subjective testing and, for widely taken national / international tests, these are often teachers familiar with the examination through preparing students to take the test. Training sessions give them the chance to become thoroughly familiar with the scale or list of criteria against which candidate performance is to be assessed.

•**sound standardization procedures.** The training of markers and examiners includes a process referred to as 'standardization', the aim being to enable them to provide a standard level of judgment of candidates' output.

•**clear marking criteria.** The scale may consist of numbers, letters or ‘labels’ (‘Good’, ‘Adequate’ etc) and contain statements of what each point on the scale refers to. These are the scale descriptors. Markers and examiners need a thorough understanding of the principles behind the particular scale(s) they are working with.

•**monitoring and evaluation of markers and examiners.** The performance of markers and examiners must also be assessed. It is necessary to set up a system of checking, monitoring and evaluating their performance and provide constructive feedback.

•**number of markers involved in each assessment**. The question of how many markers or examiners take part in each assessment affects subjective tests.