Project TERECoP

Study (Work Package 4): methodology for evaluation and assessment of the learning activities

Chapter 1: EVALUATION - A COMPLEX PROCESS

Evaluation is strongly related to education as sequence of any training or educational process.

Education includes in the same time the formal education and the continuing education and also the entire scale of non-formal and informal education possibilities available into a learning society.

It is impossible to discuss about evaluation without emphasizing, first of all, its place and role within the educational process. The chart below is suggestive enough to illustrate this context.

Any society has its own aims and objectives. Inside the macro-system called "society" the educational system is enclosed, besides other systems which are interrelated. These are self-guided based on specific goals and objectives. The objectives of the educational system are determined by those of the society, on the basis of a "demand - offer" functional relation.

This means that the society identifies its training and educational needs and makes the command towards the educational system. The educational system accomplishes the respective command, by rendering back the trained "product" to the society, society which is prepared and able to "absorb" it in its structure, positions and functions, in accordance to the qualification and to the specialization conferred by the acquired and recognized/certified skills.

The certification and the recognition are developed based on the evaluation of the acquired knowledge and of the skills formed during the educational process. Thus, the evaluation is the one which links the aims of the educational system to the aims of the society.

Definitions of some terms and concepts are of great importance for our present debate.

Evaluation (into a limited approach) represents the process of measuring a certain behaviour or feature.

Evaluation (generally speaking) is the process of formulating a value assessment (expert judgement) referring to an educational demarche.

Evaluation's role: Evaluation allows society to monitor the efficiency of the educational system, makes possible the certification of individuals' capacities and abilities, permits the aptitudinal and attitudinal characterisations and allows the selection of the best individuals.

Evaluation's aims:

- to formulate an expert judgement upon the measurement results.
- to adopt an educational decision based on conclusions of the results' interpretation and appreciation.

Evaluation's main functions:

- diagnosis (envisages tracing the gaps and the mistakes of the trainees in the view of their removal).
- prognosis (renders evidence the future performances of the trainees and supports the decision).
- selection (allows classifying the trainees).

- certification (reveals trainee's competences and knowledge at the end of a training cycle).

Observation: Inside the competitive societies, the most important functions are selection and certification, these inducing a strong "backwash" effect within the system.

Evaluation's specific functions:

- motivational function (stimulates the learning activity of the trainees and manifests itself through out a positive valorisation of the feed-back offered by evaluation, in the sense of appreciating the own activity).
- counselling and educational orientation (intervenes in choosing certain educational paths or forms).

In conclusion, evaluation means:

Measurement	**Interpretation and appreciation of the results**	**Decision adopting**
- specific procedures; - measurement instruments.	- unitary and objective criteria	- expert judgement

Chapter 2: STRATEGIES OF EVALUATION

In developing an evaluation strategy it is absolutely necessary to perform an initiating stage, in which the group designing the strategy will reflexively respond to the following questions:

Why?
Which is the aim for which the respective evaluation is realized? Which is the set of the envisaged evaluation objectives?

What?
Which exactly are the categories of the learning results following to be evaluated? Which will be the evaluation standards to be used?

Who?
Which will be the group of persons subjected to evaluation?

When?
At which moment of time and in which conditions the evaluation will be performed?

How?
How and by whom the evaluation will be performed? Who shall be paid by us for it?

What for?
What are we going to do with the evaluation results? How these results will be used, what for?

Observation: Usually, the one "ordering" for a large scale evaluation (such as national examinations) - by imposing a clear and coherent strategy - has the following claims:

1. Evaluation to be fast!
2. Evaluation to have maximum accuracy!
3. Evaluation to be cheap!

Observation: To our sadness (and to that of the organism requesting the evaluation too!), the three above mentioned conditions cannot be fulfilled at the same time!

STRATEGIC RULES OF EVALUATION

Always when a system of current evaluation is designed and when a concrete evaluation instrument is realised, the following evaluation rules have to be taken into consideration:

4 Build a positive image about evaluation! Evaluation should not be associated with failure, with penalty or control, but with the possibility to reflect upon the results. Evaluation should lead to motivating the evaluated person in order to obtain better performances!

4 The success of an evaluation must be understood, first of all, by modifying the attitude of the assessed person in relation to evaluation itself, and in forming an image as correct as possible about himself/herself, not only with his/her imperfections, but especially with the qualities he/she could capitalize in the future.

4 The good result of an evaluation should be prepared even from the moment of designing the contents; the evaluator should present too at that time which are his/her expectations concerning the persons who will be evaluated, what progress they should do in obtaining good results; an examination which is preceded by a presentation of evaluation objectives or by several evaluation criteria, will be much more efficient, will lead to better results than an examination in which the tested person has no idea about what will be evaluated!

4 Evaluation should be designed with the goal of judging the development phase of the individual's acquisitions, and thus it has to be thought as part of the learning process, even if it is realised by someone else then the person who will perform the teaching!

4 Evaluation it is not done only for the interest of the evaluated person, but also in the interest of that one who evaluates; the evaluator has at his/her disposal a privileged mean which confers him/her an objective image upon his/her action, upon the own successes or failures during the teaching-learning process!

4 Carefully design your evaluation tests and do not use evaluation in other aims than those for which it has been initially designed! Evaluation should not punish, but it must stimulate for the next learning stage! Even from a failed exam someone could learn a lot, if you taught him how to do this!

4 Evaluate what the evaluated person knows and do not try to demonstrate him/her that he/she does not know anything, because if he/she does not have the knowledge and skills, you might have part of the guilt (you did not teach him/her or you did not find a proper evaluation method)! Evaluation should specify the stage of the trainee's evolution, to clarify his/her mistakes or lacks, to correct the own mistakes of the teachers, to help and support the learner. This person should understand that the goal of evaluation is to objectively inform him/her upon what is still to be learned and to "secure" the quality of his/her acquirements and achievements. This is why evaluation should be done regularly, objectively, by preparing the trainee to be ready for the examination, and if the case is, being even preceded by a simulation of the examination!

4 Try to apply several evaluation instruments, and if it is only one tool, try to apply different categories of evaluation items inside that tool. Discover as many qualities of the tested person as possible, which could be emphasized based on his/her responses; do not use a unique criterion of evaluation (appreciation); the many and divers the criteria are, the more qualities (and not only imperfections) you will discover at each of the examined persons! There are not only the evaluation tools you are already used with and to which you cannot give up anymore! Also, not all the evaluation types should end by definitive and written sanction of the evaluated person (some of the examinations are really incompatible with this type of appreciation)! Try to "shade" and to improve the evaluation methods and also the way you present the results towards evaluated persons!

4 Do not generalize the data obtained by applying a non-standardised test, because this generalization may produce a lot of interpreting errors. This generalization is valuable at the level of the group you applied it to, or at the level of the individual; if you want to generalize you need to be sure that the evaluation instrument is valid, is reliable, is calibrated to the investigated population and it observes all the criteria of a standardized test! Applying a test should also observe own rules and general rules too, rules that suppose certain uniformity!

4 Interpreting the evaluation results should be understood as compulsory part of the evaluation, and has to be done (depending of the evaluation type) in front of the evaluated persons, the results being publicly announced as fast as possible when it is about an evaluation requested by an institution. Superficiality in evaluating or in analysing might compromise the entire evaluation demarche or might cancel its efficiency and value in the eyes of the evaluated person! Do not forget this person should receive the entire attention and respect, no matter the results he/she will obtain!

4 Do not neglect the fact that there is an efficiency curve of the evaluated person, which generally may be calculated for certain periods. If evaluation does not observe the rule of performing the evaluation in the moment which represents the optimal one for the evaluated person, then evaluation is not relevant and it will be finalised with visible lower results! If you want to measure exactly the trainees' capacity to resist at stress and to train them to successfully surpass a stress situation (such as a selection exam), simulate with them potential examination situations, use specialized psychological tests associated to your tests, but do not transform the examinations into a stress source!

Chapter 3: EVALUATION AS DIDACTIC PROCESS

Evaluation is a very complex and delicate activity, but it is compulsory in the view of acquiring a complete learning/training process. This is why the didactic evaluation has to match the needs and the style of the concerned person, making use of proper techniques, methods and tools.

Participatory evaluation (PE) is a method of inquiry within the family of participatory and action research. These traditions in research and evaluation grew out of conflicts and contradictions about how knowledge is created and used. There are at least three major traditions in participatory research and evaluation, all of which are concerned with democratizing the research process, and making the inquiry and the findings relevant and useful to the stakeholders for informing future actions.

The participatory action research model based on the Freirian theories of education (Fals-Borda, Tandon, Hall) grew out of the contradictions of using coercive, non-participatory field research methods in the largely participation-oriented field of adult education. In this tradition, issues of building power and promoting liberation and social justice are central.

The participatory action research model drawn from the action research tradition (Whyte) is based on the contradiction between management and workers in organizational decision-making. In this model, participation is aimed at increasing front-line workers' sense of empowerment, though not necessarily at changing the basic power relationships among members of the organization.

Participatory evaluation (PE) notes the contradiction between an evaluation's design and findings, and the lack of usefulness or relevance the information has for primary consumers and stakeholders (Cousins and Earl, 1992). PE draws from either or both of the previous traditions for its theoretical basis, but is distinctly evaluative in its purpose and design.

PE approaches seek to be practical, useful, formative and empowering; practical in that they respond to the needs, interests and concerns of their primary users; useful because findings are disseminated in ways in which primary users can use them; and formative because they seek to improve program outcomes. Finally, the more the project is determined, implemented and used by participants, the more empowering the experience will be.

Of course one can ask “How does PE differ from other forms of evaluation?”. We will try to answer this …

PE approaches usually are more appropriate for a formative, rather than summative evaluation. Participating organizations must understand that the goal is to provide information for program improvement or organizational development, not necessarily to make definitive statements about program outcomes. The agenda for the evaluation is not set by an outside funding source, a federal agency, or by the evaluator. Rather, in PE, both the role of the evaluator and that of the organization change. The evaluator is no longer the expert, but instead a teacher, collaborator and participant in a process. Organization members are integrally involved in establishing the questions to be asked and the methods to be used, in collecting and analysing data, and in writing up findings. Staff, clients, board members, and even interested community members, are involved in deciding whether to evaluate, what to evaluate, how to draw conclusions, how and when to disseminate findings, and how and when to implement recommendations. This means that rarely are PE findings that can be extended to other projects.

Further on we will present in brief the advantages and disadvantages of the PE approaches.

Since it is grounded in the experience of staff, clients, and participants, PE is more likely to provide information that is useful to program administrators and decision makers. PE enhances utilization of evaluation findings by changing the social construction of the organization. Rather than receiving (and resisting) an outside evaluation report, the process of participating in an evaluation gives ownership of the information to those most involved in carrying out the work of the organization: the staff, administrators, board members, clients, and

participants. PE is also viewed as more flexible and less rigid than traditional evaluation approaches. PE often results in cognitive, affective and political change within an organization-including increased communication between staff members, positive impacts on program development, and higher quality evaluations.

And now, let's see which the disadvantages of PE…are!

PE may be much more time-consuming for both the evaluator and the organization than a traditional goal-oriented evaluation where the questions to be asked and the methods to be used are set in advance, or established by the evaluator working with only one or two administrators. Staff will need to be allowed time from regular duties in order to participate effectively in the evaluation; clients and participants may need special assistance to become integrally involved in the evaluation. To assure adequate participation by all involved, rewards and consequences must be clearly spelled out.

For an entire evaluative process to be participatory, the details of the evaluation cannot be fully identified in advance (such as to a funding source). This is because specific reporting criteria or other evaluation guidelines dictated by sponsors or financers limit the participation and input of both evaluators and non-evaluators. The final result of a truly participatory process is entirely in the hands of the participants, not the evaluator or an outside source. This can empower participants, but means that in order to use PE, the organization must be committed to the endeavour and the context must be appropriate. It is always possible, however, to use some participatory methods at different stages of the evaluation process (for example when generating important evaluative questions at the beginning, or when developing conclusions based on data findings at the end), but do not commit it to an entire participatory process.

Evaluations are of different types and forms, and may be classified in accordance to certain criteria.

1st classification

Based on the criterion of ***the moment when*** evaluation is performed, we can distinguish the following forms of evaluation:

(a) ***initial*** evaluation - realised at the beginning of a training/learning sequence

(b) ***continuing*** evaluation - performed over the entire development of training process.

As methods and procedures used in continuing evaluation we can mention here:

- observation and verbal appreciation;
- oral questioning;
- written papers;
- standardized tests;
- practical works;
- projects.

A solid current evaluation system supposes the following indispensable elements:

- A set of general and operational evaluation objectives together with the associated standards;
- A database with evaluation items, adequate to the defined objectives;
- Calibrated evaluation instruments, adequate to the defined objectives and to the evaluated community.
- A system of certifying the current, and/or final, and/or summative evaluation.

(c) ***summative*** evaluation (or global, cumulative evaluation) – usually achieved after large periods of time, in the end of certain temporal or thematic sequences (such as chapter, course, training cycle, etc.)

2nd classification

In accordance to the evaluation ***target***:

(a) ***process*** evaluation – when the training/teaching process and the didactic demarche itself is evaluated;

(b) ***product*** evaluation – when the final product of the training/teaching process (the trainee/student/pupil) is evaluated;

(c) ***system*** evaluation – when the training system and its subcomponents are evaluated.

In order the evaluation functions to be active and functional, it is necessary to use in a very well balanced way all the evaluation strategies and the different types of evaluation techniques and instruments as well.

Normally, within the current educational system there are preponderantly used the traditional evaluation methods such as:

- written tests;
- oral tests;
- practical works.

Each of these traditional methods has its own advantages and disadvantages! This is why they must be combined into an optimal manner!

Thinking about evaluation, besides the traditional methods some alternative evaluation methods should be also used:

- systematically observation of the trainee's behaviour through:
  - evaluation and/or self-evaluation cards;
  - checking/testing lists;
  - classification scales.
- chat;
- questionnaire;
- interview;
- investigation;
- project;
- essay (report, review);
- portfolio, (ePortfolio) etc.

Observation: The alternative methods offer to the trainer supplementary information about the trainee's activity and about the level of skills acquiring. Thus, they are completing the data provided by the traditional methods and make the evaluated person to feel more comfortable and secure!

Chapter 4: EVALUATION ITEMS: DEFINITIONS AND EXAMPLES

THE ITEM represents the smallest identifiable component of an evaluation instrument (test) and has the following structure:

- the premises and the tasks;
- the solving pathway;
- the required level of performance.

Speaking about evaluation items, they can be classified, based on the criterion of "objectivity in assessment" into the following categories:

1. **Objective items**
2. **Semi-objective items**
3. **Subjective items (or non-objective, or open)**
4. **Non-standardized items**

In the following lines we shall present and discuss each of these types of evaluation items.

1. OBJECTIVE ITEMS

- are preponderantly used in evaluations trying to reveal the progress, especially in standardized tests.
- have a high level of objectivity in assessing the learning results.
- do not need detailed evaluation grids, because the mark (or the evaluation points) is (are) awarded only when the correct answer is ticked off.

Some types of objective items are rendered bellow.

1.1. Dual choice items: request the trainee to select one answer out of two possible answers: true/false; right/wrong; yes/no; agreement/disagreement; general/particular; smaller/bigger; statement of opinion/factual statement, etc.

Item example:

The **field** of evaluation: active citizenship – life hygiene.
Objective: the trainee must to identify the value of true or false of an enunciation (sentence).
Stem: Please read carefully the sentence below. If you appreciate the sentence is true, mark the letter **T**. If you think the sentence is false (not true), mark the letter **F**.

T	F	
☐	☐	A moral society is a healthy one.

1.2. Pair items: solicit the trainees to establish correspondences/associations between words, sentences, phrases, letters or other categories of symbols which are disposed in two columns. The elements of first columns are the "premises" and those of the second columns represent the "answers". The task to be performed is clearly formulated in the stem and the correspondence/association may be between terms/definitions, rules/examples, symbols/concepts, etc.

Item example:

The **field** of evaluation: active citizenship.
Objective: the trainee must to establish correlations between names of European institutions and their attributes.
Stem: Please write in the empty spaces of the left column (containing names of European institutions) the letters of the right column (containing the main attribute/function of each institution) which you consider that match the elements of the left column.

___ 1. Council of the European Union (Council of Ministers)	A. Council of deputies and senators
___ 2. European Commission	B. Group of individual EU experts
___ 3. Parliament of Europe	C. The main decisional forum of EU
	D. Authority having prerogatives of initiative, implementation, management and control
	E. Associations of industrialists
	F. Union of politicians
	G. Reunion of the representatives of 370 billions of EU citizens

1.3. Multiple choice items: presume the existence of a premise (enunciation) and of a list of options (possible solutions). Among these options (solutions), only one is correct and represents the "key" (the answer); the others are not correct and they are called "distractors". The trainee must choose/identify the key.

Item example:

The **field** of evaluation: consumer's education – safety in farming and gardening.
Objective: the trainee must to choose/identify certain properties of the chemical substances.
Stem: Read carefully the enunciation below, select from the offered options the one you consider as correct and mark it in (☒) the solutions grid.
"The chemicals used in farming and gardening are:

a. insoluble
b. non-pollutant
c. harmless for snails
d. dangerous for the humans"

2. SEMI-OBJECTIVE ITEMS

- request an answer which can be limited in terms of dimension, form, content by the structure of the stem/question;
- are characterised by a strongly structured task;
- reduce the trainee's freedom to re-organize the received information and to formulate the answer in the wanted form;
- solicit the trainee to prove/demonstrate not only the knowledge but also the capacity to structure/elaborate the shortest and the most correct answer.

The types of semi-objective items are rendered bellow.

2.1. Short answer items: the trainee must to formulate the answer in the form of a short sentence or phrase, 1- 2 words, a number, a symbol, an acronym, etc. The task is presented in the form of an indirect question.

Item example:

The **field** of evaluation: active citizenship.
Objective: the trainee must to render out knowledge about European Union and integration process.
Stem: Please answer the question below by writing the response in the empty space between the round brackets:
"How a country which started negotiations chapters to join the European Union is called? (_____)"

2.2. Completion items: the trainee must to express the answer in the form of one or maximum two words, which to match the stem and to be filled in the stem, thus the stem to become correct and complete. The task is presented in the form of an incomplete statement/sentence.

Item example:

The **field** of evaluation: environment protection.
Objective: the trainee must to apply laws, principles, etc. studied in real situations of the surrounding environment.
Stem: Please fill in the blank space from the following sentence, thus to become correct and complete.
"The increase of the temperature is in . . . . . proportion with the quantity of the CO2 poured in the atmosphere."

2.3. Structured questions: the trainee must to construct the response in a way that is between the objective items (closed items) and the subjective ones (open items). The item is formulated in the form of several sub-questions which might be objective, semi-objective or mini-essay, linked all together by a common element. A structured question could comprise for example:

- some materials/stimulus (texts, data, diagrams, graphics, etc.);
- sub-questions;
- supplementary data;
- other sub-questions.

Item example:

The **field** of evaluation: consumer's education.
Objective: the trainee must to analyse possible hypothesis, finding the explanation for the way certain phenomena and processes are produced.
Stem: "Fresh fruits are very perishable."

a. At room temperature they change their colour, smell and consistency. Which are the processes taking place in these conditions?

b. Why it is necessary to store fresh fruits into a cold place?

c. Give an example of method to properly preserve the fresh fruits.

3. SUBJECTIVE ITEMS (or non-objective, or open)

- represent one of the traditional evaluation forms;
- are relatively easy to be conceived;
- test the objectives envisaging the originality, the creativity and the personal feature of the answer.

The types of subjective items are rendered bellow.

3.1. Problem-situations solving: represents an activity in which the trainees are involved in developing creativity, divergent thinking, imagination, capacity to generalize, to reframe a problem, etc.

Item example:

The **field** of evaluation: active citizenship – labour market.
Objective: the trainee must to apply knowledge, skills, abilities, aptitudes, etc. to solve problems appeared in concrete situations.
Stem: "Suppose your employer just fired you out without a previous notice. You are married, with 2 children, your wife is a house keeper, kids are attending school (both of them) and you live in a rented house. You have no other financial support. How do you solve the problem of a new employment contract for you? Which is your plan? What are the steps you will take in finding a new job on the labour market?

3.2. Essay: requests the trainee to construct, to produce an independent answer in accordance to a set of imposed requirements; it valorises the ability to recall, organize and integrate the ideas; the skills to express himself/herself in written and to interpret and apply different data.

Based on the **dimension** of the expected answer, two categories of essays may be designed:

a. **Essay with limited answer (mini-essay)** in which a limit of number of words, paragraphs, line sis specified (in the stem of the essay) ;
b. **Essay with extended answer** (for which operates only the time limit for adequately solving it).

Based on the **type** of the expected answer, two categories of essays may be designed:

a. **Structured or semi-structured essay** (in which, by the help of some clues, suggestions or requests the expected answer is "sorted" and "oriented");
b. **Free essay** (proper for objectives envisaging the creative and imaginative thinking/writing, the creativity, the immagination, etc.).

Item example (structured essay):

The **field** of evaluation: consumer's education – saving energy.
Objective: the trainee must to express the ability to use knowledge, skills, abilities, aptitudes, etc. in properly managing the energy resources and in utilising them.
Stem: Please compose an essay with the title" Saving energy at home" for which to use the following plan of ideas:

a. domestic types of energy (brief presentation);
b. economical use of electric current and water in housing;
c. methods to save domestic energy;
d. motivating the others in saving energy at home.

4. NON-STANDARDIZED ITEMS

- usually used when evaluating very high abilities (recommended for example in testing the students at medicine, aviation, etc.);
- have an increased level of complexity and difficulty.

This type of items request the trainee to assess the value of true or false of two sentences (phrases) linked by the conjunction ***BECAUSE*** and to establish a relation between them.

Item example:

The **field** of evaluation: consumer's education – life hygiene.
Objective: the trainee must to appreciate the value of the statements in the stem and to identify the correlation between them.
Stem: Please read carefully the sentences below and select one of the (a, b, c, d, e) possibilities which you consider is matching for them:
a. both sentences are true and correlated each-other;
b. both sentences are true but not correlated each-other;
c. both sentences are false;
d. first sentence is true, the second sentence is false;
e. first sentences is false, the second sentence is true.
"In order to be healthy, a young child needs to wash his teeth at least two times a day BECAUSE fresh carrots contain large quantities of carotene."

Item example [EPIC classification]:

The **field** of evaluation: increasing of learning achievements via dialog in a constructivist learning environment
Objective: the trainee must to appreciate the "Domains of Learning Dispositions" in a constructivist learning environment
Stem: The **EPIC** classification shows the kind of dispositions that are considered necessary for developing pupil autonomy. The challenge for the teacher is to design a curriculum that will stimulate and necessitate their use, causing pupils to inquire, collaborate, share ideas, consider alternatives and reach conclusions when contemplating problematic situations.

Expressive

- Confidence: expressing an idea, thinking and communicating with clarity and precision
- Being curious: expressing curiosity or the desire to know more: questioning and problem posing,
- Open-mindedness: speculating, predicting, thinking aloud, remaining alert to situations…
- Responsiveness: responding with wonderment and awe, fun and enjoyment

Productive

- Exploratory: investigating, experimenting, and gathering data using all the senses
- Strategic: planning, setting goals, planning procedures, prioritizing, organising and ordering events during problem solving
- Applying: using what is known: applying previous knowledge and understanding to unfamiliar and unknown situations
- Monitoring: checking progress and thinking about thinking, reflective

Innovative

- Adventurous: handling uncertainty, taking responsible risks, sense of adventure, trying out new ways of doing things
- Flexibility: thinking flexibly, suggesting alternatives, considering options, seeing things in different ways
- Being creative: creating, generating, imagining and innovating
- Evaluating: evaluating a method or outcome, suggesting modifications, or improvements

Collaborative

- Interdependence: interacting and thinking interdependently, working together, managing impulsivity, accepting responsibility

- Resilience: persistence in negotiating ideas and reaching conclusions
- Sensitivity: listening with understanding and empathy, suspending judgement
- Coaching: scaffolding, supporting and encouraging, assisting and guiding

Levels of interaction, exploration, and engagement

Level 1
Will not want to volunteer or get involved. Practices avoidance tactics and is reluctant to engage naturally. No signs of any explorative activity and any apparent activity has an absence of cognitive demand. Passive involvement with a low-level energy input.
Level 2
Engagement is conditional. Concentration is weak and is easily distracted. Lots of non-activity and not focused enough for exploratory activity. Will attempt to engage only if asked directly by the teacher and normally in response to a fairly low-level question nor demand.
Level 3
Engagement is hesitant and explorations are routine and unimaginative. Real signs of engagement are missing. Some progress but lacking in energy and concentration. Easily distracted. Requires careful scaffolding of questions and probing to elicit an appropriate response
Level 4
Activity with intense moments: not continuous. Engagement and exploration is characterised by concentration, persistence and energy. Can manage distractions. Volunteers an answer but one that is incomplete: requires further probing to reveal a complete understanding.
Level 5
Continuous intense activity characterised by concentration, creative exploration, initiative, energy and persistence. Volunteers answers and questions naturally. Engages in problem solving and problem posing. Coherent, well reasoned, holistic responses that demonstrate understanding and completeness.

EPIC Evaluation Result Sample:

Student: **Vanessa**
Learning domain: **Sciences**

	L1	L2	L3	L4	L5
Expressive					
Confidence: expressing an idea, thinking and communicating with clarity and precision				**X**	
Being curious: expressing curiosity or the desire to know more: questioning and problem posing,					**X**
Open-mindedness: speculating, predicting, thinking aloud, remaining alert to situations…					**X**
Responsiveness: responding with wonderment and awe, fun and enjoyment					**X**
Productive					
Exploratory: investigating, experimenting, and gathering data using all the senses					**X**
Strategic: planning, setting goals, planning procedures, prioritizing, organising and ordering events during problem solving					**X**
Applying: using what is known: applying previous knowledge and understanding to unfamiliar and unknown situations					**X**
Monitoring: checking progress and thinking about thinking, reflective					**X**
Innovative					
Adventurous: handling uncertainty, taking responsible risks, sense of adventure, trying out new ways of doing things					**X**
Flexibility: thinking flexibly, suggesting alternatives, considering options, seeing things in different ways				**X**	
Being creative: creating, generating, imagining and innovating				**X**	
Evaluating: evaluating a method or outcome, suggesting modifications, or improvements					**X**

Collaborative					
Interdependence: interacting and thinking interdependently, working together, managing impulsivity, accepting responsibility					**X**

Resilience: persistence in negotiating ideas and reaching conclusions					**X**
Sensitivity: listening with understanding and empathy, suspending judgement					**X**
Coaching: scaffolding, supporting and encouraging, assisting and guiding					**X**

L1 Reluctant to engage naturally, no exploratory activity, practices avoidance tactics
L2 Engagement is haphazard, exploratory activity not focused, easily distracted
L3 Engagement is hesitant, explorations routine, requires probing and prompting
L4 Engagement is more continuous, exploration more imaginative, interactive
L5 Engagement is continuous, open-ended and reasoned, autonomous and responsible

Chapter 5: MAIN TYPES OF TESTS TO BE USED IN THE EVALUATION PROCESS

PROGRESS TESTS (tests of efficiency)

Have as goal to emphasize what has been achieved through a given program.
Depending on the mode to report the results, the progress tests may be:

- **normative tests,** in which the trainee's performances are evaluated in relation with that ones of a reference group;
- **criterial tests,** in which the trainee's performances are appreciated in relation with the objectives.

Comparison between normative tests (NT) and the criterial tests (CT)

NT presents the **relative performance** of the individual in relation with the other people of the tested group, without giving information about what the respective individual is really able to do; CT presents **the absolute performance** but does not classify the tested subjects based on their abilities.

Do not forget the fact that having the criteria of a test does not means the test is a criterial one, because the normative tests have established criteria too. Still, for the CT the criteria (objectives) should be very clear defined.

Normative test (NT)	Criterial test (CT)
Is testing a large number of issues (but through a small number of items per each issue)	Verifies only few aspects, but in deep (by a large number of items for each aspect)
A large distribution of the scores (marks) is expected	A reduced variability of the scores (marks) appears
Is recommended for selection situations in imposed conditions	Is useful in selection situations which requires minimal necessary abilities
Offers a good rate of covering the curriculum	Presents better the failure or the success (concerning each criteria, or in relation with a certain number of investigated domains)
	It is recommended especially for diagnosis and evaluation (program evaluation especially)

So, each of these type of tests has its own advantages and disadvantages, being necessary a correct use of them, correlated with the testing type. But many times a "battery" of tests is necessary, battery made of one NT and one NT. Concerning the manner to build the tests by selecting the items, item elaboration rules and also assembling methodology of the test should be observed as well.

APTITUDES TESTS

Intend to make the prognosis of the future performances. These tests emphasize certain aptitudes which trainees have at a certain moment and which may favour in future, the success into a specific subject.

DIAGNOSIS TESTS

Have as goal to underline the lacks and the mistakes of the trainees.

PERFORMANCE TESTS

Testing the performance refers to evaluating an activity beyond the scope of the simple techniques with "pen and paper".

The examples include:

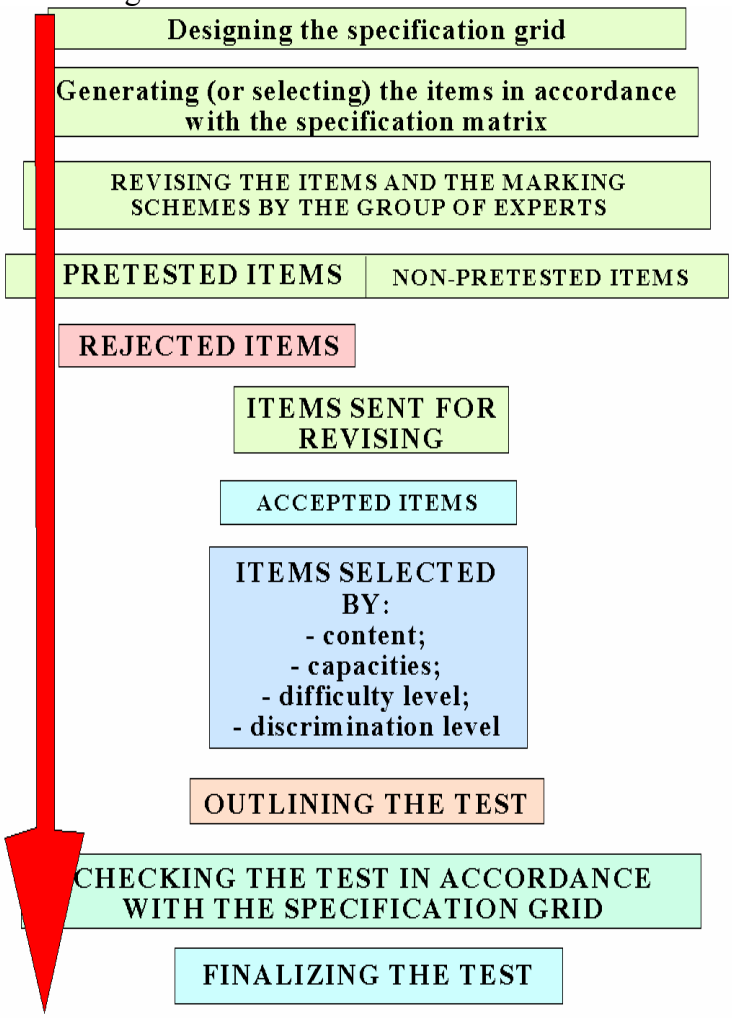
- tests for the field of sciences;
- tests for oral understanding in the field of foreign languages;
- portfolios in the artistic domains.

COMPOSING A TEST

When a test is designed, four important elements have to be taken into consideration:

- Specification grid
- Experts (for each subject separately)
- Evaluation grid
- Statisticians and their instruments for analysis and diagnosis

Generating a test observes the following flow:



Any test or evaluation instrument must be accompanied by the ***specification grid*** and by the ***evaluation grid*** in order to be complete, valid and efficient.

DEFINITION: ***Specification grid*** is a functional correlation between the type of competence that is intended to be evaluated and the type of evaluation item(s) used to asses the achievement of the respective competence.

Example:

	Dual choice items	Short answer items	Multiple choice items	….	Structured questions
Competence no. 1	3				3
Competence no. 2		9	13		
Competence no. 3	4				
Competence no. 4			13		7
	4				
Competence no. "N"	1	11	13		
Weight:	12%	20%	39%	…	10%
	Total weight : 100%				

A competence/skill/ability may be assesed through different types of items (more than one type). Vice versa, it is accepted to have a "missing" type of item when evaluating a certain competence (to do not use all item categories). Also, it is possible to assign more than one items per competence. To each kind of items, it doesn't matter their numbers, it is necessary to allocate an individual weight (which to reflect the correlation between the complexity of the competence and the difficulty of the evaluation item), thus the sum of the individual weights to be equal with 100%.

DEFINITION: ***Evaluation grid*** is the list containing:

- the correct answers or the solutions of the evaluation tasks (items);
- the points allocation for each of task (item) or the afferent marks/qualifications/scores etc;
- the time allocation for the tasks (items) contained in the evaluation instrument.

When designing an evaluation instrument certain basic qualities of the evaluation tool must be taken into consideration:

- validity
- reliability
- objectivity
- applicability

Validity refers to "the fact if the evaluation instrument really measures or not what was intended to measure" (Ausubel, 1981).

Here, several types of validity can be defined:

- content validity (the level at which the evaluation tool uniformly covers the major content elements it is testing);
- construct validity (the accuracy level at which an evaluation instrument measures a certain construct, such as the intelligence, the creativity, the success, etc.);
- concurrent validity (the concordance between the results obtained by applying that evaluation instrument and some similar behavioural criteria);
- predictive validity (the grade in which the evaluation tool makes the prognosis of the future performances).

Reliability is the quality of an evaluation instrument to provide constant results when it is applied successively.

Objectivity represents the level of concordance between the expert-judgement of independent evaluators related to a correct answer for each of the test items.

Applicability is the quality of an evaluation tool to be easily applied and processed.

Chapter 6: BASIC CONCEPTS IN ITEM AND TEST ANALYSIS

Making fair and systematic evaluations of the others' performance can be a challenging task. Judgments cannot be made solely on the basis of intuition, haphazard guessing, or custom. Teachers, employers, and others in evaluative positions use a variety of tools to assist them in their evaluations.

Tests are tools that are frequently used to facilitate the evaluation process. When norm-referenced tests are developed for instructional purposes, to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses.

Test analysis examines how the test items perform as a set. Item analysis "investigates the performance of items considered individually either in relation to some external criterion or in relation to the remaining items on the test".

These analyses evaluate the quality of items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole.

However, some best practices in item and test analysis are too infrequently used in actual practice.

It is summarised the recommendations for item and test analysis practices. These tools include item difficulty, item discrimination, and item distractors.

Item Difficulty

Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be. To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty. An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50.

A p value is basically a behavioural measure. Rather than defining difficulty in terms of some intrinsic characteristic of the item, difficulty is defined in terms of the relative frequency with which those taking the test choose the correct response. For instance, in the example below, which item is more difficult?

1. Who was Nicolae Titulescu?
2. Who was Winston Churchill?

One cannot determine which item is more difficult simply by reading the questions. One can recognize the name in the second question more readily than that in the first. But saying that the first question is more difficult than the second, simply because the name in the second question is easily recognized, would be to compute the difficulty of the item using an intrinsic characteristic. This method determines the difficulty of the item in a much more subjective manner than that of a p value.

Another implication of a p value is that the difficulty is a characteristic of both the item and the sample taking the test. For example, an English test item that is very difficult for an elementary student will be very easy for a high school student. A p value also provides a common measure of the difficulty of test items that measure completely different domains. It is very difficult to determine whether answering a history question involves knowledge that is more obscure, complex, or specialized than that needed to answer a math problem. When p values are used to define difficulty, it is very simple to determine whether an item on a history test is more difficult than a specific item on a math test taken by the same group of students.

To make this more concrete, take into consideration the following examples. When the correct answer is not chosen ($p = 0$), there are no individual differences in the "score" on that item. As

shown in Table 1, the correct answer C was not chosen by either the upper group or the lower group. (The upper group and lower group will be explained later.) The same is true when everyone taking the test chooses the correct response as is seen in Table 2. An item with a p value of .0 or a p value of 1.0 does not contribute to measuring individual differences, and this is almost certain to be useless. Item difficulty has a profound effect on both the variability of test scores and the precision with which test scores discriminate among different groups of examinees. When all of the test items are extremely difficult, the great majority of the test scores will be very low. When all items are extremely easy, most test scores will be extremely high. In either case, test scores will show very little variability. Thus, extreme p values directly restrict the variability of test scores.

Table 1
Minimum Item Difficulty Example Illustrating No Individual Differences

Group	Item Response				
				*	
		A	B	C	D
Upper group	4	5	0	6	
Lower group	2	6	0	7	

Note. * denotes correct response
Item difficulty: (0 + 0)/30 = .00p
Discrimination Index: (0 - 0)/15 = .00

Table 2
Maximum Item Difficulty Example Illustrating No Individual Differences

Group	Item Response				
				*	
		A	B	C	D
Upper group	0	0	15	0	
Lower group	0	0	15	0	

Note. * denotes correct response
Item difficulty: (15 + 15)/30 = 1.00p
Discrimination Index: (15-15)/15 = .00

It is now accepted that: items tend to improve test reliability when the percentage of students who correctly answer the item is halfway between the percentage expected to correctly answer if pure guessing governed responses and the percentage (100%) who would correctly answer if everyone knew the answer.

For example, many teachers may think that the minimum score on a test consisting of 100 items with four alternatives each is 0, when in actuality the theoretical floor on such a test is 25. This is the score that would be most likely if a student answered every item by guessing (e.g., without even being given the test booklet containing the items).

Similarly, the ideal percentage of correct answers on a four-choice multiple-choice test is not 70-90%. The ideal difficulty for such an item would be halfway between the percentage of pure guess (25%) and 100%, (25% + {(100% - 25%)/2}. Therefore, for a test with 100 items with four alternatives each, the ideal mean percentage of correct items, for the purpose of maximizing score reliability, is roughly 63%.

Tables 3, 4, and 5 show examples of items with p values of roughly 63%.

Table 3
Maximum Item Difficulty Example Illustrating Individual Differences

Group	Item Response				
				*	
		A	B	C	D
Upper group	1	0	13	3	
Lower group	2	5	5	6	

Note. * denotes correct response
Item difficulty: (13 + 5)/30 = .60p
Discrimination Index: (13-5)/15 = .53

Table 4
Maximum Item Difficulty Example Illustrating Individual Differences

Differences Group	Item Response				
				*	
		A	B	C	D
Upper group	1	0	11	3	
Lower group	2	0	7	6	

Note. * denotes correct response
Item difficulty: (11 + 7)/30 = .60p
Discrimination Index: (11-7)/15 = .267

Table 5
Maximum Item Difficulty Example Illustrating Individual Differences

Group	Item Response				
				*	
		A	B	C	D
Upper group	1	0	7	3	
Lower group	2	0	11	6	

Note. * denotes correct response
Item difficulty: (11 + 7)/30 = .60p
Discrimination Index: (7 - 11)/15 = .267

Item Discrimination

If the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. Two indices can be computed to determine the discriminating power of an item, the item discrimination index, D, and discrimination coefficients.

Item Discrimination Index, D

The method of extreme groups can be applied to compute a very simple measure of the discriminating power of a test item. If a test is given to a large group of people, the discriminating power of an item can be measured by comparing the number of people with high test scores who answered that item correctly with the number of people with low scores who answered the same item correctly. If a particular item is doing a good job of discriminating between those who score high and those who score low, more people in the top-scoring group will have answered the item correctly.

In computing the discrimination index, D, first score each student's test and rank order the test scores. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. 27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis. There need to be as many students as possible in each group to promote stability, at the same time it is desirable to have the two groups be as different as possible to make the discriminations clearer. The use of 27% maximizes these two characteristics.

The discrimination index, D, is the number of people in the upper group who answered the item correctly minus the number of people in the lower group who answered the item correctly, divided by the number of people in the largest of the two groups. It may be stated that when more students in the lower group than in the upper group select the right answer to an item, the item actually has negative validity. Assuming that the criterion itself has validity, the item is not only useless but is actually serving to decrease the validity of the test.

The higher the discrimination index, the better the item because such a value indicates that the item discriminates in favour of the upper group, which should get more items correct, as shown in Table 6. An item that everyone gets correct or that everyone gets incorrect, as shown in Tables 1 and 2, will have a discrimination index equal to zero. Table 7 illustrates that if more students in the lower group get an item correct than in the upper group, the item will have a negative D value and is probably flawed.

Table 6
Positive Item Discrimination Index D

Group	Item Response				
				*	
		A	B	C	D
Upper group	3	2	15	0	
Lower group	12	3	3	2	

Note. * denotes correct response
74 students took the test
27% = 20(N)
Item difficulty: (15 + 3)/40 = .45p
Discrimination Index: (15 - 3)/20 = .60

Table 7
Negative Item Discrimination Index D

Group	Item Response				
				*	
		A	B	C	D
Upper group	0	0	0	0	
Lower group	0	0	15	0	

Note. * denotes correct response
Item difficulty: (0 + 15)/30 = .50p
Discrimination Index: (0 - 15)/15 = -1.0

A negative discrimination index is most likely to occur with an item covers complex material written in such a way that it is possible to select the correct response without any real understanding of what is being assessed. A poor student may make a guess, select that response, and come up with the correct answer. Good students may be suspicious of a question that looks too easy, may take the harder path to solving the problem, read too much into the question, and may end up being less successful than those who guess. As a rule of thumb, in terms of discrimination index, .40 and greater are very good items, .30 to .39 are reasonably good but

possibly subject to improvement, .20 to .29 are marginal items and need some revision, below .19 are considered poor items and need major revision or should be eliminated.

Discrimination Coefficients

Two indicators of the item's discrimination effectiveness are point biserial correlation and biserial correlation coefficient. The choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.

Point biserial. The point biserial (r_{pbis}) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. The r_{pbis} tells more about the predictive validity of the total test than does the biserial r, in that it tends to favour items of average difficulty. It is further suggested that the r_{pbis} is a combined measure of item-criterion relationship and of difficulty level.

Biserial correlation. Biserial correlation coefficients (r_{bis}) are computed to determine whether the attribute or attributes measured by the criterion are also measured by the item and the extent to which the item measures them. The r_{bis} gives an estimate of the well-known Pearson product-moment correlation between the criterion score and the hypothesized item continuum when the item is dichotomized into right and wrong. The r_{bis} simply describes the relationship between scores on a test item (e.g., "0" or "1") and scores (e.g., "0", "1",..."50") on the total test for all examinees.

Distractors

Analyzing the distractors (e.i., incorrect alternatives) is useful in determining the relative usefulness of the decoys in each item. Items should be modified if students consistently fail to select certain multiple choice alternatives. The alternatives are probably totally implausible and therefore of little use as decoys in multiple choice items. A discrimination index or discrimination coefficient should be obtained for each option in order to determine each distractor's usefulness. Whereas the discrimination value of the correct answer should be positive, the discrimination values for the distractors should be lower and, preferably, negative. Distractors should be carefully examined when items show large positive D values. When one or more of the distractors looks extremely plausible to the informed reader and when recognition of the correct response depends on some extremely subtle point, it is possible that examinees will be penalized for partial knowledge.

Computing reliability estimates for a test scores to determine an item's usefulness to the test as a whole. The total test reliability is reported first and then each item is removed from the test and the reliability for the test less that item is calculated. From this the test developer deletes the indicated items so that the test scores have the greatest possible reliability.

REFERENCES

1. SOCRATES, COMENIUS C21 Project **"Dial-Connect: Using DIALogue to CONNECT learning minds"**, 118155 - CP -1-2004-1- UK - COMENIUS – C21, 2004.
2. G. Chirleşan, D. Chirleşan, I. Chelu, S. Chirilă, M. Mândruț, R. Pop, O. Rusu: "Ghid de Evaluare la Disciplina "Fizică"" - lucrare sub egida MEN - SNEE, Editura Trithemius Media, Bucureşti, 1999, ISBN: 973-98822-3-4
3. ***, Set of evaluation instruments for Adult Education and Training, Editura Universității din Pitești, 2004, ISBN 973-690-285-4;

4. ***, Guide d'evaluation pour l'Education des Adultes, Editura Universităţii din Piteşti, 2004, ISBN 973-690-284-6;
5. ***, Evaluation guide for Adult Education and Training, Editura Universităţii din Piteşti, 2004, ISBN 973-690-283-8;
6. SOCRATES, GRUNDTVIG 1 Project **ALERT - 2**, 101067-CP-1-2002-1-UK – GRUNDTVIG-G1, 2002
7. Susan Matlock-Hetzel (1997) Basic Concepts in Item and Test Analysis. Texas A&M University.
8. Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.
9. Ebel, R.L., & Frisbie, D.A. (1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.
10. Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: MacMillan.
11. Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), Educational Measurement (p. 141). Washington DC: American Council on Education.
12. Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), Educational measurement (pp. 335-366). Phoenix, AZ: Oryx Press.
13. Nunnally, J.C. (1972). Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill.
14. Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.
15. Popham, W.J. (1981). Modern educational measurement. Englewood Cliff, NJ: Prentice-Hall.
16. Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.
17. Thompson, B., & Levitov, J.E. (1985). Using microcomputers to score and evaluate test items. Collegiate Microcomputer, 3, 163-168.
18. Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: MacMillan.
19. Wiersma, W. & Jurs, S.G. (1990). Educational measurement and testing (2nd ed.). Boston, MA: Allyn and Bacon.
20. Wood, D.A. (1960). Test construction: Development and interpretation of achievement tests. Columbus, OH: Charles E. Merrill Books, Inc.