Evaluating evaluation: a study carried out on 6th grade Natural Sciences' written tests

Cláudia Rodrigues (1) and José Precioso (2)

1- School EB1/JI Quinta de S. Gens, Senhora da Hora, Matosinhos

2- Institute of Education and Psychology, University of Minho, Braga

ABSTRACT

Introduction: Several authors say that tests are the main and, sometimes, the sole mean of evaluating students. If these instruments are frequently used during the evaluation process, than it is essential that the teachers formulate them with technical quality, as on its application often depends the students' obtained grade and, consequently, their future. It was the importance of tests as a (still) privileged tool in the evaluation process and the chance to know its quality that took us to develop this thesis. **Objectives:** 1) To determine if the teacher-formulated tests are conceived to evaluate the specific skills required for scientific literacy; 2) To determine whether those tests evaluate the achievement of the educational goals proposed by Bloom's taxonomy; 3) To analyze the summative evaluation sheets for the 6th grade of Natural Sciences, in what concerns their technical requirements. **Methods:** Recurring to a process of convenience sampling, we used a sample of 24 summative evaluation sheets concerning "Human Reproduction and growth" unit and we performed a content analysis. **Results:** The data brought on by this study states that the largest percentage (90%) of the tests' questions is part of the substantive knowledge skills' category. Items that evaluate other kinds of skills are almost non-existent in the tests; 90% of all of the tests' questions concern the domains of "knowledge" (50%) and "understanding" (40%) – that asserts that the evaluation practices are often reduced to mere memory exercises.

Conclusion: So the results obtained by this study testify the need to improve the process of initial and continuous training of teachers under evaluation.

1. Introduction

Basic Education's Curricular Reorganization has introduced several modifications into the education system in Portugal. All these modifications go in the sense of establishing as the teacher's and school's main mission "not the teaching of disciplinary contents, but the personal competences' development." (Machado, 2002, *in* Perrenoud *et al,* 2002, p.137).

Understood by the 2001 Basic Education National Curriculum as "knowledge in action" or "knowledge in use", the concept of "skill" integrates knowledge, capacities, attitudes and values, and it expresses itself in a rational mobilization of knowledge so as to solve daily common problems (DGEB, 2001).

The document previously referred to encloses skills in a big group named "essential skills", since they are indispensable to the student's growth as an critical, autonomous, reflexive and social individual who needs to integrate a society constantly transforming itself and evolving (DGEB, 2001). These **essential skills** which, in their turn, comprise ten **general skills** that students should develop throughout basic education in all disciplinary and non-disciplinary curricular areas, and **specific skills** to be developed by each of those areas.

In the particular case of the Natural Sciences' area, the essential skills specific for scientific literacy, in its four domains, are designated as: knowledge (substantive, processual and epistemological), reasoning, communication and attitudes.

Although it recommends some methods and strategies to be adopted in the development of those skills, the referred publication does not contemplate, however, the process of evaluating skills and, consequently, the means to be used.

The use of tests' fame as the almost unique and exclusive formal means of evaluation has so generalized that Valadares and Graça (1998) affirm that the test "constitutes the dominant and, sometimes, almost exclusive means of evaluating students" (Valadares and Graça, 1998, p.69).

If tests are determinant in the evaluation process, it is then fundamental that teachers elaborate then with technical quality, for on its application often depend the marks given to students and, consequently, their future.

Basing ourselves on and sharing Cardoso's discourse (*in* Estrela and Nóvoa, 1999, p. 84),"if tests (…) end up determining (…) what is taught and what is learned, then it is possible to infer, via observation of the tests effectively applied, the objectives which effectively have guided teachers and students", then we considered to be pertinent to proceed to their analysis, that is, to verify whether tests elaborated by teachers evaluate the skills specific for scientific literacy and the attainment of Bloom's educational goals, in such a way as to insure the achievement of one of education's general principles – to form "citizens capable to critically and creatively judge the social environment they integrate and to engage themselves in their progressive transformation. (LBSE, article 2, point 5)

2. Study's objectives

- To determine whether the tests elaborated by teachers are built so as to evaluate the skills specific for scientific literacy, proclaimed by the Physical and Natural Sciences Curriculum in its different domains.
- To determine whether the tests elaborated by teachers evaluate the attainment of the educational goals proposed by Bloom's taxonomy;

3. Method

Sample

The sample is constituted by a set of concerning 24 summative evaluation sheets concerning the 6th grade's program unit "Human Reproduction and growth" applied in schools from Vila Nova de Gaia and Oporto. The tests' selection was performed based on a sampling system of the non probabilistic type and of convenience (also known as non intentional or accidental sampling).

The tests' sample collected and selected, an analysis of content was done, which consisted in: 1) the elaboration of analysis grills constituted by two-entry tables. In the first column the tests' numbers were placed and in the ones left the skills relevant to the scientific

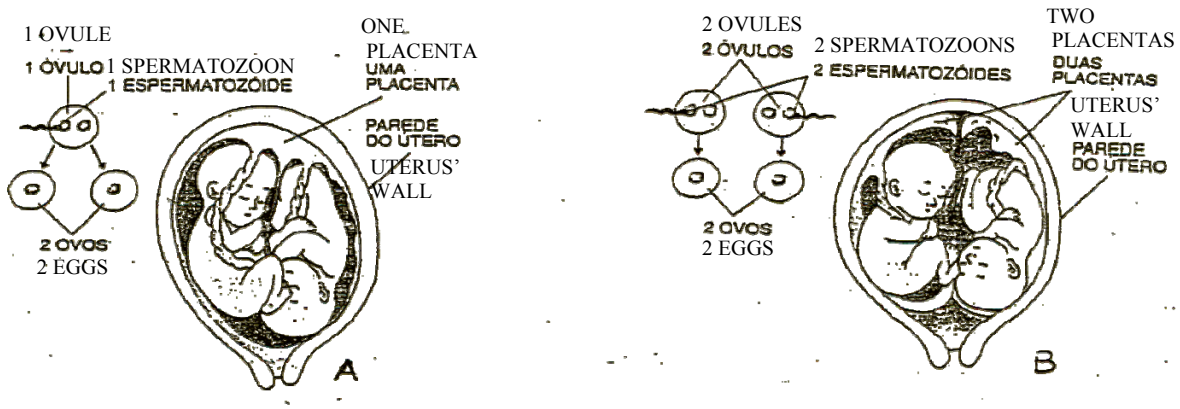
literacy (substantive knowledge, processual knowledge and epistemological knowledge, reasoning, communication and attitudes). In the second grill the columns contained Bloom's taxonomic category's objectives (knowledge, understanding, application, analysis, synthesis and evaluation); 2) the random attribution to each one of the tests of a alphabet letter, an individual analysis of each one's questions and the register, in each of the grills' categories, of the test questions fitting in it; 3) the determination of each category's absolute and relative prevalence.

The questions'/items' categorization cannot be assumed as rigorous and inflexible, nor was it an easy task, since we do not know the context in which the classes may have occurred and, consequently, the depth level of each sub-theme's exploration. The ignorance of each sub-theme's approach type may lead us to an item's wrong classification. Figure 1 presents one of the tests' questions.

This item's categorization, as it happened with others, constituted an enormous challenge. If the teacher did not cover this topic, the question might be considered as of the reasoning kind and placed in the skills grill, in the column concerning the "reasoning skills". If the teacher had covered this theme, the question might be placed in the substantive knowledge column (question/skills grill). In the second grill (question/objectives grill), if the teacher had covered the twins' formation theme, this item might be placed in the understanding or, at most, in the application column (in the Bloom's taxonomy objectives' grill); if the theme had not been developed, the question might be placed in the column of the analysis' objectives.

Figure 1
Question put in one of the sample tests

7. The following schemes refer to the formation of true twins and false twins:
7. Os esquemas seguintes referem-se à formação de gémeos verdadeiro e falsos:



Which of the schemes A or B refers to true twins?
Justify.
Qual dos esquemas A ou B se refere a gémeos verdadeiros?
Justifica.

4. Results

4.1. Tests' composition regarding the evaluation of scientific literacy skills

Through the table 1's analysis we can verify that of the 365 questions (corresponding to the sum of the number of all tests' questions): 332 are of substantive knowledge; 0 of processual knowledge; 1 of epistemological knowledge; 12 of reasoning;20 of communication and 0 of attitudes. The following are an example of substantive knowledge domain questions present in the tests evaluated: "Label the figures 1 and 2 *(feminine and masculine reproductive systems)*"; "What is the common organ to the masculine reproductive

system and the excretory system?"; "What is the name of this process *(ovulation)*?"; "Order the sentences so as to obtain the labour's stage sequence". The following are example of reasoning questions: "Relate fecundation with menstruation"; "Why is a pregnant woman advised to exercise physically?"; "Miss Berta had two twins, a girl and a boy. Will these be true or false twins? Justify your answer"; "Comment the following statement: "Even before birth a Mother ought already to take care of her baby.""

An example of a question which was put in the tests aiming to evaluate the communication skills is the following: "Try to describe how it is that, from just one cell, we become beings constituted by millions and millions of cells".

Table 1 – Tests' items' absolute and relative frequency, according to the domains of skills specific for scientific literacy.

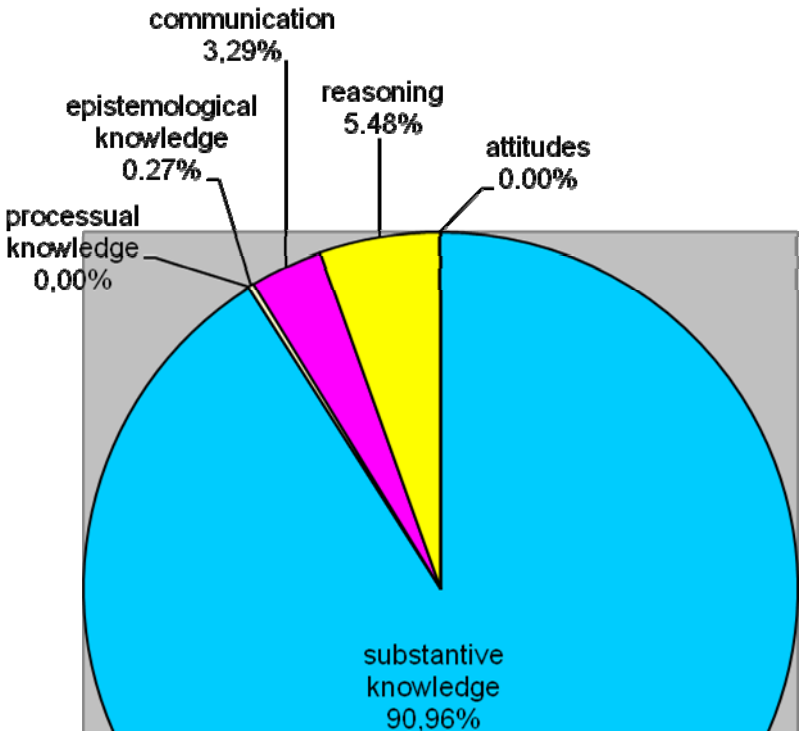
TESTS	SKILLS											
	Substantive Knowledge		Processual Knowledge		Epistemological Knowledge		Reasoning		Communication		Attitudes	
	F	%	F	%	F	%	F	%	F	%	F	%
A n= 14	14	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
B n= 20	18	**90**	0	**0**	0	**0**	1	**5**	1	**5**	0	**0**
C n= 25	24	**96**	0	**0**	0	**0**	0	**0**	1	**4**	0	**0**
D n= 13	13	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
E n= 16	14	**88**	0	**0**	0	**0**	1	**6**	1	**6**	0	**0**
F n= 10	14	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
G n=15	10	**86**	0	**0**	0	**0**	1	**7**	1	**7**	0	**0**
H n=15	15	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
I n= 13	13	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
J n= 18	16	**89**	0	**0**	0	**0**	2	**11**	0	**0**	0	**0**
K n=14	12	**86**	0	**0**	1	**7**	0	**0**	1	**7**	0	**0**
L n= 18	16	**89**	0	**0**	0	**0**	0	**0**	2	**11**	0	**0**
M n= 12	12	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
N n= 19	14	**74**	0	**0**	0	**0**	1	**5**	3	**16**	0	**0**
O n= 14	11	**79**	0	**0**	0	**0**	2	**14**	1	**7**	0	**0**
P n= 21	17	**81**	0	**0**	0	**0**	1	**5**	3	**14**	0	**0**
Q n= 20	17	**85**	0	**0**	0	**0**	1	**5**	2	**10**	0	**0**
R n=18	17	**94**	0	**0**	0	**0**	1	**6**	0	**0**	0	**0**
S n=10	9	**90**	0	**0**	0	**0**	0	**0**	1	**10**	0	**0**
T n=17	16	**94**	0	**0**	0	**0**	0	**0**	1	**6**	0	**0**
U n= 10	10	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
V n=9	7	**78**	0	**0**	0	**0**	1	**11**	1	**11**	0	**0**
X n= 14	14	**100**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**
Z n= 10	9	**90**	0	**0**	0	**0**	0	**0**	1	**10**	0	**0**
TOTAL n = 365	332		0		1		12		20		0	

n = test's number of questions

%	90,96	0,00	0,27	3,29	5,48	0,00

As we can verify in graph 1, 90,96% of the 332 belong to the "substantive knowledge" skill domain, which means that the almost totality of the tests' questions have as objective to evaluate students in what concerns scientific knowledge acquisition and concepts' and laws' interpretation. The promotion of processual knowledge (planning and performing experimental work), of epistemological knowledge (questioning Science), of reasoning, of communication and of attitudes is little significant, corresponding, respectively, to 0,00%, 0,27%, 3,29%, 5,48% and 0,00%, as graph 1 clearly shows:

Graph 1 - Tests' sample percentage of questions (items) by skills specific to the teaching of Sciences



4.2 Tests' composition regarding Bloom's taxonomy's educational objectives

Through the analysis of table 2 we can see that, of the questions' total present in tests: 191 are questions of knowledge; 146 are questions of understanding; 5 are questions of application; 2 are questions of analysis; 20 are questions of synthesis; 1 is a question of evaluation.

The following are examples of questions of the knowledge domain: "Provide two examples of sexually transmissible diseases", "Of the previous statements, indicate those which concern: a) Primary sexual characters; b) Man's secondary sexual characters; c) Woman's secondary sexual characters."

Following there are examples of questions of the understanding domain: "Make the correspondence between column A's terms and column B's expressions. *(reproductive systems' organs and their respective functions);* "Relate fecundation with menstruation."; "During pregnancy the mother ought not to smoke, drink alcohol nor take drugs. Justify this statement."; "Explain what the umbilical cord is for."; "Explain what you understand by sexually transmissible diseases."; "The seminal vesicles and the prostate produce liquids very important for spermatozoons. What are those liquids for?"; "Indicate the spermatozoon's way until it reaches the ovule (write in the boxes the number of the organs in order)."

Among the questions of the application domain we point out the following ones: "Supposing that the spermatozoon represented in scheme A was a chromosome Y carrier. What will the babies' sex be?"; "Indicate which figures represent risky AIDS transmissible behaviours.". We have selected the following as an example of analysis domain questions: "Indicate two differences between them. *(gametes)*. The following are examples of questions of the synthesis domain: "Explain what is the importance of reproduction for human being."; "Explain the difference between false twins and true twins."; "Based on the table, indicate the ingredient which should be increased during lactation. Justify your answer.". Example questions of the evaluation domain are the following: "Comment the following statement: "Even before birth the mother ought already to take care of her baby.""

Table 2 – Tests' items' absolute and relative frequency, according to the educational objectives of Bloom's taxonomy

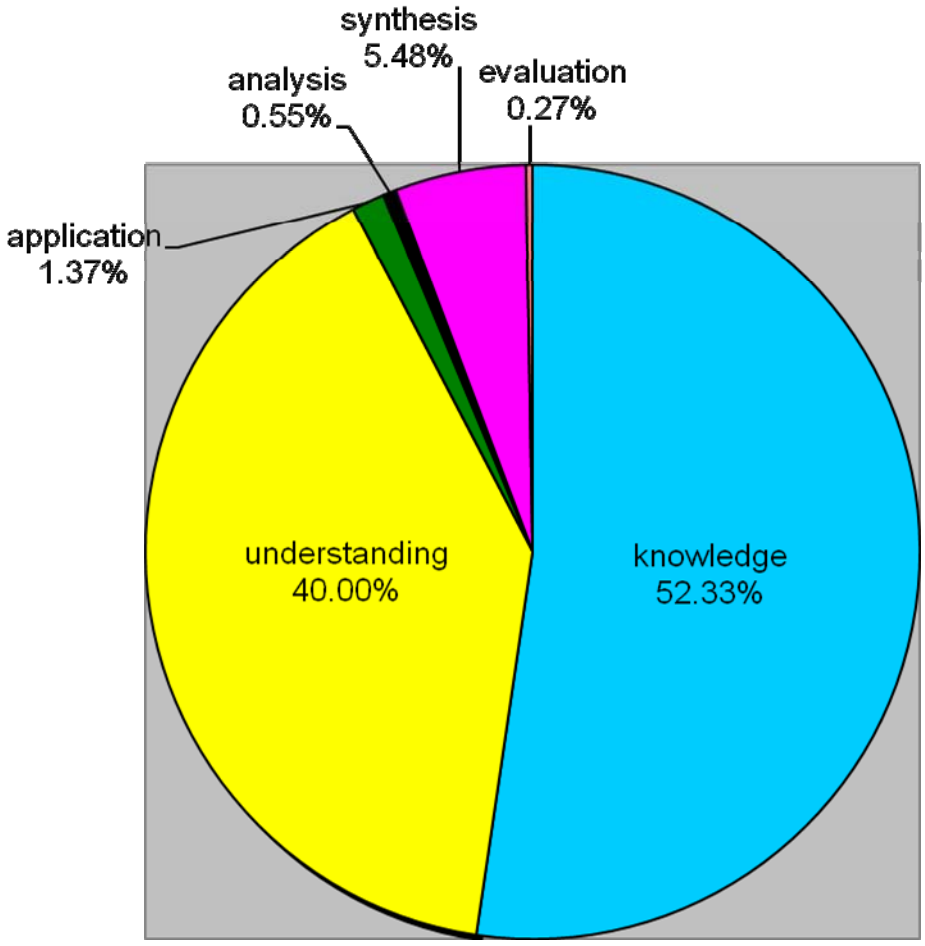
TESTS	OBJECTIVES											
	Knowledge		Understanding		Application		Analysis		Synthesis		Evaluation	
	F	%	F	%	F	%	F	%	F	%	F	%
A n= 14	11	**79**	3	**21**	0	**0**	0	**0**	0	**0**	0	**0**
B n= 20	10	**50**	9	**45**	1	**5**	0	**0**	0	**0**	0	**0**
C n= 25	11	**44**	12	**48**	0	**8**	0	**0**	2	**0**	0	**0**
D n= 13	6	**46**	6	**46**	0	**0**	0	**0**	1	**8**	0	**0**
E n= 16	6	**38**	5	**31**	1	**6**	0	**0**	4	**25**	0	**0**
F n= 10	6	**60**	4	**40**	0	**0**	0	**0**	0	**0**	0	**0**
G n=15	7	**46**	6	**40**	0	**0**	1	**7**	1	**7**	0	**0**
H n=15	10	**67**	5	**33**	0	**0**	0	**0**	0	**0**	0	**0**
I n= 13	10	**77**	3	**23**	0	**0**	0	**0**	0	**0**	0	**0**
J n= 18	10	**56**	7	**39**	0	**0**	0	**0**	1	**5**	0	**0**
K n=14	10	**71**	3	**22**	0	**0**	0	**0**	1	**7**	0	**0**
L n= 18	13	**72**	5	**28**	0	**0**	0	**0**	0	**0**	0	**0**
M n= 12	10	**83**	2	**17**	0	**0**	0	**0**	0	**0**	0	**0**
N n= 19	7	**37**	10	**53**	0	**0**	1	**5**	1	**5**	0	**0**
O n= 14	5	**36**	7	**50**	0	**0**	0	**0**	1	**7**	1	**7**
P n= 21	7	**33**	8	**38**	0	**0**	0	**0**	6	**29**	0	**0**
Q n= 20	10	**50**	8	**40**	2	**10**	0	**0**	0	**0**	0	**0**
R n=18	12	**67**	6	**33**	0	**0**	0	**0**	0	**0**	0	**0**
S n=10	6	**60**	3	**30**	1	**10**	0	**0**	0	**0**	0	**0**
T n=17	5	**29**	12	**71**	0	**0**	0	**0**	0	**0**	0	**0**
U n= 10	3	**30**	7	**70**	0	**0**	0	**0**	0	**0**	0	**0**
V n=9	5	**56**	3	**33**	0	**0**	0	**0**	1	**11**	0	**0**
X n= 14	7	**50**	7	**50**	0	**0**	0	**0**	0	**0**	0	**0**
Z n= 10	4	**40**	5	**50**	0	**0**	0	**0**	1	**10**	0	**0**
TOTAL n = 365	**191**		**146**		**5**		**2**		**20**		**1**	

n = test's number of questions

AVERAGE X	52,33	40,00	1,37	0,55	5,48	0,27

In terms of percentage, 52,33% (191) of the questions present in the tests are of the "knowledge" domain, which means that more than half the questions aim to evaluate students in what concerns the scientific knowledge's acquisition. Tests are equally conceived to evaluate phenomena's, concepts' and laws' understanding. We can see through the analysis of graph 2 that 40,00% of the questions correspond to the "understanding" domain. The evaluation of knowledge application in new situations, of analysis, synthesis and evaluation is little significant, corresponding, respectively, to 1,37%, 0,55%, 5,48% and 0,27% of the questions present in the body of tests, as it is evident in graph 2.

Graph 2 - Tests' sample percentage of questions (items) by cognitive domain's objectives, according to Bloom's taxonomy



5. Conclusions/implications

Based on this study no generalizations concerning the evaluative practices at the national level can be done (especially regarding the tests' quality), nor can be done any inferences concerning other themes' and subjects' tests. In spite of the generalization difficulties, this study's results seem to show that evaluation tests present a clear prevalence of questions aiming to evaluate substantive knowledge, there rarely being questions to evaluate the skills left. No questions related to the processual and epistemological and reasoning knowledge are made. We can acknowledge that Science questioning and evaluating attitudes are performances that teachers who have carried out the sample's tests have not yet introduced in their evaluative practices. Despite the regulations' in force defense of a teaching centered in the development of skills (knowledge mobilization to real life concrete situations), truth is that evaluative practices' analysis allows us to infer that teachers do not meet the development and evaluation of some important skills, as processual knowledge and reasoning.

Questions of the knowledge and understanding domains prevail, there being quite inferior the prevalence of questions related to the attainment of objectives of higher cognitive domains. From the analysis of the sample of tests used in this study, we can equally admit that Natural Sciences teachers continue to "privilege more the reproductive aspects and less the critical spirit" (Pacheco, 1994, p. 68), leaving to oblivion exercises which require knowledge

application in different situations, analysis, synthesis and evaluation/reflection, that is, which involve superior level of abstraction mental operations. Most of the questions placed in the tests show a concern with the attainment of low cognitive level (knowledge) objectives, leaving to second plan the themes' understanding and leaving farther the concern with objectives placed at superior levels, such as analysis, synthesis and evaluation.

These data reveal that tests and evaluative practices are very often reduced to mere exercises of remembrance or memorization (even because the "knowledge" category is also named "memorization") and of understanding of concepts ("understanding"), without the exigency of applying them to different contexts ("application"), of analyzing their organization, their elements and relations ("analysis"), of developing hypotheses, reflecting, planning ("synthesis") and of criticizing and evaluating knowledge ("evaluation").

On the other hand, tests evaluate topics and objectives which, in most of the cases, seem not to have any interest for students' lives and which are fundamentally related with contents' memorization. It is foreseeable that after the test's elaboration those should be forgotten because they had no practical interest for the student. This factor may constitute one of the school failures' causes. Learning contents without any meaning, sometimes useless for life, may lead to disinterest and, logically, to the school failure referred to.

This work seems to provide data which support the common-sense thesis that we are forming students, some of which with exemplary results, but who are incapable to respond to the challenges presented to them by society.

Furthermore, this study's results reveal the importance of teachers' initial and continuous training in what refers to evaluation. We defend the skills' thematic integration and their evaluation at any training level: initial, continuous and post-graduate. This training should be assumed as a need for, as Jordão (*in* Leite, 1995, p. 62) states "without an aware, informed and motivated faculty there is no possible change!".

We underline that evaluation should be integrated in the study plan of teachers' training courses, for the results allow the inference that this constitutes a deep lacuna in Licentiate' degree and post-graduation courses.

With this work we intend to encourage teachers to critically analyzing their evaluation means and to reformulate or conceive another type of tests, which evaluate knowledge acquired but also the problem-solving capacity, the creativity, etc…

However we would like to make clear that students' evaluation should not be materialized in tests only, because these means, on their own, do not evaluate the student in all of his/her aspects. We emphasize the need for the use of other evaluation means capable of evaluating aspects which otherwise become difficult with the exclusive use of tests.

It is up to the teacher to define strategies, to diversify and adjust different means to the characteristics of the students with whom he/she (re)builds knowledge and with whom he/she develops attitudes and values. To take into consideration and to value the student's participation, his/her interest, enterprise spirit and autonomy, his/her relation with others and with his/her material, the classroom work (elaborating reports, compositions, worksheets, self-evaluation…) and the homework, assiduity and punctuality, his/her critical sense and attitudes, behaviours and even reflections: "a reflection performed by the student in the end of a work day and its noting down in the notebook can be rich in content." are only some examples of the universe of aspects we ought to measure and that we hardly will do by recurring to a test.

Thus we suggest that teacher should develop a method which involves an active participation in tasks, that is, that works for projects and problems and which promotes the realization of portfolios which, according to Freire (2006), consists in a means which develops innumerous skills, from the organization of ideas, to systematization, problematization, argumentation and self-reflection and evaluation. Besides the portfolios we advise teachers to make use of observation grills and of registers of critical incidents to evaluate the students' diligence, interest and performance.

Bibliography

Bloom, B. (Ed.) (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: David McKay Company, Inc.

Direcção Geral Ensino Básico (2001). Currículo Nacional das Ciências Físicas e Naturais de 2001. Competências Essenciais do Ensino Básico). (consulted in 20 March 2005). Available in: http://nautilus.fis.uc.pt/spf/DTE/pdfs/competencias_essenciais_em_ciencias_fisicas_e_naturais.pdf

ESTRELA, A. e NÓVOA, A. (Orgs.) (1999). *Avaliações em Educação: Novas perspectivas*. Porto: Porto Editora.

FREIRE, L. (2006). "Portfólios. Instrumentos que melhoram as aprendizagens e o ensino". In Jornal *A Página da Educação*, November 2006, pp.43.

LBSE - Lei n.º 46/86, de 14 de Outubro (LBSE) (Consultada em 25 de Maio de 2005). Available in: http://www.sg.min-edu.pt/leis/lei_46_86.pdf

LEITE, C. *et al* (org.). (1995). *Avaliar a avaliação*. Porto: Edições Asa.

MACHADO, N. J. (2002). "Sobre a ideia de competência". In P. Perrenoud *et al*, *As competências para ensinar no século XXI*. Porto Alegre: Artmed.

PACHECO, J. (1994). *A avaliação dos alunos na perspectiva da reforma. Propostas de trabalho.* Porto: Porto Editora.

VALADARES, J. e GRAÇA, M. (1998). *Avaliando… para melhorar a aprendizagem*. Lisboa. Colecção: Plátano Universitária.