Assessment of learning outcomes in computing studies

Carmen Vizcarro Guarch, Pilar Martín Espinosa
Facultad de Psicología
Universidad Autónoma de Madrid
Madrid, Spain

Jorge E. Pérez, Edmundo Tovar, Gregoria Blanco,
Agueda Arquero, Javier García
Technical University of Madrid
Madrid, Spain

Abstract—**The assessment of learning outcomes is a key concept in the European Credit Transfer and Accumulation System (ECTS) since credits are awarded when the assessment shows the competences which were aimed at have been developed at an appropriate level. This paper describes a study which was first part of the project of the Bologna Experts Team-Spain and then developed as an independent study. It was carried out with the overall goal to gain experience in the assessment of learning outcomes. More specifically it aimed at 1) designing procedures for the assessment of learning outcomes related to these compulsory generic competences; 2) testing some basic psychometric features that an assessment device with some consequences for the subjects being evaluated needs to prove; 3) testing different procedures of standard setting, and 4) using assessment results as orienting feedback to students and their tutors. The process of development of tests to carry out the assessment of learning outcomes related to these competences, as well as some basic features regarding their reliability and validity is described and first results on the comparison of results achieved at two academic levels, will also be described at a later stage.**

Learning outcomes assessment, higher education, competence based education, assessment of learnin

I. Introduction

The assessment of learning outcomes is a key concept in the European Credit Transfer and Accumulation System (ECTS) since credits are awarded when the assessment shows the competences which were aimed at have been developed. ECTS is the credit allocation system for higher education used in the European Higher Education Area (EHEA), which involves all the countries engaged in the Bologna Process, 47 at this point in time. Its main role is to make higher education systems transparent and comparable, thus helping to bring to reality some crucial EHEA concepts such as mobility, employability or educational quality among others. Most Bologna countries have adopted ECTS by law for their higher education systems [1]. In Spain, additionally, a decree passed in 2007 [2] establishes the generic competences which any student with a university degree must have developed; these include understanding basic and gradually more advanced scientific texts; problem solving; looking for, selecting and using information to solve problems or making decisions and, finally, the capacity to learn independently, all of them in the students' specific fields of study. But more generally, basing higher education on the development of key competences and measuring them is a strong international trend in higher education as, for instance, the OECD funded AHELO study shows [3].

This paper describes a study which first was part of the project of the Bologna Experts Team-Spain (http://www.expertosbet.es/) and then evolved as an independent study. It was carried out with the main goal of gaining experience in the assessment of learning outcomes and, more specifically to: 1) designing some procedures for the assessment of learning outcomes; 2) testing some basic psychometric features that an assessment device with some consequences for the subjects being evaluated needs to prove; 3) an issue of special interest for us was testing different procedures of standard setting; finally, we were interested in using assessment results to give feedback on competence development to students, their tutors and the institution which holds them.

In order to achieve our goals, participants from degrees representative of different fields of knowledge (Biology, Psychology, Computing, Economy and History) were invited to take part in the study. In this paper we shall focus on the work done in the degrees of Computing and Mathematics & Computer Science. The sample of students includes 1st and 3rd year students from 3 different schools with a degree in computer science and we shall report specifically on one of the competences, namely understanding scientific texts at the level of university studies. We shall describe the process of development of the test and discuss some measures taken to guarantee an acceptable level of objectivity and validity of the data. At a later stage, we shall report some results of data analyses carried regarding results achieved at the two participating academic levels. Finally, we shall discuss our experience in the use of these procedures as well as the implications for the development of this kind of tests and its crucial role in higher education reform.

II. Context and Background

In the context of higher education, a competence may be understood as the combination of skills, knowledge, attitudes, values and abilities that underpin effective and/or superior performance in a professional area [1]. In this way, when we try to assess student performance, we are interested in assessing not only knowledge, as has been the case in traditional education, but also what the student is able to do (and how) using this knowledge. By how, we understand adhering to disciplinary methodological standards and values. Thus, competence or learning outcomes assessment includes the assessment of knowledge, but is not limited to it. It is normally assessed through complex, representative disciplinary tasks that imply knowledge and are often complemented with students' reflections whereby students justify the decisions they have taken on a theoretical and/or disciplinary base, and take into account their consequences or the values that inform them.

The starting point for this study were the basic transferable competences which, according to the Spanish Decree 1393/2007 [2] every higher education graduate should have developed by the end of their studies. They were selected since they are common to all degrees although every discipline is expected to further introduce its own particular coloring and nuances. For this reason, they were considered to be at the same time a good basis for independent work and also for making interesting comparisons related, for instance, to fairness. On the one hand, we could learn about the particularities of the assessment of learning outcomes regarding different disciplines; on the other, if the structure used for the tasks was similar, we could explore to which extent assessment criteria and standards were used in similar ways.

III. RESEARCH QUESTIONS

As mentioned above, our aim was to design assessment procedures to assess the basic competences which all graduates must have mastered by the end of their undergraduate academic life according to the Spanish law. This should be complemented by the development of assessment criteria that would allow enough objectivity when correcting and eventually grading students' work. We also tried to validate the tasks as appropriated for the assessment of these basic competences in various ways.

Some additional questions arose from a pilot study performed previously to the work reported in this paper. This pilot showed to us many valuable things such as the importance of correctly wording the questions, since light nuances in language can make dramatic changes in how students understand them; how test administrations procedures need to be very clear and strictly followed if we want to work together and compare or sum up results from different schools or teachers; how rating criteria for open questions need to be very carefully developed if a basic level of objectivity is to be assured. As a starting point, we deeply believed in constructed responses for the assessment of competences, since they usually represent more complex tasks. However, we were also aware that open questions are more difficult and costly to rate, so we opted for a mix of the two so their results could be summed up and eventually compared. As a means to assure some basic common conditions, we also opted for computerized tests.

IV. METHOD

A. Objetive

This paper presents the process of development of a computerized procedure to assess a transferable competence basic for learning and academic life: understanding scientific texts. It further describes how basic objectivity and validity data were assured and finally adds some results on how the two academic levels participating in the study compare. Other comparisons of interest are paper vs. computerized versions of the test, as well as closed vs. open questions.

B. Development and nature of the task

In order to develop the appropriate tasks to measure in a comprehensive way the learning outcomes associated to these transferable competences, they were in the first place analyzed it their facets or components. The various questions included in the tests were then mapped on this scheme, as can be partly seen in Table I for the competence understanding scientific texts.

In the second place, we looked for texts with specific computing content which, however, did not require highly specialized domain knowledge to be understood, since they were to be applied to first year students. Nevertheless, since they also had to be given to 3rd year students, these texts needed to be amenable to be understood and interpreted at a higher or advanced level. A text from a scientific university dissemination platform was selected dealing with the topic of wireless sensor network.

Based on this text questions were prepared regarding either information directly contained in the text or which could be derived from it if students had the necessary knowledge. In this way, the task contained two types of questions of different difficulty level. Furthermore, these questions adopted two different formats: on the one hand, 7 open questions which implied a constructive response and 11 closed questions which required to chose from 4 alternative answers. Table II contains examples of these two types of questions.

TABLE I. EXAMPLE OF TASK ANALYSIS OF THE TEST FOR TEXT COMPREHENSION USING COMPETENCE FACET ANALYSIS FOR SOME TEST ITEMS

Questions (Tasks)	Facets				
	Main idea	*Secondary ideas*	*Relationship among ideas*	*Personal/ professional Interest*	*Authors intention/ implied ideas*
3				✓	
4	✓	✓	✓		
5					✓
6	✓				
7		✓	✓		
8		✓			
9	✓				
10		✓			
11		✓			
12					✓
13	✓				

C. Pilot testing of tasks

This text was first given to a student sample participating in a pilot study. Their performance was automatically rated in the case of the closed questions (CQ) while open questions (OQ) were rated by human judges according to agreed upon assessment criteria whose development will be described below. Additionally, attention was given to questions difficult to understand (identified because they raised many questions or comments by students) or which resulted in responses very different from the intended ones; based on these observations, the text of the questions was corrected and then given a computerized format (Fig. 1).

TABLE II. SOME SAMPLES OF OPEN AND CLOSED ITEMS

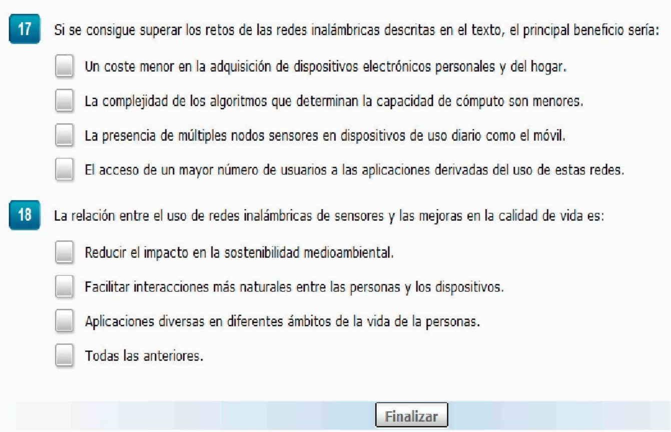
Open question:	Question 5: Based on this text describe some concrete applications of wireless sensor networks
Closed question:	Question 12. When we say that wireless sensor networks must be able to self organize we mean that: (options a, b, c & d follow)

As mentioned, this computerized test corrected closed questions automatically, but it also contained some open questions which had to be rated by human judges. The development of the criteria by which the performance to open questions was rated is described next.

D. *Developing assesment criteria*

As a first step to develop clear assessment criteria for open questions, 2 teachers prepared the best possible response to each question, discussed them and agreed on the criteria which make a highest scoring response. Then second best, third and unacceptable responses to the same questions were described. Finally, these criteria were validated against 10 exercises from the pilot sample which were corrected separately by the two judges. All disagreements were taken as a basis to either refine the criteria or, also quite often, review the text of the questions itself. Next, they validated the new criteria against a new sample of 20 exercises until perfect agreement was reached. These were subsequently considered expert ratings.

In this process, examples of each of the 4 alternative ratings for each open question were also selected and included along the assessment criteria.



17 Si se consigue superar los retos de las redes inalámbricas descritas en el texto, el principal beneficio sería:

□ Un coste menor en la adquisición de dispositivos electrónicos personales y del hogar.

□ La complejidad de los algoritmos que determinan la capacidad de cómputo son menores.

□ La presencia de múltiples nodos sensores en dispositivos de uso diario como el móvil.

□ El acceso de un mayor número de usuarios a las aplicaciones derivadas del uso de estas redes.

18 La relación entre el uso de redes inalámbricas de sensores y las mejoras en la calidad de vida es:

□ Reducir el impacto en la sostenibilidad medioambiental.

□ Facilitar interacciones más naturales entre las personas y los dispositivos.

□ Aplicaciones diversas en diferentes ámbitos de la vida de la personas.

□ Todas las anteriores.

Fig. 1. Computerized format of test

E. *Experts' judgments on content validity of the assessment task*

In order to estimate the content validity of the task, it was given to experts who had to answer two sets of questions: 1) which one of the basic competences the task measured? And 2) which facets covered by the task analysis did each questions measure? This last question was meant to address whether all relevant facets of the competence were covered by the task in a comprehensive way.

Two judges participated in this stage. Both of them were specialists in high education and reading and comprehension and had experience in competence based education.

The results of this phase were as the follows. The judges reached a 100% agreement regarding the competences we were trying to measure by means of the test. However, one of them also mentioned other competences which we thought were marginal for the task at hand. The same was true for the facets we identified in our tasks: 100% of the facets we mentioned were also found by the judges who, whoever, introduced some additional ones. We can thus conclude these results essentially validate our analysis regarding the content validity of the task.

F. *Administering the task to students*

The student sample who took the test is described in Table III.

Although the size of the sample is not very large, its varied nature should be emphasized, since it makes it more representative. However, unfortunately the size of both academic levels it is not well balanced at this point in time. As already mentioned, a different group of students participated in a pilot test of the procedures but is not described here.

TABLE III. DESCRIPTION OF THE STUDENT SAMPLE

UNIVERSITY	Gender (N, %)		Age	Year (%)	
Sample %	Male	Female	(mean)	1	3
UAM:15.63	n=5 (50)	n=5 (50)	23.80	30	70
UCLM: 28.13	n=17 (94.4)	n=1 (5.5)	20.83	61.1	38.9
UPM: 56.25	n=29 (80.6)	n=3 (8.33)	18.31	88.9	11.1
TOTAL: 60	N=51 (79)	N=9 (21)	20.98	71.8	28.1

G. *Training human judges to rate open questions*

Once the expert assessment criteria were agreed upon as described above, we needed to train the judges who would rate the students' work. Of course, this was also an opportunity to observe how objectively these criteria could be learned and used by other raters. The raters were 2 master students who, in the first place, and before any rating, answered the test in order to understand its demands. Then they started rating the same exercises used to develop the expert ratings. They first received 10 of them together with the assessment criteria they were asked to use and their ratings were compared to the expert ratings, discussing any difference that was encountered. In a second stage, they received a new set of 22 exercises and the agreement of their ratings with that of the experts was calculated using the Cohen's kappa coefficient in order to remove random agreement. The process continued until the agreement was satisfactory. This was usually reached in all tests with 20 to 30 exercises.

Following this procedure, inter-rater reliability for 2 judges and 22 exercises, using Cohen's kappa [4] was found to be 0.497. The SPSS statistics 19 software was used for the calculation of this coefficient. According to Landis and Koch [5], we can conclude this is a moderate level of agreement.

Once objectivity of ratings was achieved in this way, the judges were given the exercises of the whole student sample. Throughout the whole grading process, they were aware of the fact that some exercises, unknown to them, were randomly distributed to all judges and their reliability was being continuously monitored.

V. DATA ANALYSIS

Item 2 of the test refers to students' interest in the proposed text and as such, is not part of the test, so it is treated apart. In a Likert scale with values 0-3 the mean score for the whole sample is 1,98 (SD .77) showing a medium interest, which probably means students' involvement with the task is also medium, especially if we take into account that only 3.125% choose score 0.

A. Open and closed questions and total score

As mentioned, the test consisted of a section with constructed or open questions (OQ) and another one with closed questions (CQ). The total score of the test is the sum of the two parts. The maximum score for the whole test is 31. We found a mean of 14.16 (SD 3.79) and, as expected, this is a symmetrical distribution.

Constructed response items, or OQ, seem to reflect better the nature of the competences but they are also more costly to score in a reliable way. So, it was interesting for us to compare the OQ with the CQ and find out what each of them adds to the whole test. The OQ section includes 7 open questions which may be valued 0-3 by human judges. So, the maximum score is 21 and the minimum 0. Regarding the CQ section, it contains 10 questions with 4 alternatives with a value of 1 each; the maximum score is thus 10 and the minimum 0. Table IV shows the results we found for our sample.

TABLE IV. TABLE IV. SCORES FOR OPEN AND CLOSED SECTIONS AND TOTAL TEST (N= 64)

	OPEN QUESTIONS	CLOSED QUESTIONS	TOTAL
Mean	7,45	6,70	14,16
SD	2,81	1,64	3,79

As can be seen the mean leaves ample space for improvement in all cases but especially in the case of the OQ and Total scores. Moreover, There is an indication that OQ would seem to be more difficult, since their mean does not reach the mid-point of the score range (10,5), while the CQ questions seem easier since they overcome it (5). However, this question will have to be postponed to more specific item difficulty analyses.

B. Internal consistency

When we look at the internal consistency of the total test, that is, the way in which the items seem to measure the same construct, the value of Cronbach's alpha of .442 for the whole test (.232 for the OQ and .378 for the CQ). The meaning of this index depends on the kind of test and in our case it seems to be a medium value which seems to confirm the reliability of the test. However taking into account our test has only 17 items in total, it could be improved adding more items.

C. Comparison of results in two academic levels

When we compare the results of the 1st year students with those of the 3rd. year, our results seem to be in line with our expectations, since they seem to reflect a development from the 1st year to the 3rd year when we look at the mean total score and holds also true also when we compare the different sections of the test (see Table v).

TABLE V. TABLE V. COMPARISON OF RESULTS FOR 1ST AND 3RD YEAR STUDENTS

	1st year students	3rd year students
Total Test Score	Mean:13.67 SD: 3.67	Mean: 15.39 SD: 3.89
Open questions	Mean: 8.50 SD: 2.61	Mean: 7.04 SD: 3.11
Closed questions	Mean: 6.89 SD: 1.62	Mean: 6.63 SD: 1.71

However, when we perfomed a t test for independent samples, we could not find evidence that these differences were significative. The Levene test allowed us to assume equality of variances (Total Score, F<.0005, p= .997) and the t test yielded a value of .68, p=-.385. These results are interesting. At this stage, they may be taken to mean that the test does not have enough discriminative power. But they may also suggest that competences do not develop unless they are taken seriously and fully integrated in educational practice, including assessment. However, we hope further analyses taking into account a general ability indicator, which we are currently collecting, will help us attain a more specific conclusion.

D. Mode of presentation (computerized vs. paper and pencil)

Another interesting question is how the mode of presentation of the test (computer vs. paper) impinges on the total scores. Our results are showed in Table VI.

TABLE VI. TABLE.VI. TOTAL SCORES BY MODE OF PRESENTATION (COMPUTERIZED VS. PAPER AND PENCIL)

	Computerized test	Paper & pencil
Total Test Score	N= 32 Mean: 14,09 SD: 4,31 Score range: 4-22	N= 32 Mean: 14,22 SD: 3,26 Score range: 8-20
Open questions	N=32 Mean: 7.47 SD: 3.05	N=32 Mean: 7.44 SD: 2.59
Closed questions	N=32 Mean: 6.62 SD: 1.87	N=32 Mean: 6.78 SD: 1.38

These results, seem to suggest that both presentations yield very similar results and, indeed, when a t test was calculated for the total scores in each presentation it was not significative (T=.48, p= .6). This is interesting in that different modes of presentation may be of interest under different conditions.

VI. DISCUSSION

Competences and learning outcomes really seem to play a pivotal role in higher education reform. However, although many academics agree on their value, many also raise their worries regarding how they can and should be assessed. Now, assessment plays a nuclear role in educational practice. In this regard, the quote by Resnick and Resnick [6] "you get what you assess" seems in order. No matter how much we strive to help students develop the competences they will need in their professional lives, it is difficult to achieve them, al least in a general sense if we do not take the pain to assess them. Assessment determines the real goals that must be achieved by students to be successful and at the same time are a rich opportunity for learning if criteria are clearly understood and shared by students and can be worked upon. In other words, educational reform can be a void effort if it is not reflected in the way assessment is performed.

The main achievement of this project seems to be that indeed we have succeeded, at least in a first phase, in the development of a procedure to measure learning outcomes which in light of the present results can be considered reasonable and can be taken as a base for future developments. However, it must be acknowledged this process takes much time and effort, as shown in this paper, and is probably best approached as a multidisciplinary endeavor.

Of course also many difficulties arose along the way. Maybe the first worthwhile mentioning are the difficulties found in the administration of these tests to natural groups of students. Teachers as well as authorities did not seem to be clear about the benefits of this administration and at times simply considered it a loss of time. The practical result of this is that, despite our efforts to the contrary, they were taken mostly by students who volunteered and the sample size was below what we expected. In this sense, it may be considered that the results represent an overestimation. This is interesting considering the modest mean we found in a very basic competence of reading comprehension.

Even acknowledging the value of competence based education, many tutors just seem to prefer to avoid the specific assessment of learning outcomes and, in fact, are also not prone to leave time for this assessment. Whether this means avoiding the assessment of learning outcomes altogether or just carrying out this assessment in less controlled ways that would be desirable would be interesting questions to discuss. However, it should not be born in mind that students deserve to be assessed by means which are reliable, valid and fair. These are the basic features that any measurement with an impact on the life of the assessed person needs to prove.

No doubt the process of developing learning outcome assessment devices, as described in this paper is long and costly. However, it seems efforts of this sort need to be done in order to guarantee that students are assessed by means of procedures which have proved their objectivity and measure what they are supposed to measure. This is especially so when important decisions are taken based on this information, as is the case when they are used for certification purposes or when used for accountability

Several lines of reasoning seem to be relevant in this respect. Of course, once procedures of this type are developed and adopted by an institution they do not need to be so costly for subsequent use. Maybe the crucial question is whether they need to be developed at each institution, taking into account its costs and multidisciplinary nature, or they should be developed elsewhere, maybe for more general use at least in some specific cases (Hutchings, 2009) (7). We are aware this second option raises the question of the limitations of standardized tests vs. more open, qualitative and situated alternatives (Banta, 2007; Banta, Griffin, Flateby and Kahn. 2009; Shavelson, 2011, Shavelson, Klein y Benjamin, 2001) [8, 9, 10, 11]. It also raises the question of the fact that, if the procedures have not been developed at an institution, its members do not feel ownership over them. Indeed, some of the resistance we found in tutors seemed to be related to the use of results for external evaluation and control.

On the one hand, the experience of development and use of the procedures described was most enriching for all participants and it could be said, it was a great opportunity for teachers' professional development and it prompted them to use similar procedures for developing competences. For students alike, it was an opportunity to understand in a practical way what competences are about.

To summarize our experience to date, this work has been long and costly, but also very rewarding for those who directly participated. In fact, they readily used the tasks and others developed following this example in their daily activity. In this sense, the tasks seem to be very intuitive and stimulate educational activities geared to develop valuable competences. Finding a balance between the effort needed to develop this kind of assessment devices and the possibility of not measuring them or doing so in less reliable and valid ways is something the academic community will need to consider seriously.

ACKNOWLEDGMENT

This study was possible thanks to funds from the Bologna Experts Team (Spain), the Spanish Ministry of Education's tender for Actions Oriented to the Improvement of Student Support and Employability, reference CAIE012 (Resolution December 12, 2011) and scholarships provided to students who acted as raters by the Autonomous University of Madrid.

REFERENCES

[1] European Communities, "ECTS Users'Guide," Accessed at: http://ec.europa.eu/education/lifelong-learning-policy/ects_en.htm on 20 Oct 2012.

[2] Ministerio de Educación y Ciencia, "REAL DECRETO 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales", BOE núm. 260, de 30 octubre 2007, pp. 44037-44048.

[3] OECD, "Higher education and adult learning, Testing student and university performance globally: OECD's AHELO". Retrieved from http://www.oecd.org/education/highereducationandadultlearning/testingstudentanduniversityperformancegloballyoecdsahelo.htm.

[4] Cohen, J. (1960) A coefficient of agreement for nominal tables. Educational & Psychological Measurement, 20. 37-46.

[5] Landis J, Koch G. The measurement of observeragreement for categorical data. Biometrics 1977; 33:159-74

[6] Resnick, L. B., & Resnick, D.P. (1989). Assessing the thinking curriculum: New tools for educational reform. Washington, DC: National Commission on Testing and Public Policy.

[7] (7) Hutchings, P. (2009) The new guys in assessment town. *Change, 41*, 26-33.

[8] (8) Banta, T. (2007) A warning on measuring learning outcomes. *Inside Higher Ed.* Enero *http://www.insidehighered.com/views/2007/01/26/banta* Acceso 25 de agosto de 2001

[9] Banta, T.W., Griffin, M., Flateby, T.L. & Kahn, S. (2009) Three proimising alternatives for assessing college students' knowledge and skills. National Institute for Learning Outcomes Assessment occasional paper nº 2, December.

[10] Shavelson, R.J. (2011) On an approach to testing and modelling competence. Comunicación presentada en la recepción del E.L.Thorndike Award, en el congreso anual de la American Psychological Association, Washington, DC, 5 de Agosto.

[11] Shavelson, R.J., Klein, S. y Benjamin, R. (2001) The limitations of portfolios. Inside Higher Education, Octubre