ABSTRACT
This paper presents a brief review about the state of the art in e-learning assessment. This study concludes that there is a lack of tools for testing learners that combine all the aspects commented before, namely pedagogy, adaptation and standard. With the aim of covering this hollow, the paper also presents an authoring tool for developing items and tests taking into account all these factors.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education – computer science education and information systems education.

General Terms
Measurement, Design, Standardization.

Keywords
e-Learning, assessment, IRT, standards, pedagogy.

1. INTRODUCTION
The role of assessment in e-learning is a very important aspect. Assessment is invaluable as a way of affording students the opportunity to learn. In this respect it is certainly more significant for online courses than it is for traditional learning [14].

E-learning assessment is a multidisciplinary subject where Information and Communication Technologies (ICTs) and Pedagogy converge. Developing successful evaluations implies a deep study about each subject and the manner in which are going to be mixed [4].

From the point of view of Pedagogy, there are two main aspects to take into account: (i) pedagogical theories that explain how knowledge is acquired and how to develop questions to decide if learners have acquired that knowledge, and (ii) psychometric theories that explain how to develop adaptive assessments, one of the most important advances in assessment in the last decades that implies a lot of advantages [7, 19]. In the ICTs field, and concretely talking about e-learning, standardization is an important area that must not be forbidden.

2. COMPUTER TESTING REVIEW
This section presents a brief review about the evolution of computerized tests in e-learning systems. Concretely, the research is done from four different points of view about tests: (i) adaptation to the learner, (ii) item types used in a test, (iii) pedagogical/learning domain where the test is applied and (iv) use of standards.

The last subsection lists some systems and applications that work on e-learning assessment, enumerating the characteristics of each one regarding with the points of view of the previous subsections.

2.1 Learner Adaptation
From the point of view of adaptation, it is possible to distinguish two types of test. In one hand there are the classic tests (non-adaptive) [18] that are the oldest and the more basic ones. They consist of various items that are administered to users independently from their knowledge level or preferences. On the other hand, there are the adaptive tests, based on the Item Response Theory (IRT) [26]. Adaptive tests are able to adapt the evaluation to the learners providing tests suitable for their knowledge level. Even, it is possible to make an accurate assessment with fewer items (so, in less time) than with the classic tests.

Classic tests are divided into predefined and dynamic tests. The first ones are developed prior to starting the learning process, and are the same for every student. When it is necessary to evaluate a learner about a concept of the learning domain, the system
launches a test (previously stored in a repository) that evaluates it. The same test is used for evaluating all the students in reference to that concept, independently from their preferences or knowledge level. Opposite, dynamic tests are not developed before student starts the learning process, but when is time to take an evaluation, the system randomly selects some items (from an item bank) related to the concept it wants to evaluate and composes the test. The probability of generating distinct tests for different learners is directly proportional to the size of the item bank.

Adaptive tests are not stored in a repository. The test is given item by item, and the answer to the previous item determines the selection of the next one. The next item is chosen applying the IRT equations that supply the adaptation to the knowledge of the learner. Although the IRT was described more than 50 years ago, the mixture between computers and IRT was a decisive milestone [7]. Computers with their computational power have been able to exploit the advantages of adaptive tests. This research area is known as Computer Adaptive Testing (CAT) [8]. Although the advantages of adaptive assessment are tangible [3], the requisites to developing it are demanding. The selection of one or another type of test depends on the learning process and the learning environment in which the assessment is going to be applied. Pedagogies and domain designers must choose which type of test is better for the learning process and the acquisition of knowledge. It is possible the coexistence of the three types of tests in the same learning environment, selecting one or another in different moments of the learning process, depending on the needs.

### 2.2 Item Types

Although the well-known multiple choice item is the most used for evaluation, there are many more types that could be applied in e-learning assessment. There is a good compilation in [18].

Most systems provide multiple choice items as unique manner of assessment. This item type is complete for evaluation because assess learning at the highest level of Bloom’s Taxonomy [2], if it is well constructed (which is not easy). But there are many other item types that must not be omitted. Some of them are essays, projects, case studies and simulations [4], matching, ordering, true/false, localization and short answer [15].

### 2.3 Pedagogical Domain

Other aspect to consider in assessment is the pedagogical domain where it is applied. Available e-learning systems usually represent the domain like nodes (concepts) and links (relations between concepts) [21]. Depending on the structure of the pedagogical domain it is possible to distinguish two types. Simple pedagogical domains are those that have nodes at one level (simple nodes) with relations between these nodes. Opposite, there are complex pedagogical domains where nodes are grouped according to some criterions, obtaining composed nodes. Here, the relations could be between simple nodes, between composed nodes, from simple nodes to composed nodes or vice versa.

The more complex the pedagogical domain is, the bigger the possibilities of assessment are. In simple domains items/tests could be about nodes or about relation between nodes. But in complex pedagogical domains could be items/tests about simple nodes, about composed ones and about all the possible relations describe in the previous paragraph. This aspect enriches enormously the area of assessment.

### 2.4 Standards

The usefulness of standards has been demonstrated in many areas of our live. But thinking about e-learning, the utility of standardization is stronger because of the universal and ubiquitous character it has. A lot of papers, web pages and associations talk about this phenomenon. In fact, in the last 10 years several associations (AICC, ADL, IMS, etc.) are working at standardization about everything related to e-learning (Learning Management Systems (LMSs), Learning Objects (LOs), Packaging, Tests, and more), and they are making a lot of improvements [24].

Particularly, in the area of assessment the standard QTI (Question & Test Interoperability) [10] developed by IMS [9] is widely accepted inside the developers’ community. It is possible to say that IMS QTI is ‘de facto’ standard for assessment due to the fact that a lot of systems are adopting it [12], and also it is the only international standard specification for e-learning assessment.

The last version available of this standard (IMS QTI 2.0) enables to implement a wide range of item types: multiple choice, ordering, association (1:1), union (1:N), fill in blanks, essays, hotspots, select point, ordering and associations with hotspots, object positioning, slice, painting and more. In addition, QTI use the standard language XML for coding the items and tests. This fact allows the visualization of items/tests in different devices like desktop/laptop PCs, PDAs or mobile telephones. That could be very interesting for expanding the functionality of an e-learning system.

### 2.5 Existing Systems/Applications

Nowadays it is possible to find commercial systems for developing tests as well as system developed by researchers from different universities. This section enumerates the most important ones explaining their strength and their weakness.

Regarding to commercial systems the market offers a lot of them for developing tests as Fast TestPro, MicroCAT, DEMOCAT, METRIX Engine, ADTEST, Perception and so on. These systems allow developing various types of items and also some of them enable adaptation to the learner because of the use of IRT. But their main weakness is that are proprietary systems and do not use any standard for coding items or tests. This is a big inconvenience for integrating them into an e-learning environment. According to the pedagogical domain, all of them only allow to define items for one concept. There are no possibilities to define items for a relation between concepts, or complex pedagogical domains.

In the academic area, different universities have developed some interesting systems and tools. <e-aula> [23] is an e-learning system with a module for assessment. All the items stored in the item bank are compliant with IMS QTI (QTI Lite version). The item types possible are true/false and multiple choice (with one or more correct answers). Test Editor [22] is an authoring tool for generating items and tests that store all the items in XML. It allows storing multiple choice items (four possibilities, one correct) and can work in an adaptive way (IRT 2 parameters). The KOD Project [11] presents an architecture for defining re-usable adaptive educational content. Inside this architecture, there is a Questions & Tests Toolkit that enables the editor to define items and tests related to each concept of the ontology. Each item is stored in XML in a database.

All these systems/applications have important highlights, most of
them referring to standards and adaptation to the knowledge of learners. <e-aula> uses IMS QTI, but work in a non adaptive way. Test Editor and KOD Project allow adaptation to the student and store items and tests in XML files. But these files do not follow the rules of IMS QTI, the ‘de facto’ standard for assessment.

Regarding to pedagogy, all the systems/applications have a lack of a wide range of item types. They are capable of generating one or two types. But, actually, there are much more possibilities very important for a complete learning [4]. Furthermore, no one of the systems/applications mention any pedagogic rule or technique to develop items and tests. Only the KOD Project mentions content expert help when developing an item or test, but it is not explained how. In addition, all the systems/applications are applied into simple pedagogical domains.

A mixture between commercial and research tools are Moodle [17] and WebCT [25], both widely implanted in the academic sector. One of the main advantages of the former is that is open source software, while the latter is proprietary. Both have an assessment module that let teachers define different item types like multiple choice (both), short answer (both), true/false (Moodle), matching (both), embedded-answer questions (Moodle) or essays (WebCT). No one of these systems allow adaptation to the knowledge of the learner. Both, Moodle and WebCT, store all the items/tests in a repository without relation between them and the concepts of the pedagogical domain.

3. A NEW AUTHORING TOOL FOR E-LEARNING ASSESSMENT.
REQUERIMENTS AND DEVELOPMENT STATUS

The last section shows a lack about authoring tools for generating items and tests that mix all the factors that influence in the evaluation process. This section presents the authoring tool we are developing to cover this hollow in the e-learning assessment area.

Our aim is to implement an authoring tool for generating classic and adaptive assessments, using standards and paying special attention to Pedagogy, a science that takes a very important role in e-learning [4], and it is usually forgotten.

The next subsections explain the requirements the authoring tool should fulfill to achieve the aim, in relation with the different points of view broke down in section 2. The last paragraph of each subsection comments the progresses made on each area until today.

3.1 Adaptive Assessment

The notion of adaptation is very important in traditional learning, and much more in e-learning, where there is no constant human presence for guiding the learning process. The research group where we work [5] has large experience in adaptive hypermedia systems and IRT [6, 13, 20]. So, we are going to expand this knowledge in depth and apply it to the new item/test authoring tool we are developing.

The requirements for making a powerful and complete adaptive assessment tool are: (1) the coexistence of classic tests (predefined and dynamics) with adaptive ones and (2) the use of IRT with three parameters (difficulty, discrimination and pseudo-guessing), because the use of four parameters does not cause a big improvement in the adaptation level [1].

Achieving the second requirement is neither trivial nor immediate. At the beginning, the tool offers adaptation with 1 parameter (difficulty). Later, when the items that compose the item bank have been administered to a large amount of students, a software application estimates the three parameters of the IRT commented before, calibrates the item bank and a better adaptation with three parameters is offer from now on. This calibration process is completely transparent for teachers and there is no need to know the basis of IRT.

At the moment, the tool is prepared to create adaptive tests. For this purpose, when teachers are defining an item are requested about the difficulty of that item (in relation with the domain where the item is going to be applied). This simple data makes possible to administer CATs using IRT with 1 parameter (difficulty). Also, we are studying software applications as BILOG [16] and LOGIST [27] to calibrate the item bank.

3.2 Pedagogy

Pedagogy and e-learning are naturally linked, but so many times there are not join in the right manner or they are not linked enough. Our research works wants to take advantage of pedagogical theories and research studies to improve e-learning assessment.

The objective is to apply pedagogy for developing an item/test authoring tool that enables the possibility of creating a lot of types of items, almost the most used [9, 18]. Also, we want to include a help guide for teachers. The aim of this guide is help teachers to design good items to achieve the pedagogical objectives fixed. For example, if a teacher wants to develop a true/false item, the guide can help his/her with some suggestions, tricks and examples about how to do it. Or if he/she wants to develop an item for testing concrete abilities of the student about a concept, the guide can recommend teacher which item type select and how to develop it.

We also need pedagogy to take advantage of the structure of the pedagogical domain in which the assessment takes part. Most of the available e-learning systems represent the knowledge like simple pedagogical domains. Each item is related only with one or more concepts of the domain, but not with the relations between them. We would like to take a step ahead in this area relating items not only with concepts. Concretely, we are working on relating items with:

- Links. That is, if two nodes (concepts) are related, it is probably the teacher wants to design items to know if the student has learned that connection, not only the two concepts separately.
• Composed nodes. In the same way, if we have a complex pedagogical domain, it could be very interesting to have items for testing simple nodes that are into a composed node, but also for testing the composed node as a whole.

• Bloom’s Taxonomy [2]. Students can learn concepts developing different competences. Bloom defines six levels in the cognitive domain: knowledge, comprehension, application, analysis, synthesis and evaluation. It would be very interesting to define items that work different cognitive levels about the same concept. That is because some learners need a more deep knowledge about some concepts (from knowledge to evaluation), while for others is enough to go to the second or third level of Bloom’s Taxonomy.

This pedagogical knowledge is not yet inside the authoring tool. For the time being we are in contact with pedagogies to collect their knowledge and apply it to resolve all the pedagogical aspects commented before. Concretely, we need to know which item types work which abilities/skills, which multimedia format are better to achieve the pedagogical objectives fixed and the learner preference, and how to guide teachers for developing good assessment. We expect to have a preliminary report about these pedagogical aspects at the end of this year. Later, we will put the study into practice, providing the tool with this knowledge.

3.3 Standards Compliant
The advantages of using standards are well-known in any area. As commented before, IMS QTI is the ‘de facto’ standard for assessment in e-learning. It offers a lot of documentation and examples, everything very well explained and structured. In addition, the last version of QTI (2.0) allows the definition of multiple types of items.

Developing items and tests with this standard makes easy to export them to other systems and tools, as well as import new items and tests from other systems an authoring tools QTI compliant.

These are the main reasons for choosing IMS QTI as the standard for coding items and text in the authoring tool we propose. At the moment, all the items develop with the authoring tool are IMS QTI compliant (1.2 and 2.0 versions). Even, the tool is prepared to house all the items that IMS QTI v2.0 defines.

4. CONCLUSIONS AND FUTURE WORK
As shown in section 2, there is a hollow in tools about e-learning assessment, in the sense that they do not take into account all the factors that affect this area. We have studied commercial tools as well as research tools developed in different universities, but no one of them mix pedagogical aspects with adaptive assessment and the use of e-learning standards.

Our proposal is a new e-learning assessment authoring tool that covers this hollow. We are conscious of the importance of pedagogy in e-learning, as well as adaptive assessment and the use of standards for interchange reasons. So, we propose an authoring tool that mixes all this factors with the aim of achieving high quality e-learning assessments.

The authoring tool presented in the previous section, lets teachers working with a generic tool (applicable to any domain) for testing learners. Teachers can easily develop adaptive tests without deep knowledge about pedagogical adaptive theories (IRT), just filling a form. It is possible to create and update items and tests for fitting them to the necessities of each moment.

Also, and due to the fact that the tool is IMS QTI compliant (version 1.2 and 2.0), teachers can import and export items and tests from/to other systems and tools that follow the same standard. In addition, teachers can easily modify these items/tests that have being imported from other systems, adjusting some parameters of them if it was necessary.

The authoring tool is prepared to house multiple item types for a better adaptation to the learner knowledge and preferences. At the moment it is possible to define multiple choice, multiple response, ordering, true/false and matching (1:1) items, but we are working on other types as union (1:N), fill in blanks, essays, hotspots, select point and object positioning.

Even, the authoring tool is going to provide a guide for helping teachers to decide which type of item is better for achieving some skills and how to develop good assessments.

We are aware of the complexity of the tool. But we expect to keep it going thanks to the large experience of the GHyM research group in all the fields commented in section 2.

5. REFERENCES

[10] IMS Question & Test Interoperability Specification, 

Architecture for Defining Re-usable Adaptive Educational 
Content. In Proceedings of the Second IEEE International 
Conference on Advanced Learning Technologies (ICALT’01) (Madison, USA, August 6-8, 2001). IEEE 

[12] Kraan, W. IMS Question and Test Interoperability gets major 
make-over. CETIS, 2004 
http://assessment.cetis.ac.uk/content2/20040712004043

J. A. and Gutiérrez, J. Generation of Computerized Adaptive 
Tests in an Adaptive Hypermedia System. In Proceedings of the International Conference on Information and 
Communication Technologies in Education (ICTE’02) (Badajoz, November 13-16, 2002). A. Méndez-Vilas, J. A. 
Mesa González & I. Solo de Zaldívar (eds): Education 
technology.

[14] Macdonald, J. Developing competent e-learners: the role of 
assessment. Assessment & Evaluation in Higher Education, 

[15] Mateo, J. A. La evaluación educativa, su práctica y otras 

International. Mooresville (USA), 1990.


Development, implementation and validation of a predictive 
and prescriptive test for statewide assessment. AERA 

Hypermedia System. In Proceedings of the 7th World 
Conference on Artificial Intelligence in Education 
(AIED’95) (Washington DC, USA), Published by AACE: 
Charlottesville, Virginia (USA), (1995), 351-358.

sistemas hipermedia adaptativos. Informática y automática. 
(Dec., 1997), 45-70.

[22] Romero, C., Martín-Palomos, S., De Bra, P. and Ventura, S. 
An Authoring Tool for Web-Based Adaptive And Classic 
Tests. In Proceedings of the World Conference on E- 
Learning in Corporate, Government, Healthcare and Higher 
Education (E-learn’04) (Washington D.C., USA, November 
1-5, 2004), 174-177.

[23] Sancho, P. and Fernández-Manjón, B. <e-aula>: Entorno de 
aprendizaje personalizado basado en estándares educativos. 
In Proceedings of the VII Jornadas de Ingeniería del 
Software y Bases de Datos (El Escorial, Madrid, Spain, 

[24] Sanz-Santamaría, S. Estándares de Representación de 
Conocimiento para LMSs. Technical Report UPV- 
EHU/LSI/TR 1-2006, University of the Basque Country, San 
Sebastian, Spain, 2006.


Advances in Educational and Psychological Testing, Kluwer 
Academic Publishers: Dordrecht (The Netherlands), (1990), 
69-95.

[27] Wingersky, M. S. LOGIST: A Program for Computing 
Maximum Likelihood Procedures for Logistic Test Models. 
Applications of Item Response Theory, R. K. Hambleton 
(Eds.), Educational Research Institute of British Columbia, 
Vancouver (Canada), (1983), 45-56.