

Mixing Standards, IRT and Pedagogy for Quality e-Assessment

Silvia Sanz-Santamaría*, José Á. Vadillo Zorita, and Julián Gutiérrez Serrano

Group of Hypermedia and Multimedia, Dept. of Computer Languages and Systems, University of the Basque Country (UPV-EHU), Pº Manuel de Lardizabal 1, 20018 San Sebastián, Spain

This document presents a new authoring tool for e-assessment that lets teachers testing students efficiently. The combination of different research fields regarding assessment, namely learner adaptation, standards and pedagogy, makes possible a better and complete evaluation adapted to learners' knowledge and preferences.

Keywords e-assessment; user adaptation; standards; pedagogy

1. Introduction

E-Assessment is a relatively new concept use to designate the evaluation process done using computer assistance. The assessment could be carried out within a complete e-learning system or thanks to an independent tool dedicated to this purpose. In any case, assessment is a very important part in the learning process that must be understood not as a mere step to go to the next lesson or course, but as a way to continue learning. In this respect it is more significant for online learning than it is for traditional one [1].

There are multiple factors that affect e-assessment. Mainly, it is important to pay attention to Pedagogy and Standards. The first one supplies different psychometric techniques to adapt assessments to learners' level knowledge and abilities [2, 3], as well as other pedagogical techniques to achieve quality assessments [4]. The latest provides rules to design items and tests, and lets teachers exchange them with other mates, which enrich the assessment process of each teacher.

This paper presents a new tool for creation and delivery of adaptive assessment in any domain. All the factors commented before have being taken into account to develop an easy-to-use tool that lets teacher to define items and test as well as import/export them from/to other systems/tools. Then, once the items and tests are stored, the tool is prepared to deliver different types of assessment.

The next section introduces some pedagogical details that should be applied when building up a complete e-assessment tool, as well as some notions about assessment standards. It also comments some available tools/systems that work the assessment process, highlighting their strength and weakness. Section four explains the necessity of a new assessment tool and enumerates the advantages of the proposed tool, focusing on Pedagogy and Standards. Finally, section five presents the conclusions.

2. Characteristics of a quality e-assessment tool

This section presents a brief review about the characteristics that, from the point of view of authors, a quality e-assessment tool must offer. Concretely, this review is done from two points of view: Standards and Pedagogy. The last one is a very vast area, so the section focus the discussion in three important sub-areas related to computer testing: adaptation to the learners' abilities and knowledge, types of items used in the assessment and the pedagogical domain where the assessment takes part. The next four subsections explain all these characteristics and the fifth section enumerates some available systems/applications that work with assessment, emphasizing their pros and cons regarding these characteristics.

* Silvia Sanz-Santamaría: e-mail: sanz@si.ehu.es, Phone: +34 943 01 51 11

2.1 Standards

The usefulness of standards has been demonstrated in many areas of our lives. But thinking about e-learning, the utility of standardization is stronger because of the universal and ubiquitous character it has. In fact, in the last 10 years several associations (AICC, ADL, IMS, etc.) are working at standardization about everything related to e-learning (LMSs, LOs, Packaging, Tests, and more), and they are making a lot of improvements [9].

Particularly, in the area of assessment the standard QTI (Question & Test Interoperability) [10] developed by IMS [11] 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].

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, object positioning, 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.

2.2 Learners' 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) [5] 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) [6]. 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. 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 learner's knowledge. Although the IRT was described more than 50 years ago, the mixture between computers and IRT was a decisive milestone [3]. This research area is known as Computer Adaptive Testing (CAT) [7].

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. Adaptive and classic tests can coexist in the same learning environment. Pedagogues and domain designers must choose which type of test is better for the learning process and the acquisition of knowledge. So, a complete e-assessment tool should include all this types of tests in order to cover all the necessities.

2.3 Types of items

Although the well-known multiple choice item is the most used for evaluation, there are other types that could be applied in e-learning assessment. There is a good compilation in [5].

Most systems provide multiple choice items as unique manner of assessment. This item is complete for evaluation since it could assess different competences. But this is only possible if the item is well constructed, which is not easy. Besides, and due to e-assessment can benefit from computers power, there are many other item types that can enrich an evaluation. Some of them are essays, projects, case studies and simulations [4], matching, ordering, true/false, localization and short answer [8].

2.4 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) [13]. 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. A good e-assessment tool should be prepared to define items and tests for both simple pedagogical domains and complex ones.

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 or ADTEST. These systems allow developing various types of items and also 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 inconvenient 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> [14] is an e-learning system with a module for assessment. All the items stored in the item bank are compliant with IMS QTI (Lite version). The item types possible are true/false and multiple choice. Test Editor [15] is an authoring tool for generating items and tests that store all the items in XML. It allows storing multiple choice items and can work in an adaptive way (IRT 2 parameters). The KOD Project [16] 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 the use of 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 (allow only one or two types) and are applied into simple pedagogical domains. 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.

A mixture between commercial and research tools are Moodle [17] and WebCT [18], both widely implanted in the academic sector. The former is open source software while the latter is proprietary. Moodle allows the definition of true/false, embedded-answer questions, multiple choice, short answer and matching item types, and WebCT allows the last three types and also essays. Both systems are non-adaptive and items/tests are stored in repositories without relation between them and the concepts of the pedagogical domain.

3. The e-assessment generation/delivery tool. The proposal.

As shown in previous section, there is a lack of e-assessment systems/tools that combine all the characteristics commented before. The objective is to cover this gap implementing a tool for developing and delivering e-assessments, using standards and paying special attention to Pedagogy.

Next subsections explain the details of the proposed tool regarding the use of standard and the different pedagogical sub-areas broke done in section two: learners' adaptation, types of items and learning pedagogical domain.

3.1 Standards

IMS QTI is the 'de facto' standard for assessment in e-learning. Developing items and tests with this standard makes easy to import/export them from/to other systems and tools QTI compliant.

At present all the items and tests developed with the authoring tool explained in this paper are IMS QTI compliant (1.2 and 2.0 versions). Teachers do not need to know anything about QTI. The tool automatically stores items/tests according to the standard rules. Now, we are working in developing a software application for checking that imported items/tests are QTI compliant.

3.2 Pedagogy

Regarding learners' adaptation, our aim is to develop a tool for generating adaptive assessment using 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 [19]. The tool also must offer teachers the possibility to develop classic test (predefined and dynamic). At the moment, we have developed an interface that let teachers develop adaptive items only filling a simple form.

This tool wants to take advantage of other pedagogical theories and research studies to improve e-learning assessment. Concretely, the aim 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 [5, 8]. At this moment it is possible to define multiple choice, multiple response, ordering, association (1:1), union (1:N) and fill in blanks items.

Furthermore, the tool will include a help guide for teachers. The objective 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, the guide can recommend teacher which item type select and how to develop it.

Also, Pedagogy is needed for the tool to benefit from the structure of the pedagogical domain in which the assessment takes part. Authors would like to take a step ahead in this area relating items not only with concepts (which is the typical). Specifically, we are working on relating items with: (1) 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; (2) Composed nodes, 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; and (3) Bloom's Taxonomy [20], due to 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. In this respect, there is no pedagogical rule/technique inside the authoring tool yet. We are in contact with pedagogues to collect their knowledge and apply it to resolve all the pedagogical aspects commented in this paper.

4. Conclusions

As shown in section 2, there is a gap 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 re-

search 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 gap. We are conscious of the importance of pedagogy in e-learning, as well as adaptive assessment and the use of standards for exchange 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. Teachers can easily modify these items/tests that have being imported from other systems, adjusting some parameters of them if it was necessary.

The 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 with pedagogical tips and ticks for helping teachers to decide which type of item is better for achieving some skills and how to develop good assessments.

References

- [1] J. Macdonald, *Journal of Assessment & Evaluation in Higher Education*, **29-2**, 215 (2004).
- [2] J.B. Olsen, A. Cox, C. Preece and M. Strozski, Development, implementation and validation of a predictive and prescriptive test for statewide assessment, AERA Meeting, San Francisco, (1989).
- [3] R.K. Hambleton, H. Swaminathan, H.J. Rogers, *Fundamentals of Item Response Theory*, 2, (Kluwer Academic Puclisher, Norwell, Massachusetts, USA, 1991).
- [4] T. Govindasamy, *Journal of Internet and Higher Education*, 4, 287 (2002)
- [5] J. Muñiz, *Teoría Clásica de los Tests*, (Pirámide, Madrid, 1992).
- [6] D.J. Weiss, M.E. Yoes, *Advances in Educational and Psychological Testing*, 69, 1990.
- [7] R.K. Hambleton, J.N. Zaal, J.P.M. Pieters, *Advances in Educational and Psychological Testing*, 341, 1990.
- [8] J.A Mateo, *La evaluación educativa, su práctica y otras metáforas*, (ICE-Horsori, Barcelona, 2000).
- [9] S. Sanz-Santamaría, *Estándares de Representación de Conocimiento para LMSs*. Internal Report: UPV-EHU/LSI/TR 1-2006
- [10] IMS Question & Test Interoperability Specification, <http://www.imsglobal.org/question/index.html>.
- [11] IMS Global Learning Consortium, Inc. <http://www.imspjroject.org>.
- [12] W. Kraan, IMS Question and Test Interoperability gets major make-over. CETIS, <http://assessment.cetis.ac.uk/content2/20040712004043> (2004)
- [13] T.A. Pérez, J. Gutiérrez, P.Lopistéguy, *Journal Informática y Automática*, December, 45 (1997)
- [14] P. Sancho, B. Fernández-Manjón, <e-aula>: Entorno de aprendizaje personalizado basado en estándares educativos, Proceedings of VII Jornadas de Ingeniería del Software y Bases de Datos, El Escorial (Madrid), Spain, 18-22 November 2003.
- [15] C. Romero, S. Martín-Palomos, P. De Bra, S. Ventura, An Authoring Tool for Web-Based Adaptive and Classic Tests, Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education, Washington D.C.,USA, 1-5 November 2004, pp. 174-177.
- [16] C. Karagiannidis, D. Sampson, F. Cardinali, Proceedings of the Second IEEE International Conference on Advanced Learning Technologies, Madison, USA, 6-8 August, pp.21-24.
- [17] Moodle, <http://moodle.org>
- [18] WebCT, <http://www.webct.com>
- [19] M.A. Barton, F.M. Lord, An upper asymptote for the three parameter logistic item-response model, *Research Bulletin* 81-20, (Educational Testing Service, Princeton, New Jersey, 1981).
- [20] B.S. Bloom, *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*, (David McKay Company Inc., New York, USA, 1956).