Interactive Learning Environments
Vol. 19, No. 3, June 2011, 267–283

Routledge
Taylor & Francis Group

Multi-criteria adaptation in a personalized multimedia testing tool based on semantic technologies

Fotis Lazarinis*, Steve Green and Elaine Pearson

School of Computing, University of Teesside, Middlesbrough, Tees Valley, TS1 3BA, UK

(*Received 1 August 2008; final version received 20 December 2008*)

In this article, we present the characteristics and the design of a modular personalized multimedia testing tool based fully on XML learning specifications. Personalization is based on the characteristics of the individual learners, thus the testing paths are tailored to their needs and goals. The system maintains learner profiles rich in content from which diverse information can be elicited and presented to educators to help them understand their learners. At the end of the article, specific use cases are discussed and the educational advantages are discussed on the basis of an evaluative study.

Keywords: adaptive testing; e-learning; interoperability; re-usability; educational hypermedia; personalization

Introduction

The advent of educational technology has led to a complete rethink on both the curricula offered at school level and their assessment. Fixed-length and adaptive computerized tests exist for many years (Kingsbury & Houser, 1993; Mills et al., 2002). Adaptive testing systems are computerized tests aiming to accurately reflect the instructional level of each student with a shorter number of queries tailored to the characteristics of the individual learner (van der Linden & Glas, 2000; Wainer, 2000). The existing adaptive systems have proved to be beneficial as the assessments tend to be shorter and more accurate. However they base their adaptation mainly on the learner's performance using various statistical models. On the basis of adaptation strategy with other learner characteristics, such as their aims and their preferences, would help educators gain a better understanding of their students and students to form a clear idea of their knowledge and difficulties.

Adaptive testing and learning systems maintain several data about the learning progress of learners and the educators' approaches in the profiles that the systems maintain (Brusilovsky, 2001; Kobsa, 2007). These data could be utilized in making useful assumptions available to learners and educators for further enhancing the learning process. Data however should be stored in such a way that could efficiently support their efficient utilization and alternative presentations. This could be achieved by coding the data in standardized XML structures.

*Corresponding author. Email: f.lazarinis@scm.tees.ac.uk

ISSN 1049-4820 print/ISSN 1744-5191 online

DOI: 10.1080/10494820902909459
http://www.informaworld.com

In educational technology, the exchange of data among applications is of crucial importance as it allows applications to re-use testing items and online lessons and to share the knowledge about learners accumulated in various applications (Duval, 2004; Fallon & Brown, 2002; Robson, 2000). To support these demands several co-operating organizations are working to develop learning standards. Learning standards refer to the standardization of XML structures which are used to describe various aspects of the learning procedure. IEEE (Institute of Electrical and Electronics Engineers – www.ieee.org) and IMS Global Learning Consortium (Instructional Management Systems project – www.imsglobal.org) are two examples of organizations and consortiums which developed XML learning standards for describing, among others, e-lessons, user profiles, e-portfolios and testing data. A promising but still open research issue is the efficient utilization of learning standards for storing and semantically enriching the various categories of data (e.g. topics, user profiles, testing data, results) in a testing system.

In this article, we discuss the design decisions and we present the components of a modular adaptive testing tool based on the Topic Maps, IMS LIP and IMS QTI XML standards. Personalization is based on a set of rules which concern the knowledge, performance, goals and preferences of learners. The system is composed of separate modules so as to allow the system to be extensible. Each module codes its data into a distinct XML e-learning standard to promote reusability and interoperability of the information. At the end of the article, specific use cases are discussed and the anticipated advantages are presented through an authentic testing process.

Computer-based assessment

Computerized testing is increasingly being viewed as a practical alternative to paper-and-pencil testing (Kingsbury & Houser, 1993; Mills et al., 2002). A traditional fixed-length computerized exam presents the same number of questions to each test taker, without considering the previous knowledge or other characteristics of each learner. The score from this type of test usually depends on the number of questions answered correctly. Adaptive testing systems are computerized tests aiming to accurately reflect the instructional level of each student with a shorter number of queries tailored to the characteristics of the individual learner (van der Linden & Glas, 2000; Wainer, 2000). They are based on Item Response Theory (IRT) (Hambleton, Swamination, & Rogers, 1991) and are used mainly as skill meters presenting the overall learner's score on a subject and a pass/fail indication. More specifically, test items dynamically adjust to a student's performance level, and as a result, tests are usually shorter and test scores tend to be more accurate (Thissen & Mislevy, 2000).

A number of computerized adaptive testing (CAT) tools have been implemented by academic institutions, e.g. SIETTE (Conejo et al., 2004), and international companies for specific examinations (GRE Exam, 2006; Microsoft CAT, 1999), and have been extensively utilized in recent years. The main advantage of a CAT over a traditional computerized test design is efficiency. The CAT systems can determine a person's score with fewer questions, sometimes reducing the length of the test by 60% or more. Thus, the estimation of the current learner knowledge has been proven quite reliable in CAT tools. Nevertheless, some limitations of CAT systems have been identified in the literature. For example, it is impossible to feed an operational

adaptive test with brand-new, unseen items; all items must be pre-tested with a large enough sample to obtain stable item statistics (Wainer & Mislevy, 2000). Additionally, in a simulated evaluation it was shown that some student stereotypes may not be appropriately assessed in CAT systems (Abdullah Chua & Cooley, 2002).

Other types of adaptivity in assessment systems can be found in QuizPACK (Brusilovsky & Sosnovsky, 2005) and QuizGuide (Sosnovsky, 2004). These tools support self-assessment of programming knowledge with the aid of Web-based individualized dynamic parameterized quizzes and adaptive annotation support (Brusilovsky, 2001). The tools described in these papers are domain dependent and are basically possible in domains such as mathematics, physics, and programming. The ideas of a rule based testing system are discussed in (Tzanavari, Retalas, & Pastelis et al., 2004). However, the presented system does not fully conform to learning specifications. For example, the limited predefined rules and the topics are not coded in an XML standard. Also, the user profiles combine elements from different standards thus making more difficult the exchange of data with other educational tools. Its centralized architecture prevents the straightforward integration of new modules and the pedagogical implications of their work are not discussed or evaluated. Another adaptation technique, adaptive questionnaires (Kehoe & Pitkow, 1996), has been used mainly in computer-assisted Web surveys. This method causes the generation of a dynamic sequence of questions depending on learner's responses reducing the number and complexity of questions presented to users. They have been used to assess web users' attitudes in a computer-assisted testing and evaluation system for the World Wide Web (Chou, 2000).

Assessments may be either formative or summative (Harlen & James, 1997; Sadler, 1989). Formative assessment is often performed at the beginning or during a program, thus providing the opportunity for immediate evidence for student learning in a particular course or at a particular point in a program. Summative assessment is comprehensive in nature, provides accountability and is used to check the level of learning at the end of the program. Tailoring the assessment procedure only to the performance of a learner, as happens in the previous systems, is a constrained approach suitable only for summative assessments. From a pedagogical perspective, assessments should additionally consider the preferences and goals of the learners. Educational adaptive hypermedia learning systems build a model of the goals, preferences and knowledge of each individual user and use this model throughout the interaction with the user, to adapt to the needs of that user (Brusilovsky, 2001). Similarly, adaptive testing tools should tailor the testing paths to the preferences and goals of the learners in addition to their performance.

Adaptive testing and learning systems maintain several data about the learning progress of learners and the educators' approaches in the profiles that the systems maintain (Brusilovsky, 2001; Kobsa, 2007). However, most of them use proprietary formats to store this information. Consequently, the knowledge cannot be shared across different educational tools (De Bra, Aroyo, & Cristea, 2004). E-learning standards provide fixed XML data structures and communication protocols for e-learning objects and cross-system workflows. This enables interoperability between applications. Currently, there are a number of XML e-learning specifications for user profiles, learning content, assessments and metadata from various organizations. An interesting research path would be the utilization and combination of e-learning standards in an adaptive testing system in an attempt to promote interoperability.

Research questions

The Advanced Distributed Learning initiative (ADL – www.adlnet.gov) defines a set of 'abilities' for e-learning tools and technologies. These abilities are reusability, accessibility, interoperability, adaptability, durability, and affordability. On the basis of these set of 'abilities' and on some of the shortcomings of the existing adaptive systems discussed in the previous section, the questions driving our research are:

(1) How can we improve adaptive testing systems to include the preferences, goals, previous educational experiences, and knowledge of the learners?
(2) How can we use learning standards to promote interoperability and reusability of the learners' data?
(3) How can we create an extensible testing tool where additional modules could be promptly integrated?

These questions concern the reusability, interoperability, adaptability, and durability goals set in educational tools. Building a system with these goals in mind is beneficial both pedagogically and practically. From a pedagogical perspective learners would have the opportunity to test their knowledge based on their performance, on their goals and on their difficulties and educators will be able to gain a better understanding of their learners' achievements and difficulties. From a practical view, educational technologists and administrators will be able to share information between their systems and form alternative views of the existing data for learners.

Description of the system

Figure 1 shows the components of our adaptive testing tool. The system maintains user profiles (models) for learners and educators. In these models, several attributes describing the users are maintained. These attributes vary from personal details, to user knowledge and to user preferences. The testing items are questions in textual or in a multimedia form (sound, image, video, java applet, etc). Each test is associated with a specific topic of a domain (e.g. Electromagnetism in Physics).

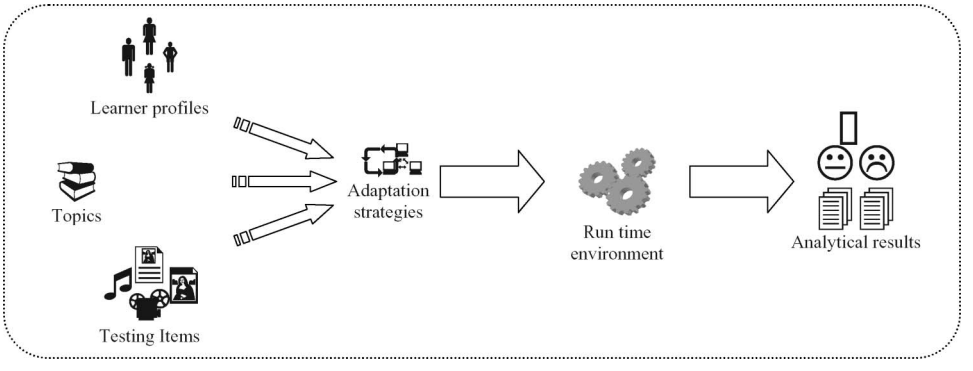

Figure 1. Components and flowchart of the adaptive testing procedure.

The adaptation strategies rely on the sequence of the testing items and on rules which are based on the performance of a learner, their preferences and their goals. For example, a learner might want to test her/his knowledge on items of increased difficulty, so the system has to present only these items. The run time environment takes as input all the aforementioned and executes the test, adapting the sequence of the presented items according to the tester's replies and the adaptation rules. Finally, testers are presented with their score and analytic descriptions and conclusions about their progress.

Topics

To manage the topics a visual tool which graphically represents the topics and their dependencies is developed (see Figure 2). The topic dependencies may form a hierarchy or a graph editable by educators with administrative privileges. Educators who prepare tests could then visually select their appropriate topic and associate it with their testing material. The topics' data need to be represented in an XML standard to promote the goal of interoperability set in a previous section. Suitable e-learning technologies for coding the topics' data are concept maps and topic maps.

Concept mapping is a technique for representing knowledge in graphs. Knowledge graphs are networks of concepts (Lawson, 1994). Networks consist of nodes (points/vertices) and links (arcs/edges). Nodes represent concepts and links represent the relations between concepts.

A more appropriate technology is topic maps. Topic maps are an ISO standard for the representation and interchange of knowledge, with an emphasis on how easy it is to find the right information (Topic Maps, 2002). A topic map can represent information using topics, associations, and occurrences. They are thus similar to semantic networks and both concept and mind maps in many respects. In loose usage all those concepts are often used synonymously, though only topic maps are standardized. Topic maps have a standard XML based interchange syntax called XML Topic Maps (XTM), as well as a de facto standard API called Common Topic Map Application Programming Interface (TMAPI), and query and schema

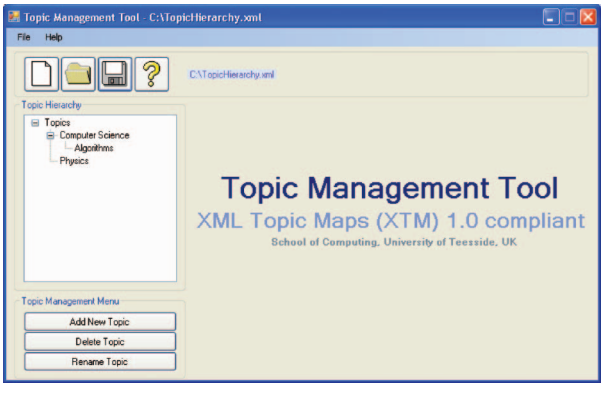

Figure 2. Topic map management tool.

languages are being developed within ISO. Topic maps for e-learning are proposed in other studies as well (Dichev et al., 2004).

Because topic maps are standardized they were selected to code the topic dependencies in our adaptive testing tool. A graphical tool was created for this purpose (see Figure 2). This tool supports the creation and update of topics and presents the dependencies in a graphical mode. The output of the tool is validated against the DTD (Document Type Definition) of topic maps (http://www.topicmaps.org/xtm/1.0/). The purpose of a DTD is to define the legal building blocks of an XML document. It defines the document structure with a list of legal elements and attributes. So, the produced XML files should be validated using the DTD schema.

Learner profiles

Learner profiles consist of a set of attributes which describe the personal data of learners (e.g. name, date of birth), their formal education (e.g. degrees), their goals (e.g. to achieve a high mark), their preferences (e.g. to be presented with analytical explanations at the end of test or to allow their profiles to be anonymously viewed (Bull & McKay, 2004) and their evaluation data (e.g. activities performed in the testing system). In adaptive education hypermedia systems, topics are overlaid in the learner profile (De Bra et al., 2004). That is, the knowledge of the learner is associated and measured with respect to specific topics of the topic hierarchy, usually called domain model.

IEEE PAPI (Public and Private Information – http://edutool.com/papi) and IMS LIP (Learner Information Package – http://www.imsglobal.org/profiles) are well known XML learning standards for learner profiles. As IMS mentions, IMS LIP utilizes and expands most of the features of IEEE PAPI. Therefore IMS LIP is more general. It has also been proposed to combine elements from the two standards to provide an expanded user profile (Dolog, Gavriloaie, Nejdl, & Brase, 2003). However, combination of the two standards compromises data sharing as other educational tools abiding by one of the two standards would not be able to import data from the combined model.

IMS LIP contains a large number of attributes capable of coding all the information for learners. Studying the e-learning standard thoroughly it was decided to use the following attributes to code the desired information: <learnerinformation>, <goal>, <accessibility>, <qcl>, <competency>, <transcript>, and <activity>. Using these data, we can code the desired learner data which as explained are personal data, knowledge, qualifications, achievements, and activities of the learners.

The learner profiles management tool supports initialization of profiles, import from other sources, update and display of profiles in alternative modes (see Figure 3). Initialization is performed either per user or in batches, based on specific learner pre-defined stereotypes as in older tools (Kay, 1990). At the moment, these stereotypes are beginner, intermediate and advanced and concern primarily the knowledge level of learners to specific topics. Through the iLM (interoperable Learner Modelling kit) shown in Figure 3, both learners and educators can log in and based on their assigned rights to edit or view their profiles. Educators can additionally create or update new accounts. The output of the iLM tool validated using the IMS LIP v1.0 DTD.

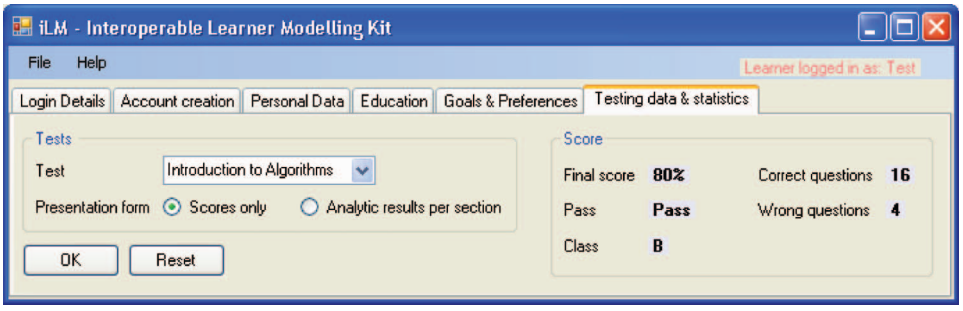

Figure 3. Interoperable learner modelling kit.

Testing items

IMS QTI (Question and Test Interoperability) standard (http://www.imsglobal.org/question) structures material into assessments, sections, and items. An item is the formal name for a question and assessment is the terminology used for a test. The latest version of IMS QTI supports simple and composite questions and question repositories.

To manage assessments and questions an easy to use visual tool was developed. Figure 4 shows the interface of the tool which allows the creation of various question types (e.g. true/false, multiple choice, gap match, association, etc). Educators are able to associate their assessments to specific topics and view their assessment in a hierarchical mode. The editor allows creation, import, and modification of QTI items. The output of the editor conforms to QTI v2.1.

As seen in Figure 4, for each question educators can provide feedback for each reply. Similarly in Multiple Choice questions, educators can provide feedback for each of the available choices. Feedback may be simple text or reference to an external resource. The existence of hyperlinks to Web resources is also possible. Apart from the feedback, educators are able to define the maximum number of attempts, the score of the question, the penalty for each try, the difficulty level of the questions and the maximum allowed time to answer the question. They are also able to associate multimedia data with their questions (e.g. images). Thus, educators are able to semantically enrich their questions by associating a rich number of attributes with each question. These attributes are supported by the latest version of QTI and are encoded in XML format.

Adaptation strategies

Adaptation of the testing procedure relies on the performance of a learner and on their preferences and goals. Educators are able to force advancement of the procedure in case a learner has passed a threshold, set in specific questions. For example, an educator will be able to define rules like:

- On question i of section j check the learner's score.
- If score is greater than k move to question l of section m [Else move to next question $i + 1$ of section j].
- If question i of section j is wrongly answered then show a similar question.

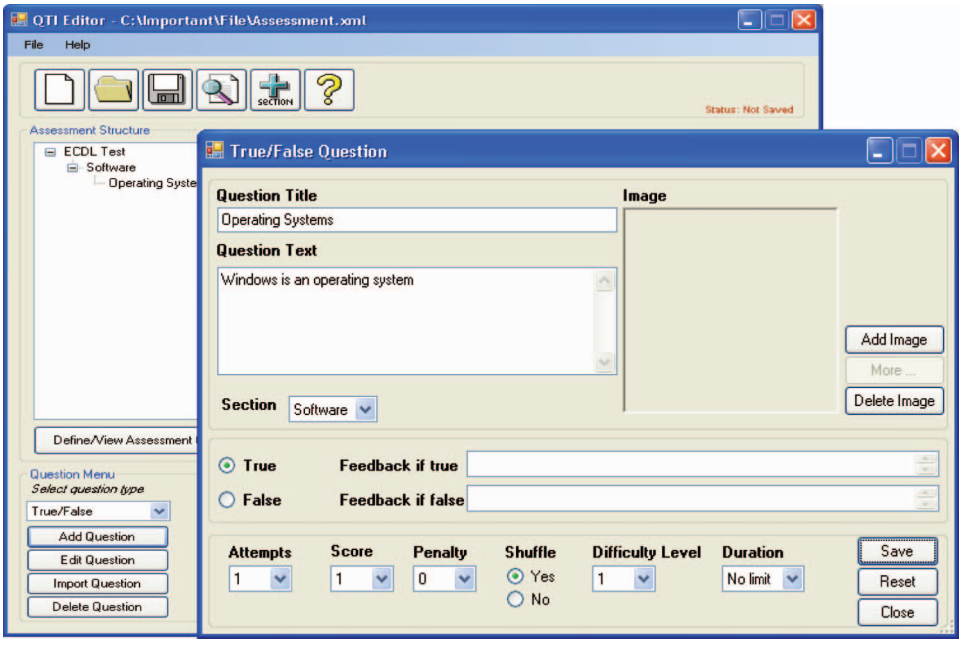

Figure 4. Editor of QTI assessments.

Basing the adaptation on the preferences and goals of users could be achieved by employing rules like:

- Show questions of medium or higher difficulty only.
- Show all questions and give immediate feedback.
- Show a similar question to the current question.

As explained in the introductory section, the described system is designed both for summative and formative evaluations. Summative assessments have primarily to rely on the performance of the learner, so customization should be left to the rules set by the educators. Formative assessments aim at helping learners realize what they do know and what they do not. So they should be based on the goals of the learners. If learners feel that they are well prepared they could ask for difficult questions. Or if they do not understand some questions, testers could ask for more questions on this sub-topic, if they are available of course.

To be able to support this kind of behaviour educators need to provide alternative questions for certain testing items. As seen in Figure 5, assessments look like a two dimensional array. Each section will consist of a number of questions and for each question a number of alternative items could be defined. The creation of alternative questions is not compulsory and the number and format of the alternative questions are not pre-defined and it is left to the educator to decide. Thus, educators can define alternative material for the questions which they feel that are difficult and decide on the quantity and type of the alternative questions.

Our tool apart from being flexible could provide educators with ample data to understand their students. They can see the scores of their students, the questions that haven't managed to answer, the topics for which they had to try more than one question, their adaptation paths, the questions that were rarely presented to users, etc. These data could be used to enhance their teaching as well. For example, they could understand in which topics their students are not confident and provide more examples and study resources. Also they could see in which questions their students have tried more similar questions, in which points the adaptation rules triggered off, if their adaptation strategies really help students completing the tests faster and with less questions based on their characteristics, etc.

The adaptation management tool allows educators to visually customize predefined rules and apply them to certain questions (see Figure 6). The latest version of QTI supports adaptive items and rules which are used to store the adaptation data in our tool. The <responseProcessing> QTI element and its sub-elements such as <responseCondition>, <responseIf>, <and>, <not>, are used, among other QTI XML elements, to code our rules. To code efficiently more complex rules in the future, IMS Simple Sequencing (http://www.imsglobal.org/simplesequencing/) could be integrated to the system.

Each adaptation rule will have an action, a criterion and a trigger point. Trigger points are set on specific questions. The supported set of actions is:

(1) Start assessment on question.
(2) Start assessment on section.
(3) Move to previous/next question.
(4) Move to previous/next section.
(5) Re-try specific questions.
(6) Re-try assessment.
(7) Show questions of specified difficulty level.

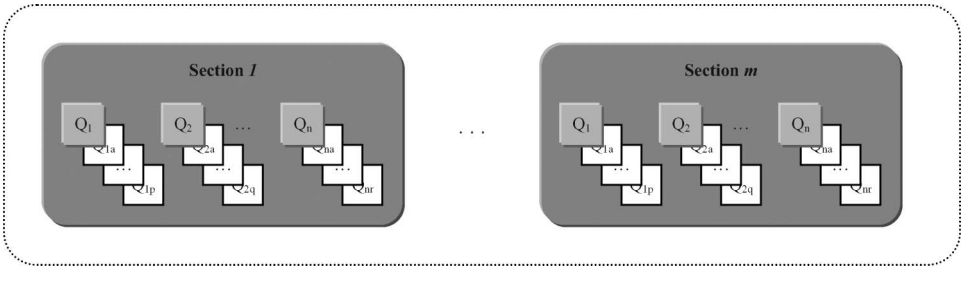

Figure 5. Visualization of the structure of assessments.

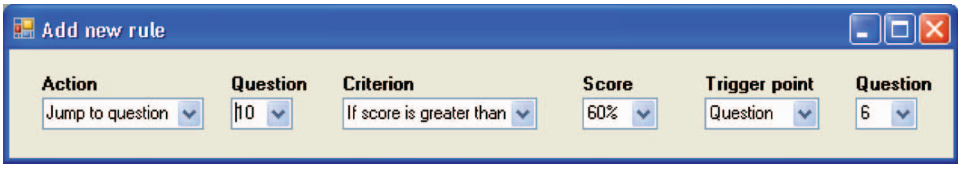

Figure 6. Rule customization.

The criteria are based on the following factors:

(1) The performance.
(2) The previous knowledge level.
(3) The formal qualifications.
(4) The difficulty level of the questions a learner wishes to try.

Run time environment

The run time environment takes as input the data of the previous phases and presents the user with the adapted sequence of questions (see Figure 7). As explained, tests are split in sections and each user is presented sequentially one question at a time. On the basis of the response in the current question and on the data of the previous questions the next question is decided. Adaptation decisions are taken when test takers reach specific threshold set by educators or meet a specific criterion related to the difficulty of the questions. On the basis of learners' responses and actions the learner profile is updated and the knowledge of the learner on the test topics is updated.

At the end of the assessment procedure learners are presented with analytical statistics and inferences about their progress and they can anonymously view the assessment results of other learners who took the same exam (see Figure 3). Educators may have named access to all the data of their learners and inferences automatically made by the system will be presented to them.

Usage scenarios

The previous sections presented the ideas and the technical design decisions of the testing system and also discussed its implementation details. To help readers understand the system's pedagogical gains, apart from its obvious data sharing ability, in this section we discuss some usage scenarios and in the subsequent section we analyze a complete assessment.

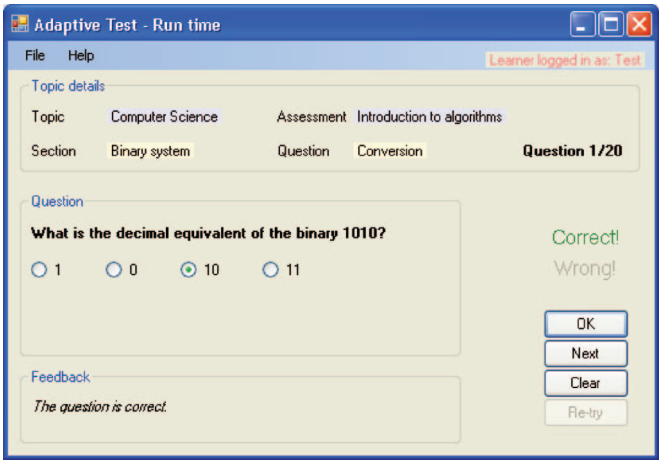

Figure 7. Run time environment of the adaptive testing tool.

Scenario A

A student wants to practice for the final exam but s/he wants only to try difficult questions as s/he is confident that s/he is well prepared. In a fixed-length computerized test, users have to answer all questions and thus assessments are unnecessarily longer and they are based only on the performance of the learners. In our testing environment testers could ask the system to show them only the difficult questions. The system will be able to adapt the items presented to the user according to the specified preferences. In this way the goals of the users are taken into account in the testing procedure.

Scenario B

A teacher wants to prepare a test for a diverse student population. She/he can use the previous knowledge and formal qualifications attributes of the learner profiles to adapt the testing procedure. Teachers will be able to define rules like, 'if the previous knowledge of a learner on the topic is high or if the learners hold a higher degree on the topic then show only questions of medium or greater difficulty level to the learner'. Or they could use a rule like 'start the testing procedure at a specific question'. Thus educators would be able to harmonize the testing procedure based on their learners' previous knowledge and qualifications and not on their performance only as in CAT systems.

At the end of the testing procedure a number of conclusions could be made based on the adaptation decisions made by the system. Educators will be able, for example, to see how users with previous knowledge did and if their knowledge level increased or decreased after all.

Scenario C

Educators usually present more difficult questions to students with higher knowledge and an augmented set of easier questions to students with lower performance to motivate both of these student categories. Our system is able to simulate this kind of educators' intelligence by applying rules like 'show only questions that match the knowledge level'. More importantly, different educators may define different adaptation rules on the same student population and on the same test. In all cases, the adaptation is based on the previous knowledge the system maintains for the learners. In CAT systems, educators are able to define the questions only and cannot intervene on how these questions will be presented to the learners.

Running a complete assessment

Structure of the adaptive assessment and student sample

The described system offers a rich set of personalization options and thus extensive evaluations are needed to realize its full potential. As a first, evaluation attempt we run a 3-section formative assessment each consisting of 20 single choice questions related to algorithmic concepts. The duration of the assessment was 45 min. The first section contains easy questions, the second one contains questions of medium difficulty and the last set of questions is of increased difficulty. This formative

assessment was run in January 2008 after a 3-week introductory course related to algorithms. The participants of this course were first year undergraduate college students. Thirty students agreed to participate in our experiment. Feedback for each reply was given only if it was available. The default feedback for each question was Correct/Wrong and it was presented in case teacher defined feedback was not available (see Figure 7).

The main aim of this assessment is to help students understand their current knowledge on the algorithmic topics they have been taught and to help the teacher see the misconceptions of their students. Also, we wanted to see what the students believe about their level of knowledge and whether their beliefs are true or not. Therefore, after having initialized their profiles we let the users decide their level of knowledge on the introductory algorithmic concepts.

Our adaptive system allows us to create a number of rules so as to tailor the testing procedure to the individual. Questions and question choices were automatically shuffled within each section to increase utilization of each item. In the current assessment we employed the following rules:

Rule 1: Ask the student to decide whether s/he wants to start with the easy, the medium, or the difficult questions and then show the respective questions.
Aim: The aim of this rule is to involve learners in the formative assessment by letting them decide which questions they want to try. This way the testing procedure can adapt the procedure on their goals.

Rule 2: If learner's previous knowledge on the topic is high, then start the testing procedure on question 21 (the first question of medium difficulty).
Aim: The aim of this rule is to help users who declared that they have a high previous knowledge on the tested topic to complete their assessment faster. Also the results of this formative assessment will help both students and educators understand their true knowledge level. This rule is activated only if does not collide with the previous rule.

Rule 3: (In every section) If a learner has correctly answered all the first 15 questions of the current section then move on to the next section.
Aim: This rule aims to help students with a high performance to proceed faster than the other students.

Rule 4: (In every section) At the end of the section if the score of a learner is less than 70% present again the erroneous questions to the user.
Aim: This rule aims to help students re-attempt the questions that they answered incorrectly.

The purpose of the above rules is to allow users with a high knowledge level or with a high performance to proceed faster and to give a second chance to learners with lower performance. Also it takes into account the goals of the learners by asking them to decide on the difficulty level of the questions. All the activities of the users were recorded to their learner profiles and were shown later to both the learner and the educator in textual mode in an attempt to help them their learning progress.

Data analysis

As mentioned the adaptive formative assessment was taken by 30 adult students. During the initialization of their profile 21/30 (70%) of the learners stated that their knowledge level is high in the respective topic. The rest of the group (9/30–30%) indicated that their knowledge level is average on the introductory algorithmic concepts.

The first adaptive rule concerns the setting of the assessment's starting point based on the user-preferred difficulty level. Although the majority of the students had initialized their models to high knowledge, most of them (17/30–56.67%) wanted to start the assessment on a lower difficulty level (see Table 1). For example, the nine of the 21 (42.85%) learners who initialized their knowledge level as high on the introductory algorithmic concepts preferred to start on the medium difficulty questions. Four of the 21 (19.05%) high knowledge users preferred to start from the easy questions, and only the 8/21 (38.10%) test participants of this knowledge level decided to let the system start the assessment based on their initial knowledge level. 4/9 (44.45%) of the nine learners with average knowledge level decided to set themselves the starting difficulty level. The second rule was initiated for the rest 13/30 users (43.33%).

Table 2 shows that, finally after the activation of the first two adaptive rules, the starting point of the assessment was on different questions based either on the system's beliefs or on the goals of the learners. These statistics indicate that self tailoring testing paths are important especially in formative assessments, as they support more effectively the needs of the test participants.

The third rule triggered in 11 cases (see Table 1). Seven of these cases concerned the advancement from the easy questions to the average difficulty questions and in the other four cases, the testing procedure advanced to the last section from the medium difficulty questions. This rule helped students with high performance to proceed faster. On the activation of this rule, users were asked whether they really wanted to proceed to the next level or to try the rest of the questions of the current section. This flexibility was given to learners in order to better accomplish their

Table 1. Statistics related to the activation of the adaptive rules.

Rule	Times the rule activated
Rule 1 (Selection of the starting difficulty level by the users)	17
Rule 2 (Selection of the starting difficulty level by the system)	13
Rule 3 (Advancement to the next section)	11
Rule 4 (Retry of the erroneous questions)	9

Table 2. Assessment's starting point after the initiation of the adaptive rules 1 and 2.

Starting point	No. of learners
Section 1 – question 1	8
Section 2 – question 1	14
Section 3 – question 1	8

goals. Two users had chosen not to proceed to the next level but to answer all the questions of the section.

Figure 8 shows the number of questions that the students have attempted based on the adaptive rules that have been activated. As it can be seen most of the students were shown different number of questions, although some of them started on the same section. We have three different starting points, based on their prior knowledge level or on the learner goals. Group A started on section 1 and it was shown 60 questions. Groups B and C attempted 40 and 20 questions, respectively. Within each group the number of questions was differentiated according to the performance of the test participants (i.e. activation of rule 3). For example, seven participants of Group A moved faster to the next level because they answered correctly the 15 questions of section 1.

The last rule activated nine times. Five of these cases concerned users of the last group; that is users who started on the last set of questions. Some of these users had a high error percentage and the erroneous instances were presented to them in an attempt to give them a second chance. However, the final score was calculated on the basis of the first examination round to reflect their real knowledge.

From the previous analysis, we can conclude that the prior knowledge and the user goals influenced the testing procedure more than the performance of the testers and tailored the testing procedure to the users' educational goals. From a pedagogical point of view this has many advantages. Learners who believed that they have a high knowledge on the tested topics tested on their questions which matched to their knowledge level. They also had the chance to re-attempt the erroneous items in case they had many errors. By presenting the erroneous items to the learners, they were able to realize their misconceptions and appreciate their true knowledge. Students with higher knowledge completed the test faster and were presented only with the questions matching their knowledge level. The results shown to the educators helped them in identifying the most problematic topics based on the difficulties the learners faced. They were able to see which questions had the most erroneous replies and to utilize this knowledge in enhancing their teaching and possibly rephrasing these questions.

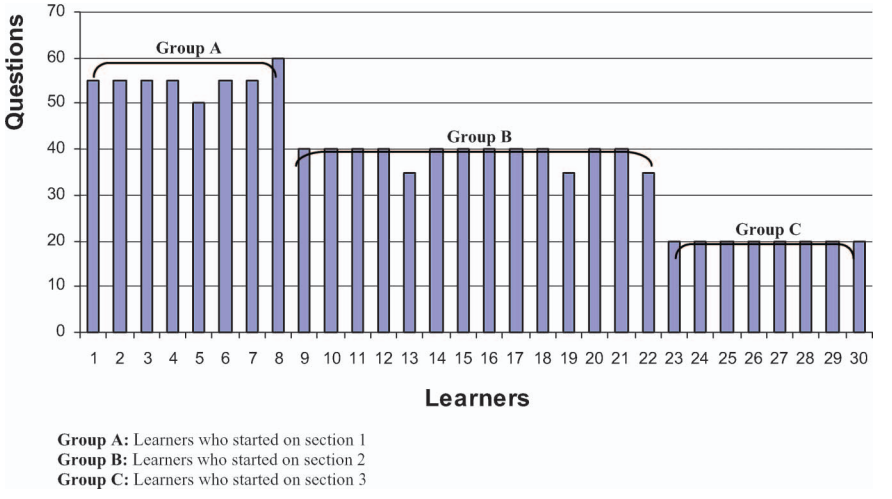

Figure 8. Number of questions per group of users.

Discussion and future research

This article presents a multimedia testing system with adaptive behaviour driven by the set of research aims presented in a previous section. The presented system offers different rules for adapting the testing procedure. Most of them are typical approaches originating from how teachers behave in real classrooms. They re-state their questions to help students better understand the underlying concept, they re-ask the same question in case of an erroneous answer, they ask questions of increased difficult to students who have higher previous knowledge and easier questions to students with lower knowledge level in order to motivate them. Thus one of the innovations of the adaptive assessment tool is the ability to imitate, up to a point, the intelligence of a tutor. On the contrary, CAT systems based on IRT theory rely on statistical models to adapt the testing procedure to the performance of the students and to reliably estimate their knowledge on mainly summative assessments. In an essence our system is complimentary to the previous techniques as it is more oriented on different factors than solely the performance of the learners. This has an obvious effect in learning as it will allow learners to estimate their true knowledge and misconceptions and educators to understand the difficulties of their students.

The modularized architecture of the tool allows easy authoring of adaptive and non adaptive tests and the straightforward integration of additional modules. Testing data conform to IMS QTI and thus they can be re-used and shared across different testing systems. Adaptation could be customized based on a set of rules. A number of actions and criteria are offered to test creators. The system maintains interoperable user profiles conforming to IMS LIP which contain a number of attributes for recording the learning progress of each learner and which could be presented to learners and educators graphically and textually. The second innovation of the tool is that all the data conform to learning standards and thus they could be shared with other learning management systems. Coding all the data categories in a distinct XML standard has the advantage of the independent utilization of the data by applications with different pedagogical aims, thus maximizing content re-usability.

The evaluative study showed that, compared with fixed length computerized tests, in our testing tool assessments are shorter and customized in a number of ways. Existing computerized adaptive tests have a high percentage of unused items and all new items must be pre-tested with a large enough sample to obtain stable item statistics. In our system, item utilization would be higher and new items and tests are directly integrated into the testing procedure. Most importantly adaptation is not only based on performance as in CAT systems, but also on factors such as prior knowledge or educational goals.

Learners are able to try additional items related to concepts which they consider difficult. Thus, the multimedia testing tool presented in this article efficiently supports the educational aims of various learner categories. The supporting rules allow backtracking to wrongly answered questions providing students with a second chance. The analytic results presented at the end of the assessment allow both learners and educators to realize the learning difficulties and misconceptions of the testers. These options are especially important in formative assessments, but could be utilized in summative assessments as well. In any case more tests are needed to realize its full potential and possibly add new functionalities based on the suggestions of the users.

To realize the full potential of the tool and its pedagogical added value, we need to run more tests designed by educators of various disciplines. This will allow us to

realize which adaptation rules are more useful and which they need to be modified. An augmented base of test participants would direct the future modifications and additions to the system. Comparisons with similar paper based and non-adaptive tests are needed so as to realize how accurate and efficient are the adaptive tests. Post-survey questionnaires are already designed in order to reflect the student and teacher experiences with our adaptive tool.

Notes on contributors

Fotis Lazarinis is a part-time lecturer at a Technological Educational Institute in Greece. He has authored over 45 refereed papers in international or national conferences, journals and research handbooks. He has also published several Computer Science educational books in Greek and served as a review member for conferences and workshops.

Steve Green is Principal Lecturer in Digital Media in the School of Computing at the University of Teesside, England. He is the section head for Digital Media and the Web and also the Director of the RIME (Research into Interactive Multimedia Education) research centre. He researches and publishes in the areas of multimedia design, learning technologies and the use of computers with people with special needs.

Elaine Pearson is Director of the Accessibility Research Centre at the University of Teesside, England. She has research interests in all aspects of computing to support the needs and preferences of disabled students. Since the award of a study abroad fellowship with the Leverhulme Trust in 2000, Elaine has been a Visiting Research Fellow at the University of New South Wales, Sydney, Australia. She has published and presented nationally and internationally on accessibility issues and is a member of the conference review committees for EDMEDIA and ALT and a referee for a number of journals including ALT-J and IJEL.

References

Abdullah Chua, S., & Cooley, R. (2002). Using simulated students to evaluate an adaptive testing system. *IEEE Proceedings of the International Conference on Computers in Education* (pp. 614–618). Los Alamitos, CA: IEEE Computer Society.

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, *11*, 87–110.

Brusilovsky, P., & Sosnovsky, S. (2005). Individualized exercises for self-assessment of programming knowledge: An evaluation of QuizPACK. *Journal on Educational Resources in Computing (JERIC)*, *5*(3). Retrieved December 1, 2008, from: http://doi.acm.org/10.1145/1163405.1163411

Bull, S., & McKay, M. (2004). An open learner model for children and teachers: Inspecting knowledge level of individuals and peers. In J.C. Lester, R.M. Vicari, & F. Paraguaçu (Eds.), *Intelligent tutoring systems: Seventh international conference* (pp. 646–655). Berlin, Heidelberg: Springer-Verlag.

Chou, C. (2000). Constructing a computer-assisted testing and evaluation system on the world wide web – the CATES experience. *IEEE Transactions on Education*, *43*(3), 266–272.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J., & Ríos, A. (2004). SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, *14*(1), 29–61.

De Bra, P., Aroyo, L., & Cristea, A. (2004). Adaptive web-based educational hypermedia. In M. Levene & A. Poulovassilis (Eds.), *Web dynamics, adaptive to change in content, size, topology and use* (pp. 387–410). Heidelberg, Germany: Springer.

Dichev, C., Dicheva, D., & Aroyo, L. (2004). Using topic maps for web-based education. *International Journal of Advanced Technology for Learning*, *1*(1), 1–7.

Dolog, P., Gavriloaie, R., Nejdl, W., & Brase, J. (2003). Integrating adaptive hypermedia techniques and open rdf-based environments. *Proceedings of 12th International World Wide Web Conference*. New York, NY: ACM.

Duval, E. (2004). Learning technology standardization: Making sense of it all. *International Journal on Computer Science and Information Systems*, *1*(1), 33–43.

Fallon, C., & Brown, S. (2002). *e-Learning standards: A guide to purchasing, developing, and deploying standards-conformant e-learning*. Boca Raton, FL: CRC Press.
GRE Exam (2006). *Kaplan GRE exam, 2007. Edition: Premier Program*. New York, NY: Kaplan Education.
Hambleton, R.K., Swamination, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.
Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy and Practice*, *4*(3), 365–379.
Huang, S. (1996). A content-balanced adaptive testing algorithm for computer-based training systems. *Intelligent Tutoring Systems, 3rd International Conference* (pp. 306–314). Springer.
Kay, J. (1990). *UM: A user modelling toolkit*. Paper presented at the Second International User Modelling Workshop, (p. 11), Hawaii.
Kehoe, C.M., & Pitkow, J. (1996). Surveying the territory: GVU's five WWW user surveys. *World Wide Web Journal*, *1*(3), 77–84.
Kingsbury, G., & Houser, R. (1993). Assessing the utility of item response models: Computer adaptive testing. *Educational Measurement: Issues and Practice*, *12*, 21–27.
Kobsa, A. (2007). Generic user modeling systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 136–154). Heidelberg, Germany: Springer Verlag.
Lawson, M.J. (1994). Concept mapping. In T. Husén & T.N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., Vol. 2, pp. 1026–1031). Oxford: Elsevier Science.
Microsoft CAT (1999). *Microsoft unveils innovative testing technology to simulate work environment*. Retrieved December 1, 2008, from: http://www.microsoft.com/presspass/press/1999/jan99/innovativepr.mspx
Mills, C., Potenza, M., Fremer, J., & Ward, W. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.
Pitkow, J., & Recker, M. (1995). Using the web as a survey tool: Results from the second WWW user survey. *Computer Networks ISDN Systems*, *27*(6), 809–822.
Robson, R. (2000). Report on learning technology standards. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)* (pp. 971–976). Chesapeake, VA: AACE.
Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Journal Instructional Science*, *18*(2), 119–144.
Sosnovsky, S. (2004). Adaptive navigation for self-assessment quizzes. *Lecture Notes in Computer Science*, *3137*, 365–371.
South, J., & Monson, D. (2000). A university-wide system for creating, capturing, and delivering learning objects. In D.A. Wiley (Ed.), *The instructional use of learning objects: Online version*. Retrieved December 1, 2008, from: http://reusability.org/read/chapters/south.doc
Thissen, D., & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. (pp. 101–133). Mahwah, NJ: Lawrence Erlbaum Associates.
Topic Maps (2002). *ISO/IEC 13250 Topic Maps* (2nd ed.). Retrieved December 1, 2008, from: http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf
Tzanavari, A., Retalas, P., & Pastelis, P. (2004). Giving adaptation flexibility to authors of adaptive assessments. In *Adaptive Hypermedia and Adaptive WEb-Based Systems Conference* (pp. 340–343). LNCS 3137. Berlin/Heidelberg: Springer.
van der Linden, W.J. & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.
Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed). Mahwah, NJ: Lawrence Erlbaum Associates.
Wainer, H., & Mislevy, R.J. (2000). Item response theory, calibration, and estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.