Online testing, current issues and future trends

Gennaro Costagliola, Vittorio Fuccella

Dipartimento di Matematica e Informatica, Università degli Studi di Salerno

{gencos, vfuccella}@unisa.it

Keywords: online testing, assessment, test, multiple choice, quiz

Abstract

In this paper we analyze the state of art of the assessment of students' knowledge through online tests. In particular, we describe the main functionalities currently implemented in the main existing online testing systems. Furthermore, we present a literary review of the most recently proposed techniques for using online tests in different application ambits, such as: Computerized Adaptive Testing (CAT), automatic question generation, log data analysis, m-learning and educational gaming.

1 Introduction

In recent years, the means for knowledge evaluation have evolved in order to satisfy the necessity to assess a big mass of learners in little time and to better track their learning. For this reason, objective tests have gained more importance in the assessment process.

Several terms allow us to refer, with a more or less specific meaning, to the use of the computer, and in particular of Web-based systems, when administering tests to learners (Wikipedia, 2009). For instance, the term Computer-Assisted Assessment or Computer-Aided Assessment (CAA) is generically referred to the use of the computer in the assessment process. The term Computer-Based Assessment (CBA), instead, is more specifically referred to the automatic evaluation of the responses provided by the students. Online Testing is the administration of structured tests through the Net. Lastly, in order to generically refer to the assessment through ICT, we use the term of E-Assessment.

There are several question types which can be used in online tests, including: multiple choice, true/false, multiple response, matching, ordering, fill in the blanks and so on. The Multiple choice question type is extremely popular, since, among other advantages, a large number of tests based on it can be easily corrected automatically. Furthermore, it is possible to evaluate the quality of multiple choice questions (also called items), in terms of difficulty and discriminative capacity, through statistical models, such as the Item Analysis and Item Response Theory (IRT). Multiple choice items are composed of a stem and a list of options. The stem is the text that states the question. The only correct option is called the key, whilst the incorrect options are called distractors (Woodford & Bancroft, 2005).

Questions types can be classified on the basis of stimulus and response types. The stimulus induces the students to express their knowledge (i.e. the outline of an essay, the stem of an item and so on). It is open when the learner is free to interprete what s/he is asked to do, closed when s/he has some constraints on the performance (length, ordering of the concepts to exhibit, etc.). The response is open when the learner can feel free to elaborate it in a personal way, closed when s/he must choose it among a list of options. Most of the questions included in online tests are characterized by closed stimulus and response. Tests including only these questions are called objective tests and have the advantage of being free from distortional effects, such as emotional judgments and so on. They have, in particular when well formulated, the further advantage of lending themselves well to the verification of knowledge, comprehension and achievement of application objectives. Nevertheless, they do not allow the tutor to verify the expressive capacity and the ability to organize the answers. Furthermore, test construction, especially when using multiple choice questions, can require a long time.

Online tests allow us to assess the students both with formative and summative objectives. In the former case, they are administered during the learning process, give information on the learning state of each learner and, thus, allow the tutor to improve it. In the latter case, instead, they are employed at the end of the learning process (of a learning unit or a temporally bound learning process) and are used for expressing a judgement of the learning state of each learner. For a more comprehensive discussion on the concepts related to objective tests, the reader should refer to the book by Frignani and Bonazza (2003).

Several commercial and Open Source software systems are available for managing and administering online tests. At present, most online testing software modules are part of general purpose Learning Management Systems (LMS). Online testing systems can be evaluated from the point of view of the support of a list of desirable functionalities. In the sequel, we describe these functionalities and verify their support in several LMS. Furthermore, we present a state of art analysis of the most recently proposed techniques for using online tests in different application ambits, such as: Computerized Adaptive Testing (CAT), automatic question generation, log data analysis, m-learning and educational gaming.

2 Online Testing Systems

Online testing systems enable the composition and the administration of online tests. Many of them are integrated in LMSs. Some of them are designed with summative purposes, some others with formative purposes, most of them with both. Formative systems should include the possibility of inserting tutor feedback during the test execution in case of wrong response. The systems designed with summative purposes, instead, should be equipped with tools (in particular for security concerns) for executing proctored laboratory exams.

According to their objectives, such systems should have several desirable features, analyzed in the next sub-sections.

2.1 Question Repository

Most of the online testing systems make use of a question repository in which questions can be inserted and successively selected for composing tests. Question selection can occur by choosing them explicitly when constructing tests or by randomly selecting them among a set of questions satisfying some conditions (difficulty, subject, etc.) at test execution time. The repositories often support the insertion of several question types and some of them support the definition of new types.

In question repository based systems the questions are often organized by subject. A good organization of the repository can help avoiding question replication or the insertion of very similar questions.

Another desirable feature is the possibility of composing questions through advanced editors which enable the use of multimedia and equations.

2.2 Support of Standards

Standards have been introduced in e-learning mainly with the objective of improving interoperability among systems, defined as their capacity of exchanging information. The main specifications introduced in the standardization process define interchange formats. In particular, a common format has been defined for exchanging Learning Objects (LO) and their metadata, in order to launch LO produced with different authoring tools on different LMSs.

More closely related to online testing is the Question and Test Interoperability (QTI) specification, produced by IMS, which boosts the exchange of data related to tests. In particular, it defines a data model for representing tests, questions and the results achieved by the students. The model is based on a large data set. Thus, a specification defining a reduced data set, called QTI Lite, has been introduced. Furthermore, QTI defines an XML data binding and has several extension points, which can be used to define specialized or proprietary extensions to the data model.

Another specification concerning online testing is Computer Managed Instruction (CMI), defining a standard environment in which the LOs can be launched and can exchange data with the LMS. The adoption of this specification is desirable in online testing systems in order to support the tracking of students interactions during the execution of the test.

The adoption of standard functionalities is not always an easy achievement, due to the difficulties in implementing the specification, which are often outdated by newer versions of themselves.

2.3 Assessment, Reports and Item Analysis

According to the question types, the assessment can be automatic, as with multiple choice questions, or require the intervention of the tutor, as in the case of short essays. The most advanced systems allow the tutor to revise and, possibly, modify the marks given by the system. A certain flexibility is desired for establishing a suitable marking strategy: some systems support the definition of rules for calculating the final mark, by assigning different weights to test items and by using penalty and bonus factors for wrong and right responses, respectively.

The systems generally allow the tutor to analyze the results of students, group of students and of the entire class through a dedicated report section. In particular, in this section it is possible to evaluate the improvements achieved through time. More detailed reports enable the evaluation of the gaps of the students across the subjects. The most advanced systems allow the tutor to evaluate question quality by exploiting the already cited IA and IRT statistical models. Several studies, such as the one performed in (Stage, 1999), regard both of them as effective and have identified their pros and cons. Both models are based on the interpretation of statistical indicators calculated on test outcomes. The most important of them are the difficulty indicator and the discrimination indicator, which represents the information of how well an item discriminates between strong and weak students.

2.4 Analisys of Existing Systems

An analysis of several online testing systems in respect to the support of the above features has been carried out: seven products out of the most popular LMSs, accompanied either with an Open Source or a commercial license, have been included in the survey. The analysis is summarized in table 1. The table shows the supported features for each LMS. Each cell in the table reports the supported features for each product and for each feature. To elaborate, the following features have been evaluated in the survey:

- Question Types: number of question types available in the system; support of custom question types;
- Random Items: possibility of randomly selecting questions from repository to compose tests;
- Multimedia: support of multimedia elements in the questions;
- Equations: support of equations in the questions;
- Feedback: possibility for the students of receiving immediate feedback during self-assessment tests;
- Proctored Tests: support of tools for executing laboratory exam;
- Test Analysis: availability of statistics on tests and questions;
- Standards: support of standard functionalities for online testing (QTI and/or CMI)

TABLE 1
Support of main online testing functionalities in LMSs

LMS	Functionalities							
	Types	Multimedia	Random Selection	Feedback	Equations	Lab Test	Test Analysis	Standard
ANGEL LMS V7.2 http://www.angellearning.com	9+custom	√	√	√	√	√	√	QTI; CMI (SCORM)
ATutor 1.5.3.2 http://www.atutor.ca/	6	√	√			√	√	
Blackboard LS EL7 http://www.blackboard.com	10	√	√	√	√	√	√	CMI (SCORM)
Claroline 1.8.1 http://www.claroline.net	4	√	√	√				QTI; CMI (SCORM)
Desire2Learn 8.2 http://www.desire2learn.com/	9+custom	√	√	√	√	√	√	QTI; CMI (AICC, SCORM)
Moodle 1.6.1 http://moodle.org/	6	√	√	√				QTI; CMI (SCORM)
Sakai 2.3 http://sakaiproject.org/	8	√	√	√	√	√	√	QTI

By observing the table, we gather that the support of the described functionalities is rather spread in the surveyed LMSs. We also gather some difficulties in the support of the standard functionalities.

3 Main Research Trends

Some features are at present a topic of research and are rarely present in commercial or popular online testing systems. In particular, we are taking into account the following application ambits, analyzed in the next sub-sections:

- Computerized Adaptive Testing (CAT);
- Automatic question generation;
- Automatic correction to open response questions;
- Log data analysis.

3.1 Computerized Adaptive Testing

CAT is a special case of computer-based testing, in which each examinee takes a unique test that is tailored to his/her ability level. Generally, the ability estimate is updated after each response and the next item is selected such that it has optimal properties according to the new estimate (van der Linden & van Krimpen-Stoop, 2003). The adoption of CAT functionalities has pros and cons. Advantages include (Triantafillou, 2008):

- Possibility for the students of any level to take tests: the students do not need to reach a sufficient knowledge level before starting to take tests, since the test is adapted to their level;
- A greater uniformity in the assessment compared to traditional tests, which have good discriminative capacity only for average levels of knowledge.
- Reduced test size: half the number of questions used in traditional tests can be enough to obtain significant assessment results.

Among the advantages, the following have been reported (Eggen, 2001): the necessity of performing a preliminary item calibration to fix their difficulty and the impossibility for the students to revise the responses. The latter is due to a trick the students can adopt in order to obtain a high score on an easy test (compared to their knowledge level): they intentionally give a wrong response to the questions and revise the response subsequently.

3.2 Generation / Automatic Correction of Questions

Automatic question generation and automatic correction to open response questions are two sectors of the research which have gained the interest of researchers in Natural Language Processing (NLP) techniques.

Automatic question generation can be performed in a completely automatic way, or in a semi-automatic one. The automatic systems generate the items, while the semi-automatic ones assist the user in their generation. In general, the human intervention is anyhow necessary for verifying the good sense of the items before using them in a test.

An example of automatic question generation is the system proposed by Mitkov and Ha (2003), which generates multiple choice items by selecting some sentences from an input text. The text in a declarative form is transformed in a question. Some words inside the input text are removed and used as a key option. The distractors are generated by using concepts semantically close to the key option. In the experiment performed by the authors, 43% of the produced items was discarded after a manual verification of their quality. Then, IA was used to evaluate the difference in the quality between the generated items and others produced without the use of the system. The results were satisfactory, since the generated items showed a greater discriminative capacity (0.40) than those produced manually (0.25).

Among the semi-automatic systems, Hoshino & Nakagawa (2008) propose a process based on NLP aimed at the generation of multiple choice items to be used for the assessment of medicine students' knowledge. The tutor intervenes in the question generation process by choosing the most plausible distractors among those suggested by the system.

As for the automatic correction of essays, we can cite the system proposed by Kakkonen and Sutinen (2004), which uses Latent Semantic Analysis (LSA, a commonly known information retrieval technique) to compare the conceptual similarity between the essays and selected text passages from the course material covering the essay assignment-specific subject matter. Their experiment shows a high correlation between the scores given by the system and a human grader.

3.3 Log Data Analysis

An advantage of online tests versus traditional paper-and-pencil testing is the availability of more information other than the list of the final responses and the time spent on the test: in particular, the complete list of the responses and the time spent on each item are available. Despite its availability, the potential of this information has, however, not been completely exploited. An example of its use is the discovery of the aberrant responses, meaning responses acquired in such ways as cheating and guessing (van der Linden and van Krimpen-Stoop, 2003).

A method to visually analyze these data has been proposed by Costagliola et al. (2009) in order to discover the strategies used by the students to complete the test. The method exploits the graphical representation of the salient events occurring during the execution of a test. The chart (figure 1) represents the test of a single student: the horizontal axis reports the time elapsed from the beginning of the test; the items are reported on the vertical axis. A horizontal segment represents the engagement of the student on an item for a given time interval. A blue circle is for a correct response, while a red circle is for an incorrect response. From the example in the picture, we note that the student has executed the test in two phases: a first one, spanning about 11 minutes, in which s/he has analyzed and answered all of the questions, and a second one in which s/he has revised the responses, and modified two of them (item 4 and 8).

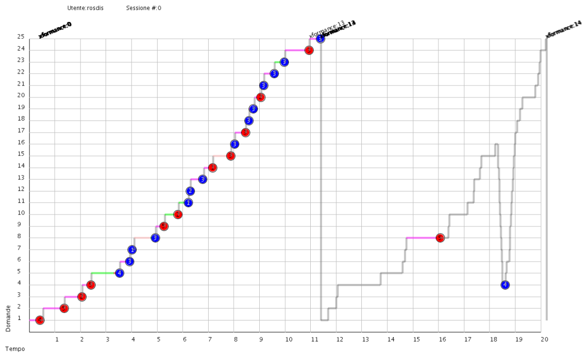
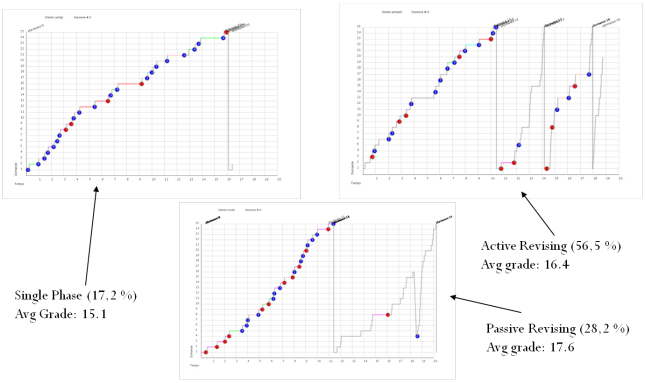


Fig. 1 Log data visualization.

From the inspection of the charts obtained through an experiment within a university course exam, the following most exploited strategies have been discovered (see fig. 2), with a few exceptions:



Single Phase (17,2 %)
Avg Grade: 15.1

Active Revising (56,5 %)
Avg grade: 16.4

Passive Revising (28,2 %)
Avg grade: 17.6

Fig. 2 Strategies exploited by the students to complete the test.

- Single Phase. This strategy is composed of just one phase. The time available to complete the test is organized by the learner in order to browse all of the questions just once;
- Passive Revising. This strategy is composed of two or more phases. During the first one, the student views and answers all the questions in a time shorter than the one available. The remaining time is used for one or more revising phases, in which some responses are changed, if necessary.

- Active Revising. Similar to the previous. The difference lies in that in the first phase the student intentionally skips some items, postponing the response to the subsequent phases.

By analyzing the data of the experiment, it came out that the most frequently adopted strategy is Active Revising, which was used by 40 learners out of 71 (56.5%), followed by the Passive Revising strategy (20 learners out of 71, 28.2%) and by the Single Phase one, used only in 9 cases out of 71 (12.7%). Only two learners adopted an atypical strategy, which cannot be classified in any of the previously described patterns.

The best results have been achieved by learners who adopted the Passive Revising strategy, with an average score of 17.6 exact responses out of the 25 test questions. With Active Revising, on the other hand, an average score of 16.4 has been achieved. Lastly, the Single Phase strategy turned out to be the worst one, showing an average score of 15.1. Therefore, it appears that a winning strategy is one of using more than one phase, and this is confirmed by the slightly positive linear correlation (0.14) observed between the number of phases and the score achieved on the test.

3.4 Educational Gaming and M-Learning

The attention of the researchers on the use of computer games in education is growing. Nevertheless, there are few examples of the use of games to evaluate learners. One of them is the work of Ramani et al. (2008). They describe a system to administer online tests “clothed” as computer games. In particular, the test is administered to the students through the metaphor of a cricket match. Following the principle that users engage not only in playing, but also in designing the games (in fact video game companies encourage ‘modders’, those users who modify the games), they also allow students to create questions to be answered by the opponent team. Unfortunately, no comparison with traditional online testing system has been performed in terms of students’ levels of engagement and learning.

As for m-learning (mobile learning), we wonder which would be the usability level of the online testing systems running on PDA. Segall et al. (2004) compared the usability effectiveness, efficiency, and satisfaction of a PDA-based quiz application to that of standard paper-and-pencil quizzes in a university course, finding a greater efficiency in the PDA-based quiz, that is, students completed it in less time than they needed to complete the paper-and-pencil quiz.

Conclusions

In this paper we have analyzed online testing systems, describing both the currently implemented features and the research features. In particular, we have verified the support of the former in the most popular LMSs and we have presented an analysis of the latter in different application ambits which recently have attracted the interest of the researchers.

BIBLIOGRAPHY

Ims Question & Test Interoperability specication. http://www .imsglobal.org/question/, Verificato il 06/10/2009.

Costagliola G., Fuccella V., Giordano M., Polese G. (2009), *Monitoring online tests through data visualization*. IEEE Trans. on Knowl. and Data Eng., 21(6):773-784.

Eggen T. (2001), *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports, 1.

Frignani P., Bonazza V. (2003), *Le prove oggettive di profitto*. Strumenti docimologici per l'insegnante, Roma, Carocci, 2003 .

Hoshino A., Nakagawa H. (2007), *Advances in web based learning*, In: ICWL 2007, 6th international conference, Edinburgh, uk, august 15-17, 2007, revised papers. In H. Leung, F. W. B. Li, R. W. H. Lau, and Q. Li, editors, ICWL, volume 4823 of Lecture Notes in Computer Science. Springer, 2008 .

Kakkonen T., Sutinen E. (2004), *Automatic Assessment of the Content of Essays Based on Course Materials*. In Proceedings of the International Conference on Information Technology: Research and Education 2004 (ITRE 2004), pages 126-130, London, UK.

Mitkov R., Ha L. A. (2003), *Computer-aided generation of multiple-choice tests*. In Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing, pages 17-22, Morristown, NJ, USA, 2003.

Ramani S., Sirigiri V., Panigrahi N. L., Sabharwal S. (2008), *Games as skins for online tests*. In DIGITEL '08: Proceedings of the 2008 Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, pages 90-92, Washington, DC, USA, 2008. IEEE Computer Society.

Segall N., Doolen T. L., Porter J. D. (2004), *A Usability Comparison of PDA-Based Quizzes and Paper-and-Pencil Quizzes*, volume 45. Elsevier Science Ltd., Oxford, UK, UK, 2004.

Triantafillou E., Georgiadou E., Economides A. A. (2008), *The design and evaluation of a computerized adaptive test on mobile devices*. Comput. Educ ., 50(4):1319-1330, 2008.

Van der Linden W. J., Glas C. A. W. (2003), *Computerised adaptive testing: theory and practice*. Kluwer Academic Publishers, Norwell, MA, USA .

Wikipedia. *E-Assessment*. http://en.wikipedia.org/wiki/E-assessment. Verificato il 06/10/2009