Methods and Tools for On-Line Objective Testing

Gennaro Costagliola, Filomena Ferrucci, Vittorio Fuccella*

Prof. Gennaro Costagliola

Dipartimento di Matematica e Informatica
Università degli Studi di Salerno
Via Ponte Don Melillo
84084 Fisciano, Salerno, Italy
phone: +39 089 963319
fax: +39 089 963303
e-mail: gencos@unisa.it

Prof. Filomena Ferrucci

Dipartimento di Matematica e Informatica
Università degli Studi di Salerno
Via Ponte Don Melillo
84084 Fisciano, Salerno, Italy
phone: +39 089 963374
fax: +39 089 963303
e-mail: fferrucci@unisa.it

Dr. Vittorio Fuccella*

Dipartimento di Matematica e Informatica
Università degli Studi di Salerno
Via Ponte Don Melillo
84084 Fisciano, Salerno, Italy
phone: +39 089 963390
fax: +39 089 796438
e-mail: vfuccella@unisa.it

Methods and Tools for On-Line Objective Testing

Keywords: On-Line Testing, Computer Assisted Assessment, Assessment, Learning Management System, LMS, Learning Object, LO, Standardization, Interoperability, Multiple Choice, Item Analysis, IA, Distractor.

INTRODUCTION: ON-LINE TESTING

On-Line Testing, also known as *Computer Assisted Assessment* (*CAA*), is a sector of e-learning aimed at assessing learner's knowledge through e-learning means. In recent years, the means for knowledge evaluation have evolved in order to satisfy the necessity of evaluating a big mass of learners in strict times: *objective tests*, more rapidly assessable, have gained a heavier weight in the determination of learners' results.

Multiple Choice question type is extremely popular in *objective tests*, since, among other advantages, a large number of tests based on it can be easily corrected automatically. These items are composed of a *stem* and a list of *options*. The *stem* is the text that states the question. The only correct answer is called the *key*, whilst the incorrect answers are called *distractors* (Woodford & Bancroft, 2005).

Several commercial and Open Source software systems are available for managing and administering on-line tests. At present, most *on-line testing* systems are part of a more general purpose e-learning system, often called *Learning Management System* (*LMS*) or *Course Management System* (*CMS*). These products offer a complete set of functionalities for e-learning, both for on-line learning and for *blended learning*. They are primarily used for administering on-line learning material, commonly referred to as *Learning Objects* (*LOs*). *On-line testing* systems can be evaluated from the support of a list of desirable features, analyzed in the sequel.

In *on-line testing* it is important to administer tests composed of good quality questions (*items*). By the term "quality" we intend the potential of an *item* in effectively discriminating between strong and weak learners and in obtaining tutor's desired difficulty level. There are statistical models which can help tutors in understanding whether their *multiple choice items* have good performances or not. Statistics can be displayed or used in *on-line testing* systems for determining question quality.

Another important aspect of e-learning, which has also been applied to *on-line testing*, is the standardization of e-learning systems. Standardization efforts in e-learning are mainly aimed at achieving *interoperability* among *LMS* and *LO* authoring tools. For *on-line testing* it can be important to share test data and to track learners' interaction during test execution. This is valuable information for understanding the learner's behavior when taking a test: in the past, several experiments have been carried out to this extent.

The rest of this article is organized as follows: the next section describes some basic principles of assessment, gives some basic definitions and introduces *objective tests*; the subsequent section focuses on item quality and its management in *on-line testing* systems; then, another section, called "*On-line Testing* and Standardization", is devoted to describe the standardization process and its application to *on-line testing*; the subsequent section describes *on-line testing* systems, presenting a parade of the most desirable features for these systems and a survey on their support in some of the most popular *LMSs*; before concluding, a description of some experimental features treated in research literature is presented.

ASSESSMENT AND OBJECTIVE TESTS

In the last years *objective tests*, as an integrating part of the learning process, have aroused a growing interest in educators. Nevertheless, *objective tests* are often designed with superficiality, ignoring the indications that *docimology* has achieved through research activity and experiments.

Assessment can be *formative* and *summative* (Frignani and Bonazza, 2003). *Formative* assessment occurs during the learning process, gives information on the learning state of each learner and allows the tutor to decide the most suitable learning path for her/him. *Summative* assessment, instead, occurs at the end of the learning process (of a learning unit or a temporally bound learning process) and is used for the evaluation of the learning state of each learner.

In many traditional learning contexts assessment is regarded as a single process, without distinguishing between the verification and the grading phases. Research, instead, considers assessment as composed of two different phases: the *measuring* and the *grading* ones. The *measuring* phase consists of gathering information on the learners; the *grading* phase lies in expressing a judgement on the information obtained in the previous phase.

Grading can be performed by using one of following criteria: in *absolute grading* the passing threshold is established a priori; in *relative grading* the passing threshold is established only after the results of all the learners have been recorded.

Tests can be classified on the basis of *stimulus* and *answer* types. The *stimulus* is the part of the measuring phase which induces the learners to express their knowledge (i.e. the outline of an essay, the *stem* of an item and so on). *Stimulus* is *open* when the learner is free of interpreting what s/he is asked to do, *closed* when s/he has some constraints on the performance (length, ordering of the concepts to exhibit). The *answer* is *open* when the learner can feel free to elaborate the answer in a personal way, *closed* when s/he must choose the answer among a list of options. The classification of some test types is summarized in table 1 on the basis of the possible values for *stimulus* and *answer*. *Objective tests* have *closed stimulus* and *answer*.

Table 1: Test Classification on the Basis of Stimulus and Answer

		Answer	
		Open	**Closed**
Stimulus	**Open**	Oral examination, essays	Typical case: while the tutor is teaching, asks a learner for an approval sign (it can be useful for monitoring learners' attention during lessons)
	Closed	Short answer, summary	Objective tests, grammatical exercises, mathematical expressions

Objective tests are considered *objective* since it is possible to establish the grade to give to right and wrong answers of each question a priori (at the time the test is constructed); the correction of the test, and, consequently, its grading, is independent from the corrector/grader; lastly, the learners are all in the same condition, since they are all required to perform the same task, in the same time interval, in the same environmental conditions. Objectivity makes tests free from several distortional effects, such as emotional judgements and so on.

Nevertheless, *objective tests* have some disadvantages: i.e., 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.

I T E M Q U A L I T Y

The experience gained by educators and the results obtained from several experiments provide some guidelines for writing good *multiple choice* items, such as: "use the right language", "avoid a big number of unlikely *distractors* for an item", etc. Furthermore, it is possible to evaluate the effectiveness of the items, through the use of several statistical models, such as *Item Analysis* (*IA*) and *Item Response theory* (*IRT*). Several studies, such as the one performed in (Stage, 1999), make a comparison between the two models, often concluding that they can both be effective in evaluating the quality of the items. They are both based on the interpretation of statistical indicators calculated on test outcomes. The most important of them are the *difficulty* indicator, which measures the difficulty of the items, and the *discrimination* indicator, which represents the information of how well an item discriminates between strong and weak students. More statistical indicators are related to the *distractors* of an item.

IA is still preferred over *IRT* in *on-line testing* systems since it needs a smaller sample size for obtaining statistically significant indicators and its

results are easier to read for humans. As mentioned above, its main indicators are *difficulty* and *discrimination*. The former is calculated as the proportion of learners who get the item incorrect; the latter is calculated as the *point biserial* correlation coefficient between the score obtained on the item and the total score obtained on the test. A good value for *discrimination* is greater than 0.5. A positive value lower than 0.2 indicates an item which does not discriminate well. This can be due to several reasons, including: the question does not assess learners on the desired knowledge; the *stem* or the *options* are badly/ambiguously expressed; etc. If *difficulty* is too high (>0.85) or too low (<0.15), there is the risk of not correctly evaluating on the desired knowledge. This is particularly true when such values for *difficulty* are sought together with medium-low values for *discrimination*.

Several *on-line testing* systems provide the tutors with a feedback on the quality of their questions by using *IA* or *IRT*.

ON-LINE TESTING AND STANDARDIZATION

In recent years, great efforts have been made to define standards, reference models, and guidelines for e-learning. These efforts are mainly aimed at obtaining a stronger interoperability among e-learning systems. In the context of these systems, the term "interoperability" refers to the possibility of running *LOs* produced with any authoring tool on any *LMS* compliant to the standard specifications. At present, the main specifications are focused on the proposal of common formats for *LO* metadata and for resource interchange. The main standard and guidelines producers are *AICC*, *IEEE LTSC*, *ADL SCORM* and *IMS*. Another important set of specifications defines a standard environment in which the *LOs* can be launched and can exchange data with the *LMS*. This model is currently proposed in several specification documents, such as *AICC CMI Guidelines for Interoperability* (2004), *SCORM Run Time Environment* (RTE, 2004) and *IEEE CMI* (2002). We will refer to the functionalities proposed in these documents using the acronym *CMI*, abbreviation of *Computer Managed Instruction*.

The most important standard specifications concerning *on-line testing* are *Question and Test Interoperability*, produced by *IMS*, and the already mentioned *CMI*. The former describes a data model for representing questions and test outlines and outcomes. The specification boosts the data exchange between *LMSs* and authoring tools. Furthermore, it defines an XML data binding. Lastly, it has several extension points, which can be used to define specialized or proprietary extensions to the data model. There is also a specification defining a reduced data set, called *QTI Lite*.

CMI functionalities are crucial for obtaining full interoperability, since a *LO* has the necessity of exchanging data with the *LMS* in all the different phases of its execution (*launch*, *suspension*, *resumption*, *dismissal*). The specifications define some standard rules with which the *LO* must be launched and perform the communication with the *LMS* and, lastly, the data

set on which *LMS* and *LO* must agree and on which their communication must be based. Several data models have been proposed so far. These sets of data often include, but are not limited to, information about the learner, interactions that the learner has had with the *LO*, *objectives*, *success status* and *completion status* of the *LO*. Particularly important in *on-line testing* is the part of *CMI* regarding the interaction tracking. Part of the data model proposed by all the document issuers is devoted to the learner interaction tracking during the execution of tests. The data model generally includes the following elements:

- Timestamp of the response given to an item
- The weight of the item in the test
- The right response and the one given by the learner
- The evaluation of the response
- Time needed to give a response

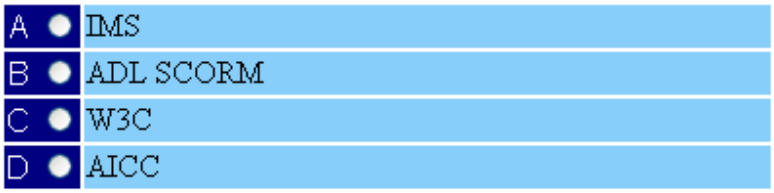
The support of standard functionalities in *on-line testing* systems is discussed in the next section.

ON-LINE TESTING SYSTEMS

With *on-line testing* systems it is possible to construct and administer on-line *objective tests*. Most of the systems are integrated in Web-based *LMSs*. Some of them are designed with *summative* purposes, some others with *formative* purposes, most of them with both. According to their objectives, such systems should have several desirable features. In any case, the support of a large set of question types is very important. A list including the most common question types, with an explanation of their meaning, an example of their use and their possible implementation in Web-based interfaces is presented in the following tables. The examples have been produced with the *QuizFaber* v2.10 system.

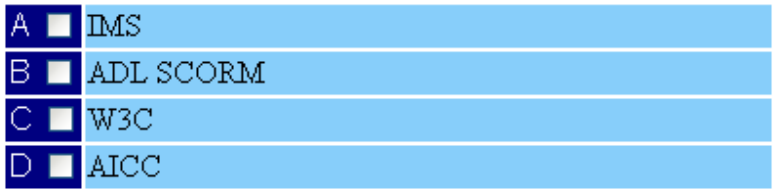
Table 2: The most common question types.

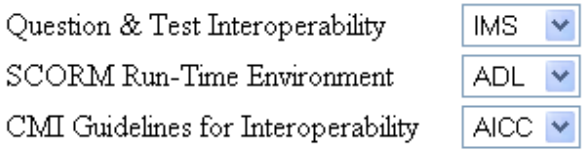
Question Type	**Multiple Choice (traditional)**
Description	A traditional multiple choice item is one in which a student chooses one answer from a number of options supplied.
On-line testing Implementation	It can be implemented through a simple HTML form. **Interaction Type**: Each option has an interactive form element whose type is *radio button*. **Response**: The response can be represented through a number, that is the index of the chosen option.

Example	**Which of the following organizations does not produce e-learning standards?** 

Question Type	**True/False**
Description	A variation of a *multiple choice* question with two only options whose text is “true” and “false”, respectively.
On-line testing Implementation	It can be implemented through a simple HTML form. **Interaction Type**: Each option has an interactive form element which is a *radio button*. **Response Type**: The response can be represented through a Boolean value (or an integer number), that is the index of the chosen option.
Example	**Does W3C produce e-learning standards?** 

Question Type	**Multiple Response**
Description	A variation of *multiple choice* in which more than one option can be selected as correct by the student.
On-line testing Implementation	It can be implemented through a simple HTML form. **Interaction Type**: Each option has an interactive form element whose type is *checkbox*. **Response Type**: The response can be represented through a vector of integer numbers, that is the index of the checked options. Alternatively, a vector of Boolean values can be used. The i-th value of the vector is true if the i-th option has been checked.

Example	**Which of the following organizations does produce e-learning standards?** 

Question Type	**Matching**
Description	The question presents two lists of phrases (or images) on two columns placed side by side. The learner must match each phrase in the left column with one in the right column.
On-line testing Implementation	It can be implemented through a simple HTML form. The left column is composed of text items, the right one of *list* elements. **Interaction Type**: Each item in the right column must be chosen inside a *list* type form element. **Response Type**: The response can be represented by a vector of integer numbers. The i-th value of the vector represents the index of the chosen option in the i-th list.
Example	**Match the specification with the standard producer!** 

Question Type	**Fill in the Blanks**
Description	The Fill in the Blanks question requires a learner to complete a blank or more blanks within a brief piece of text, using words, symbols or numbers. The values are usually chosen from a number of choices supplied.
On-line testing Implementation	It can be implemented through a simple HTML form. **Interaction Type**: Each blank has an interactive form element whose type is *list* or *text*. **Response Type**: The response can be represented by a vector of integer numbers, that is the index of the chosen options. If a *text* element is used for the blanks, the response can be represented by a vector of strings.

Example	**Fill in the blanks** IMS produces QTI specification while AICC produces CMI specification

Formative systems should include the possibility of inserting tutor feedback during the test execution and, eventually, of including links to the lecture material for deepening. The systems designed with *summative* purposes, instead, should be equipped with tools for executing laboratory exams: it should be possible to define security settings, such as: time intervals to access the test, the maximum time limit to give the test, the maximum number of attempts on the test. Furthermore, it should be possible to limit test access by username and password or IP address and disable some browser functionalities.

Most of the assessment systems make use of a pre-populated question repository from which questions can be chosen explicitly when constructing tests or randomly at test execution time. In question repository based systems, the challenge is to give a good organization to the repository, to avoid question replication, and to use a good question selection procedure in order to assess learners' skills on the desired subjects.

The assessment is generally completely automatic (except for those question types, such as *essay* homework, which require the judgement of a human). The most advanced systems allow the tutor to revise and, eventually, modify the marks given by the system. A certain flexibility is desired for establishing the *marking strategy*: some systems allow to define rules for calculating the final mark, by giving different weights to the items and using *penalty* and *bonus factors* for wrong and right responses, respectively.

A list of other desirable features include:

1. Support for multimedia and equations;
2. Support for standard specifications, such as *QTI* and *CMI*;
3. Reports on the knowledge improvement of the class and on the quality of the items (in the case of multiple choice items).

An analysis on the support of the above features has been carried out: ten 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 3. 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. The following features have been evaluated in the survey:

1. **Question Types**: number of question types available in the system; possibility to define custom question types;
2. **Random Items**: possibility of randomly selecting questions from question repository to compose tests;
3. **Multimedia**: possibility of inserting multimedia elements in the questions;

4. **Feedback**: possibility of giving immediate tutor feedback during self-assessment tests;
5. **Equations**: possibility of inserting equations in the questions;
6. **Proctored Tests**: presence of tools for executing laboratory exam;
7. **Test Analysis**: availability of statistics on tests and questions;
8. **Support of Standards**: support of standard specifications for *on-line testing* (*QTI* and/or *CMI*)

Table 3: On-Line Testing Features Support in LMSs

LMS	Features							
	Question Types	Multimedia	Random Items	Feedback	Equations	Proctored Test	Test Analysis	Support of Standards
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	√	√	√				CMI (SCORM)
Sakai 2.3 http://sakaiproject.org/	8	√	√	√	√	√	√	QTI
QuizFaber v2.10 http://www.lucagalli.net/	6	√		√	√	√		

MAIN RESEARCH TOPICS

Some features are at present topic of research and are rarely present in commercial or popular *on-line testing* systems. The main research topics are the following:

- Automatic question generation or question sequence generation;
- Automatic correction to *open response* questions;
- Item quality management and improvement;
- Interpretation of learner behaviour during tests.

Automatic question generation is rather a challenging issue. Many experimental systems employ *Natural Language Processing* (*NLP*) techniques. An example is the system proposed by Mitkov & Ha (2003), which generates *fill in the blanks* questions by isolating sentences from an

input text and by removing some words from them. Hoshino & Nakagawa (2007) propose a system for learning English grammar and vocabulary which semi-automatically makes questions on a given input text. Other systems, instead of dynamically generate questions, try to change question sequence according to learners' responses: the *ASSISTment* system (Feng et al., 2006), tries to provide instructional assistance in the process of assessing learners by providing hints and further question (called *scaffolding questions*) on a topic if the learner gets the *original item* wrong.

Another challenging research topic in *on-line testing* is the automatic correction of essays. Kakkonen and Sutinen (2004) use *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 human grader.

As for item quality management, most *on-line testing* systems generate and show item statistics but they do not interpret them, so they do not advise or help the tutor in improving item quality. The *eWorkbook* system (Costagliola et al., 2007a) detects defective question items and, when possible, provides the tutors with advice to improve their quality. The system detects defective items by firing rules based on *IA* indicators. This work is an improvement of a previous work, presented in (Hung et al., 2004).

Learners' behaviour during tests has been analyzed in several experiments (Bath, 1967; McClain, 2003) in order to obtain information on the strategies used by the learners to complete tests. These experiments have been carried out in traditional 'papery' testing by exploiting the "think out loud" method. Costagliola et al. (2007b) have analyzed the behavior of the learners in on-line tests by exploiting Information Visualization techniques. Learners' interactions with the *on-line testing* system interface are recorded and then graphically displayed in an ad hoc defined chart.

C O N C L U S I O N

In this article several methods and tools for *on-line testing* have been described. The article focuses above all on the *on-line testing* systems for administering *objective* tests. The main topics treated have been assessment and *objective tests*, item quality, standardization and *on-line testing* systems features. Lastly, a survey on the most important research topics in *on-line testing* has been presented.

R E F E R E N C E S

AICC (2007), Aviation Industry CBT Committee, http://www.aicc.org/

AICC CMI (2004), CMI Guidelines for Interoperability AICC rev. 4.0, http://www.aicc.org/docs/tech/cmi001v4.pdf

Bath, J.A. (1967) Answer-changing Behaviour on objective examinations. *The Journal of Educational Research*, 61, 105-107.

Costagliola, G. Fuccella, V. Giordano, M. and Polese, G. (2007a). A Web-Based E-Testing System Supporting Test Quality Improvement, In *Proceedings of The 6th International Conference on Web-based Learning*, 272-279.

Costagliola, G. Ferrucci, F. Fuccella, V. (2007b). A Web-Based E-Testing System Supporting Test Quality Improvement, In *Proceedings of The 6th International Conference on Web-based Learning*, 190-197.

Feng, M. Heffernan, N. T. and Koedinger, K. R. (2006). Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses. *Proceedings of the 15th International Conference on World Wide Web*. ACM Press, 307-316.

Frignani, P. and Bonazza, V. (2003). *Le Prove Oggettive di Profitto. Strumenti Docimologici per l'Insegnante*. Carocci.

Hoshino, A. and Nakagawa, H. (2007). A Cloze Test Authoring System and its Automation. In *Proceedings of 6th International Conference on Web-Based Learning*, 174-181.

Hung, J.C. Lin, L.J. Chang, W.C. Shih, T.K. Hsu, H.H. Chang, H.B. Chang, H.P. Huang, K.H. (2004). A Cognition Assessment Authoring System for E-Learning. In *Proceedings of 24th International Conference on Distributed Computing Systems Workshops*, 262-267

IEEE LTSC (2007), IEEE Learning Technology Standards Committee, http://ieeeltsc.org/

IEEE LTSC WG11 (2002), IEEE LTSC, WG11, Computing Managed Instruction, http://ltsc.ieee.org/wg11/index.html

IMS (2007), IMS Global Learning Consortium, http://www.imsglobal.org/

Kakkonen, T. and Sutinen, E. (2004). Automatic Assessment of the Content of Essays Based on Course Materials. In *Proceedings of The 2nd IEEE International Conference on Information Technology: Research and Education* 2004, 126-130.

McClain, L. (2003). Behavior during examinations: A comparison of "A", "C" and "F" students. *Teaching of Psychology*, 10 (2), 69-71

Mitkov, R. and Ha L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, 17-22.

SCORM. Advanced Distributed Learning – SCORM, http://www.adlnet.gov/scorm/index.cfm

SCORM RTE. The Scorm Run-Time Environment ver 1.3.1, http://www.adlnet.org/scorm/history/2004/documents.cfm

Stage, C. (1999). *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest READ*, http:// www. umu. se/ edmeas/ publikationer/ pdf/ enr3098sec.pdf

T E R M S A N D D E F I N I T I O N S

On-Line Testing: sector of e-learning aimed at assessing learner's knowledge through e-learning means.

Objective Tests: tests composed of questions which have some constraints on the performance (closed stimulus) and whose answer must be chosen by the learner among a list of options (closed answer).

Multiple Choice Item: question in which a student chooses one answer from a number of choices supplied.

Distractors: the incorrect answers in the list of options of a multiple choice item.

Item Discrimination: statistical indicator from *Item Analysis* model which expresses the information of how well a multiple choice item discriminates between strong and weak students.

Interoperability (among software systems): the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units. (ISO/IEC 2382-01).

Learning Management System (LMS): the software platform for delivering, tracking and managing training. The main features of an LMS include: course management, learners enrollment, on-line activity tracking, etc.