A library for items construction in an adaptive evaluation system

Eduardo Guzmán, Juan Andrés Riveros & Ricardo Conejo

Departamento de Lenguajes y Ciencias de la Computación.
E.T.S.I. Informática, Universidad de Málaga. Apdo. 4114, Málaga 29080. SPAIN
e-mail: {guzman,conejo}@lcc.uma.es

Abstract. Traditional computer tests usually offer items (questions) where examinees can only select one or more answers from a set of possible answers. On the other hand, in classical test-based evaluation systems, there is a fixed number of items to be posed to students. This implies that all students must answer to the same number of questions, independently of their knowledge level. By contrast, adaptive testing systems are able to make more accurate predictions of student's knowledge level with shorter tests, by choosing the most adequate item to ask next, depending on the current estimation of student's knowledge level. In this paper, a library of templates for the construction of more sophisticated items is presented. These type of items are used in SIETTE, an adaptive web-based system for evaluation of knowledge by means of tests.

Keywords: assessment, cognitive diagnosis, psychometrics, Computerized Adaptive Tests, Item Response Theory.

1 Introduction

Evaluation is an important part in the process of education. Teachers need to measure the knowledge acquired by the students. Tests are one of the most extended mechanisms for evaluation. They have been widely used in intelligent tutoring systems and generally in computer educational systems.

There are great deal of commercial test-based tools [1, 2, 3, 4]. These systems offer powerful interfaces to teachers and students. One of the main lacks of these systems is that the evaluation process is not well-founded. In most cases they only present to examinees a set of fixed questions, and after, they give the number of items that have been correctly answered, without regarding the different difficulty of the questions posed.

A *Computerized Adaptive Test* (CAT) [5] is a test where the selection of each item and the finalization of the test are dynamically adopted, based on the current estimation of the student's knowledge level.

SIETTE is an efficient adaptive web-based system for evaluation by using tests. The mechanisms used to carry out the selection of the most suitable question (*item*) to pose and the test finalization criterion are based on a psychometric theory called *Item Response Theory* (IRT).

In this paper we present a library for the automatic construction of items by means of templates that will be integrated into the SIETTE system. These templates intend to be a complete collection of all generic exercises that teachers can pose to students. The use of this library provides the capability of posing exercises and tests questions in the same evaluation session.

In the next section we present the main characteristics of SIETTE system. Later, the different types of questions offered by SIETTE will be discussed. In section 4, the components of the library will be explained in detail. Finally, the main contributions of our work will be summarized.

2 The SIETTE system

SIETTE (this stands for *System of Intelligent Evaluation using Tests for Teleeducation* in Spanish) [6] has been designed to be used through WWW. By means of a navigation application, teachers can create tests and examinees can evaluate their knowledge in certain subject.

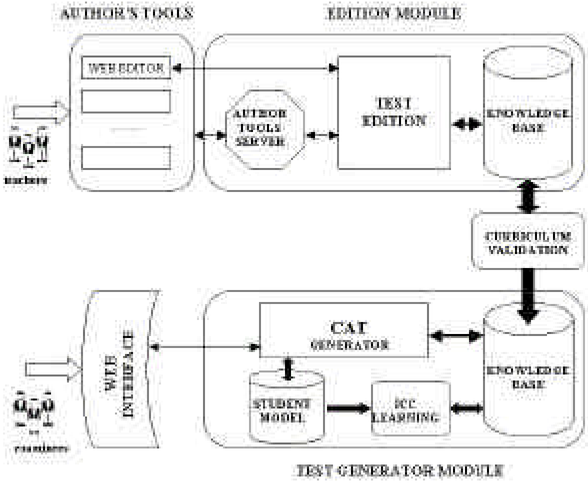


Fig. 1. The architecture of SIETTE

The architecture of SIETTE is depicted in Fig. 1. It is mainly organized in six modules:

? *Knowledge base,* composed by the domain (*curriculum*), tests specifications and the item pools.
? *CAT generator*, that selects the most suitable item to ask next.
? *Author's editors,* that allow teachers to insert and modify contents of the knowledge base.
? *The student model,* that collects all the information about the student.
? *Curriculum validation module,* that checks the validity and consistency of tests specifications.
? *ICC Learning module*, that is used to learn (calibrate) the parameters using the information obtained from the student´s that take the test [7].

2.1 Student's knowledge level estimation

IRT [8], is based on the hypothesis that the answer given to each item of the test depends probabilistically on certain *latent trait (?)* that can be measured by means of an unknown fixed numerical value. This theory has been successfully applied to student's knowledge level estimation in adaptive tests. In educational environments the *latent trait* is the *knowledge level* of the student. In this theory, conditional probabilities of the successful answer to the item by a student with a certain *knowledge level* must be previously known for each item. This probability is expressed by means of a function $f : ?? ? , ? ? ?? \; ?0,1?$ that is called *Item Characteristic Curve* (ICC). The calculus of the ICC can be accomplished by several models. SIETTE uses a model of three parameters based on the logistic function [9]:

$$P_i(?) ? P?\mathrm{u_i} ? 1|??? \mathrm{c_i} ? (1 ? \mathrm{c_i})\frac{1}{1 ? \mathrm{e}^{?1.7\mathrm{a_i}(? ?\mathrm{b_i})}} \tag{1}$$

The three parameters of this ICC model are:

? *Difficulty* (b_i): It corresponds to the knowledge level in which the probability of success is the equal to the probability of failure.
? *Discrimination factor* (a_i): It is proportional to the slope of the curve. A high value indicates a high probability of success for students with knowledge higher than the difficulty of the item.
? *Guessing factor* (c_i): It is the probability that a student with no knowledge gives a correct answer to the item, and can be computed as the proportion of right answers over the total number of answers.

The value of *?* is estimated using the response to each item. There are several methods to get this value. In SIETTE a Bayesian method [10] is used. In this method, the probability distribution of the student's knowledge level is calculated applying Bayes' rule. It is also assumed that the latent trait *?* can only take *K* discrete values (from *0* to *K-1*) in order to decrease the computational cost of the calculus. Therefore, ICCs are given by vectors of *K* components $P_i(u|?) = (p_i(u|?=0),\ p_i(u|?=1), \ldots\ p_i(u|?=K\text{-}1))$. Thanks to this consideration, the Bayesian method can be simplified to a vectorial product of ICC vectors by the *a priori* normalized density vector:

$$\overline{\mathrm{P}(?\,|\mathrm{u})}\ ?\ \left\| \overset{n}{\underset{i?1}{?}}\ \overline{\mathrm{P}_{\mathrm{i}}(?)}^{\mathrm{u_i}}(\bar{1}\ ?\ \overline{P_i(?)})^{(1?u_i)}\,\overline{\mathrm{P}(?)} \right\| \qquad \textbf{(2)}$$

2.1 Item selection algorithms and termination criterion

An adaptive test is an iterative algorithm that begins with an initial estimation of the knowledge level of the examinee and has the following steps:

1. All the questions in the knowledge base (that have not been administered yet) are examined to determine which is the best item to ask next according to the current estimation of the examinee's knowledge level.
2. The question is asked, and the examinee answers.
3. According to the response, a new estimation of the knowledge level is computed.
4. Steps 1 to 3 are repeated until the stopping criterion defined is met.

In SIETTE, teachers must indicate the item selection criterion of tests. There are three alternatives:

? *Bayesian criterion.* Starting from the distribution of the estimated knowledge level of the student, the selected item is that one which minimizes the sum of the *a posteriori* variances resulting from a correct/incorrect answer to the item.

? *Difficulty-based criterion.* This method selects the items whose difficulty is closer to the estimated knowledge level of the student.

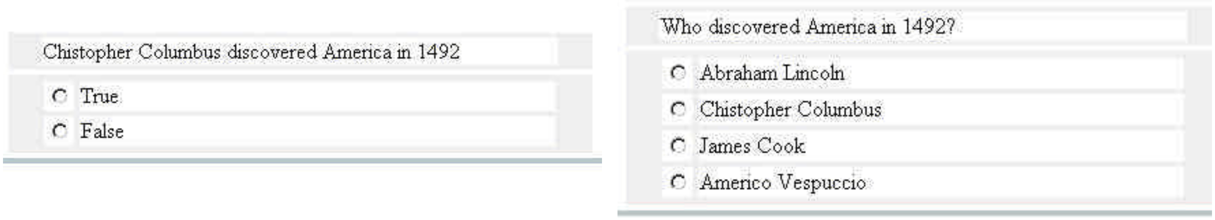
? *Random criterion.* This method randomly selects an item. Certainly, this is not an adaptive criterion.

The finalization criterion is also configured in each test. First, in the edition phase, the teacher must indicate a minimum and a maximum number of items. When this maximum number is reached, the current estimation of student's knowledge level becomes the final estimation. Additionally SIETTE offers the following adaptive finalization criteria: a) the most likely value of the estimated knowledge distribution is bigger than a certain threshold; or b) the variance of the estimated knowledge distribution is lower than a certain value.

3 Types of items in SIETTE

SIETTE allows teachers to propose different types of items. All these items can be combined into the same test. The types of items are:

✍ *Dichotomous items*: This is the most simple item. It can only have two answers: true or false.



Chistopher Columbus discovered America in 1492

○ True

○ False

(a)

Who discovered America in 1492?

○ Abraham Lincoln

○ Chistopher Columbus

○ James Cook

○ Americo Vespuccio

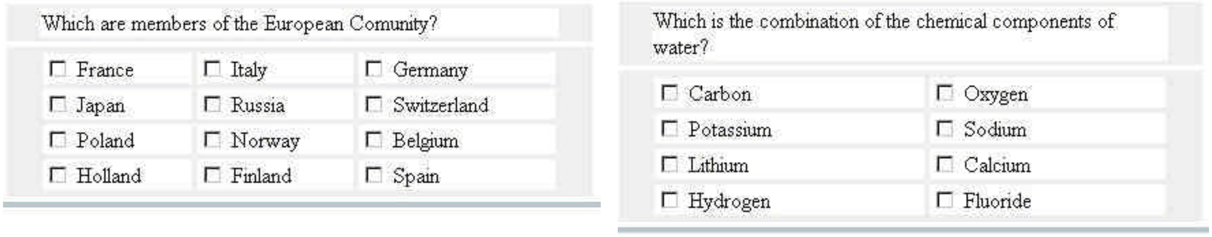
(b)

Fig. 2. (a) A dichotomous item. (b) A multiple-choice item

✍ *Multiple-choice items*: These kind of items present several alternatives, and the correct one has to be chosen.

✍ *Polytomous items*: These items are multiple-choice items that have more than one correct answer. Examinees must select all correct answers in order to pass the question. This kind of items can be also classified into:

a) *Items with independent answer.* Answers are mutually independent. This type of item is equivalent to a set of dichotomous items. For instance, the item shown in Fig. 3 (a) is equivalent to the following dichotomous items: *Is France member of the European Community?, Is Italy member of the European Community?,* and so on.



Which are members of the European Comunity?

□ France	□ Italy	□ Germany
□ Japan	□ Russia	□ Switzerland
□ Poland	□ Norway	□ Belgium
□ Holland	□ Finland	□ Spain

Which is the combination of the chemical components of water?

□ Carbon	□ Oxygen
□ Potassium	□ Sodium
□ Lithium	□ Calcium
□ Hydrogen	□ Fluoride

(a) *(b)*

Fig. 3. Polytomous items: (a) with independent answer, (b) with dependent answer

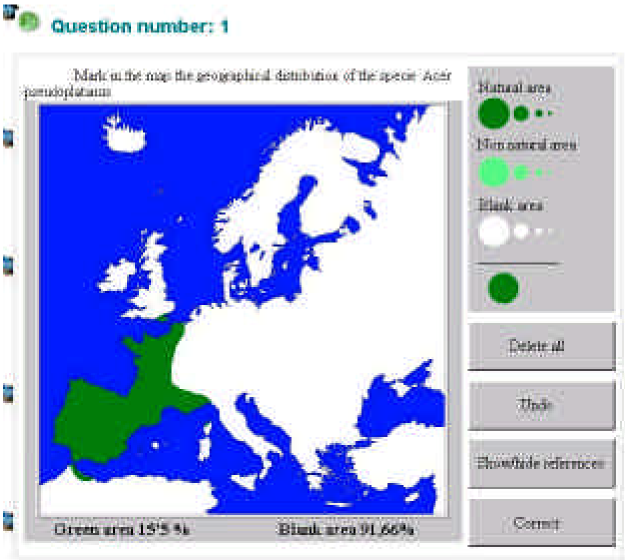
b) *Items with dependent answer:* The right answers are combinations of the set of possible answers. For the answer to be correct, all correct alternatives must be selected. This kind of items is equivalent to a set of dichotomous items asking if certain combinations of answers are correct.

✍ *Items controlled by Java applets.* Items interfaces are HTML documents, so Java applets can be added to the stem or to any of the alternatives to enhance presentation, keeping the evaluation mechanism unchanged. SIETTE provides another kind of items where the evaluation mechanism is accomplished by an applet itself. This type of questions does not offer a list of possible responses. In this case the student must interact with a program. The student actions are recorded and processed by the program to determine if the answer is right or wrong. Several specific tests have been developed in SIETTE using this kind of items, like a Piagetian test for cognitive ability estimation [11, a test of European trees geographical distribution, etc.

Fig. 4 shows an example of a question from the *European trees geographical distribution* test. The goal of this question is, by means of a paint brush, to select the European regions where certain species of tree can be found. Once the examinee has selected a region in the map, he must click on the "*Correct*" button. Then the applet will compare the region selected by the student with the correct region (allowing certain degree of error). When the applet has classified the answer of the student in terms of its correctness, it gives the result of this evaluation to SIETTE.

Fig. 5 shows an example of this evaluation mechanism (see [11] for a complete description). These kind of items can be also classified as dichotomous or multiple choice items, depending on the number of possible responses which the applet uses to assess the student's answer.

Thanks to this type of items, SIETTE offers the possibility to include virtually any kind of item (as long as it can be implemented by means of a Java applet). Certainly, this possibility is restricted to test developers with some programming skills.



Question number: 1

Mark in the map the geographical distribution of the specie Acer pseudoplatanus

Fig. 4. An item controlled by means of a Java applet

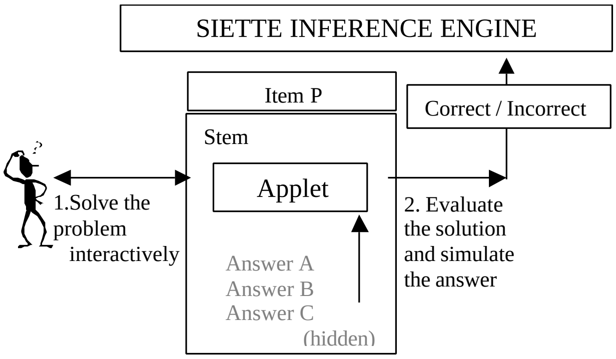


Fig. 5. Evaluation process by an applet

4 The library of templates

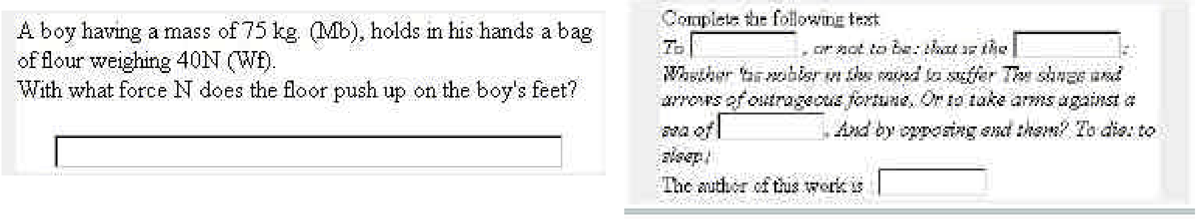
Additionally to the items shown in the previous section, SIETTE includes a library of templates for items. This library allows teachers to pose different kinds of exercises. It does not only offer typical questions of selection like the items shown in Fig. 2 and Fig. 3., but intends to be a complete collection of all different types of exercises that usually appear in text books.

All these templates have been implemented by means of Java applets, but in this case, this library allows teachers to automatically construct items for any kind of test. They only have to properly instantiate one of the templates of the library. This task can be easily accomplished by using some of the test editors provided with SIETTE.

To use this library teachers do not need to have any programming skills. Also, all types of items can be used in the same test.

The items that can be constructed by means of the library are the following:

? *Fill-in-the-blanks items*, in which the examinee has to fill some blanks in a text. The correctness of the answer is checked by means of regular expressions (in the case that the blank needs to be filled with text) or formulas that allow a selectable percentage of error (in the case that the answer is a number), or a combination of both. Examples are shown in Fig. 6.



A boy having a mass of 75 kg. (Mb), holds in his hands a bag of flour weighing 40N (Wf).
With what force N does the floor push up on the boy's feet?

(a)

Complete the following text
To ▭ *, or not to be: that is the* ▭*:*
Whether 'tis nobler in the mind to suffer The slings and arrows of outrageous fortune, Or to take arms against a sea of ▭*, And by opposing end them? To die: to sleep;*
The author of this work is ▭

(b)

Fig. 6. Fill-blank items: (a) numeric (b) text

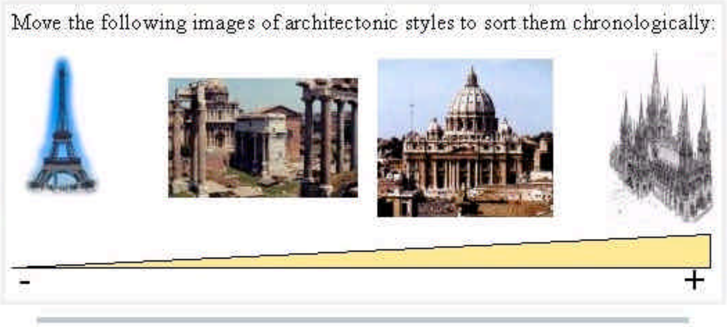
The solution of question 6(a) can be introduced by the teacher using the regular expression:

```
#(Mb+Wf/9.81)3% ([Kk][Gg]|Kilograms) |
#(9.81*Mb+Wf)3% ([Nn][Ww]|Newtons)
```

While the student's answer might be any valid combination (*e.g.* `79 Kg`).

These questions are transformed into *n* dichotomous items, where *n* is the number of blanks in the composed item.

? *Sorting items.* In these items students have to sort a set of *n* elements by a drag-and-drop mouse operation. Elements can be text or images. These items are transformed to multiple-choice items with *n!* possible answers.



Move the following images of architectonic styles to sort them chronologically:

Fig. 7. Sorting item

? *Inset items.* In these kind of items examinees must select the elements which fulfill a condition. These elements must be inserted into a set. These items are equivalent to polytomous items with independent answers. Fig. 8(a) shows an example of this type of items. This example is equivalent to the example of Fig. 4, but it offers a more attractive interface.

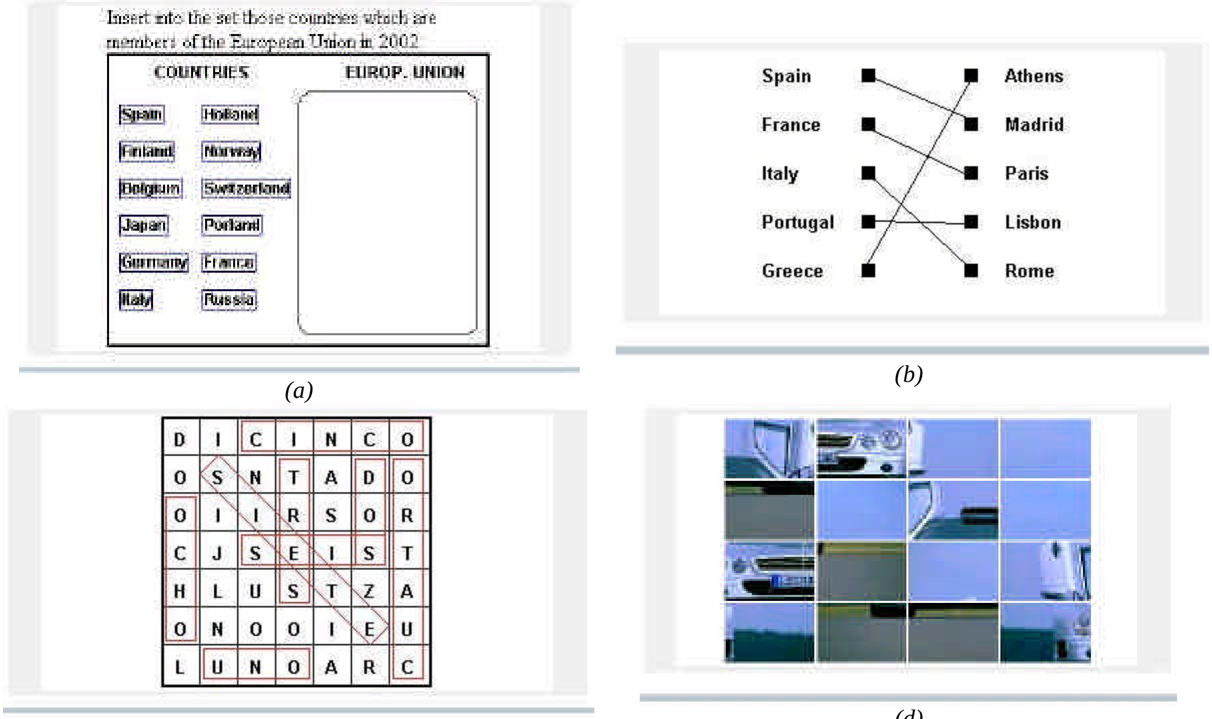
? *Connection items.* Two columns of *n* elements are presented, text or graphics. The students must link each element of the left column with one element of the right column. These items can be transformed into a multiple-choice item with *n!* choices. The objective of the example of Fig. 8(b) is to link each country with its capital. In this case the item is equivalent to the following collection of items: an item asking about the capital of Spain, where the optional answers are all the components of the second column (Rome, Lisbon, etc.); the same for the capital of France, and so on.

As it has been shown, each one of the different types of items that can be constructed by using these templates has an equivalent among the items presented in the previous section. Therefore the addition of this library has not involved significant modifications in SIETTE architecture.

In order to improve the presentation of items, SIETTE offers some other templates in the style of pastimes. Some of these types of templates are:

? *Word search items*. These items consist of locating the missing word in a matrix of letters. In the example shown in Fig. 8(c) the students have to locate the numbers (1 to 8) in Spanish.

? *Puzzle items*. This is the classical Frank Lloyd's puzzles. In these kind of items the students must order the pieces of an image to get the correct shape of the object represented in the image.



Insert into the set those countries which are members of the European Union in 2002

(a)

(b)

D	I	C	I	N	C	O
O	S	N	T	A	D	O
O	I	I	R	S	O	R
C	J	S	E	I	S	T
H	L	U	S	T	Z	A
O	N	O	O	I	E	U
L	U	N	O	A	R	C

(c)

(d)

Fig. 8. (a) Inset item. (b) Connection item. (c) Words search item. (d) Puzzle item.

4.1 Temporized items

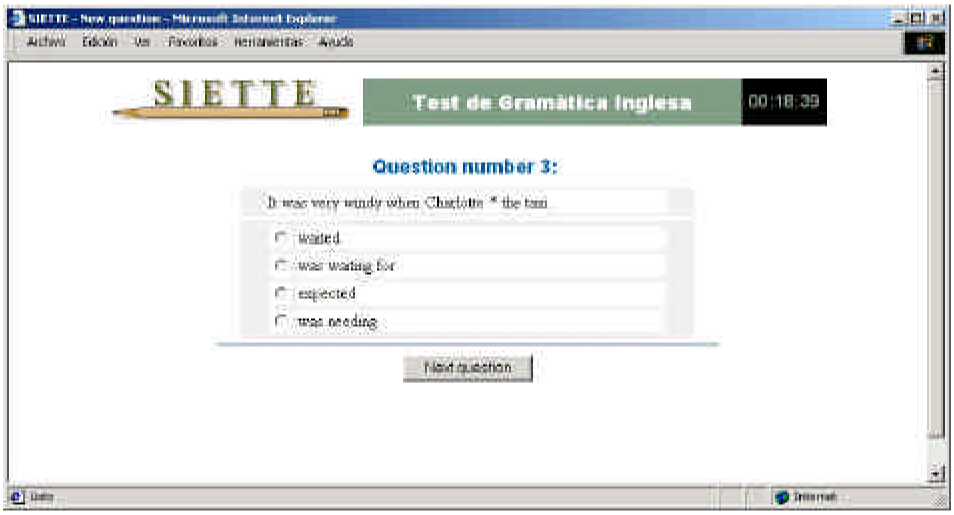
The supporting library also provides the capability of making temporized tests. Response time has been deeply studied in IRT [12, 13]. SIETTE is used to estimate the *knowledge level*, not any psychometric *latent trait*, so it uses a simplified approximation. It does not include time as a parameter of the ICC, but provides a mechanism to add time constraints. There are three possibilities:

? *Non-temporized tests*. In these tests the examinees have as much time as they need to finish the test. It is recommended that adaptive tests use this option.

? *Temporized-item tests*. In this type of tests, each item is assigned a maximum time (in the edition stage). If the does not answer the item on time, the generator will automatically show the next one. To calculate the new estimated knowledge level of the student, the generator assumes that the student has not answered this item.

? *Temporized tests*. These tests have a maximum global time to answer to all questions.

This feature has been implemented using a clock applet with the same mechanism used in the implementation of the templates of the supporting library. The clock begins to count down when the HTML page, with the item, has been loaded. In a temporized test, if the student clicks on the *Next* button, the clock stops and the time value is passed as an initial value for the next item. If the time finishes before the student answers the item, the clock applet will force the finalization of the availability of the item, firing the final estimation of the knowledge level of the student, and therefore the test will end.


SIETTE - New question - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
SIETTE
Test de Gramática Inglesa
00:18:39
Next question
Listo
Internet

Question number 3:

It was very windy when Charlotte * the taxi

○ waited
○ was waiting for
○ expected
○ was needing

Fig. 9. A temporized item

In temporized-item tests, if the student clicks on the *Next* button, the clock will stop but there is no need to store the remaining time, because each item has its own independent clock time.

Teachers can design tests where some items are temporized and other items are not. They only have to select, in the test editor, the temporal option for each item. Thanks to the dynamic addition of this temporal characteristic, items can be easily converted into temporized items and vice versa. Equally, complete tests can be easily set or unset as temporized by means of the test editor options. Of course the ICCs change if a different maximum answering time is selected. There is no special feature yet implemented to deal with time varying ICCs.

5 Conclusion

Classical evaluation systems based on tests presents to the examinees a fixed number of items, not making any distinction between students in terms of their ability. Advanced and inexperienced students take the same tests. This feature enforces teachers to design tests for students with a medium level of knowledge. Also, all these items are usually posed in the same order. Therefore, students may get bored/disappointed while taking the tests because questions are too easy/difficult for them.

Fortunately, adaptive tests-based evaluation systems can improve these drawbacks. The number of items in a test depends on the ability of the student. The system estimates, after the presentation of each item, the new knowledge of the student and the selection of the next item is done according to this level. As a result, the next item posed to the examinee will be more informative. Also, when the system considers that it has enough information to correctly evaluate the student, the test finishes.

The introduction of the new library of item templates in the SIETTE system further extends its capabilities. Now teachers without programming skills can design tests with different kinds of questions and exercises. Tests are not limited to classical questions where students must select one or more answers from a set of possible options. As a result, tests can be richer and more amusing.

The library implemented is an *open* library, in the sense that it allows the addition of new templates. This feature is available thanks to the implementation of these templates by means of Java applets, which can be inserted into HTML pages. Any test developer with programming skills can add new templates to the library using test editors.

Currently we are developing other libraries for the SIETTE system. These libraries are specific for subjects like *Logic*, *Compilers*, etc.

SIETTE can be accessed and tested at http://www.lcc.uma.es/SIETTE .

References

1. Intralearn Soft.ware Corp. (2002). Intralearn SME. http://www.intralearn.com (Accesses Jan 6, 2002).
2. Drake Kryterion Inc. (2002). Webassessor. http://www.webassessor.com/webassessor (Accesses Jan 6, 2002).
3. WebCT Inc.(2002). WebCT. http://www.webct.com (Accesses Jan 6, 2002).
4. WBT Systems (2002). TopClass. http://topclass.uncg.edu/ (Accesses Jan 6, 2002).
5. Wainer, H. (ed.). (1990). *Computerized Adaptive Testing: a Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.
6. Ríos, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L. & Conejo, R. (1999). Internet Based Evaluation System. In *Proceedings of AIED'99*. Amsterdam: IOS Press.
7. Conejo, R., Millán, E., Pérez-de-la-Cruz, J. L. & Trella, M. (2000). An Empirical approach to on-line learning in SIETTE. In *Proceedings of the ITS 2000*, Montreal. Springer-Verlag.
8. Van der Linden, W. & Hambleton, R. (eds). (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag.
9. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's mental ability. In Lord, F. M. & Novick, M.R. (eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
10. Owen, R. J. (1975). A Bayesian sequential procedures for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association 70*, 351-356.
11. Arroyo, I., Conejo, R., Guzmán, E. & Woolf, B.P. (2001). An Adaptive Web-based Component for Cognitive Ability Estimation. In J.D. Moore, C. Luckhardt-Redfield, W. Lewis Johnson (eds.), *Artificial Intelligent in Education: AI-ED in the Wired and Wireless Future.* IOS Press. Amsterdam.
12. Verhelst, N.D., Verstralen, H.H.F.M. & Jansen, M.G.H. (1997). A logistic model for Time-limit Tests. In W.J. Van der Linden & R.K. Hambleton (eds.), *Handbook of Modern IRT*. New York: Springer.
13. Roskam, E.E. (1997). Models for Speed and Time-Limit Tests. In W.J. Van der Linden & R.K. Hambleton (eds.), *Handbook of Modern IRT*. New York: Springer.