Adaptive evaluation through different types of items

Eduardo Guzmán & Ricardo Conejo

Departamento de Lenguajes y Ciencias de la Computación
E.T.S.I. Informática, Universidad de Málaga. Apdo. 4114, Málaga 29080. SPAIN
Phone n.: +34 952 13 28 63
e-mail: {guzman,conejo}@lcc.uma.es

Abstract. SIETTE is a general purpose tool to create and make web-based *Computerized Adaptive Tests* (CAT). Its adaptive mechanisms are given by a Bayesian procedure based on a discrete version of the *Item Response Theory* (IRT). The use of this theory is very extended in CATs, mainly to estimate the student's knowledge level. The classical approach of IRT generally restricts its use to evaluate dichotomous (true/false) questions. In SIETTE, this and other kind of questions (items) are available, thanks to the use of extensions of the classical IRT. In this paper, all types of items available in this system are described, the advantages of their use, as well as how they are used to adaptively estimate the student's knowledge level.

Keywords: AI in education, Computerized Adaptive Tests, Item Response Theory

Section: paper track

Adaptive evaluation through different types of items

Eduardo Guzmán & 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. SIETTE is a general purpose tool to create and make web-based *Computerized Adaptive Tests* (CAT). Its adaptive mechanisms are given by a Bayesian procedure based on a discrete version of the *Item Response Theory* (IRT). The use of this theory is very extended in CATs, mainly to estimate the student's knowledge level. The classical approach of IRT generally restricts its use to evaluate dichotomous (true/false) questions. In SIETTE, this and other kind of questions (items) are available, thanks to the use of extensions of the classical IRT. In this paper, all types of items available in this system are described, the advantages of their use, as well as how they are used to adaptively estimate the student's knowledge level.

1 Introduction

Most of test-based evaluation systems on Internet are like the classical paper-and pencil tests. They propose for all the students the same questions. As a result, these systems do not take into account the initial knowledge level of the students.

A *Computerized Adaptive Test* (CAT) [1] is a test where the presentation of each question and the finalization of the tests are dynamically decided, based on the student's proficiency. Therefore, these systems are adapted to the characteristics of each student. All the questions are proposed according to the estimated student's knowledge level. As a result, the number of questions to pose to students is reduced, because the system is able to estimate faster the student's knowledge level asking him the most adequate question every time.

SIETTE is an efficient implementation of CATs. The mechanisms used to carry out the selection of the most suitable question (*item*) to pose to the student, as well as the test finalization criterion, are based on a psychometric theory called *Item Response Theory* (IRT).

Classical IRT was defined for only dichotomous items, *i.e.* true/false questions. In SIETTE, different types of items have been introduced to improve the capabilities of the system. The answers to these items affect to the student's knowledge level in a different way, therefore the use of some extensions of IRT must be used.

In this paper, all types of items provided by SIETTE are presented, mainly focusing on how they make influence in the estimation process of the student's knowledge level.

In the next section, the architecture of SIETTE as well as its operation mode are described. In section 3, the adaptive estimation of the student's knowledge level using the IRT is approached, especially focusing on how it is carried out in SIETTE. Section 4 analyzes the types of items provided in SIETTE and its influence in the estimation of the knowledge level. At last, the contributions given by this work will be drawn.

2 The SIETTE system

SIETTE (System of Intelligent Evaluation using Tests for Teleeducation) [2] has been designed to be used through World Wide Web. By means of a navigation application, teachers can create and modify tests, and examinees can evaluate their knowledge about different subject.

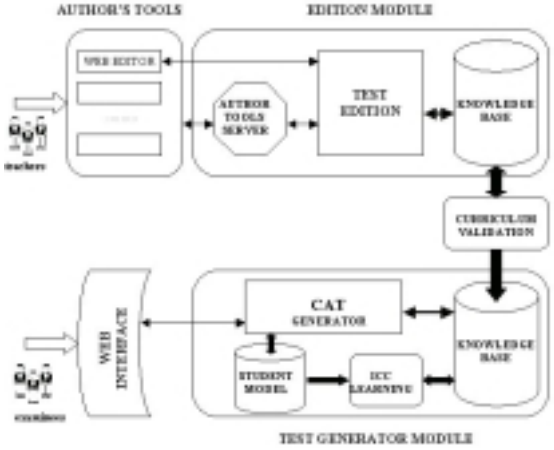

Fig. 1. The architecture of SIETTE

Its architecture, Fig. 1, comprehends the main components of an adaptive test. The following parts can be pointed out:

- *Knowledge base.* It is composed by the concept domain (*curriculum*), the specifications of the tests and the item pool [3].
- *CAT generator*. It poses the most suitable items every time in each test session.
- *Edition module and author's tools.* [4] They allow teachers to accomplish insertions and modifications in the knowledge base.
- *The student's model.* It collects all the information about the student needed by the test generator. It is dynamically updated after each response.
- *Curriculum validation module.* It ensures that all the teacher's specifications of tests are coherent and their data are consistence, and therefore they can be used by examinees.
- *ICC learning module.* When teachers add new items to the knowledge base, they must estimate some configuration parameters. These parameters are only an

initial approximation, so a calibration process is required. This module uses the information obtained from the student's model [5].

2.1 The operation mode

CAT systems follow an iterative algorithm. It begins with an initial estimation of the examinees' knowledge level and it 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. Depending on the response and on the kind of item, the student's model is updated computing a new estimation of his knowledge level.
4. Steps 1 to 3 are repeated until the finalization criterion defined is met.

The finalization criterion is configured in each test. First, in the edition phase, the teacher must indicate a minimum and a maximum number of items. These values set bounds to the number of items that may be posed to the students. If a student has taken the maximum value of items, the final estimation process of his knowledge is forced. Additionally, SIETTE offers the following adaptive finalization criteria: The most likely value of the estimated knowledge distribution of the examinee, upper to a certain threshold; or the variance of the estimated knowledge distribution is lower than a certain value.

3 Adaptive estimation of the knowledge level

IRT [6], is based on the hypothesis that the answer given to each item of the test, probabilistically depends on certain *latent trait (θ)*, that can be measured by means of an unknown fixed numerical value. This theory has been successfully applied to the item selection mechanisms and student's knowledge level estimation in CATs. In this domain the *latent trait* is the *knowledge level* of the student about the topics involved in the test.

The value of θ is estimated using the response to each item. There are several methods to get this value. In SIETTE, a Bayesian method [7] is used. In this method, the *a posteriori* probability distribution of the student's knowledge level ($P(\theta|u)$) is calculated by the Bayes' rule. Also, it is assumed that the latent trait θ can only take *K* discrete values (from *0* to *K-1*) because of the high computational cost of the calculus. As a result, the estimation process is simplified to the following product:

$$\overline{P(\theta|u)} = \left\| \prod_{i=1}^{n} \overline{P_i(u_i = 1|\theta)}^{u_i} (1 - \overline{P_i(u_i = 1|\theta)})^{(1-u_i)} \overline{P(\theta)} \right\| \quad (1)$$

where $\overline{P(\theta)}$ represents the *a priori* normalized student's knowledge level distribution. 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 $f:(-\infty,+\infty)\rightarrow[0,1]$, named *Item Characteristic Curve* (ICC). In SIETTE, ICC is given by a vector of *K* components $P_i(u|\theta) = (p_i(u|\theta=0), p_i(u|\theta=1),... p_i(u|\theta=K-1))$. The initial estimation of the ICC can be accomplished by several models. SIETTE uses a model of three parameters based on the logistic function [8]:

$$P_i(\theta) = P(u_i = 1|\theta) = c_i + (1-c_i)\frac{1}{1+e^{-1.7a_i(\theta-b_i)}} \quad (2)$$

Forth, to simplify, let us use $P_i(\theta)$ instead of $P(u_i = 1|\theta)$. The three parameters of ICC are:

- *Discrimination factor* (a_i): It is proportional to the slope of the curve. A high value indicates that the probability of success by students with a higher knowledge than the difficulty of the item, is higher.
- *Difficulty* (b_i): It corresponds to the knowledge level in which the probability of success is the same as fail.
- *Guessing factor* (c_i): It is the probability that a student without any knowledge answers correctly to the item. It represents the case in which student answers randomly. It is equal to the portion of right answers over the total number of answers.

The initial values of the ICC vector require from the teacher the estimation of only two values: the difficulty of the item and the discrimination factor, or even only the first one, since the discrimination factor can be assumed a low value (*e.g.* 0.5). This will be later adjusted in SIETTE by the on-line learning mechanism [5].

4 Types of items in SIETTE

SIETTE allows teachers to propose different types of items. All this kind of items can be combined into the same test. In this section, these items are going to be approached. A description of each kind of item will be given, as well as how its ICC is initially estimated. Also, it will be analyzed how the student's model is updated depending on the kind of item posed.

4.1 Dichotomous items

These items can only have two answers: true or false. They state sentences and examinees must answer in terms of their correction.

The estimation of the student's knowledge level is carried out following the classical approach of IRT summarized in equation (1). After the item response, the knowledge level distribution curve of the student is updated with the ICC ($P_i(\theta) = P(u_i = 1|\theta)$) if the answer is correct, and otherwise with the inverse of the ICC ($P(u_i = 0|\theta) = 1 - P_i(\theta)$). In this case, the guessing factor is initially considered 0.5, since only two possible responses are allowed.

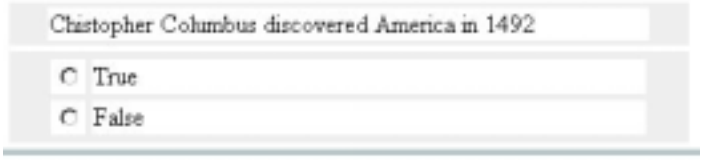

Fig. 2. A dichotomous item

4.2 Multiple-choice items

This kind of items presents more than two possible responses (options). The examinees may select one of these responses, or even none of them. The classical dichotomous IRT model does not support these items, although some extended models have been developed [9, 10, 11].
In SIETTE, each possible option has an ICC associated, the *Option Characteristic Curve* (OCC) [11], which represents the *conditional probability* of selecting this option given certain knowledge level. Note that examinees may not select any option. This state or *latent response* is often called *don't know* option [9] and must be represented by an additional OCC.

Fig. 3. A multiple-choice item

The initial OCCs for the all the options is calculated from the discretization of formula (2), where the initial estimation of the guessing parameter c_r is *1/N*, being *N* the number of choices. Parameter a_r, *(discrimination)* can be left to a reduced value and parameter b_r, *(difficulty*) is given by the teacher.
Given an item with *N* possible options $\boldsymbol{u_i} = (u_{i1}, u_{i2}, \ldots, u_{iN})$, consider u_{ir} like the right response, $1 \le r \le N$, and u_{i0} like the *don't know* option.

$$P(\mathsf{u}_\mathsf{i} = \mathsf{u}_{\mathsf{ir}} \mid \theta) = P_i(\theta) \tag{3}$$

$$P(\mathsf{u}_\mathsf{i} = \mathsf{u}_{\mathsf{ij}} \mid \theta) = \frac{1 - P_i(\theta)}{N} \qquad \forall j, 0 \le j \le N, \mathsf{j} \ne \mathsf{r} \tag{4}$$

As a result, in the estimation process, the only modification to accomplish, is the substitution of the ICC vector for the corresponding OCC vector in equation (1) for each multiple choice item which form the test session.

Initially all OCCs for incorrect options are considered to be equals. The on-line learning procedure will modify the curves as the test goes on according to the examinees' answers [5]. The same can be said of the characteristic curves of the remaining kinds of items.

4.3 Polytomous items

They are multiple-choice items where the right answer is one or more combinations of options. Examinees must select one combination in order to successfully answer to the question. These items are used because they are more informative and reliable than dichotomously scored items [12]. This kind of items can be classified into:

a) ***Items with independent options*:** In this case, the correction of one option does not have any influence in the correction of the remaining options. That is, if in item of Fig. 4 the option *France* is selected, this response can be considered partially right.

Fig. 4. A polytomous item with independent options

The mechanism of initial estimation of the ICCs implies an extension of the classical IRT model. In this case, there are an OCC for each option, but the way in which it is initially inferred, as well as the mechanism of knowledge level estimation differ.

Formally, consider that item i has N options $\boldsymbol{u_i} = (u_{i1}, u_{i2}, \ldots, u_{iN})$. Examinees can select or not each option. As a result, each option may have two different OCCs which represent these states. Additionally, if one option is selected, it can be right o wrong. Suppose that this item has R right responses $\boldsymbol{r_i} = (r_{i1}, r_{i2}, \ldots, r_{iR})$. These right options are members of $\boldsymbol{u_i}$. OCCs are initially calculated from equation (2), with a guessing parameter of c=0.5; a low discrimination factor; and a difficulty parameter given by the teacher, assumed equal for all options.

$$P(\mathsf{u}_i = \mathsf{u}_{ij} \mid \theta) = P_i(\theta) \qquad \forall \mathsf{u}_{ij} \in \mathsf{r}_i \tag{5}$$

$$P(\mathsf{u}_i = \mathsf{u}_{ij} \mid \theta) = 1 - P_i(\theta) \qquad \forall \mathsf{u}_{ij} \notin \mathsf{r}_i \tag{6}$$

Equation (5) collects the initial estimation for the case of selecting a right option, and equation (6), for the case of selecting a wrong answer. An additional OCC is required to initially estimate the case that option j of item i has not been selected. This last curve equals $1 - P_i(\theta)$.

The evaluation of these items will be as much successful as the number of the right selected options. The effect of posing one of this kind of items is the same of proposing a set of dichotomous items one by one. The number of dichotomous items is equal to the number of choices. Each one of these items will ask if the corresponding answer belongs to the set of answers that satisfy the stem. For instance, the item shown in Fig. 4 is equivalent to the following dichotomous items: *Is France member of the European Union in 2002?, Is Italy member of the European Union in 2002?, Is Germany member of the European Union in 2002?, etc.* Formally, the response to an item *i*, the calculus of the *a posteriori* student's knowledge distribution is accomplished following equation (7). This equation is a modification of equation (2) particularized for an only one item *i*:

$$P(\theta \mid u) = \left[\prod_{j=1}^{N} P(u_{ij} \mid \theta)\right] P(\theta) \tag{7}$$

b) ***Items with dependent options:*** The appearance of these items is just like the previous, but they are semantically different. In this case, the correct answers are subsets of the set of possible combination. The correction of an individual option does have direct influence in the correction of the question. That is, in Fig. 5, only if the combination of the options *Germany* and *Japan*, or the combination of *Germany*, *United Kingdom*, *Russia* and *USA* are selected, the answer is evaluated as correct. In general, more than one correct combinations might be allowed.

This kind of items is equivalent to a multiple choice item with 2^N possible options, where N is the number of real options (*e.g.*, in Fig. 5, $N=6$).

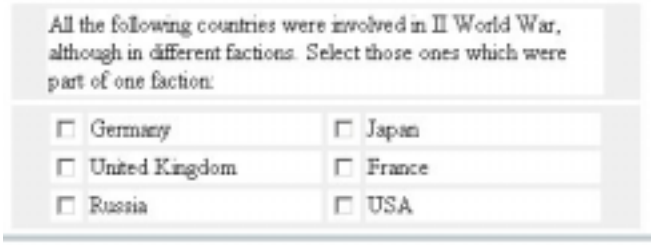

Fig. 5. A polytomous item with dependent options

From the point of view of the generation of the ICCs, these items have a set of OCCs, each one associated with one of the 2^N combinations of options.

The initial estimation of the OCC of each combination is similar to the case of multiple choice items, where the set of options $\boldsymbol{u_i} = \wp(\{u_{i1}, u_{i2}, ..., u_{in}\})$. Let us consider that this item has R right combinations of options $\boldsymbol{r_i} = (r_{i1}, r_{i2}, \ldots, r_{iR})$, where r_{ij} represents a combination of options and $\forall r_{ij}, 1 \le j \le R, r_{ij} \in r_i, r_{ij} \in u_i$. Equations (8) and (9) summarize how the initial estimation procedure of OCC for each combination of options is done. The guessing parameter used to calculate the value of expression (2) is now $1/2^N$.

$$P(u_i = u_{ij} \mid \theta) = \frac{P_i(\theta)}{R} \qquad \forall u_{ij} \in r_i \tag{8}$$

$$P(u_i = u_{ij} \mid \theta) = \frac{1 - P_i(\theta)}{2^N - R} \qquad \forall u_{ij} \notin r_i \tag{9}$$

The estimation process of the student's knowledge level, is carried out just like equation (1) indicates, but substituting the ICC for the OCC of the combination of options selected by the student.

4.4 Items controlled by Java applets

Items interfaces are HTML document, so Java applets can be added to the stem or to any option of the answer 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 little program. The

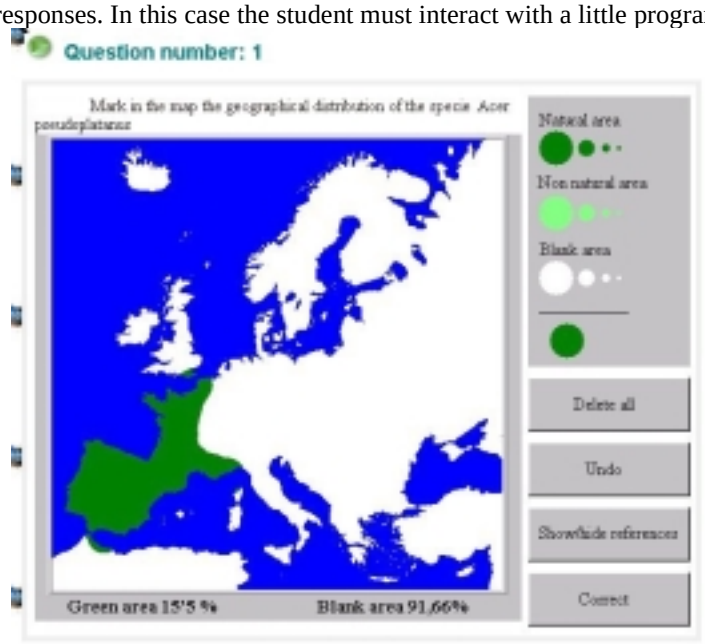

Fig. 6. An item controlled by a Java applets

student's actions should be caught 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 [13], a test of European trees geographical distribution, etc.

Fig. 6 shows an example of a question from the *European trees geographical distribution* test. The objective 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 push the "*Correct*" button. At this moment, the applet will compare the region selected by the student with the right region, taking into account certain degree of error. When the applet has classified the answer of the student in terms of its correction, it gives the result of this evaluation to SIETTE (see [13] for a complete description). These kind of items can be also classified as dichotomous, multiple choice items, etc. depending on the number of possible responses which the applet internally uses to assess the student's response. Therefore, the mechanism of evaluation varies depending of the type of the item. The initial estimation of the ICC is accomplished according to the type of item too.

5 Conclusion

SIETTE offers the same functionally as other commercial test-based tools [14, 15, 16, 17], and provides well-founded mechanisms of evaluation and item selection. These mechanisms are based on IRT. Items posed to examinees are dynamically decided in terms of their estimated knowledge level. The number of questions of the test is not fixed. The adaptive inference engine decides when the test must finish.

SIETTE offers several types of items: dichotomous, multiple-choice, polytomous. Classical IRT was defined to evaluate dichotomous items. Therefore in SIETTE, some extensions of IRT have been used to evaluate the response of the students when these new kinds of items are posed. As a result, the student's temporary model is dynamically updated, according to each kind of item.

All the characteristic curves used by the different types of items are initially inferred from the ICC of the item. This ICC only requires a parameter, the difficulty, since the guessing factor depends on the number of item options, and the discrimination factor can be assumed to a low value. Thanks to the on-line learning module provided in the architecture of SIETTE, this initial estimation is only an approximation given by the criteria of the teacher. The dynamical calibration process of the ICCs will fit this curves properly according to the previous responses to the items, stored in the knowledge base.

The introduction of these new items makes richer the information provided to the student's model by each item. For instance, using polytomous items with independent options, teachers can pose an item which is equivalent to pose multiple dichotomous items. This may reduce the number of items posed to the student to obtain a final precise estimation of his knowledge level.

Thanks to the items controlled by Java applets, SIETTE offers the possibility to include virtually any kind of item which could be implemented by means of a Java applet. Certainly, this possibility is restricted to test developers with some programming skills.

Additionally, a library of templates for the automatic construction of items has been developed [18]. 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. From the point of view of the student's knowledge estimation, items generated by the

components of this library are equivalent, to the items shown in this paper. The advantage of this library is that the generated items have a powerful and amusing interface and also it can be easily included in tests by teachers without programming skills.

The system can be tested at http://www.lcc.uma.es/SIETTE

References

1. Wainer, H. (ed.). (1990). *Computerized Adaptive Testing: a Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.
2. Ríos, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Conejo, R. (1999). Internet Based Evaluation System. In *Proceedings of the 9th AIED*. Amsterdam: IOS Press.
3. Guzmán, E. & Conejo, R. (2002). Simultaneous evaluation of multiple topics in SIETTE. *Proceedings of ITS Conference 2002*. Biarritz (France). (to be published).
4. Guzmán, E., Conejo, R. & Domínguez, F. J. (2002). Una herramienta de autor autónoma para un sistema de tests adaptativos a través de Internet. *III Congreso Internacional de Interacción Persona-Ordenador*, Madrid. 90-96.
5. 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.
6. Van der Linden, W. & Hambleton, R. (eds). (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag.
7. 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.
8. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's mental ability. In Lord, F. M. & Novick, M.R. (ed.) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
9. Thissen, D. & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika*, 49.501-519.
10. Thissen, D. & Steinberg, L. (1997). A response model for multiple-choice items. In W.J. Van der Linden & R.K. Hambleton (eds.): *Handbook of Modern IRT*. New York: Springer.
11. Revuelta, J. (2000). *A psychometric model for multiple choice items*. Ph. Thesis. Autonoma University of Madrid. Madrid.
12. Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for psychologists*. Lawrence Erlbaum Associates. New Jersey.
13. 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.
14. Intralearn Soft.ware Corp. (2002). Intralearn SME. http://www.intralearn.com (Accesses Jan 6, 2002).
15. Drake Kryterion Inc. (2002). Webassessor. http://www.webassessor.com/webassessor (Accesses Jan 6, 2002).
16. WebCT Inc.(2002). WebCT. http://www.webct.com (Accesses Jan 6, 2002).
17. WBT Systems (2002). TopClass. http://topclass.uncg.edu/ (Accesses Jan 6, 2002).

Guzmán, E., Riveros, J.A. & Conejo, R. (2002). A library for items construction in an adaptive evaluation system. *Proceedings of the "Creating Diagnostic Assessments" Workshop of the ITS Conference 2002*. Biarritz, France. (to be published).