User Model User-Adap Inter
DOI 10.1007/s11257-006-9018-1

ORIGINAL PAPER

Adaptive testing for hierarchical student models

Eduardo Guzmán · Ricardo Conejo ·
José-Luis Pérez-de-la-Cruz

Received: 31 October 2005 / Accepted in revised form: 1 July 2006
© Springer Science+Business Media B.V. 2006

Abstract This paper presents an approach to student modeling in which knowledge is represented by means of probability distributions associated to a tree of concepts. A diagnosis procedure which uses adaptive testing is part of this approach. Adaptive tests provide well-founded and accurate diagnosis thanks to the underlying probabilistic theory, i.e., the *Item Response Theory*. Most adaptive testing proposals are based on dichotomous models, where the student answer can only be considered either correct or incorrect. In the work described here, a polytomous model has been used, i.e., answers can be given partial credits. Thus, models are more informative and diagnosis is more efficient. This paper also presents an algorithm for estimating question characteristic curves, which are necessary in order to apply the Item Response Theory to a given domain and hence must be inferred before testing begins. Most prior estimation procedures need huge sets of data. We have modified preexisting procedures in such a way that data requirements are significantly reduced. Finally, this paper presents the results of some controlled evaluations that have been carried out in order to analyze the feasibility and advantages of this approach.

Keywords ITS · Student diagnosis · Adaptive testing · Item response theory · Statistical kernel smoothing

This paper (or a similar version) is not currently under review by a journal or conference, nor will it be submitted to such within the next three months.

E. Guzmán (✉)· R. Conejo · J-L. Pérez-de-la-Cruz
Dpto. de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Bulevar Louis Pasteur, 35, Campus de Teatinos, 29071 Málaga, Spain
e-mail: guzman@lcc.uma.es

R. Conejo
e-mail: conejo@lcc.uma.es

J-L. Pérez-de-la-Cruz
e-mail: perez@lcc.uma.es

1 Introduction

Some of the features which determine the quality of an intelligent tutoring system (ITS) are the suitability of its student model and the accuracy of the corresponding diagnosis mechanisms, since information stored in student models has a high impact on training and knowledge (Papanikolaou et al. 2003). However, several researchers such as Self (1994) have stressed the inherent difficulty of student model construction and initialization (Zukerman and Albrecht 2001).

All measurement tools must fulfill some requirements to ensure scientific adequacy:

- *Validity*: Validity does not depend on the tool, but on how it is applied. A measurement tool has this property when it actually measures what it is supposed to. Therefore, procedures to check validity are based on establishing relationships between results and other observable facts in direct relation to the kind of capability being measured.
- *Reliability*: This feature refers to the accuracy of measurement, independently of the features being measured. A tool must exhibit stability and consistency, i.e., when performing the same procedure with the same individual in similar circumstances twice, similar results are obtained.
- *Objectivity*: A tool has this feature when results are independent of the observer's opinion or personal perspective. Frequently, objectivity leads to reliability.

When inferring student models it would be desirable to use mechanisms which ensure that these requirements are fulfilled. In this sense, the *evidence-centered design* (ECD) proposal is a framework for designing, producing and delivering educational assessments (Mislevy et al. 1999, 2000). ECD models incorporate representations of what a student knows and does not know, in terms of the results of his/her interaction performance (evidences) with assessment tasks (Shute et al. 2005). ECD is basically composed of the following three models: (a) A student proficiency model, which collects the skills or other attributes to be assessed. (b) Evidence models, i.e., the performances which reveal targeted performances, and the connection between these performances and the student model variables. (c) A task model, that is, the tasks which elicit those performances. The use of the ECD framework in assessment design should contribute to achieve effective and efficient diagnoses (Mislevy et al. 1999).

One of the most popular solutions for student diagnosis (and that contemplated in this paper) is testing. A *test* is an assessment exam composed of a set of questions (called in this context *items*). From an abstract point of view, each item is composed of a *stem* (the question or situation posed to the student) and a set of *choices* (the possible answers). It is widely accepted that tests have some desirable features such as generic applicability (i.e., they can be applied to a broad class of domains) and instructional efficiency (i.e., not too much effort is required in order to administer and correct a test).

Test-based diagnosis systems in the real world tend to use heuristic solutions to infer student knowledge. However, there is a type of tests, adaptive tests (a.k.a. *Computerized Adaptive Tests*) (van der Linden and Glas 2000), which are based on a sound statistical grounding, namely the *item response theory* (IRT) (Embretson and Reise 2000). In this way, these kinds of tests, when properly constructed and administered, can contribute to fulfill the requirements stated above.

The most popular implementations of adaptive tests make use of dichotomous IRT models. This means that an answer to a question can be evaluated as either correct or incorrect, i.e., no partial credit can be given. However, other models can be and have been defined within the IRT framework (*polytomous models*) that can attribute partial credit to item answers (Embretson and Reise 2000). These are more powerful, since they make better use of the answers provided by the students and, as a result, student knowledge estimations can be obtained more quickly and more accurately. A lot of polytomous models have been proposed in the literature (e.g., Samejima 1997; Bock 1997; Muraki 1992; Thissen and Steinberg 1997), but they are seldom applied to adaptive tests.

The reason for this lies in the increased difficulty of *item calibration* for polytomous models. The relationship between answers and knowledge states is expressed by probability distributions (response curves or *item characteristic curves*, ICCs) which must be learnt before administering adaptive tests. This is the item calibration procedure, and the calibration of each curve usually requires lots of information (the results of students who have taken the test previously). Since polytomous models are defined by a greater number of response curves, the amount of data needed for calibration grows even more and becomes infeasible.

Even assuming that adaptive testing is appropriate for student diagnosis, a fundamental problem must be solved before using it to this end in ITS environments: the consideration and handling of information about *multiple knowledge factors* (i.e., concepts, skills, ...) There are theoretical proposals inside IRT that define and make use of multidimensional representations (e.g., Embretson 1991; Tam 1992; Segall 1996); however, these proposals quickly become unrealistic, due to the huge amount of data and computation needed. On the other hand and from a practical point of view, inside the student modeling community there are several proposals which apply adaptive testing techniques (e.g., Huang 1996; Collins et al. 1996); they make use of heuristic approaches, thereby making adaptive tests partially lose their solid founding.

In this paper we present a cognitive student modeling approach based on the following assumptions and principles, which will be discussed and justified in Sect. 3:

- The domain model is a concept tree.
- Student models are obtained by attaching a discrete knowledge level to each node of the domain model.
- The system maintains a probability distribution which at each moment, estimates the student model at that point in time.
- Diagnosis is carried out by means of adaptive testing.
- Testing is based on a discrete polytomous IRT model.
- Item calibration is based on an efficient and partially new algorithm.

The next section summarizes the theoretical basis of this work, i.e., the adaptive testing theory and IRT. Section 3 describes in detail our proposal for student modelling. The corresponding procedure for knowledge diagnosis is described and discussed in Sect. 4. Section 5 presents a partially new algorithm for learning characteristic curves. The algorithm has fewer requirements than the solutions usually found in the literature. Section 6 makes a summative evaluation of this proposal in order to analyze its advantages and feasibility. In Sect. 7, some related works are described, focusing specially on their differences and similarities with ours. Finally, in Sect. 8, the contributions of this work are discussed and some conclusions are drawn.

2 Preliminaries

2.1 Computerized adaptive testing

The final goal of an adaptive test is to quantitatively estimate a student knowledge level expressed by means of a numerical value. To this end, items are posed sequentially, one at a time. The presentation of each item and the decision to finish the test are dynamically adopted based on the student's answers. In general, an adaptive test applies an iterative algorithm which starts with an initial estimation of the student's knowledge level and has the following steps: (1) All the items in the item pool (that have not yet been administered) 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 item is asked and the examinee answers; (3) According to the answer, a new estimation of his/her knowledge level is computed; (4) Steps 1 to 3 are repeated until some finalization criterion is met.

Selection and finalization criteria can be theoretically determined according to the required assessment accuracy, and are controlled by certain parameters. The number of items is not fixed and each examinee usually takes a different number and sequence of items.

In this way, the basic elements in an adaptive testing system are:

- *Response model*: It describes how examinees answer the item depending on their knowledge level, thus providing the probabilistic foundations of adaptive testing.
- *Item pool*: It contains a certain number of correctly calibrated items at each knowledge level.
- *Item selection method*: Adaptive tests select the next item to be posed depending on the estimated knowledge level of the examinee (obtained from the answers to items previously administered). However, there are several procedures to take into account at this level.
- *Termination criterion*: Different criteria can be used to decide when the test should finish, depending both on the desired accuracy and on the intended use of the information gathered.

The main advantage of adaptive testing is that it reduces the number of questions needed to estimate the student knowledge level and as a result the time spent on establishing it. This results in an improvement in examinees' motivation (van der Linden and Pashley 2001). The accuracy of the estimation is much higher than the estimation achieved by randomly picking the same number of questions (Conejo et al. 2004). However, adaptive tests have some drawbacks. They require the availability of huge item pools and techniques to control item exposure and detect compromised items. Also, item parameters must be calibrated. To accomplish this task, a high number of examinees' performances are required and these are not always available. However these considerations could be relaxed somewhat for merely formative purposes, e.g., to support learning, as is the case with ITS.

2.2 Item response theory

The response model is the central element of the adaptive testing theory (usually based on IRT (Lord 1980)). It provides a probabilistic, well-founded theoretical background. By virtue of the response model, the following issues can be theoretically determined: (i) how the student knowledge is inferred; (ii) which is the most suitable item which must be posed to each student in the next step; and (iii) when the test must be finished.

IRT is based on two principles: (a) Student performance in a test can be explained by means of his/her knowledge level; (b) the performance of a student with a certain knowledge level answering an item, can be probabilistically predicted and modeled by means of certain functions called characteristic curves. There are hundreds of IRT-based models, and different classification criteria for them. One of these addresses to how the models update the estimated student knowledge in terms of his/her response. Thereby, IRT-based models can be:

- *Dichotomous models*: Only two possible scores are considered, i.e., either correct or incorrect. A characteristic curve is enough to model each item, the *item characteristic curve* (ICC). It expresses the probability that a student with a certain knowledge level will answer the item correctly.
- *Polytomous models*: The former family of models does not make any distinction in terms of the answer selected by the student. No partial credit is given. This means information loss. To overcome this problem, in this family of models each possible answer has a characteristic curve. These curves express the probability that a student with a certain knowledge level will select this answer. These kinds of models also allow the blank response to be modeled by means of a characteristic curve.

Polytomous models usually require a smaller number of items per test than the dichotomous ones. Nonetheless, dichotomous models are most commonly used in adaptive testing environments. The main reason is that the calibration process is harder in polytomous models. Instead of calibrating one curve per item, a set of characteristic curves must be learnt per item. This means that the set of previously done test sessions must be higher. While a test of dichotomous items requires several hundreds of prior test sessions, a test of polytomous items requires several thousands.

3 A proposal for hierarchical modeling and assessment

The proposal presented in this paper can be viewed as a trade-off between several conflicting requirements. From a theoretical point of view, diagnosis methods should be sound and well-founded on generally accepted formal theories. On the other hand, from a practical perspective, the knowledge engineering effort, the amount of data needed to calibrate the model and the computational burden imposed on the system must be affordable. Finally, from a pedagogical point of view, domain and student modeling should be accurate enough to support effective adaptive instruction.

In ITS literature, the representation of the knowledge to be communicated is usually called *Expert Module* (Holt et al. 1994), that corresponds to the *proficiency model* in Mislevy's ECD framework. We will assume that knowledge can be structured into a tree of *concepts*; for this reason, the expert module will be called *Conceptual Model* throughout this paper. This model is described in Sect. 3.1. A *Student Model* is the representation of the student state at a certain step of the instructional process. Overlay models are used in our proposal. They represent the student knowledge as a subset of the Expert Module. Issues of student modeling are discussed in Sect. 3.2.

The student model update is performed by the *Diagnosis Module*. This is the main problem addressed by this paper and solved by means of a general procedure based

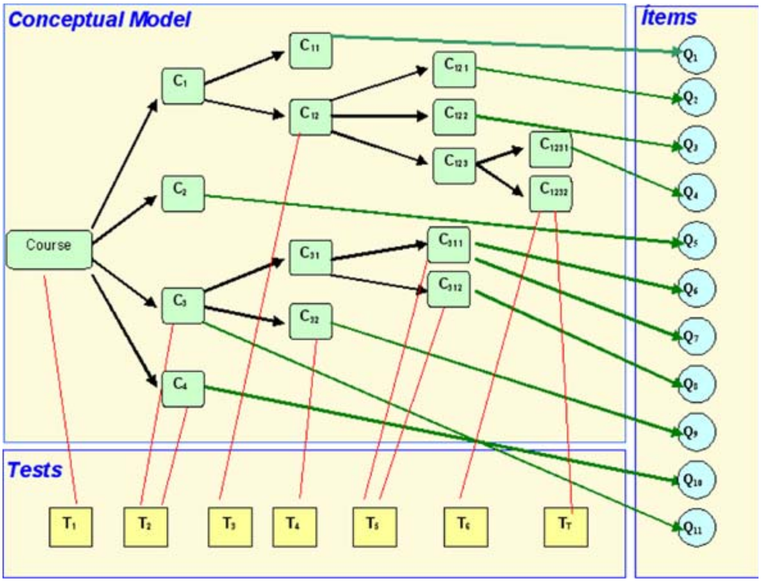

Fig. 1 Relationship among modules and elements

on IRT. The procedure is fully described in Sect. 4. However, for every domain, this general procedure must resort to specific domain knowledge, concerning the possible questions to be administered to the student and the dependences between questions and concepts. This knowledge is structured in an *item pool*, whose description is addressed in Sect. 3.3. But the update may not be needed for the whole student model however; for this reason, the definition of several *test specifications* (described in 3.4) is also considered.

Thus, the knowledge provided by the course developer can be seen as a triplet (Ω, Π, Θ) composed of a concept tree Ω, a set of items Π and a set of test specifications Θ. Notice that this knowledge is qualitative in nature and can be mostly expressed in terms familiar to course developers: what are the parts and subparts of this course? Which part of the course does this test item belong to? Which part of the course do you want to diagnose? On the other hand, quantitative information needed by the diagnosis procedure will be collected and computed automatically when calibrating the domain (see Sect. 5). Figure 1 displays a graphical representation of the relations between Ω, Π, and Θ.

By applying the diagnosis procedure to this domain knowledge, the system will be able to interact with a student, select the most suitable questions, process the student's answers and stop when a certain state is reached.

3.1 Conceptual model

In traditional teaching, it is customary to structure the contents of a course into parts, which are in turn divided into subparts, and so on. In this way, a hierarchy of variable granularity, called *curriculum* (Greer and McCalla 1994), is obtained.

Curricula are often represented in ITSs by *semantic networks*, i.e., by directed acyclic graphs whose nodes are the pieces originated by the partition of the course, and whose arcs represent relationships among them. In ITS literature, a huge set of proposals exist (e.g., Schank and Cleary 1994) in which those parts have different names depending on their level in the hierarchy, e.g. topics, concepts, entities, chapters, sections, definitions, etc.

In our proposal, all nodes in the hierarchy will be generically called *concepts*. As Reye (2002) states, concepts are curriculum elements which represent knowledge pieces or cognitive skills acquired (or not) by students. From the point of view of student diagnosis, concepts are those elements susceptible to being assessed. Note that final nodes (leaf nodes) correspond to single concepts or to a set of them indiscernible from the assessment perspective. Even the root node, which will be called *course*, can also be considered a concept.

Regarding relationships, it is assumed, in our proposal, that concepts from one level are related to the concepts from the previous and the subsequent levels by means of aggregation relations ("*part-of*"). That is, the knowledge of a sibling concept will be part of the knowledge of its parent concept. We will say an inclusion relation exists between those concepts. This is the only relationship considered in our model. Other relations such as "prerequisite-of" are not taken into account, i.e., we assume that the knowledge of a concept is completely independent of the knowledge of the other concepts at the same level. In this way, the domain model is just a tree of concepts. This is a real limitation of our approach. However, researchers such as Collins (1996), have pointed out that prerequisite relationships do not make student model values more accurate, but using them, diagnosis may require fewer items.

We will say that there is a *direct inclusion relation between two concepts* C_i and C_j ($C_i, C_j \in \Omega$) when there is an arc in the graph which comes from C_j and goes to C_i, i.e., if there is an aggregation relation between both. This will be denoted by $C_i \in \wp(C_j)$. For instance, taking the conceptual model of Fig. 1, there is a *direct inclusion relation* between concepts C_1 and C_{12}, or $C_{12} \in \wp(C_1)$.

We will say that there is an *inclusion relation between two concepts* C_i and C_j ($C_i, C_j \in \Omega$) when there is a directed path in the graph (with at least one arc) which comes from C_j and goes to C_i, i.e., if there is a chain of one or more aggregation relations between both. This will be denoted by $C_i \in \wp^+(C_j)$. Notice that the relation of inclusion is the transitive closure of the direct inclusion relation.

We will say that there is an *indirect inclusion relation between two concepts* C_i and C_j ($C_i, C_j \in \Omega$) when they are related by the inclusion relation but not by the direct inclusion relation, i.e., when there is a path in the graph (with two or more arcs) which joins them. This will be denoted by $C_i \in \wp^{++}(C_j)$.

Obviously, the inclusion relation (and the indirect inclusion relation) between concepts is irreflexive, asymmetric and transitive.

From the assessment perspective, when $C_i \in \wp^+(C_j)$ and a student knows C_i, he/she will also have a certain degree of knowledge of C_j, since the knowledge of C_i is (at least part of) the knowledge of C_j.

3.2 Student model

In ITS literature, we can find proposals to model a student by means of comprehensive structures taking into account, for example, affective states (Conati 2002) or learning preferences and styles (Carver et al. 1999). These proposals, as interesting as they can be, pose a set of additional issues when adopting an evidence-centered approach. For this reason, they are not contemplated in this paper; we only focus on *cognitive* *models*. Moreover, mental models (Gentner and Stevens 1983), misconceptions and bug libraries (Burton 1982) are not considered in our proposal; plain overlay models are used. In this way, a student model is a subset of the conceptual model described above. More specifically, a student model is approximated by means of a set of discrete probability distributions (one for each concept in the conceptual model). Other proposals such as (Paek and Chickering 2007) also use probability distributions to represent user models.

The rationale for such a drastic simplification is clear: it has been possible to define and implement sound and efficient procedures to update and handle these overlay models. These procedures are presented in Sect. 4 and tested in Sect. 6.

On the other hand, in psychometric literature, the student is often modeled by just a real number θ. It is clear that just a real number will seldom be a powerful model for tutoring; even for assessment tasks, the increasing interest in formative assessment creates the "... challenge of converting each examinee's test response pattern into a *multidimensional* student profile score report detailing the examinee's skills learned and skills needing study" (our emphasis) (Stout 2002). Nevertheless, proposals inside IRT defining and making use of multidimensional representations (e.g., Embretson 1991; Tam 1992; Segall 1996), although theoretically sound, quickly become impractical, due to the huge amount of data and computation needed in order to calibrate and handle the models. However, as shown in Sect. 6.4—at least for the simple cases tested—concept trees yield a feasible approach whose predictive power is comparable to that of multidimensional IRT.

3.3 Item pool

Items used in adaptive testing are stored in an item pool. According to Barbero (1999), the notion of an item pool has changed through time, although the underlying idea remains unchanged: a set of items, which measure the same trait or ability, and which are stored in such a way that, when required, the item which best fits student needs can be selected.

In our model, items are used as tools for diagnosing student knowledge, i.e., for determining which part of the concept tree is mastered by the student. Items are the entities which provide evidence about what he/she knows. Consequently, they are not restricted to only multiple-choice items or other classical formats used in paper-and-pencil tests. In our proposal, an item could be any provider of evidences about the student knowledge. To simplify, we will consider that the output of an item could be captured as a choice, allowing not only its correct or incorrect evaluation, but also partial credits, i.e., allowing a polytomous item modeling (Guzmán and Conejo 2004a).

The author of the course must assign each test item to a concept, i.e., he/she must determine which concept students must know, in order to answer the item. As a consequence, each concept will have an item pool assigned. In literature, one of the most popular approaches for assigning items to concepts is the *Q-Matrix* (Tatsuoka 1985). It is a matrix of binary values through which the course author indicates the cognitive tasks (in our case, concepts) needed to answer each item. Accordingly, the number of rows of this matrix is equal to the number of cognitive tasks involved in assessment, and the number of columns corresponds to the items available for diagnosis. We will use an equivalent approach described in the following paragraph.

First, the author must define the *association of an item to a concept*. If $Q_i \in \Pi$ and $C_j \in \Omega$, item Q_i is associated to concept C_j (or item Q_i *directly evaluates* concept C_j) if the author has stated that the probability of solving correctly Q_i depends on the knowledge of C_j. That is, the answer selected by a student for that item allows us to make inferences about his/her knowledge level in that concept. To represent this relation, we define the function $E_D : \Pi \times \Omega \rightarrow \{0,1\}$, $E_D(Q_i, C_j) = 1$ if Q_i is associated to C_j, otherwise $E_D(Q_i, C_j) = 0$. In Fig. 1, the association relation between an item and a concept has been represented by a line which joins both of them. For instance, item Q_1 is associated to concept C_{11}, i.e., $E_D(Q_1, C_{11}) = 1$. It was the course developer who directly added item Q_1 to the pool of C_{11}.

However, despite the fact that each item directly evaluates one and only one concept, it can *indirectly evaluate* several concepts (Guzmán and Conejo 2002), by taking into account the relations among concepts of the tree. Given an item Q_i and a concept C_j, the indirect evaluation function of a concept from an item, $E_I : \Pi \times \Omega \rightarrow \{0,1\}$, is defined as follows: $E_I(Q_i, C_j) = 1$ if $E_D(Q_i, C_j) = 0$, and exists $C_l \in \Omega$ such that $E_D(Q_i, C_l) = 1$ and $C_l \in \wp^{++}(C_j)$; otherwise $E_I(Q_i, C_j) = 0$. That is, Q_i indirectly evaluates C_j when there is another concept C_l evaluated directly by Q_i and between them there is an indirect inclusion relation. Notice that the item Q_i is not associated to the concept, i.e., does not belong directly to its item pool.

An item Q_i *evaluates* a concept C_j when Q_i evaluates C_j either directly or indirectly. The corresponding function will be $E : \Pi \times \Omega \rightarrow \{0,1\}$, $E(Q_i, C_j) = E_D(Q_i, C_j) + E_I(Q_i, C_j)$.

For example, in Fig. 1, item Q_6 directly evaluates concept C_{311}. It also supplies evidence about the student knowledge on the concept preceding C_{311}, i.e., on C_{31}. Applying the same reasoning, Q_6 also provides evidence about the parent of C_{31}, i.e., about C_3. Finally, as mentioned before, the whole course, including all its children, is considered a concept. Thus, Q_6 also provides evidence about the course. Items could directly evaluate either leaf, intermediate concepts or even the whole course. If they directly evaluate leaf or intermediate concepts, they also indirectly evaluate some other concepts. However, when an item directly evaluates an intermediate concept, knowledge about their descendants is not evaluated indirectly.

Until now, the description of the model has remained at a qualitative level. Let us describe now the quantitative aspects of our proposal. They are expressed by the response curves or characteristic curves which represent the association between an item and (the knowledge of) a concept.

We propose a response model with the following features:

- *Discrete*: Most IRT-based response models are continuous, i.e., the knowledge level (or generically, the latent trait) is usually measured in the real number domain. However, discrete models are more efficient from a computational perspective, since they do not use iterative algorithms (e.g. Newton–Raphson) to compute knowledge levels. Discretization entails that the knowledge level scale will be composed by K knowledge levels, from zero (absence of knowledge) to $K - 1$ (full knowledge). Thus, student knowledge distributions and characteristic curves are vectors of pairs of knowledge level/probability. Course developers can determine any number of knowledge levels. However, a constraint is imposed on our model: the number must be the same for every item and concept in a course.
- *Non-parametric*: Most response models use parametric approaches for modeling characteristic curves. The most commonly used are logistic functions. As a

drawback, they require more prior information to calibrate the curves. For this reason, we have adopted a non-parametric approach (Junker and Sijtsma 2001) to model characteristic curves. In general, the use of non-parametric approaches leads to relax the strong parametric model assumptions (Domshlak and Joachmis 2007). From the response model's perspective, non-parametric models economize the requirements of prior information and make calibration easier.

- *Polytomous*: One item will have one different characteristic curve per choice. We call them *Choice Characteristic Curves* (CCC). They represent the probability of selecting one choice given the student's knowledge level. To simplify the presentation, we will assume that each item has just one correct choice, and therefore the ICC will be equal to the CCC of the correct choice. We also consider a CCC for the blank response.
- *Hierarchical*: Each item will have a set of CCCs associated for each concept it evaluates. Let us assume an item Q_i such as is associated to a concept C_j. Since we allow evaluation of all the ascendants of a concept C_j, if n is the depth of C_j in the curricular tree, then there will be n sets of CCCs, one for each concept evaluated directly or indirectly by the item Q_i.

3.4 Test specifications

The third module provided by the course developer is a set of test specifications Θ. A test specification describes how assessment sessions will be generated. Its final goal is to obtain an estimation of the student knowledge in one or more concepts. When test (or course) developers specify tests they must answer the following questions:

1. *What to assess?*, that is, which elements of the concept tree will be scrutinized with respect to student knowledge.
2. *What is the student's initial state?* This information will be contained in his/her student model.
3. *How to assess?*, that is: (*a*) Which criterion is going to be used, i.e., how the student score is inferred from his/her performance in the test. (*b*) The level of detail: in how many knowledge levels are students going to be assessed. (*c*) The scope or the concepts involved in the test. And finally, (*d*) How are the assessment elements (items) sequenced?, that is, the item selection criterion used.
4. *When does the assessment end?*, since adaptive assessment criteria require an intended level for the accuracy of the estimation of student knowledge to be stated.

The answers to these questions are materialized in (i) a set of test configuration parameters; and (ii) a set of concepts to be assessed. However, as a collateral effect, and due to the hierarchical structure of the conceptual model and to the relationships among items and concepts, a test may also assess other concepts. For this reason we will define five relations between a test and a concept: direct evaluation of a test on a concept, indirect downward evaluation of a test on a concept, indirect upward evaluation of a test on a concept, indirect evaluation of a test on a concept and evaluation of a test on a concept. Notice that there is no direct relationship between tests and items. This relationship is established through the concepts of the conceptual model.

Let T_s be an assessment test, $T_s \in \Theta$, and C_j a concept, $C_j \in \Omega$. The *direct evaluation function of a test on a concept*, $\Phi_D : \Theta \times \Omega \rightarrow \{0, 1\}$, is defined as follows:

$\Phi_D(T_s, C_j) = 1$ if C_j is one of the concepts selected by the teacher to take part in the test T_s, $\Phi_D(T_s, C_j) = 0$ otherwise. For instance, in Fig. 1, test T_3 directly evaluates the concept C_{12}. In this figure, the connection is described by means of a line joining the test with the concepts it directly assesses.

We will impose the following restriction: in a test, several concepts can be assessed directly and simultaneously, but no aggregation relationship is allowed among these concepts. That is, for all $C_j, C_k \in \Omega$, if $\Phi_D(T_s, C_j) = 1$ and $\Phi_D(T_s, C_k) = 1$ then $C_j \notin \wp^+(C_k)$ and $C_k \notin \wp^+(C_j)$.

Let T_s be an assessment test ($T_s \in \Theta$) and C_j a concept ($C_j \in \Omega$). The *indirect downward evaluation function of a test on a concept*, $\Phi_{I\downarrow} : \Theta \times \Omega \to \{0, 1\}$, can be expressed as follows: $\Phi_{I\downarrow}(T_s, C_j) = 1$ if there exist $C_h \in \Omega$ such that $\Phi_D(T_s, C_h) = 1$ and $C_j \in \wp^+(C_h)$, $\Phi_{I\downarrow}(T_s, C_j) = 0$ otherwise.

Consequently, a test will indirectly evaluate downward all the concepts which descend from the concepts evaluated directly in this test. For instance, test T_3 of Fig. 1 evaluates indirectly downward the concepts C_{121}, C_{122}, C_{123}, C_{1231} and C_{1232}. Thus, for a student taking T_3, his/her knowledge level could be inferred simultaneously in each one of these concepts.

Given T_s and C_j, the *indirect upward evaluation function of a test on a concept* can be expressed by the function $\Phi_{I\uparrow} : \Theta \times \Omega \to \{0, 1\}$, $\Phi_{I\uparrow}(T_s, C_j) = 1$ if exists $C_h \in \Omega$ such that $\Phi_D(T_s, C_h) = 1$ and $C_h \in \wp^+(C_j)$, $\Phi_{I\uparrow}(T_s, C_j) = 0$ otherwise.

A test will indirectly evaluate upward all the concepts which are ascendants of those directly evaluated in the test. For example, test T_3 of Fig. 1 indirectly evaluates upward concept C_1 and the whole course simultaneously.

Given T_s and C_j, these two functions can be generalized in the *indirect evaluation function of a test on a concept*, $\Phi_I : \Theta \times \Omega \to \{0, 1\}$, $\Phi_I(T_s, C_j) = \Phi_{I\downarrow}(T_s, C_j) + \Phi_{I\uparrow}(T_s, C_j)$ Therefore, a test will indirectly evaluate a concept when this is done either downward or upward.

All previous functions can be generalized in the *evaluation function of a test on a concept*, $\Phi : \Theta \times \Omega \to \{0, 1\}$, defined as $\Phi(T_s, C_j) = \Phi_D(T_s, C_j) + \Phi_I(T_s, C_j)$. Thus, a test will evaluate a concept when this is done either directly, or indirectly upward or indirectly downward.

Following the terminology of Wang and Chen (2004), our proposal allows us to administer *between-item multidimensional tests*, i.e., tests where multiple concepts are assessed simultaneously. This kind of multidimensionality is achieved when unidimensional items assessing different concepts are administered in the same test. On the other hand, there are IRT multidimensional models for *within-item multidimensional tests*. In this kind of test, there are items whose ICCs are multidimensional. As will be shown in Sect. 6, our proposal is able to approximate the behavior of these items.

The next section describes in detail how the diagnosis procedure is carried out with our proposal. As mentioned before, when a test developer creates a test, he/she must supply some information about how the diagnosis phases are going to be carried out. He/she must configure: (a) Concepts directly evaluated; (b) the test evaluation mode (its alternatives are described in Sect. 4.1); (c) how the student knowledge level is inferred from his/her distributions (explained in Sect. 4.1.1); (d) item selection criterion (approached in Sect. 4.2); (e) the procedure used to initialize the student model (illustrated in Sect. 4.1.2); (f) when the test must finish (Sect. 4.3); and finally (g) the number of knowledge levels used to grade examinees.

4 Knowledge diagnosis procedure

From a general point of view, the algorithm for diagnosis is composed of the following steps (Guzmán and Conejo 2004b):

1. *Test item compilation*: Given a test specification T_s, the item pool used Ψ_s ($\Psi_s \subseteq \Pi$) is equal to the union of all pools from those concepts involved in T_s. More specifically, an item $Q_i \in \Pi$ belongs to Ψ_s when exists C_j ($C_j \in \Omega$) such that $E(Q_i, C_j) = 1$ and $\Phi(T_s, C_j) = 1$.
2. *Student cognitive model creation and initialization*: If not constructed yet, the diagnosis procedure creates and initializes a void instance of the student cognitive model that contains nodes representing the student knowledge of concepts involved in T_s. Our model provides several techniques which can be used to this end. For instance, creating a normal distribution centered on the average knowledge level, or generating a constant distribution. Test developers must decide which mechanism will be used to initialize the student models. Note that, for each node, the model keeps a discrete probability distribution.
3. *Adaptive test administration*: Finally, the student is administered the test.

This final stage is a generalization of the adaptive testing algorithm described in Sect. 2.1. This generalization is illustrated in Fig. 2 and its phases are enumerated below:

1. From set Ψ_s, the item which best contributes to the estimate of student knowledge is chosen. This selection leads to a double choice process. Firstly, the concept with the least accurate student current knowledge estimation is selected. Next, from the set of items evaluating this concept, the most informative one is chosen, i.e., the item which after being administered makes the most accurate knowledge estimation. Probabilistic distributions of student knowledge are used to carry out this selection procedure. When there is more than one most informative item, the selection process is accomplished randomly.
2. The selected item is removed from the test pool and it is posed to the student.
3. In terms of student response pattern, his/her knowledge distributions in corresponding concepts are updated according to the test assessment criterion.
4. The student knowledge level is inferred in the distribution updated in the previous step.
5. Steps 1 to 4 are repeated until test finalization criterion is satisfied for all the concepts involved in the test.

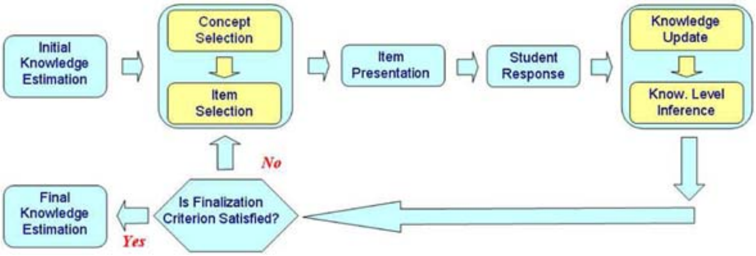

Fig. 2 Diagnosis model functioning diagram

4.1 Student knowledge estimation

During test administration, the distribution of student knowledge is updated each time he/she answers an item. This is usually done by using some variant of the bayesian technique created by Owen (1969).

In our proposal, several assessment modes have been defined depending on the concepts involved in the knowledge inference process. Consequently, in terms of the scope of the diagnosis procedure, the following assessment modes can be defined:

- *Aggregated assessment*: It is used when the goal is only to infer the student knowledge in those concepts directly evaluated in T_s, that is, those concepts $C_t \in \Omega$ such that $\Phi_D(T_s, C_t) = 1$. Thus, once the student has answered the i-th item Q^i, his/her knowledge distributions are updated according to the following formula:

$$P(\theta_t|u_1,\ldots,u_i) = \begin{cases} \|P(\theta_t|u_1,\ldots,u_{i-1})P_i(u_i|\theta_t)\| & \text{if } E(Q^i, C_t) = 1 \wedge \\ & \Phi_D(T_s, C_t) = 1 \\ P(\theta_t|u_1,\ldots,u_{i-1}) & \text{otherwise} \end{cases} \quad (1)$$

  where u_i represents the answer selected for ith item, and θ_t his/her knowledge level in concept C_t. CCC for this response pattern and for C_t is $P_i(u_i|\theta_t)$. $P(\theta_t|u_1,\ldots,u_{i-1})$ is the prior knowledge distribution in C_t, i.e., the distribution before he/she selected the answer for item ith. Double vertical lines indicate that the result must be normalized in order to ensure that the sum of all values is equal to one.

  For example, considering again Fig. 1, if test T_2 is administered under the aggregated assessment model, the student model would be formed by just two distributions: one for concept C_3 ($P(\theta_3|\overrightarrow{\mathbf{u_i}})$), and the other for C_4 ($P(\theta_4|\overrightarrow{\mathbf{u_i}})$), where $\overrightarrow{\mathbf{u_i}} = \{u_1,\ldots,u_{i-1},u_i\}$ is the response pattern matrix.
- *Complete assessment*: This assessment mode permits student knowledge inference in those concepts evaluated directly or indirectly downward. Once he/she answers an item, this response is evidence about his/her knowledge not only in the concept C_t directly evaluated in the test, but in those which are descendent from C_t. Thus, evidence is propagated from C_t to the concepts in the path between C_t and the concept C_r associated to the item ($C_r \in \wp^+(C_t)$), both included. Generally, for this assessment mode, the update process can be formally expressed as stated below:

$$P(\theta_t|u_1,\ldots,u_i) = \begin{cases} \|P(\theta_t|u_1,\ldots,u_{i-1})P_i(u_i|\theta_t)\| & \text{if } E(Q^i, C_t) = 1 \wedge \\ & \Phi_D(T_s, C_t) + \\ & \Phi_{I\downarrow}(T_s, C_t) = 1 \\ P(\theta_t|u_1,\ldots,u_{i-1}) & \text{otherwise} \end{cases} \quad (2)$$

  Considering again Fig. 1, the student model for test T_2 under the complete assessment model would comprise the following six knowledge distributions: those of concepts C_3, C_4, C_{31}, C_{32}, C_{311} and C_{312}, i.e., $P(\theta_3|\overrightarrow{\mathbf{u_i}})$, $P(\theta_4|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{31}|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{32}|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{311}|\overrightarrow{\mathbf{u_i}})$ and $P(\theta_{312}|\overrightarrow{\mathbf{u_i}})$.
- *Complete assessment with backpropagation*: Analogously to the former assessment mode, the student knowledge update can be extended to also affect the concepts which are ancestors of those directly evaluated in the test. Accordingly, evidence provided by the student item response is propagated not only to descendants, but also to ancestors. That is, to all those concepts which are ancestors of the one directly

evaluated in the test. Formally it can be expressed as follows:

$$P(\theta_t|u_1,\ldots,u_i) = \begin{cases} \|P(\theta_t|u_1,\ldots,u_{i-1})P_i(u_i|\theta_t)\| & \text{if } E(Q^i,C_t)=1\wedge \\ & \Phi(T_s,C_t)=1 \\ P(\theta_t|u_1,\ldots,u_{i-1}) & \text{otherwise} \end{cases} \quad (3)$$

If test T_2 of Fig. 1 is in complete assessment mode with backpropagation, the student model would comprise seven knowledge distributions: those for concepts $C_3, C_4, C_{31}, C_{32}, C_{311}, C_{312}$ and the concept which represents the whole course, i.e., $P(\theta_3|\overrightarrow{\mathbf{u_i}})$, $P(\theta_4|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{31}|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{32}|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{311}|\overrightarrow{\mathbf{u_i}})$, $P(\theta_{312}|\overrightarrow{\mathbf{u_i}})$ and $P(\theta_{\text{Course}}|\overrightarrow{\mathbf{u_i}})$.

This last assessment model is the most exhaustive, since it affects the greatest number of concepts. Despite this advantage, it must be managed with caution, because estimations are biased for the ancestors. When student knowledge is updated in concepts evaluated indirectly downward, this task is carried out for all concepts at the same level in the curricular hierarchy, because all of these concepts are descendants of those directly evaluated in the test. Otherwise, when knowledge distributions are updated in concepts evaluated indirectly upward, evidence is only obtained from one branch of the conceptual tree; consequently, the information inferred is partial and estimations might be biased.

This assessment mode can be useful as a starting point for a more accurate estimation with a balanced content. For instance, let us suppose that our student model is used in an ITS. Let us also consider the curricular structure of Fig. 1 and assume that each concept has assigned a sufficient number of items to properly accomplish the diagnosis of the student knowledge, i.e., each concept has more items assigned than those depicted in the figure. If a tutor is teaching the student in different concepts, once the instruction in, for example, concept C_{11} is finished, our diagnosis proposal will proceed to update his/her student model in this concept to help the instructional planner to obtain better learning strategies. Diagnosis of C_{11} will be done by means of a test on this concept. If this test performs complete assessment with backpropagation, this means that knowledge distributions in C_1 and in the whole course will also be updated. At this point, the assessment of C_1 and the course might be partial. Later on, after instruction in C_{12}, a test of this concept will be administered to update the student model. This last test might be done under complete assessment with backpropagation and using initial student knowledge distributions for C_1 and the course results obtained from the former test. As a consequence, after the test of C_{12}, estimation of C_1 will not be partial. Therefore, using the assessment model, it is not necessary to administer an additional test to diagnose the student knowledge in C_1. Furthermore, after administering tests on concepts C_2, C_3, etc., the global estimate of the course becomes more accurate.

4.1.1 Estimated knowledge inference

Once the student knowledge distributions have been updated, his/her knowledge level can be inferred using the two most popular techniques in adaptive testing, that is:

- *Expectation a posteriori (EAP)*, where the value corresponding to the student knowledge level is the expectation (or expected value) of probability distribution.

Formally, it can be expressed as follows:

$$EAP(P(\theta_t|\overrightarrow{\mathbf{u_n}})) = \sum_{k=0}^{K-1} kP(\theta_t = k|\overrightarrow{\mathbf{u_n}}) \tag{4}$$

– *Maximum a posteriori (MAP)*, when the value corresponding to the student knowledge level is the one with the greatest probability, i.e., the mode of the distribution. Formally, it can be expressed as shown below:

$$MAP(P(\theta_t|\overrightarrow{\mathbf{u_n}})) = \max_{0\leq k<K} P(\theta_t = k|\overrightarrow{\mathbf{u_n}}) \tag{5}$$

Estimations will be used by adaptive item selection criteria to determine the next item to be administered in the test; and by finalization criteria, to check if estimations are accurate enough.

4.1.2 Initial knowledge estimation

At the beginning of the diagnosis procedure, before answering any item, knowledge distributions must be initialized. If there is no additional information about the student knowledge, our proposal takes constant distributions, where all knowledge levels have the same probability.

When the model is used by an ITS, our system allows its initialization with a numerical value for the student knowledge in a concept (provided by the ITS). From this value, the system will discretize a normal probability distribution centered at this value.

Finally, as mentioned before, if the student had previously done any test about this concept, the system could use this result as a starting point for the diagnosis.

4.2 Item selection criteria

Unlike many adaptive testing-based models, our proposal is able to assess, in the same test, more than one concept simultaneously. To this end, the item selection procedure is carried out in two stages: *concept selection* and, from the items evaluating this concept, the *choice* of the one which best contributes to achieving a more accurate student knowledge estimation. Thus, the goal of adaptive item selection criteria is to minimize the number of items required to accurately estimate the examinee knowledge.

Throughout this section, the adaptive selection criteria of our proposal are described. Note that they are adaptations of standard methods (traditionally used for single concept tests) to multiconceptual tests. They are applied in a different way, in terms of the assessment mode used.

4.2.1 Maximum expected accuracy-based Bayesian method

This method is inspired by the proposal put forward by Owen (1975). He applied it to single concept assessment using dichotomous items. In our proposal, this criterion is extended to consider item selection in multiconceptual tests. Furthermore, this is an extension adapted to our polytomous response model.

The goal is, as in its original definition, to select that item which minimizes the expectation of the posterior student knowledge distribution variance. Let us suppose a student is doing a test which assesses a set φ of t concepts, $\varphi = \{C_1, C_2, \ldots, C_t\}$, where $\varphi \subseteq \Omega$. Let us also consider that he/she has answered previously $i-1$ items, and that $\overrightarrow{\mathbf{u_{i-1}}} = \{u_1, u_2 \ldots u_{i-1}\}$ is the student response pattern matrix. To calculate which is the next Q_j that must be posed in i-th position, for each item from the test pool, the expectation of the posterior knowledge distribution variance is computed, assuming that the selected item is Q_j. At the end, the item selected is the one leading to the least expectation value.

Formally, if we consider that each item has $W+1$ choices $\{u_{i0}, u_{i1}, \ldots, u_{iW}\}$, the item that must be administered next is the one that fulfills:

$$\min_{Q_j \in \Psi\_} \sum_{s=1}^{t} \sum_{w=0}^{W} \sigma^2[\rho_w(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}, u_j)]\upsilon_{jsw} \tag{6}$$

where

$$\rho_w(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}, u_j) = \begin{cases} \|P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}})P_{jw}(u_w|\theta_s)\| & \text{if } E(Q_j, C_s) = 1 \\ P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}) & \text{otherwise} \end{cases} \tag{7}$$

and

$$\upsilon_{jsw} = \begin{cases} P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}) \cdot P_{jw}(u_w|\theta_s) & \text{if } E(Q_j, C_s) = 1 \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

$\Psi\_$ is the set of items from the pool ($\Psi\_ \subseteq \Pi$) not administered yet and σ^2 the variance. u_j is the response pattern that the student might choose. $P(\theta_s|\overrightarrow{u_{i-1}})$ is his/her prior knowledge distribution in C_s, i.e., before answering the new item; and $\rho_w(\theta_s|\overrightarrow{u_{i-1}}, u_j)$ his/her posterior knowledge distribution in C_s (after administering the candidate item Q_j), assuming he/she will select the wth response pattern (u_w). Finally, note that υ_{jsw}, when item Q_j evaluating C_s, is equal to the scalar product between the prior knowledge distribution and the CCC of the w-ith response pattern.

In Owen's original proposal (*op.cit.*), the set of response patterns were only 0 (incorrect) and 1 (incorrect). However, our method takes into account the different response patterns the student could select. Likewise, this new reformulation considers that the response is able to infer the student knowledge in more than one concept.

The test assessment mode will condition the set of concepts involved in the test, and will therefore influence the test items. Consequently, in terms of the concepts which belong to φ, three different modalities of this model can be defined:

a) *Aggregated Bayesian selection*, where test T_s evaluates directly t concepts, i.e.:

$$\forall C_j, \quad C_j \in \varphi \;\Rightarrow\; \Phi_D(T_s, C_j) = 1 \tag{9}$$

b) *Complete Bayesian selection*, where t concepts are evaluated directly or indirectly downward, that is:

$$\forall C_j, \quad C_j \in \varphi \;\Rightarrow\; \Phi_D(T_s, C_j) = 1 \vee \Phi_{I\downarrow}(T_s, C_j) = 1 \tag{10}$$

c) *Complete with backpropagation Bayesian selection*, where the t concepts are those evaluated, either directly or indirectly:

$$\forall C_j, \quad C_j \in \varphi \;\Rightarrow\; \Phi(T_s, C_j) = 1 \tag{11}$$

4.2.2 Difficulty-based method

This method is an adaptation of the difficulty-based criterion of Owen (1975), carried out for our proposal. It consists of turning it into a two phase method. In the first phase, the concept with the least accurate knowledge estimation is selected, and subsequently, the item whose difficulty is nearest to the student level in the concept is chosen. Estimation accuracy is evaluated in terms of distribution variance. The larger the variance, the larger the distribution dispersion. Formally, the procedure accomplished by this selection mechanism can be expressed as stated below:

(1) *Selection of concept C_s*:

$$\max_{C_s \in \varphi} \sigma^2(P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}})) \tag{12}$$

(2) *Selection of item Q_j*:

$$\min_{Q_j \in \Psi_-} d(b_j, N), \quad \exists C_s, \quad C_s \in \varphi, \quad E(Q_j, C_s) = 1 \tag{13}$$

where

$$N = EAP(P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}))$$

or

$$N = MAP(P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}}))$$

depending on the student knowledge level inference mechanism used in the test; and

$$d(a, b) = |a - b|$$

The item *difficulty* (b_j) is one of the parameters which characterizes ICCs of dichotomous response models. Despite there being several definitions of this term, we will assume the one provided by IRT. Difficulty is the knowledge level for which the probability of answering an item correctly is the same as answering it incorrectly, in addition to the guessing factor. That is, the knowledge level whose probability is the ICC average value. Formally, it can be computed according to the following expression:

$$b_j = \min_{0<k<K} \left| ICC_i(\theta = k) - \frac{ICC_i(\theta = K-1) - ICC_i(\theta = 0)}{2} \right| \tag{14}$$

Observe that analogously to the former item selection method, this one has three different modalities in terms of the test assessment mode:

a) *Aggregated difficulty-based selection*, if concepts of set φ satisfy the condition expressed in 9;
b) *Complete difficulty-based selection*, when concepts fulfill condition 10;
c) *Complete with backpropagation difficulty-based selection*, in the case where they satisfy the constraint 11.

4.2.3 Maximum information-based method

This technique is based on calculating the item whose information function is maximum for the current student knowledge level. This criterion is the most popular both in dichotomous and polytomous models (Hontangas et al. 2000). One of the reasons for this popularity is that it is easy to use, since information functions can be computed a priori for all the items. As a consequence, if all these functions are calculated before starting the test, applying this criterion is only substituting the student knowledge level in the item information function, and selecting the one with the highest result.

There are several information function-based criteria for polytomous items. In our approach, we compute it following an adaption of the proposal by Dodd et al. (1995). As in difficulty-based criterion, this method is unable by itself to make a content-balanced item selection (in multiple concept tests). For this reason, this criterion must be applied in two stages: Firstly, the concept C_s with the least student knowledge estimation accuracy is selected. Secondly, the criterion selects the item evaluating this concept, whose information function for student knowledge level in C_s has the highest value.

In our approach, the definition of information function must be modified in order to take into account that an item can assess more than one concept. Consequently, a different information function is defined not only for an item, but also for the concept it assesses. Formally, the selection process can be formulated as follows:

(1) Concept C_s selection:

$$\max_{C_s \in \varphi} \sigma^2(P(\theta_s|\overrightarrow{\mathbf{u_{i-1}}})) \tag{15}$$

(2) Item Q_j selection:

$$\max_{Q_j \in \Psi_-} I_{js}(\theta_s) \tag{16}$$

where θ_s is the value, according to the student current knowledge level estimation in concept C_s; and where the function information $I_{js}(\theta_s)$ of item j for C_s can be computed in the following manner:

$$I_{js}(\theta_s) = \begin{cases} \sum_{w=0}^{W} \dfrac{P'_{jw}(u_w|\theta_s)^2}{P_{jw}(u_w|\theta_s)} & \text{if } E(Q_j, C_s) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

being $P'_{jw}(u_w|\theta_s)$ the derivative of the CCC corresponding to the response pattern u_w of Q_j and C_s.

As in the remaining item selection criteria, in terms of the test assessment mode, three different versions of this criterion can be defined:

a) *Aggregated maximum information-based method*, when concepts from set φ satisfy the condition expressed in 9;
b) *Complete maximum information-based method*, when the condition indicated in 10 is fulfilled;
c) *Complete with backpropagation maximum information-based method*, when the condition satisfied is 11.

4.3 Test finalization criteria

Finalization criteria determine when the item administration must finish. In adaptive testing, the most suitable finalization criterion is the one which ensures accurate assessment employing the least number of items. In our proposal, we have defined two adaptive criteria. In both of them student knowledge distributions are analyzed after each response. The goal is to determine whether finalization conditions are satisfied for all student probability distributions in concepts involved in the test.

- *Minimum Probability-based Criterion*: It considers a test must finish when probability of the student knowledge level goes beyond a certain threshold. Formally, let us consider an assessment test T_s of a certain course ($T_s \in \Theta$), which evaluates a set φ of t concepts $\varphi = \{C_1, C_2, \ldots, C_t\}$. Let δ and $\overrightarrow{\mathbf{u_i}}$ also be the threshold and the response pattern given to administered items, respectively. The constraint which must be fulfilled can be expressed as shown below:

$$\forall C_j, \quad C_j \in \varphi, \quad P(\theta_j = MAP(P(\theta_j|\overrightarrow{\mathbf{u_i}}))|\overrightarrow{\mathbf{u_i}}) > \delta \tag{18}$$

  where MAP is inferred from Eq. 5. Observe that this criterion is satisfied when, for all knowledge distributions, the maximum probability is greater than the threshold δ.
- *Estimation Accuracy-based Criterion*: Its goal is to achieve knowledge estimations with minimum variance. Note that analytically, the lesser the variance, the more peaked the distribution is. Accordingly, for lesser variance values, there is one knowledge level whose probability is considerably greater than the others. Therefore, when knowledge distribution variance is lesser than a certain threshold, this finalization criterion is met.

  This method can be formally expressed in the following manner: Given T_s, the constraint that must be fulfilled is the following:

$$\forall C_j, \quad C_j \in \varphi, \quad \sigma^2(P(\theta_j|\overrightarrow{\mathbf{u_i}})) < \delta \tag{19}$$

These two dynamic finalization criteria are convergent, i.e., ensure test finalization, when item pools are constructed properly. Additionally, it is recommendable to combine one of the former criterion with another static one such as:

- *Maximum number of item criterion*: It states the test must finish when the number of posed items is greater than a certain threshold.
- *Time limit criterion*: It is based on determining a time limit to complete a test. When this time limit is reached, finalization is forced.

The combination of a static criterion with a dynamic one provides a mechanism to avoid item overexposure and it prevents tests from consuming a lot of time.

5 Item calibration

Characteristic curve calibration is an important issue when defining a response model. To be able to administer adaptive tests, it is necessary to have a procedure available for inferring these curve values. Without an efficient calibration algorithm, adaptive tests are infeasible.

Calibration algorithms use information resulting from administering (non adaptively) tests to students with the items whose characteristic curves are not calibrated. This means all students take a test which is the same size and has the same items. Initially, they will be assessed with a heuristical assessment criterion, such as the percentage of items successfully answered.

To calibrate non-parametric response models, regression methods are often used (Habing 2001). Most of them are based on the following principle: given a set of observations X and a function m, the set of observations next to one point x, should contain information about the value of m in x. Accordingly, to estimate the value of $m(x)$ it is possible to use some kind of weighted average of the data closest to x (Simonoff 1996).

In keeping with this, one of the statistical techniques most frequently used because of its simplicity is *kernel smoothing* (Härdle 1992; Wand and Jones 1995). Using kernel smoothing to calibrate CCCs, the inference of each value is made by weighting the neighboring values. These weighted values are computed using a density function with a scale parameter. This value controls the influence of the neighboring values in the estimation of a certain value. It depends on the number of prior test sessions used in the calibration. The higher this number, the lower the value of scale parameter. The function and the parameter are called *kernel function* and *smoothing (or bandwidth) parameter*, respectively.

Ramsay (1991) is responsible for making this technique popular for IRT-based response model calibration. He proposed a simple calibration technique (Junker and Sijtsma 2001). In our proposal, taking this algorithm as a starting point, we have made some modifications improving it (Guzmán and Conejo 2005b). After exhaustive empirical studies, some stages of the original proposal have been modified. These modifications lead to better results. This new version of the algorithm is applied to calibrate the CCCs, for each item-concept pair. Consequently, the procedure for calibrating the CCCs of all the items which assess a certain concept C has the following steps:

1. *Prior student session compilation*: From all test sessions available, those involving the concept C are collected. The information for these sessions required for calibration is the answer that each student selected per item. Information from any other item not involving concept C is purged.
2. *Score computation*: For each student, his/her score is computed. This is done heuristically, since it is useful just for ordering the students' performance in the test. For instance, one of the ways of doing this is by calculating the percentage of items successfully answered.
3. *Score transformation*: The percentage obtained in the former phase is transformed into a temporary estimation of the student knowledge level. It is done by calculating the corresponding quartile in a standard normal distribution. After that, this value undergoes a linear transformation on the discrete scale used to represent the knowledge level.
4. *Smoothing Application*: Let N be the number of test sessions, the CCC of choice u_j from item i is computed as follows:

$$\forall k, \quad k \in \{0, 1, 2, \ldots, K-2, K-1\}, \quad P_{ij}(u_j|\theta = k) = \sum_{a=1}^{N} w_{ak} u_{ija} \tag{20}$$

   where $u_{ija} = 1$ indicates that a-th student selected the option u_j for item i. Otherwise, $u_{ija} = 0$. Furthermore, each weight w_{ak} is calculated as shown below:

$$w_{ak} = \frac{\kappa(\frac{\theta_a - \theta_k}{h})}{\sum_{b=1}^{N} \kappa(\frac{\theta_b - \theta_k}{h})} \tag{21}$$

   where θ_a is the ath student knowledge level calculated in the former step, θ_k the knowledge level whose probability is being computed, κ the kernel function and h the smoothing parameter. The underlying idea about the value of h is to minimize the mean square error of estimation (Douglas and Cohen 2001). There are a lot of studies about the most suitable value for the smoothing parameter (e.g., Ramsay 1991). According to (Guzmán 2005), for this proposal, the best value for h is around 0.8.

Regarding the kernel function, Ramsay (1991) proposes three alternatives: (a) *Gaussian function*: $\kappa(x) = e^{-x^2/2}$. (b) *Uniform function*: $\kappa(x) = 1$. (c) *Quadratic function*: $\kappa(x) = 1 - x^2$. In our proposal, after several empirical studies, we have found the best results are provided by the first one (Guzmán 2005).

5. *Students' knowledge level inference*: Using recently calibrated CCCs, the students' knowledge level is inferred. For this purpose, a maximum likelihood-based

   approach is used:

$$P(\theta|\overrightarrow{\mathbf{u_n}}) = \prod_{i=1}^{n} P_i(u_i|\theta) \tag{22}$$

   Once this procedure is applied, a knowledge distribution is obtained. The student knowledge level is inferred using MAP.

Example Let us assume that we want to calibrate an item i with three choices, where the correct one is the first. To simplify, let us also consider that five students have taken a conventional test (fixed number of items and evaluation expressed by means of percentage of success).

Let us assume a knowledge level scale with three knowledge levels, i.e., 0, 1 and 2. Students have been numbered with identifiers from 1 to 5. Student 1 gets 55% successful in the test, student 2 45%, and the other students: 25%, 80% and 10%, respectively. We also have the following information: students 1 and 4 selected the first choice (the correct one), student 2 the second one, student 3 the third, and finally, student 5 left the item blank (i.e., selected a fourth virtual choice).

Table 1 collects the results of the first three steps of the calibration algorithm. The first column contains the student identifier. Students have been ordered according to their performance in the test (represented in the second column). The third column contains the quartile corresponding to the score in a standard normal distribution. Finally, the last column shows the result of applying a linear transformation from the continuous knowledge level scale (i.e., $[-2.5, +2.5]$) to the discrete scale used in the example (i.e., $\{0, 1, 2\}$). Note that although we use a discrete scale in this step, the value obtained is not discretized.

Table 2 shows the different steps followed to compute the weights used to infer the CCCs. The first group of three columns (labeled with $\triangle\theta/h$) shows the difference between the knowledge level of the corresponding student (i.e., the value in the fourth column of Table 1) and the knowledge level indicated by the column (0 for the first

Table 1 Results of the different steps of the calibration algorithm for the example data (I part)

Student id	Score (%)	Quartile	Knowledge level
5	10	−0.967	0.601
3	25	−0.430	0.811
2	45	1.01E-07	0.980
1	55	0.430	1.149
4	80	0.967	1.359

Table 2 Results of the different steps of the calibration algorithm for the example data (II part)

Student id	$\triangle\theta/h$			$\kappa(\triangle\theta/h)$			Weight		
	0	1	2	0	1	2	0	1	2
5	0.751	−0.498	−1.748	0.754	0.883	0.216	0.312	0.186	0.094
3	1.014	−0.235	−1.485	0.597	0.972	0.331	0.247	0.205	0.145
2	1.225	−0.024	−1.274	0.471	0.999	0.443	0.195	0.210	0.194
1	1.436	0.186	−1.063	0.356	0.982	0.567	0.147	0.207	0.248
4	1.699	0.449	−0.800	0.235	0.903	0.725	0.097	0.190	0.317
				2.415	4.742	2.286			

Table 3 Calibrated CCCs for the item of the example

Choice	Knowledge levels		
	0	1	2
1	0.245	0.397	0.565
2	0.195	0.210	0.194
3	0.247	0.205	0.145
4	0.312	0.186	0.094

column and so on), divided by the smoothing parameter (assumed 0.8 for this example). The second group of three columns (labeled with $\kappa(\triangle\theta/h)$) shows the result of applying the kernel function (for this example, the Gaussian function). The values in the last row values are the sum of the values for all students. Finally, the last group of three columns (labeled 'weight') shows the weights which will be used to compute the CCCs. Weights have been computed by dividing the corresponding value in the second group of columns by the sum of all the values of the column. For example, the weight for student 5 (first row) for knowledge level 0, has been computed as follows: $0.754/2.415 = 0.312$.

Finally, Table 3 shows the calibrated CCCs. Each row contains the values of the characteristic curve of one of the choices of the item. Note that, as we mentioned before, the first choice is the correct one. The final row corresponds to the *don't know* virtual choice. Each value of this table has been computed according to equation 20. Remember that in equation 20 $u_{ija} = 1$ when student a selected choice j of item i. For instance, the first value of the first row ($P_{i1}(u_1|\theta = 0)$), has been calculated as follows: Two students (1 and 4) selected choice 1. As a result, the value of $P_{i1}(u_1|\theta = 0)$ is equal to the sum of the weights w_{10} and w_{40}, that is, $0.147 + 0.097 = 0.245$.

The estimated knowledge levels obtained in the fifth step of the algorithm can be used as feedback to recalibrate the CCCs. Accordingly, the former procedure would be repeated from the smoothing application step, until the values of the student knowledge levels and the CCCs values remain unchanged. To gauge the change between iterations, the mean square error between each CCC and its former estimation is used. A threshold is also used to determine when calibration must finish. If the sum of all errors is below this threshold (whose value is usually 0.0001), the calibration process stops. This procedure must be repeated for all the concepts of the curriculum. Once all of the CCCs have been calibrated, any time they are used (now in adaptive tests), they can be updated with these new test session results. As a result, this process could be repeated, automatically or on demand, getting more accurate estimations of the characteristic curves.

6 Evaluation

According to Scriven (1967), system or model evaluation can be conducted in two different ways, namely, formative or summative. In the first case, systems under development are studied to identify potential problems and to orient the modifications. This kind of evaluation makes sense during the design of a project or during the initial development stages. The goal is to improve the design of a system and/or its behavior.

Summative evaluation is conducted to question the construction, behavior or outputs of a certain system or proposal. Its challenge is to prove the adequacy of applied formalisms and techniques. Formal experiments are mainly used in summative evaluations, where the goals are to evaluate the effectiveness of a whole system (Twidale 1993).

We have conducted several empirical studies to make a summative evaluation of our proposal. All these experiments can be grouped in two sets. In the first one (Sects. 6.1, 6.2 and 6.3), curricular structure is composed by a single concept and simulated students are used. Some preliminary aspects are studied in such a simplified setting. First (Sect. 6.1) we have studied which is the best item selection criterion according to our model. This study is necessary, since the best criterion is used in the rest of the experiments. Next (Sect. 6.2), our discrete proposal is compared to the classical 3PL approach in terms of accuracy, number of items required for diagnosis and computational cost. An important part of our proposal is the calibration algorithm, since it is essential to make the knowledge engineering effort feasible. In the corresponding experiment (Sect. 6.3) we will try to quantify the performance of our algorithm.

The second set of experiments is more directly related to our proposal for hierarchical student modeling. First (Sect. 6.4) we compare an assumed 'accurate' bidimensional IRT-based model and a 'simplified' hierarchical model. The comparison is made in terms of the predictive power of each model, i.e., the capability of predicting student's success rate in a test. After that, in Sects. 6.5 and 6.6 we compare the performance of a one-dimensional IRT-based model and a hierarchical model. Note that this last study will be carried out using both simulated bidimensional students (Sect. 6.5) and real students (Sect. 6.6).

Our simulated students are software artifacts which try to emulate the behavior of real students when being administered tests. The use of simulated students is one of the strategies proposed by Murray (1993) for intelligent system evaluation and has been adopted by many researchers in the field (e.g. (VanLehn et al. 1998; Millán and Pérez de la Cruz 2002)). In our work, a simulated student will be described by its knowledge levels. Each student is generated with a prior knowledge level per leaf concept, which we call 'real' knowledge level in this concept. Note that this 'real' knowledge level is what adaptive tests will diagnose. These generated values follow normal distributions according to the knowledge level scale considered in the corresponding experiment.

On the other hand, it is necessary to generate items and tests whose administration will be also simulated. For each item, its ICC is generated following a three parameter logistic (3PL) function (Birnbaum 1968). This function is often used to model ICC in IRT, and particularly by student modeling researchers who apply IRT-based adaptive testing techniques (e.g. (Huang 1996; Millán and Pérez de la Cruz 2002)):

$$ICC(\theta) = c + (1 - c)\frac{1}{1 + e^{-1.7a(\theta - b)}} \tag{23}$$

Their parameters have the following meaning:

- *Discrimination* (a): This value is proportional to the slope of the curve. The greater the value, the higher the capability of the item will be to discern between students with higher knowledge levels and student with lower levels.
- *Difficulty* (b): It corresponds to the knowledge level at which the probability of answering correctly is the same as answering incorrectly, in addition to the guessing factor. The range of values allowed for this parameter is the same as the one allowed for the knowledge levels.
- *Guessing* (c): It is the student probability of answering the item correctly (by choosing a response randomly), when he/she has no knowledge at all.

These three values have been employed in simulations to make available pools where items have different features. They are also simulation input parameters. Once the ICC is determined, the CCC of correct choice is matched with the ICC. The characteristic curves of the other (incorrect) choices, are generated, for each knowledge level θ, according to the formula below:

$$P_i(\overrightarrow{u_{ij}}|\theta) = \frac{c_j}{1 + e^{-1,7a_j(\theta - b_j)}} \tag{24}$$

Values of parameters a_j, b_j and c_j, for jth choice are generated following normal distributions. These distributions are very similar to those used for ICC parameter generation. Note that in Eq. 24, discrimination (a_j) always takes negative values to ensure curves decrease monotonically.

The last curve models the blank answer and must be computed from the other CCCs. An item is itself a probabilistic space. This means that the sum of all CCCs must be equal to a vector of ones, and therefore this last curve is computed by subtracting from a vector of ones the sum of all incorrect CCCs and the correct CCC. If any value of the last curve is negative, then the curves are discarded and generated again. This procedure of generating CCCs leads to curves similar to those modeled by Thissen and Steinberg (1997).

The above process describes the generation of the characteristic curves of an item for the concept that it evaluates directly. CCCs for concepts evaluated indirectly are obtained after a calibration process, as will be explained later.

After generating an item pool, a test is constructed according to the values of certain parameters that indicate item selection criterion, assessment method, finalization criterion and its thresholds. In addition, concepts involved in the test must be specified.

When using the hierarchical model, the knowledge level of non leaf concepts must be inferred. To this end, a conventional test is administered to a simulated student (i.e., all items are administered to the student and his/her performance is measured by means of the percentage of success), using all the items which evaluate (either directly or indirectly) the corresponding concept. Student behavior in this test is determined by his/her 'real' knowledge level in the descendent leaf concepts. This procedure is repeated for the immediate upper hierarchy level until his/her 'real' knowledge had been inferred in all the concepts. For example, if we consider the curriculum of Fig. 1 initially the student knowledge is generated randomly for concepts: C_{11}, C_{121}, C_{122}, C_{1231}, C_{1232}, C_{311}, C_{312}, C_{32} and C_4. Once this step has been completed, knowledge in concept C_{123} will be computed administering a test of all items associated to concepts C_{1231} and C_{1232}. After this, knowledge in C_{12} is calculated from items associated to concepts C_{121}, C_{122} and C_{123}. This process is repeated to infer the knowledge level in the upper level concepts of the hierarchy. As a collateral effect, CCCs of items for concepts evaluated indirectly are also inferred during the calibration process.

Student behavior during a test is determined by means of their 'real' knowledge level and by the CCC values. Accordingly, when a student a is administered an item Q_i with $W+1$ choices $\{\overrightarrow{u_{i0}},\overrightarrow{u_{i1}},\ldots,\overrightarrow{u_{iW}}\}$, the answer he/she selects is determined as follows. First of all, his/her 'real' knowledge level (θ^r_{aj}) in the concept (C_j) the item Q_i assesses directly is obtained. Next, a random value between 0 and 1 is generated (v). After that, the answer selected (s) in the one which fulfills the condition below:

$$\min_{0\leq s\leq W}\sum_{w=0}^{s}P_i(\overrightarrow{u_{iw}}|\theta_j=\theta^r_{aj})\geq v \tag{25}$$

where $P_i(\overrightarrow{u_{iw}}|\theta_j=\theta^r_{aj})$ is the probability value of the wth CCC of item Q_i for the knowledge level θ^r_{aj}.

6.1 Study on diagnosis accuracy and efficiency

In our proposal we have extended some item selection criteria, usually applied to dichotomous IRT-based models to polytomous ones. For this reason, the challenge of this study was to compare these new polytomous item selection criteria in order to determine, under our model, which one is the most efficient. This comparison was made in terms of the number of items required for student diagnosis and the percentage of students whose knowledge was inferred correctly. In all these experiments curriculum was composed of one single concept.

6.1.1 Experiment 1: Comparison between bayesian and information function-based selection criteria in terms of item properties

In this experiment a sample of 100 simulated examinees were administered a test. The pool was composed of 300 items. The test assessed students in 12 knowledge levels. The 'real' students' knowledge level was generated randomly according a normal distribution centered in 5 (average value of knowledge level scale). The test finalization method was based on maximum expected accuracy, where the threshold was set at 0.001. The knowledge level inference process was carried out applying MAP criterion. Constant student knowledge distributions were assumed at the beginning of the test.

6.1.2 Results

Table 4 shows the results corresponding to different simulations. The first two columns are discrimination and guessing values used for characteristic curve generation. When a numerical value is shown, this means that all curves were generated with this value. When the word *"unif."* is shown, this means that guessing was generated randomly according to a normal distribution centered in 0.5. Difficulty was generated randomly according to a normal distribution whose center is the central value of the knowledge level scale, i.e., five.

Table 4 Comparison between bayesian and information-based selection criteria

Parameters		Bayesian				Information			
Discrim.	Guess.	Number of items administered	Standard deviation	Students correctly diagnosed (%)	Standard deviation deviation (%)	Number of items administered	Standard Standard	Students Correctly diagnosed (%)	Standard deviation (%)
0.2	0	12.73	2.91	99	0.7	37.39	3.38	99	0.8
0.2	0.25	23.8	5.47	99	0.7	52.04	4.90	99	1
0.2	0.5	61.40	8.81	97	1	87.85	6.17	97	1
0.2	0.75	162.67	10.45	87	3	174.76	6.49	89	2
0.2	unif.	36.38	7.30	98	1	72.34	7.33	98	1
0.7	0	13.61	2.95	99	0.4	43.29	2.68	98	1
0.7	0.25	26.23	7.18	98	1	56.72	6.08	97	1
0.7	0.5	60.54	9.08	97	1	89.77	6.76	97	1
0.7	0.75	166.08	9.61	91	2	180.18	6.66	93	1
0.7	unif.	39.08	6.29	97	1	68.93	7.71	98	0.9
1.2	0	9.64	0.66	98	1	27.28	1.66	99	0.5
1.2	0.25	16.75	4.38	98	1	33.50	1.71	99	0.5
1.2	0.5	36.23	6.79	98	1	55.06	4.58	98	0.9
1.2	0.75	130.18	12.02	94	3	146.92	8.77	94	2
1.2	unif.	19.57	2.28	99	1	44.36	3.25	98	1
1.9	0	6.81	0.32	99	1	20.55	1.38	99	0.5
1.9	0.25	12.48	0.70	99	0.6	22.61	1.90	98	0.8
1.9	0.5	21.93	2.60	99	0.7	36.50	2.44	99	0.6
1.9	0.75	84.11	9.44	96	1	103.34	10.91	96	2
1.9	unif.	11.70	0.58	98	0.9	28.37	1.57	99	0.7

Results illustrate that, for both criteria, diagnosis is highly accurate (the knowledge level was inferred correctly for at least 90% of individuals). The best criterion in terms of the item required for diagnosis is the bayesian. Observe that the greater the guessing, the higher the number of items needed. Additionally, the greater the discrimination, the lower that number will be.

6.1.3 *Experiment 2: Comparison between Bayesian and difficulty-based criteria in terms of test accuracy*

The previous experiment revealed Bayesian selection criterion is better than information function-based. In this second experiment, our Bayesian polytomous criterion was compared to the polytomous difficulty-based. According to their author (Owen 1969), both methods are very similar in their original dichotomous forms. In fact, the difficulty-based one is computationally more efficient than the Bayesian, but its performance is lower. The objective of this experiment is to quantify the differences in performance of both methods after being extended for our proposal.

Simulation conditions were analogous to those used in the former experiment. The difference rested on characteristic curve parameters. Guessing always took a zero value and discrimination was randomly generated according a normal distribution centered at 1.2. We use normal distributions centered at this value because a study of Kingsbury and Weiss (1979) demonstrates that, in an item pool, the mean discrimination factor is around 1.2. Difficulty was generated in the same way as in the former experiment.

6.1.4 *Results*

Table 5 shows the results obtained from the different simulations. Several simulations have been done in terms of the test finalization threshold (table first column). The third and fourth columns contain the average number of items administered per individual, and the fifth and sixth the percentage of examinees successfully diagnosed. Rows labelled with "std.dev." contain the standard deviation of the value placed in the previous row.

These results suggest bayesian criterion behaves better than difficulty-based. In addition, as can be seen, the minimum number of items required to achieve a good

Table 5 Comparison between Bayesian and difficulty-based selection criteria in terms of test accuracy

Threshold		Num. of items administered		Correct diagnoses	
		Bayesian	Difficulty-Based	Bayesian (%)	Difficulty-based (%)
0.1		1.44	3.34	39	45
	std.dev.	0.21	0.10	7	5
0.01		4.74	9.09	98	98
	std.dev.	0.17	0.36	1	2
0.001		8.79	14.01	99	99
	std.dev.	0.24	0.63	1	0.9
0.0001		14.50	18.13	100	99
	std.dev.	1.10	1.44	0	0.3
0.00001		20.87	22.51	100	100
	std.dev.	2.29	1.63	0	0

percentage of individuals correctly diagnosed (i.e., 99%) is only about nine items with the best item selection criterion.

6.2 Comparison versus 3PL continuous model

This set of experiments compares our proposal with the classical 3PL response model, i.e., the 3PL continuous model. For this reason, we consider again a curriculum composed of a single concept.

6.2.1 Experiment 1: Comparison in terms of number of items per test and diagnosis accuracy

In this analysis, 100 simulated students were administered an adaptive test whose items where extracted from a pool of 300 items. The goal was to diagnose the students' knowledge in a single concept using a scale of 12 knowledge levels. Each student was designated a 'real' knowledge level generated randomly from a normal distribution centered in 5 (i.e., the average value of knowledge level scale).

All items had three choices. Their characteristic curves were generated using normal distributions centered at the following values: 0 for the guessing factor, 1.2 for the discrimination factor and 5 for the difficulty.

Two different tests were administered. Student knowledge initial distributions were assumed constant. Both of them used the bayesian selection criteria and the test finalization method was based on maximum expected accuracy, where the threshold varied as will be seen in the results. The first test used our discrete and polytomous model and the second one the classical dichotomous, continuous and 3PL-based approach for administering adaptive tests.

6.2.2 Results

Table 6 shows the results of several simulations carried out varying the threshold of the test finalization criterion. As can be seen, the number of items is very similar (sometimes even less for our proposal) for lower thresholds. The most significant difference corresponds to a value of 0.00001. This can be easily explained since discretization entails a loss of precision and, as expected, the lesser the threshold, the greater the item number required for diagnosis with our proposal. However, diagnosis accuracy is considerably good (99.92%) with a mean of only 5.74 items per test.

6.2.3 Experiment 2: Comparison in terms of computing time

This study compares our proposal with the classical 3PL response model in terms of computing time. To this end, the time spent on item selection, knowledge update and finalization criterion checking was measured for both proposals. A computer with a 2 GHz Intel Pentium IV processor was used to carry out the simulations.

Hundred simulated students were administered a test with a pool of 300 items. Students' knowledge was measured on a scale of 12 levels. Student knowledge initial distributions were constant. Characteristic curve parameters were generated according to normal distributions centered at 1.2 (discrimination), 5 (difficulty) and 0.25 (guessing).

Table 6 Number of items and percentage of success diagnosing student knowledge

Threshold		Num. of items administered	Standard deviation	Students correctly diagnosed	Standard deviation
0.1	discr.	1.73	0.22	99.13	0.20
	cont.	2.40	0.05	98.79	0.06
0.01	discr.	5.74	0.20	99.92	0.03
	cont.	6.08	0.08	99.08	0.03
0.001	discr.	9.20	0.10	99.99	0.01
	cont.	9.37	0.24	99.04	0.06
0.0001	discr.	14.41	2.03	100.00	0.00
	cont.	13.05	0.26	99.45	0.03
0.00001	discr.	20.01	5.44	100.00	0.00
	cont.	17.29	3.04	99.98	0.03

Table 7 Computing time of item selection, knowledge update and finalization criterion checking in milliseconds

Number of knowl.levels		Continuous			Discrete		
		Item selection	Student knowledge update	Finaliz. criterion checking	Item selection	Student knowledge update	Finaliz. criterion cheking
2	time	371.71	0.20	0.88	4.73	0.04	0.17
	std.dev.	4.59	0.02	0.07	0.07	0.007	0.01
3	time	386.51	0.21	0.78	4.80	0.03	0.18
	std.dev.	4.91	0.01	0.04	0.05	0.004	0.006
6	time	379.02	0.21	0.69	5.14	0.02	0.19
	std.dev.	12.33	0.007	0.01	0.11	0.001	0.008
12	time	380.23	0.20	0.63	5.69	0.2	0.19
	std.dev.	9.39	0.002	0.01	0.17	0.001	0.008
24	time	382.89	0.19	0.61	6.62	0.02	0.27
	std.dev.	6.2	0.001	0.009	0.34	0.001	0.02
48	time	377.18	0.19	0.60	8.20	0.023	0.31
	std.dev.	4.48	0.0009	0.006	0.001	0.001	0.01
100	time	383.76	0.19	0.61	13.13	0.024	0.38
	std.dev.	3.08	0.001	0.007	1.58	0.0009	0.02

6.2.4 Results

Several simulations were made modifying the knowledge level scale (for our discrete proposal). Tests were administered using estimation accuracy-based finalization criterion (with a threshold of 0.001) and the bayesian item selection method. The knowledge level inference mechanism was MAP.

As can be seen in Table 7, time required for item selection was always notably higher when 3PL was used. In fact, our proposal reduced this time by about 97%. Note this reduction was not particularly sensitive to the number of knowledge levels used to diagnose the students' knowledge state. On the other hand, even though updates and finalization checking required less time with our proposal, in these cases, differences were not so significant.

These results are important in terms of scalability. Our goal is to implement this proposal in a web-based testing system. This system must be able to assess simultaneously hundreds (or even thousands) of students. In these situations, it is vital to provide a good performance avoiding delays which might cause students to become stressed. Thus, the computational efficiency has been one of the reasons for choosing a discrete approach.

6.3 Calibration algorithm performance

The goal of this experiment was to study the calibration algorithm behavior. In order to verify its accuracy, different size simulated student samples were used to calibrate a set of items. Their behavior in (non adaptive) tests was determined by means of their 'real' knowledge level and by the real CCCs. Once students took tests, items were calibrated, and accordingly their CCCs were inferred. Goodness of fit was measured in terms of mean square error. To validate these results, adaptive tests were administered to a new student sample using the learnt CCCs.

50 items were calibrated, each one of them with three choices and another one for the *don't know* answer. This means that a total of 200 CCCs were calibrated. The knowledge level scale was set to six. Real CCCs were generated using 3PL where parameters were generated randomly following normal distributions. The one for discrimination was centered at 1.2, difficulty at 2 and guessing at 0.15.

After calibration, adaptive tests were administered to a new sample of 100 students. Student knowledge initial distributions at the beginning of this test were constant. The item pool of these tests was only made up of the 50 calibrated items. In these tests, bayesian and estimation accuracy-based criteria were used for item selection and test finalization, respectively. The finalization threshold was 0.001 and the knowledge level inference mechanism was MAP.

6.3.1 Results

Table 8 collects the results of simulations done with different student sample sizes (first column). The second one is the number of students (from the calibration sample) whose knowledge was correctly inferred at the end of the learning process. The third column contains the number of students from the calibration sample whose knowledge was incorrectly inferred with an error of just one knowledge level (±1). The fourth column includes the sum of mean square errors of all CCCs. Finally, the last column is the percentage of students from the validation sample whose knowledge was correctly inferred by means of adaptive tests using calibrated CCCs.

The results suggest that even with a reduced sample of just 20 students, posterior diagnosis is acceptably accurate (around 90% of individuals correctly diagnosed). Note that in this posterior diagnosis, the remaining 10% of students were diagnosed with an error of ±1. As a consequence, by this algorithm with a reduced sample it is feasible to calibrate items and then administer adaptive tests.

6.4 Comparison between muldimensional IRT modeling versus the hierarchical approach

In this study we evaluate what happens when items are normally multidimensional, i.e., they depend on two or more latent trait (or concepts). We will assume a curric-

Table 8 Calibration algorithm performance

Number of students	Students correctly diagnosed	% of students correctly diagnosed	Students wrong diagnosis by one level	MSE	Posterior correct diagnosis (%)
20	14	70	6	19.62	95
50	42	84	8	15.37	98
100	89	89	11	13.23	98
300	236	78	64	12.65	95
500	424	84	76	11.63	99
1000	884	88	116	11.37	99
5000	4895	97	105	10.01	100
10000	9646	96	354	9.33	100

ulum structured into three concepts: one root C and two children C_1 and C_2. A set of 100 multidimensional items has been generated. We assumed that the response for each item depends on the knowledge of both children concepts and to this end item characteristic curves were generated following a multidimensional IRT model based on an extension of 3PL (Segall 1996). This model requires two discrimination factors (one per dimension), two difficulties and a guessing parameter for each item. These parameters were randomly generated according to normal distributions centered at 1.2 (discriminations), 2 (difficulties) and 0.25 (guessing factors). In this experiment, we considered a scale of 6 knowledge levels. To simplify, we assumed dichotomous items, i.e., items were only evaluated as either correct or incorrect. Once the item pool was constructed, we generated a set of 1000 simulated students, each of them with two 'real' knowledge levels θ_1 and θ_2 about C_1 and C_2, respectively.

The set of 1000 simulated students was administered all the items of the pool. The behavior of each student was determined according to Eq. 25 using the multidimensional item characteristic curves. Consequently, for each student a we obtained a vector of response patterns $\vec{u_a}$ and a success rate S_a^0 (i.e., number of items successfully answered over the total number of items).

The experiment tries to compute how good the estimation of the success rate is, when using not a truly bidimensional IRT description for each student, but a hierarchical description, assuming each item depends either on C_1 or on C_2. An initial control experiment was conducted by assigning items randomly to C_1 or C_2. The objective of this division was to study what happens if we simply ignore the multidimensional intrinsic behavior of the items without any source of information. By using the response patterns $\vec{u_a}$ previously obtained, we calibrated the items with the algorithm of Sect. 5, assuming that each item evaluated directly either C_1 or C_2. Once items were calibrated, a new test session was simulated using the same students previously generated with their known C_1 and C_2 values but simulating their answers to the items according to the corresponding unidimensional characteristic curves obtained from calibration. In this way a new set of success rates S_a^1 was obtained. Finally, we compared both sets of success rates and computed their correlation. The correlation coefficient was low: within the interval $(-0.065, 0.058)$ with $P = 0.05$.

Then a more interesting experiment was performed. Instead of assigning items randomly to C_1 or to C_2, the partition was done in terms of the discrimination factor of multidimensional characteristic curves. Consequently, for each item if discrimination for dimension C_1 was greater than discrimination for dimension C_2, such item was assigned to concept C_1 and vice versa. After this redistribution, we calibrated the items again and simulate a new test to obtain a new set S_a^2 of success rates. The correlation between the 'real' set S_a^0 and S_a^2 was very high: in the interval $(0.966, 0.973)$ with $P = 0.05$.

We can conclude that when items are actually multidimensional, a suitable hierarchical structuring of knowledge can be used that accurately reflects student behavior. On the other hand, if the partition of items is done at random, there is no correlation between predicted and 'real' behavior.

6.5 Comparison between 2-unidimensional IRT modeling versus the hierarchical approach

In this experiment, we assume that items are intrinsically unidimensional, but they alternatively depend on one or the other of the children concepts. The goal of this study is to establish whether in this case a hierarchical organization of concepts and items provides any advantage compared to considering all the items being assigned to a single root concept.

To this end, let us consider a curriculum like the one described in the former experiment (i.e., the three concepts C, C_1 and C_2). We will also use 1000 simulated students, whose knowledge level is generated randomly for concepts C_1 and C_2 between 0 and 5 (i.e., a knowledge scale of 6 levels). Once the students were generated, a pool of 200 items was constructed, 100 were assigned to concept C_1 and the others to concept C_2. All the items had three choices (only one of which was correct) and they allowed the (virtual) blank choice. Their CCCs were generated as explained at the beginning of Sect. 6. Parameters assumed for CCCs of correct choices were based on normal distributions centered at 1.2 (discrimination), 2 (difficulty) and 0.25 (guessing).

After generating students and items, students were administered all the items, obtaining their response patterns and success rates for concept C_1 and concept C_2. The sum of these two results is the success rate S_a^0 for C.

Using these data, items were calibrated with our algorithm in two situations. First (situation a), we applied the algorithm assuming that items evaluated directly either concept C_1 or concept C_2. We simulated a test session, using the newly calibrated curves and considering the same student knowledge levels to predict their responses. After this, we computed the success rate predicted for each student S_a^1, and compared it to the ratio initially generated. The correlation was within the interval $(0.955, 0.965)$ with $P = 0.05$. This is the benchmark to which the following result must be compared.

Then (situation b), all items were assumed to evaluate directly concept C. Again the calibration algorithm was applied, a session was simulated with the calibrated curves and a new set of success rates S_a^2 was computed. The correlation was now within the interval $(0.880, 0.905)$ with $P = 0.05$.

We may conclude that when the items are intrinsically unidimensional, but they alternatively depend on one or the other of the children concepts, the consideration of a single dimension predicts student behavior, although some predictive power is lost.

Another experiment could be carried out, considering the case of truly unidimensional items. Obviously if this is the case, the results of considering that the items are split into two item groups would be exactly the same as considering a single root concept, because there is a single latent trait.

6.6 Experiments with real students

The latter experiment was also carried out for two different real student populations. Data were collected in both cases through a web-based assessment tool, the SIETTE system (Conejo et al. 2004; Guzmán and Conejo 2005a). Items were evaluated as either 'correct' or 'incorrect' (this is equivalent to a polytomous approach where all items have only two choices.) Curriculum structure in both experiments was the same, i.e., a root concept C with two children concepts C_1 and C_2.

In this case, we did not know for sure if the items were intrinsically unidimensional or bidimensional, but they were assigned by the test developers to one of the leaf concepts in the three concept hierarchy. The previous experiments showed that in any case, considering two unidimensional leaf concepts will give better predictions than using a single root concept. We are empirically testing that this is also the case with real student data.

The first experiment was conducted using data from students studying *Botany* in the Polytechnic University of Madrid (Spain). In this experiment, the root concept was the global knowledge about *Botany*, whereas concepts C_1 and C_2 were *Angiospermae* and *Gymnospermae*, respectively. A conventional test (i.e., all items were posed to all students and their performance was measured in terms of ratio success) with a total of 20 items was administered. The test developer considered that 7 of those items evaluated directly the *Angiospermae* concept and the others the *Gymnospermae*.

A total of 172 students took this test. From these evidences we carried out the calibration in the same manner as in the former experiment. That is, first all items were calibrated taking into account the knowledge structuring done by the test developer (note that this is what we called *situation a* in the former experiment). Thus, items of concept C_1 were calibrated from the success ratio in such concept and the same for the items of concept C_2. Once the calibration was done, by using the results of this process (i.e., the calibrated CCCs and the student's knowledge level in each children concept), we tried to predict students' answer in the items of the test. For this purpose, we considered the student population as a group of simulated individuals and predict their answer accordingly. As a result, we obtained the success rate. This information was compared to the original success ratio and the correlation index was computed: it is within the interval $(0.59, 0.75)$ with $P = 0.05$.

The same was done for *situation b*. That is, we calibrated the items considering that all of them were assigned to concept C. After that, we predicted the behavior of students and also calculated the correlation index: it is within the interval $(0.62, 0.77)$ with $P = 0.05$.

It can be seen that the single concept model and the hierarchical model give similar results. There is no improvement by using the proposed hierarchical model, but at least, there is no significant loss of information.

This experiment was repeated using other real input data. We collected information on students who took a test of a LISP course. It was structured into a root concept with two children: *Functions* (C_1) and *Environments and Iteration* (C_2). A total of 93 students of Computer Science Engineering at the University of Málaga (Spain) were administered a conventional test of 12 items. The test developer of this course considered that 6 of these items evaluated directly concept C_1 and the other concept C_2. The same procedure was applied to these data. The results we obtained are the following. The correlation index for *situation a* was within the interval $(0.76, 0.89)$ with $P = 0.05$, and for *situation b* it was within the interval $(0.62, 0.82)$ with $P = 0.05$.

In this case, the results indicate that the hierarchical structure behaves better than considering just a single concept.

Unfortunately the results of the experiments with real students are not statistically significant at 95% confidence level due to the number of students involved, but at least they do not contradict experiments carried out with simulated students.

7 Related work

There are many systems which use testing for student knowledge inference. Most of them (e.g. DCG (Vassileva 1997), ELM-ART (Weber and Brusilovsky 2001) or ActiveMath (Melis et al. 2001)) use heuristic-based testing approaches. These heuristics can sometimes yield not completely reliable student models. On the contrary, there are other proposals which use IRT-based adaptive testing. Some of them use this kind of test just as it is, such as Lilley et al. (2004). Other approaches have tried to solve the problems of these tests for student modeling. For instance, the CBAT-2 algorithm (it stands for *Content-Balanced Adaptive Testing*) (Huang 1996) applies a mechanism which guarantees content-balanced selection in multiconceptual tests. Teachers must manually indicate the percentage of items which must be posed per concept. As a consequence, this strategy can lead to estimations which are not entirely accurate. Its response model is based on a dichotomous approach which uses the 3PL function for ICC modeling.

Bayesian Adaptive Tests (Millán and Pérez de la Cruz 2002) is a proposal which combines Bayesian Networks with adaptive testing. Student models are based on Bayesian Networks, where the student's knowledge state is represented by variables describing his/her knowledge level in a concept. Diagnosis is carried out by adaptive tests on the leaves of the network, whose results are propagated to other nodes. One of the disadvantages of this proposal is that the learning of network conditional probabilities is not considered by the model.

Another similar proposal is the Granularity–Bayes model (Collins et al. 1996) which combines CBAT-2 with a student model also based on Bayesian Networks. This approach inherits the disadvantages of CBAT-2 and in addition, their authors indicate that the use of Bayesian Networks makes this proposal computationally intensive. Certainly, nowadays Bayesian Networks technologies have evolved considerably. Now efficiency is not a problem thanks to the use of an approximate propagation algorithm (see (Castillo et al. 1997)) and dynamic Bayesian Networks (e.g. Mayo and Mitrovic 2001). Unfortunately, the authors of Granularity-Bayes appear not to have continued with their proposal.

Other authors (Desmarais and Pu 2005) have compared IRT and Bayesian Networks. They have developed a Bayesian Network-based proposal, called POKS, where items are linked with each other. These relations have been established without requiring any knowledge engineering effort, but are based on statistical information. According to the studies done by their authors, the performance of POKS in comparison to an IRT-based 2PL model (it is equivalent to the 3PL model, but assumes that guessing is always equal to zero) is comparable when classifying the students in two levels (master or non-master).

Finally, ACED (Shute et al. 2005) is an adaptive e-learning system for student diagnosis developed under the ECD framework (mentioned earlier). ACED is a promising prototype developed for students of middle school mathematics with and without visually disabilities. This system also employs a Bayesian Network for inferring students' models. Additionally, it uses an adaptive algorithm for task selection based on computing the one whose expected weight of evidence is maximum.

8 Conclusions and future work

In this paper we have presented a proposal for student modeling and diagnosis in conceptual domains structured in trees. Student knowledge is represented by means of probability distributions, one for each concept. The contributions of our work may be placed in two different fields, i.e., IRT and student modeling.

From the IRT perspective, we have presented a new IRT-based model for adaptive testing-based diagnosis of student knowledge. This proposal uses a feasible polytomous response model. Although there exist lots of polytomous models, none of them seem to be feasible because they have a lot of prior requirements. However, these kinds of models are able to extract more information from student answers than dichotomous ones (which only take into account whether the answer is correct or incorrect). Hopefully, this feature makes diagnosis more efficient regarding the number of items required, given a certain accuracy threshold. In fact, simulations performed suggest that this is so. For instance, for a threshold of 0.001, the diagnosis procedure requires fewer than nine items, and their results have a success ratio of 99%. Our proposal also includes an extension of item selection criteria to polytomous response modeling. Experiments have revealed that among all these criteria the best one in terms of diagnosis accuracy and number of item required is the Bayesian one.

Our IRT model is also discrete. This feature considerably reduces the computational cost, especially for the item selection stage. As we remarked before, this is a very important issue for us, since we have implemented this model as a web-based system and we want to use it massively (i.e., with huge sets of students simultaneously). In addition, when comparing our discrete response model with the continuous 3PL model, experimental data suggest that ours is more efficient computationally. This issue is particularly significant for item selection where the time is reduced by 97%. While 3PL requires on average around 380 ms., our proposal needs less than 14 ms.

This proposal also includes a calibration algorithm, based on kernel smoothing, with which the number of prior student sessions can be reduced, still obtaining reasonable estimations. Simulations again show that this algorithm is efficient, since with just a sample of only 20 test sessions, calibration results of 200 curves can be considered acceptable and useful for diagnosing correctly (a success rate of 95%).

From the point of view of user model structure, in this special issue, several proposals can be found, such as Bayesian Networks (Horvitz and Paek 2007), influence diagrams (Chickering and Paek 2007) (i.e. generalizations of Bayesian Networks) or other hierarchical networks (Nanas and Uren 2007). For our student models, we have suggested the use of hierarchically structured curricula. Some experiments show that this structuring can produce a more accurate diagnosis, in comparison to considering only one single concept. The degree of improvement depends on the goodness of the item partition.

Our diagnosis procedure also simultaneously evaluates the student knowledge in several concepts by administering just one test, obtaining accurate enough estimations in all concepts. This can be done by adding a new phase to the item selection. In this phase the concept whose estimation is less accurate is selected before the item is chosen. Note that one of our selection criteria (the bayesian one) does not need this previous stage, since it uses the accuracy of posterior knowledge distribution to make the item selection. Therefore, we provide *inter-item multidimensional tests*.

In addition, regarding multidimensional IRT, the experiments conducted have revealed that our hierarchical model is very close to multidimensional IRT in terms of prediction of student behavior.

Certainly, we are aware of the fact that our proposal for student modeling lacks many desirable features (e.g. it only considers aggregation relationships, no misconceptions are modeled, etc.). The main reason for this simplification is that we have tried to achieve a trade-off between student modeling and diagnosis based on well-founded formal theories.

Concerning future work, we plan to intensively apply our system to the construction of diagnosis modules of real ITSs. For instance, at the beginning of instruction, it could be useful to initialize the student model by means of a pretest; during the instruction, to update the student model; and at the end of the instruction, to provide a global snapshot of the state of the knowledge. We are also currently working on the development of a task model based on the library of sophisticated items provided by our system SIETTE (Guzmán and Conejo 2004a). Furthermore, the use of a polytomous model allows us to extract statistical information about the choices which have been selected incorrectly by students. Using this information, we could perhaps model misconceptions.

Finally, we must mention that all features described in this paper have been implemented in our Web-based diagnosis tool SIETTE (http://www.lcc.uma.es/siette).

Acknowledgements We would like to thank to those who have contributed directly or indirectly to the implementation of SIETTE and to those who have used and evaluated the prototypes. Special thanks go to the anonymous reviewers for their helpful comments and to Valerie Shute for her suggestions and for the bibliographical references she has supplied to us.

References

Barbero, M.I.: Gestión informatizada de bancos de ítems. In: Olea, J., Ponsoda, V., Prieto, G. (eds.) Tests Informatizados: Fundamentos y aplicaciones, pp. 63–83. Pirámide (1999)

Birnbaum, A.: Some latent trait models and their use in inferring an examinee's mental ability'. In: Statistical Theories of Mental Test Scores. Addison-Wesley: Reading, MA (1968)

Bock, R.D.: The nominal categories model. In: van der Linden, W.J., Hambleton, R.K., (eds.) Handbook of Modern Item Response Theory, pp. 33–49. Springer Verlag New York (1997)

Burton, R.: Diagnosing bugs in a simple procedural skill. In: Sleeman, D., Brown, J. (eds.) Intelligent Tutoring Systems, pp. 57–183. Academic Press (1982)

Carver, C.A., Howard, R.A., Lane, W.D.: Enhancing student learning through hypermedia courseware and incorporation of student learning styles. IEEE Trans. Educ. **42**(1), 33–38 (1999)

Castillo, E., Gutiérrez, J.M., Hadi, A.S.: Expert Systems and Probabilistic Network Models. Monographs in Computer Science. Springer-Verlag (1997)

Chickering, D.M., Paek, T.: Personalizing influence diagrams: applying online learning strategies to dialogue management. User Modeling and User-Adapted Interaction (in this issue) (2007)

Collins, J.A.: Adaptive testing with granularity. Master's thesis, University of Saskatchewan (1996)

Collins, J.A., Greer, J.E., Huang, S.X.: Adaptive assessment using granularity hierarchies and bayesian nets. In: Frasson, C., Gauthier, G., Lesgold, A. (eds.) Proceedings of the 3rd International Conference on Intelligent Tutoring Systems. ITS 1996. Lecture Notes in Computer Science, No. 1086, pp. 569–577. Springer Verlag, New York (1996)

Conati, C.: Probabilistic assessment of user's emotions in educational games. App. Artif. Intell. **16**, 555–575 (2002)

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez de la Cruz, J.L., Ríos, A.: SIETTE: a Web-based tool for adaptive testing. J Artif. Intell. Educ. **14**, 29–61 (2004)

Desmarais, M.C., Pu, X.: a bayesian inference adaptive testing framework and its comparison with item response theory. Int. J Artif. Intell. Educ. **15**, 291–323 (2005)

Dodd, B.G., De Ayala, R.J., Koch, W.R.: Computerized adaptive testing with polytomous items. Appl. Psychol. Measure. **19**(1), 5–22 (1995)

Domshlak, C., Joachmis, T.: Efficient and non-parametric reasoning over user preferences. User Model. User-adap. Interac. (in this issue) (2007)

Douglas, J., Cohen, A.: Nonparametric item response function estimation for assessing parametric model fit. Appl. Psychol. Measure. **25**(3), 234–243 (2001)

Embretson, S.E.: A multidimensional latent trait model for measuring learning and change. Psychometrika **56**(3), 495–515 (1991)

Embretson, S.E., Reise, S.P.: Item Response Theory for Psychologists. Lawrence Erlbaum, Mahwah, NJ (2000)

Gentner, D., Stevens, A.L. (eds.): Mental Models. Lawrence Erlbaum, Hillsdale, NJ (1983)

Greer, J.E., McCalla, G.: Granularity-based reasoning and belief revision in student models. In: Greer, J.E., McCalla, G. (eds.) Student Modelling: The Key to Individualized Knowledge-Based Instruction, vol. 125, pp. 39–62. Springer Verlag, New York (1994)

Guzmán, E.: Un modelo de evaluación cognitiva basado en Tests Adaptativos Informatizados para el diagnóstico en Sistemas Tutores Inteligentes. Unpublished doctoral dissertation, Dpto. Lenguajes y Ciencias de la Computación. E.T.S.I. Informática., Universidad de Málaga (2005)

Guzmán, E., Conejo, R. Simultaneous evaluation of multiple topics in SIETTE. In: Cerri, S., Gouardres, G., Paraguacu, F. (eds.) Proceedings of the 6th International Conference on Intelligent Tutoring Systems (ITS 2002). Lecture Notes in Computer Science, No. 2363, pp. 739–748. Springer Verlag, New York (2002)

Guzmán, E., Conejo, R.: A library of templates for exercise construction in an adaptive assessment system. Technol. Instruct. Cogn. Learning (TICL) **2**(1–2), 21–43 (2004a)

Guzmán, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. In: Lester, J.C., Vicari, R.M., Paraguau, F. (eds.) Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004). Lecture Notes in Computer Science, No. 3220, pp. 12–21. Springer Verlag, New York (2005b)

Guzmán, E., Conejo, R.: Self-assessment in a feasible, adaptive web-based testing system. IEEE Trans. Educ. **48**(4), 688–695 (2005a)

Guzmán, E., Conejo, R.: Towards efficient item calibration in adaptive testing. In: Ardissono, L., Brna, P., Mitrovic, A. (eds.) Proceedings of the 10th International Conference on User Modeling (UM 2005). Lecture Notes in Artificial Intelligence, No. 3538, pp. 414–418. Springer Verlag, New York (2005b)

Habing, B.: Nonparametric regression and the parametric bootstrap for local dependence assessment. Appl. Psychol. Measure. **25**(3), 221–233 (2001)

Härdle, W.: Applied Nonparametric Regression. University Press, Cambridge (1992)

Holt, P., Dubs, S., Jones, M., Greer, J.: The state of student modelling. In: Greer, J.E., McCalla, G. (eds.) Student Modelling: The Key to Individualized Knowledge-Based Instruction, vol. 125, pp. 3–35. Springer Verlag, New York (1994)

Hontangas, P., Ponsoda, V., Olea, J., Abad, F.: Los test adaptativos informatizados en la frontera del siglo XXI: una revisión. Metodología de las Ciencias del Comportamiento **2**(2), 183–216 (2000)

Horvitz, E., Paek, T.: Complementary computing: policies for transferring callers from dialog systems to human receptionists. User Model. User-Adapted Interact. (in this issue) (2007)

Huang, S.X.: A content-balanced adaptive testing algorithm for computer-based training systems. In: Frasson, C., Gauthier, G., Lesgold, A. (eds.). Lecture Notes in Computer Science 1086. Proceedings of the 3rd International Conference on Intelligent Tutoring Systems. ITS 1996, pp. 306–314. Springer Verlag New York (1996)

Junker, B.W., Sijtsma, K.: Nonparametric item response theory in action: an overview of the Special Issue. Appl. Psychol. Measure. **25**(3), 211–220 (2001)

Kingsbury, G.G., Weiss, D.J.: An adaptive testing strategy for mastery decision. Psychometric Method Program Research Report 79-5, Department of Psychology. University of Minnesota (1979)

Lilley, M., Barker, T., Britton, C.: The development and evaluation of a software prototype for computer-adaptive testing. Comp. Educ. **43**, 109–123 (2004)

Lord, F.M.: Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates, Hillsdale, NJ (1980)

Mayo, M., Mitrovic, A.: Optimising ITS behaviour with bayesian networks and decision theory. Int. J. Artif. Intell. Educ. **12**, 124–153 (2001)

Melis, E., Andres, E., Bndenbender, J., Frischauf, A., Goguadze, G., Libbrecht, P., Pollet, M., Ullrich, C.: ACTIVEMATH: a generic and adaptive web-based learning environment. Int. J. Artif. Intell. Educ. **12**, 385–407 (2001)

Millán, E., Pérez de la Cruz, J.L.: Diagnosis algorithm for student modeling diagnosis and its evaluation. User Model. User-adapted Interact. **12**(2–3), 281–330 (2002)

Mislevy, R.J., Steinberg, L.S., Almond, R.G.: On the roles of task model variables in assessment design. Technical Report CSE Technical Report 500, Education Testing Service (1999)

Mislevy, R.J., Steinberg, L.S., Almond, R.G.: Evidence-centered assessment design. Technical Report A Submission for the NCME Award for Technical or Scientific Contributions to the Field of Educational Measurement, Education Testing Service (2000)

Muraki, E.: A generalized partial credit model: Application to an EM algorithm. Appl. Psychol. Measure. **16**, 159–176 (1992)

Murray, T.: Formative qualitative evaluation for 'Exploratory' ITS research. Int. Artif. Intell. Educ.: Special Issue on Evaluation **4**(2/3), 179–207 (1993)

Nanas, N., Uren, V.: Exploting term dependencies for multi-topic information filtering with single user profile. User Model. User-Adapted Interact. (in this issue) (2007)

Owen, R.J.: A Bayesian approach to tailored testing. Research Report 69-92, Educational Testing Service (1969)

Owen, R.J.: A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. J. Am. Stat. Associ. **70**(350), 351–371 (1975)

Paek, T., Chickering, D.M.: Improving command and control speech recognition on mobile devices: using predictive user models for language modeling. User Model. User-Adapted Interact. (in this issue) (2007)

Papanikolaou, K.A., Grigoriadou, M., Kornikolakis, H., Magoulas, G.D.: Personalizing the interaction in a web-based educational hypermedia system: the case of INSPIRE. User Model. User-Adapted Interact. **13**, 213–267 (2003)

Ramsay, J.O.: Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika **56**, 611–630 (1991)

Reye, J.: A belief net backbone for student modelling. In: Cerri, S.A., Gouardres, G., Paraguacu, F. (eds.) Proceedings of the 6th International Conference on Intelligent Tutoring Systems (ITS 2002). Lecture Notes in Computer Science, No. 2363, pp. 596–604. Springer Verlag, New York (2002)

Samejima, F.: The graded response model. In: van der Linden, W.J., Hambleton, R.K., (eds.) Handbook of Modern Item Response Theory, pp. 85–100. Springer Verlag, New York (1997)

Schank, R.C., Cleary, C.: Engines for Education. Lawrence Erlbaum Associates, Hillscale, NJ (1994)

Scriven, M.: The methodology of evaluation. In: Stake, R.E., (ed.) Perspectives of Curriculum Evaluation. Rand McNally, Chicago, IL (1967)

Segall, D.O.: Multidimensional adaptive testing. Psychometrika **61**(1), 331–354 (1996)

Self, J.A.: Formal approaches to student modeling. In: Greer, J.E., McCalla, G. (eds.) Student Modeling: The Key to Individualized Knowledge-Based Instruction, vol. 125, pp. 295–352. Springer Verlag, New York (1994)

Shute, V.J., Graf, E.A., Hansen, E.G.: Technology-Based Education: Bringing Researchers and Practicioners Together, Chapt. Design Adaptive, Diagnostic Math Assessments for Sighted and Visually Disabled Students, pp. 169–202. Information Age Publishing (2005)

Simonoff, J.S.: Smoothing Methods in Statistics. Springer-Verlag, New York (1996)

Stout, W.: Psychometrics: from practice to theory and back. Psychometrika **67**, 485–518 (2002)

Tam, S.S.: A comparison of methods for adaptive estimation of a multidimensional trait. Ph.D. thesis, Graduate School of Arts and Science, Columbia University. Order number: 9221219 (1992)

Tatsuoka, K.: A probabilistic model for diagnosing misconceptions in the pattern classification approach. J. Educ. Stat. **12**(1), 55–73 (1985)

Thissen, D., Steinberg, L.: A response model for multiple choice items. In: van der Linden, W.J., Hambleton, R.K. (eds.) Handbook of modern item response theory, pp. 51–65. Springer Verlag, New York (1997)

Twidale, M.: Redressing the balance: the advantages of informal evaluation techniques for intelligent learning environments. Int. J. Artif. Intell. Educ. Special Issue on Evaluation **4**(2/3), 155–178 (1993)

van der Linden, W.J., Glas, C.A.W.: Computerized Adaptive Testing: Theory and Practice. Kluwer Academic Publishers (2000) van der Linden, W.J., Pashley, P.J.: Item selection and ability estimation in Adaptive Testing. In: van der Linden, W.J., Glas, C.A.W. (eds.) Computerized Adaptive Testing: Theory and Practice, pp. 1–26. Kluwer Academic Publisher, Dordrecht

VanLehn, K., Niu, Z., Siler, S., Gertner, A.S.: student modeling from conventional test data: a Bayesian approach without priors. In: Goettl, B., Redfield, C.L., Halff, H.M., Shute, V.J. (eds.) Proceedings of 4th International Conference on Intelligent Tutoring Systems. ITS'98. Lecture Notes in Computer Science, vol. 1452. pp. 434–443 (1998)

Vassileva, J.: Dynamic course generation on the WWW. In: du Bolay, B., Mizoguchi, R. (eds.) Knowledge and Media in Learning Systems. Proceedings of the 8th World Conference on Artificial Intelligence in Education AIED'97. pp. 498–505 (1997)

Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)

Wang, W.C., Chen, P.H.: Implementation and measurement efficiency of multidimensional computerized adaptive testing. Appl. Psychol. Measure. **28**(5), 295–316 (2004)

Weber, G., Brusilovsky, P.: ELM-ART: an adaptive versatile system for web-based instruction. Int. J. Artif. Intell. Educ. **12**, 351–383 (2001)

Zukerman, I., Albrecht, D.: Predictive statistical models for user modeling. User Model. User-Adapted Interact. **11**, 5–18 (2001)

Authors' vitae

Dr. Eduardo Guzmán is Assistant Professor of Computer Languages and Systems at the University of Málaga. He received his M.Sc. and Ph.D. degrees in Computer Science from the same University. His primary interests lie in the areas of intelligent tutoring systems, student modeling and adaptive testing. This paper is based on his Ph.D. and his ongoing research on adaptive testing.

Dr. Ricardo Conejo is Associate Professor of Computer Languages and Systems at the University of Málaga. He is the director of the research group IAIA (Research and Applications of Artificial Intelligence). Dr. Conejo received a Ph.D. degree in Civil Engineering from the Polytechnic University of Madrid. He has worked in several areas of artificial intelligence, including fuzzy sets, model-based diagnosis, planning and problem solving, and agents. His current interests lie in the areas of intelligent tutoring systems, student modeling and adaptive testing.

Dr. José-Luis Pérez-de-la-Cruz is Associate Professor of Computer Science and Artificial Intelligence at the University of Málaga. He received master and doctoral degrees in Engineering from the Polytechnic University of Madrid and a master degree in Law from the Universidad Nacional de Educación a Distancia. He has worked in several areas, including heuristic search, multiagent systems, and engineering and educational applications of AI. He has authored about one hundred technical papers and edited a book on Multiagent Systems.