# A Model for Student Knowledge Diagnosis Through Adaptive Testing*

Eduardo Guzmán and Ricardo Conejo

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

**Abstract.** This work presents a model for student knowledge diagnosis that can be used in ITSs for student model update. The diagnosis is accomplished through *Computerized Adaptive Testing* (CAT). CATs are assessment tools with theoretical background. They use an underlying psychometric theory, the *Item Response Theory* (IRT), for question selection, student knowledge estimation and test finalization. In principle, CATs are only able to assess one topic for each test. IRT models used in CATs are dichotomous, that is, questions are only scored as correct or incorrect. However, our model can be used to simultaneously assess multiple topics through content-balanced tests. In addition, we have included a polytomous IRT model, where answers can be given partial credit. Therefore, this polytomous model is able to obtain more information from student answers than the dichotomous ones. Our model has been evaluated through a study carried out with simulated students, showing that it provides accurate estimations with a reduced number of questions.

## 1 Introduction

One of the most important features of Intelligent Tutoring Systems (ITSs) is the capability of adapting instruction to student needs. To accomplish this task, the ITS must know the student's knowledge state accurately. One of the most common solutions for student diagnosis is testing. The main advantages of testing are that it can be used in quite a few domains and it is easy to implement. Generally, test-based diagnosis systems use heuristic solutions to infer student knowledge. In contrast, *Computerized Adaptive Testing* (CAT) is a well-founded technique, which uses a psychometric theory called *Item Response Theory* (IRT). The CAT theory is not used only with conventional paper-and-pencil test questions, that is, questions comprising a stem and a set of possible answers. CAT can also include a wide range of exercises [5]. On the contrary, CATs are only able to assess a single atomic topic [6]. This restricts its applicability to structured domain models, since when in a test more than one content area is being assessed, the test is only able to provide one student

* This work has been partially financed by LEActiveMath project, funded under FP6 (Contr. Nº 507826). The author is solely responsible for its content, it does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

J.C. Lester et al. (Eds.): ITS 2004, LNCS 3220, pp. 12–21, 2004.

knowledge estimation for all content areas. In addition, in these multiple topic tests, the content balance cannot be guaranteed.

In general, systems that implement CATs use dichotomous IRT based models. This means that student answers to a question can only be evaluated as correct or incorrect, i.e. no partial credit can be given. IRT has defined other kinds of response models called polytomous. These models allow giving partial credit to item answers. They are more powerful, since they make better use of the responses provided by students, and as a result, student knowledge estimations can be obtained faster and more accurately. Although in literature there are a lot of polytomous models, they are not usually applied to CATs [3], because they are difficult to implement.

In this paper, a student diagnosis model is presented. This model is based on a technique [4] of assessing multiple topics using content-balanced CATs. It can be applied to declarative domain models structured in granularity hierarchies [8], and it uses a discrete polytomous IRT inference engine. It could be applied in ITS as a student knowledge diagnosis engine. For instance, at the beginning of instruction, to initialize the student model by pretesting; during instruction, to update the student model; and/or at the end of instruction, providing a global snapshot of the state of knowledge.

The next section is devoted to showing the modus operandi of adaptive testing. Section 3 supplies the basis of IRT. Section 4 is an extension of Section 3, introducing polytomous IRT. In Section 5 our student knowledge diagnosis model is explained. Here, the diagnosis procedure of this model is described in detail. Section 6 checks the reliability and accuracy of the assessment procedure through a study with simulated students. Finally, Section 7 discusses the results obtained.

## 2 Adaptive Testing

A CAT [11] is a test-based measurement tool administered to students by means of a computer instead of the conventional paper-and-pencil format. Generally, in CATs questions (called “items”) are posed one at a time. The presentation of each item and the decision to finish the test are dynamically adopted, based on students' answers. The final goal of a CAT is to estimate quantitatively student knowledge level expressed by means of a numerical value. A CAT applies an iterative algorithm that starts with an initial estimation of the student’s knowledge level and has the following steps: 1) all the items (that have not been administered yet) are examined to determine which is the best item to ask next, according to the current estimation of the student’s knowledge level; 2) the item is asked, and the student responds; 3) in terms of the answer, a new estimation of his knowledge level is computed; 4) steps 1 to 3 are repeated until the defined test finalization criterion is met. The selection and finalization criteria are based on theoretically based procedures that can be controlled with parameters. These parameters define the required assessment accuracy. The number of items is not fixed, and each student usually takes different sequences of items, and even different items. The basic elements in the development of a CAT are: 1) The *response model associated to each item*: This model describes how students answer the item depending on their knowledge level. 2) The *item pool*: It may contain a large number of correctly calibrated items at each knowledge level. The better the quality of the item pool, the better the job that the CAT can perform . 3) *Item* *selection method*: Adaptive tests select the next item to be posed depending on the student's estimated knowledge level (obtained from the answers to items previously administered). 4) *The termination criterion*: Different criteria can be used to decide when the test should finish, in terms of the purpose of the test.

The set of advantages provided by CATs is often addressed in the literature [11]. The main advantage is that it reduces the number of questions needed to estimate student knowledge level, and as a result, the time devoted to that task. . This entails an improvement in student motivation. However, CATs contain some drawbacks. They require the availability of huge item pools, techniques to control item exposure and to detect compromised items. In addition, item parameters must be calibrated. To accomplish this task, a large number of student performances are required, and this is not always available.

## 3 Item Response Theory

IRT [7] has been successfully applied to CATs as a response model, item selection and finalization criteria. It is based on two principles: a) Student performance in a test can be explained by means of the knowledge level, which can be measured as an unknown numeric value. b) The performance of a student with an estimated knowledge level answering an item *i* can be probabilistically predicted and modeled by means of a function called *Item Characteristic Curve* (ICC). It expresses the probability that a student with certain knowledge level $\theta$ has to answer the item correctly. Each item must define an ICC, which must be previously calibrated. There are several functions to characterize ICCs. One of the most extended is the logistic function of three parameters (3PL) [1] defined as follows:

$$P(\mathrm{u}_\mathrm{i}=1\,|\,\theta)=\mathrm{c}_\mathrm{i}+(1-\mathrm{c}_\mathrm{i})\frac{1}{1+\mathrm{e}^{-1.7\mathrm{a}_\mathrm{i}(\theta-\mathrm{b}_\mathrm{i})}} \tag{1}$$

where $u_i=1$ represents that the student has successfully answered item *i*. If the student answers incorrectly, $P(u_i=0|\theta)=1-P(u_i=1|\theta)$. The three parameters that determine the shape of this curve are:

- *Discrimination factor* ($a_i$): It is proportional to the slope of the curve. High values indicate that the probability of success from students with a knowledge level higher than the item difficulty is high.
- *Difficulty* ($b_i$): It corresponds to the knowledge level at which the probability of answering correctly is the same as answering incorrectly . The range of values allowed for this parameter is the same as the ones allowed for the knowledge levels.
- *Guessing factor* ($c_i$): It is the probability of that a student with no knowledge at all will answer the item correctly by randomly selecting a response.

In our proposal, and therefore throughout this paper, the knowledge level is measured using a discrete IRT model. Instead of taking real values, the knowledge level takes K values (or latent classes) from 0 to K-1. Teachers decide the value of *K* in terms of the assessment granularity desired. Likewise, each ICC is turned into a probability vector $p(u_i=1|\theta=0), p(u_i=1|\theta=1), p(u_i=1|\theta=2), ..., p(u_i=1|\theta=K-1)$.

### 3.1 Student Knowledge Estimation

IRT supplies several methods to estimate student knowledge. All of them calculate a probability distribution curve $P(\theta|\boldsymbol{u})$**,** where $\boldsymbol{u}=u_1, .., u_n$ is the vector of items administered to students. When applied to adaptive testing, knowledge estimation is accomplished every time the student answers each item posed, obtaining a temporal estimation. The distribution obtained after posing the last item of the test becomes the final student knowledge estimation. One of the most popular estimation methods is the Bayesian method [9]. It applies the Bayes theorem to calculate student knowledge distribution after posing an item *i*:

$$\mathrm{P}(\theta|\mathrm{u}_1,..,\mathrm{u}_\mathrm{i}) \propto P(\mathrm{u}_\mathrm{i}=1\,|\,\theta)^{u_i}(1-P(\mathrm{u}_\mathrm{i}=1\,|\,\theta))^{(1-u_i)}P(\theta\,|\,u_1,..u_{i-1}) \qquad (2)$$

where $P(\theta|u_1, ..,u_{i-1})$ represents temporary student knowledge distribution before posing *i*.

### 3.2 Item Selection Procedure

One of the most popular methods for selecting items is the Bayesian method [9]. It selects the item that minimizes the expectation of a posteriori student knowledge distribution variance. That is, taking the current estimation, it calculates the posterior expectation for every non-administered item, and selects the one with the smallest expectation value. Expectation is calculated as follows:

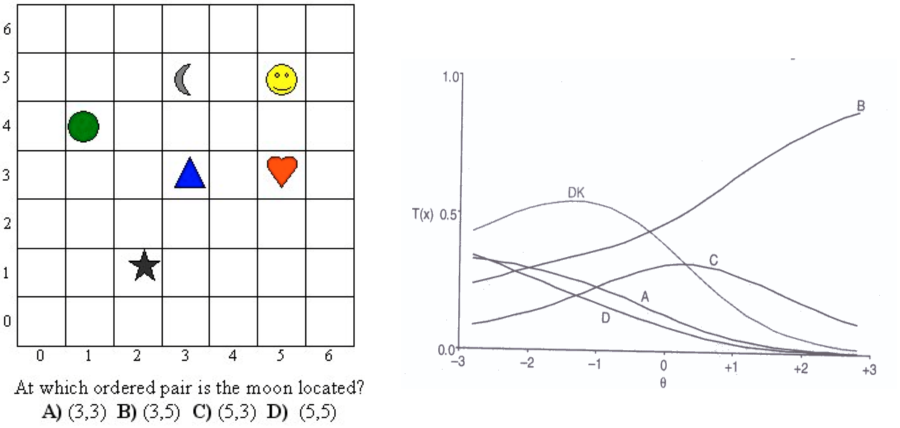
$$\begin{aligned}&\mathrm{P}'(\theta|u_1,...,\mathrm{u}_\mathrm{i}=\mathrm{r})=\mathrm{P}(\mathrm{u}_\mathrm{i}=\mathrm{r}\,|\,\theta)\cdot P(\theta\,|\,u_1,...,u_{i-1})\\&\mathrm{E}\left[\sigma^2(\mathrm{P}(\theta|\mathrm{u}_1,...,\mathrm{u}_\mathrm{i}))\right]=\sum_r \mathrm{P}'(\theta|u_1,...,\mathrm{u}_\mathrm{i}=\mathrm{r})\,\sigma^2(\mathrm{P}(\theta\,|\,u_1,...,\mathrm{u}_\mathrm{i}=\mathrm{r}))\end{aligned} \qquad (3)$$

where *r* can take value *0* or *1*. It is *r=1-*, if the response is correct, or *r=0* otherwise. $P'(\theta|u_1,...,u_i=r)$ is the scalar product between ICC (or its inverse) of item *i* and the current estimated knowledge distribution.

## 4 Polytomous IRT

In dichotomous IRT models, items are only scored as correct or incorrect. In contrast, polytomous models try to obtain as much information as possible from the student’s response. They take into account the answer selected by students in the estimation of knowledge level and in the item selection. For this purpose, these models add a new type of characteristic curve associated to each answer, in the style of ICC. In the literature these curves are called *trace lines* (TC) [3], and they represent the probability that certain student will select an answer given his knowledge level.

To understand the advantages of this kind of model, let us look at the item represented in Fig. 1 (a). A similar item was used in a study carried out in 1992 [10]. Student performances in this test were used to calibrate the test items. The calibrated TCs for the item of Fig. 1 (a) are represented in Fig. 1 (b). Analyzing these curves, we see that the correct answer is *B*, since students with the highest knowledge levels have high probabilities of selecting this answer. Options *A* and *D* are clearly wrong, because students with the lowest knowledge levels are more likely to select these answers. However, option *C* shows that a considerable number of students with medium knowledge levels tends to select this option. If the item is analyzed, it is evident that for option *C*, although incorrect , the knowledge of students selecting *it* is higher than the knowledge of students selecting *A* or *D*. Selecting *A* or *D* may be assessed more negatively than selecting *B*. Answers like *C* are called *distractors*, since, even though these answers are not correct, they are very similar to the correct answers. In addition, polytomous models make a difference between selecting an option or leave the item blank. Those students who do not select any option are modeled with the DK option TC. This answer is considered as an additional possible option and is known as *don't know* option.



At which ordered pair is the moon located?
A) (3,3) B) (3,5) C) (5,3) D) (5,5)

**Fig. 1.** (a) A multiple-choice item, and (b) its trace lines (adapted from [10])

## 5 Student Knowledge Diagnosis Through Adaptive Testing

Domain models can be structured on the basis of subjects. Subjects may be divided into different topics. A topic can be defined as a concept regarding which student knowledge can be assessed. They can also be decomposed into other topics and so on, forming a hierarchy with a degree of granularity decided by the teacher. In this hierarchy, leaf nodes represent a unique concept or a set of concepts that are indivisible from the assessment point of view. Topics and their subtopics are related by means of aggregation relations, and no precedence relations are considered. For diagnosis purposes, this domain model could be extended by adding a new layer to include two kinds of components: items and test specifications. This extended model has been represented in Fig. 2. The main features of these new components are the following:

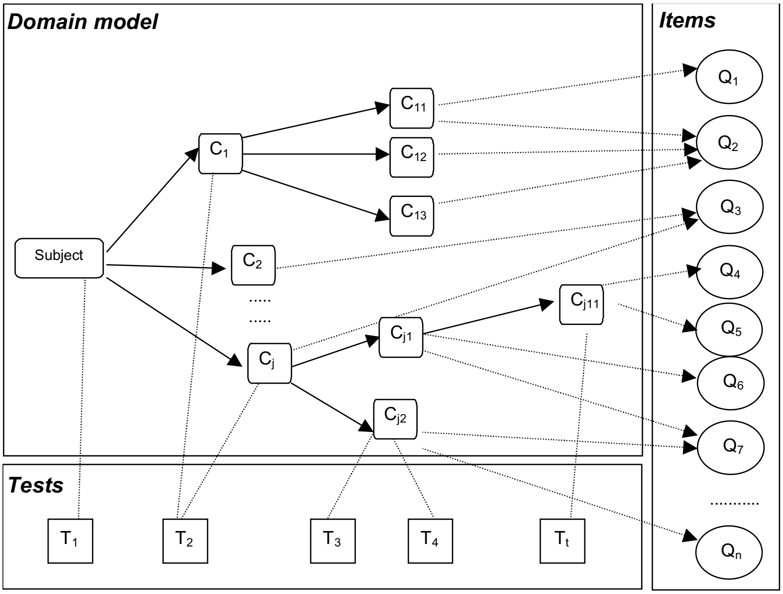


**Fig. 2.** A domain model extended for diagnosis

**Items.** They are related to a topic. This relationship is materialized by means of an ICC. Due to the aggregation relation defined in the curriculum, if an item is used to assess a topic *j*, it also provides assessment information about the knowledge state in topics preceding *j*, and even in the whole subject. To model this feature, several ICCs have been associated to each item , one for each topic the item is used to assess. These curves collect the probability of answering the item correctly given the student knowledge level in the corresponding topic. Accordingly, the number of ICCs of an item is equal to the number of topics, in different levels of the hierarchy, which are related to the item including the subject. This means that for item $Q_5$ (Fig. 2), the ICCs defined are: $P(u_5=1|\ \theta_{j11})$, $P(u_5=1|\ \theta_{j1})$, $P(u_5=1|\ \theta_j)$ and $P(u_5=1|\ \theta_{subject})$.

**Tests.** They are specifications of adaptive assessment sessions defined on topics. Therefore, after a student takes a test, it will diagnose his knowledge levels in the test topics, and in all their descendant topics. For instance, let us consider test $T_2$ (Fig. 2). Topics of this test are $C_1$ and $C_j$. After a testing session, the knowledge of students in these topics will be inferred. Additionally, the knowledge in topics $C_{11}$, $C_{12}$, $C_{13}$, $C_{j1}$, $C_{j2}$ and $C_{j11}$ can also be inferred That is, if $\boldsymbol{u}= u_1, ..., u_n$ is the set of items administered, the following knowledge distributions could be inferred: $P(\theta_1|\boldsymbol{u})$, $P(\theta_j|\boldsymbol{u})$, $P(\theta_{11}|\boldsymbol{u})$, $P(\theta_{12}|\boldsymbol{u})$, $P(\theta_{13}|\boldsymbol{u})$, $P(\theta_{j1}|\boldsymbol{u})$, $P(\theta_{j2}|\boldsymbol{u})$ and $P(\theta_{j11}|\boldsymbol{u})$.

As mentioned earlier, even though CATs are used to assess one single topic, in [4] we introduce a technique to simultaneously assess multiple topics in the same test, which is content-balanced. This technique has been included in a student knowledge diagnosis model that uses the extended domain model of Fig. 2. The model assesses through adaptive testing, and uses a discrete response model where the common dichotomous approach has been replaced by a polytomous one. Accordingly, the relationship between topics and items is modified. Now, each ICC is replaced by a set of TCs (one for each item answer), that is, the number of TCs of an item *i* is equal to the product of the number of answers of *i*, with the number of topics assessed using *i*. In this section, the elements required for diagnosis have been depicted. The next subsection will focus on how the diagnosis procedure is accomplished.

### 5.1 Diagnosis Procedure

It consists of administering an adaptive test to students on ITS demand. The initial information required by the model is the test parameters to be applied, and the current knowledge level of the student in test topics. An ITS may use these estimations to update the student model. The diagnose procedure comprises the following steps:

- *Test item compilation*: Taking the topics involved in the test as the starting point, items associated with them are collected. All items associated to their descendant topics at any level are included in the collection.
- *Temporary student cognitive model creation*: The diagnosis model creates its own temporary student cognitive model. It is an overlay model, composed of nodes representing student knowledge in the test topics. For each node, the model keeps a discrete probability distribution.
- *Student model initialization*: If any previous information about the state of student knowledge in the test topics is supplied, the diagnosis model could use this information as a priori estimation of student knowledge. In other cases, this model offers the possibility of selecting several values by default
- *Adaptive testing stage*: The student is administered the test adaptively.

### 5.2 Adaptive Testing Stage

This testing algorithm follows the steps described in Section 2, although item selection and knowledge estimation procedures differ because of the addition of a discrete polytomous response model. Student knowledge estimation uses a variation of the Bayesian method described in Equation 2. After administering item *i*, the new estimated knowledge level in topic *j* is calculated using Equation 4.

$$\mathrm{P}(\theta_j \big| u_1,\ldots,u_{i-1},\mathrm{u_i}) \propto P(\mathrm{u_i} = \mathrm{r} \mid \theta_j)P(\theta_j \mid u_1,\ldots,u_{i-1}) \tag{4}$$

Note that the TC corresponding to the student answer, $\mathrm{P(u_i{=}r|\theta_j)}$, has replaced the ICC term. Being *r* the answer selected by the student, it can take values between 1 to the number of answers *R*. When *r* is zero, it represents the *don't know* answer.

Once the student has answered an item, this response is used to update student knowledge in all topics that are descendents of topic *j*. Let us suppose test $\mathrm{T_2}$ (Fig. 1(b)) is being administered. If item $\mathrm{Q_5}$ has just been administered, student knowledge estimation in topic $\mathrm{C_j}$ is updated according to Equation 4. In addition, item $\mathrm{Q_5}$ provides information about student knowledge in topics $\mathrm{C_{j1}}$ and $\mathrm{C_{j11}}$. Consequently, the student knowledge estimation in these topics is also updated using the same equation.

The item selection mechanism modifies the dichotomous Bayesian one (Equation 3). In this modification, expectation is calculated from the TCs, instead of the ICC (or its inverse), in the following way:

$$\mathrm{E}\left[\sigma^2(\mathrm{P}(\theta_j|\mathrm{u}_1,..,\mathrm{u}_\mathrm{i}))\right]=\sum_{r=0}^{n}\left[P(u_i=r\,|\,\theta_j)\cdot P(\theta_j\,|\,u_1,..,u_{i-1})\right]\sigma^2(\mathrm{P}(\theta_j\,|\,\mathrm{u}_1,..,\mathrm{u}_\mathrm{i}=\mathrm{r})) \qquad (5)$$

$\theta_j$ represents student knowledge in topic *j*. Topic *j* is one of the test topics. Let us take test $T_2$ again. Expectation is calculated for all (non-administered) items that assess topics $C_1$, $C_j$ or any descendent. Note that Equation 5 must always be applied to knowledge distributions in test topics (i.e. $C_1$ and $C_j$), since the main goal of the test is to estimate student knowledge in these topics. The remaining estimations can be considered as a collateral effect. Additionally, this model guarantees content-balanced tests. The adaptive selection engine itself tends to select the item that makes the estimation more accurate [4]. If several topics are assessed, the selection mechanism is separated in two phases. In the first one, it will select the topic whose student knowledge distribution is the least accurate. The second one selects, from items of this topic, the one that contributes the most to increase accuracy.

## 6 Evaluation

Some authors have pointed out the advantages of using simulated students for evaluation purposes [12], since this kind of student allows having a controlled environment, and contributes to ensuring that the results obtained in the evaluation are correct. This study consists of a comparison of two CAT-based assessment methods: the polytomous versus the dichotomous one. It uses a test of a single topic, which contains an item pool of 500 items. These items are multiple-choice items with four answers, where the *don't know* answer is included. The test stops when the knowledge estimation distribution has a variance that is less than $10^{-5}$. The test has been administered to a population of 150 simulated students. These students have been generated with a real knowledge level that is used to determine their behavior during the test. Let us assume that the knowledge level of the student John is $\theta_j$. When an item *i* is posed, John's response is calculated by generating a random probability value *v* . The answer *r* selected by John (r ∈ {0, 1, 2, 3, 4}) is the one that fulfils,

$$\sum_{m=0}^{r} P\,(\mathrm{u}_\mathrm{i}=\mathrm{r}\,|\,\theta_j) >= v$$

Using the same population and the same item pool, two adaptive tests have been administered for each simulation. The former uses polytomous item selection and knowledge estimation, and the latter dichotomous item selection and knowledge estimation. Different simulations of test execution have been accomplished changing the parameters of the item curves. ICCs have been generated (and are assumed to be well calibrated), before each simulation, according to these conditions. The correct answer TC corresponds to the ICC, and the incorrect response TCs are calculated in such a way that their sum is equal to *1-ICC*. Simulation results are shown in Table 1.

In Table 1 each row represents a simulation of the students taking a test with the features specified in the columns. Discrimination factor and difficulty of all items of the pool are assigned the value indicated in the corresponding column, and the guessing factor is always zero. When the value is "uniform", item parameter values

**Table 1.** Simulation results of polytomous testing versus dichotomous testing

| Assessment & item selection | Item discrim. | Item difficult. | Item number average | Estimation variance average | Success rate |
|---|---|---|---|---|---|
| polytomous | 0,4 | uniform | 67,77 | 0,000577216 | 100% |
| dichotomous | 0,4 | uniform | 146,51 | 0,000366659 | 98% |
| polytomous | 0,7 | uniform | 40,35 | 0,000007113 | 100% |
| dichotomous | 0,7 | uniform | 50,56 | 0,000186682 | 100% |
| polytomous | 1,2 | uniform | 20,30 | 0,000006597 | 100% |
| dichotomous | 1,2 | uniform | 18,57 | 0,000007863 | 100% |
| polytomous | uniform | uniform | 20,60 | 0,000007476 | 100% |
| dichotomous | uniform | uniform | 28,97 | 0,000028836 | 100% |

have been generated uniformly along the allowed range. The last three columns represent the results of simulations. “Item number average” is the average of items posed to students in the test; “estimation variance average” is the average of the final knowledge estimation variances. Finally , “success rate” is the percentage of students assessed correctly. This last value has been obtained by comparing real student knowledge with the student knowledge inferred by the test. As can be seen, the best improvements have been obtained for a pool of items with a low discrimination factor. In this case, the number of items has been reduced drastically. The polytomous version requires less than half of the dichotomous one, and the estimation accuracy is only a bit lower . The worst performance of the polytomous version takes place when items have a high discrimination factor. This can be explained because high discrimination ICCs get the best performance in dichotomous assessment. In contrast, for the polytomous test, TCs have been generated with random discriminations, and as a result, TCs are not able to discriminate as much as dichotomous ICCs. In the most realistic case, i.e. the last two simulations, item parameters have been calculated uniformly. In this case, test results for the polytomous version is better than the dichotomous one, since the higher the accuracy, the lower the number of items required. In addition, the evaluation results obtained in [4] showed that the assessment of multiple topics is simultaneously able to make a content-balanced item selection. Teachers do not have to specify, for instance, the percentage of items that must be administered for each topic involved in the test.

## 7 Discussion

This work proposes a well-founded student diagnosis model, based on adaptive testing. It introduces some improvements in traditional CATs. It allows simultaneous assessment of multiple topics through content-balanced tests. Other approaches have presented content-balanced adaptive testing, like the CBAT-2 algorithm [6]. It is able to generate content-balanced tests, but in order to do so, teachers must manually introduce the weight of topics in the global test for the item selection. However, in our model, item selection is carried out adaptively by the model itself. It selects the next item to be posed from the topic whose knowledge estimation is the least accurate. Additionally, we have defined a discrete , IRT-based polytomous response model. The evaluation results (where accuracy has been overstated to demonstrate the strength of the model) have shown that, in general, our polytomous model makes more accurate estimations and requires fewer items.

The model presented has been implemented and is currently used in the SIETTE system [2]. SIETTE is a web-based CAT delivery and elicitation tool (http://www.lcc.uma.es/siette) that can be used as a diagnosis tool in ITSs. Currently, we are working on TC calibration techniques. The goal is to obtain a calibration mechanism that minimizes the number of prior student performances required to calibrate the TCs.

## References

1. Birnbaum, A. Some Latent Trait Models and Their Use in Inferring an Examinee's Mental Ability. In : Lord, F. M. and Novick, M. R, eds. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley; 1968.
2. Conejo, R.; Guzmán, E.; Millán, E.; Pérez-de-la-Cruz, J. L., and Trella, M. SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education* (forthcoming).
3. Dodd, B. G.; DeAyala, R. J., and Koch, W. R. Computerized Adaptive Testing with Polytomous Items. *Applied Psychological Measurement*. 1995; 19(1):pp. 5-22.
4. Guzmán, E. and Conejo, R. Simultaneous evaluation of multiple topics in SIETTE. *LNCS, 2363. ITS 2002*. Springer Verlag; 2002: 739-748.
5. Guzmán, E. and Conejo, R. A library of templates for exercise construction in an adaptive assessment system. *Technology, Instruction, Cognition and Learning (TICL)* (forthcoming).
6. Huang, S. X. A Content-Balanced Adaptive Testing Algorithm for Computer-Based Training Systems. *LNCS, 1086. ITS 1996*. Springer Verlag; 1996: pp. 306-314.
7. Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
8. McCalla, G. I. and Greer, J. E. Granularity-Based Reasoning and Belief Revision in Student Models. In: Greer, J. E. and McCalla, G., eds. *Student Modeling: The Key to Individualized Knowledge-Based Instruction*. Springer Verlag; 1994; 125 pp. 39-62.
9. Owen, R. J. A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing. *Journal of the American Statistical Association*. 1975 Jun; 70(350):351-371.
10. Thissen, D. and Steinberg, L. A Response Model for Multiple Choice Items. In: Van der Linden, W. J. and Hambleton, R. K., (eds.). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag; 1997; pp. 51-65.
11. van der Linden, W. J. and Glas, C. A. W. *Computerized Adaptive Testing: Theory and Practice*. Netherlands: Kluwer Academic Publishers; 2000.
12. VanLehn, K.; Ohlsson, S., and Nason, R. Applications of Simulated Students: An Exploration. *Journal of Artificial Intelligence and Education*. 1995; 5(2):135-175.