A Bayesian Diagnostic Algorithm for Student Modeling and its Evaluation

EVA MILLÁN and JOSÉ LUIS PÉREZ-DE-LA-CRUZ
Departamento de Lenguajes y Ciencias de la Computación E.T.S.I. Informática, Universidad de Málaga. Apdo. 4114, Málaga 29080. Spain, E-mail: eva.perez@lcc.uma.es
(Received 5 September 2000; accepted in revised form 29 January 2001)

Abstract. In this paper, we present a new approach to diagnosis in student modeling based on the use of Bayesian Networks and Computer Adaptive Tests. A new integrated Bayesian student model is defined and then combined with an Adaptive Testing algorithm. The structural model defined has the advantage that it measures students’ abilities at different levels of granularity, allows substantial simplifications when specifying the parameters (conditional probabilities) needed to construct the Bayesian Network that describes the student model, and supports the Adaptive Diagnosis algorithm. The validity of the approach has been tested intensively by using simulated students. The results obtained show that the Bayesian student model has excellent performance in terms of accuracy, and that the introduction of adaptive question selection methods improves its behavior both in terms of accuracy and efficiency.

Keywords: adaptive testing, Bayesian networks, student modeling.

1. Introduction

New technologies have provided the Education field with innovations that allow significant improvements in the teaching/learning process. Their introduction not only reduces the effective cost of the application of pedagogical theories, but also opens up the possibility of exploring models from very different fields, facilitating their interaction and integration. One of the main innovations introduced since the first Computer Aided Learning programs are the so-called Intelligent Tutoring Systems (ITS), that, in contrast to traditional programs, have the ability to adapt to each individual learner. It is precisely this ability to adapt to each student that allows these programs to improve the teaching/learning process, as it has already been shown that the best learning method is individualized learning (Bloom, 1984).

Therefore, if the key characteristic of an ITS is its ability to adapt to each student (Shute, 1995), the key component of such a system is the student model, where all the information about the student is stored, including his/her cognitive state about the subject domain. The cognitive state is generated from student behavior during interaction with the system, that is, it is inferred by the system from the information available; previous data answers to questions posed by the system,
instructional episodes, etc. The process that consists of inferring the cognitive state of the student from observable data is called diagnosis. Diagnosis is without doubt the most complicated process in an ITS since, besides the inherent difficulty of any inference process, it involves the treatment of information that in many cases is uncertain and/or imprecise. In addition, although it has been shown that a student model can be useful even without being very accurate (Stern, Beck and Woolf, 1996), it is clear that the more accurate it is, the better the job it can do. However, when it comes to the diagnosis process, the great knowledge engineering effort involved in developing an ITS is such that many designers prefer to develop their own heuristics instead of using Approximate Reasoning techniques available within the Artificial Intelligence field. The problem is that, in some cases, the lack of theoretical foundations of such heuristics can make the system’s behavior inadequate or unpredictable, yielding results different from the ones originally expected.

In this way, the main goal of our work is to improve the accuracy and efficiency of the diagnosis process in an ITS. To this end, we have explored the possibility of using Approximate Reasoning techniques, with special emphasis on simplifying their application as much as possible to encourage their use among ITS researchers. The proposed solution is founded on the definition of a new integrated student model based on Bayesian Networks (BNs), and on the application of Computer Adaptive Tests theory to improve the efficiency and accuracy of the diagnosis process. This new Bayesian student model allows measurement of a student’s knowledge at different levels of granularity (that is, the subject domain is curriculum-structured), as well as substantial simplifications when defining the BN (nodes, links, and parameters). It also accounts for the possibility of lucky guesses (giving the right answer to a question even when the student has not mastered the related concepts) and of having an unintentional error or slip (giving the wrong answer to a question even when the student knows all the related concepts).

Both the Bayesian student model alone and its combination with Adaptive Testing techniques have been tested intensively by using simulated students. The main advantage of using simulated students is that the cognitive state obtained as a result of the diagnostic algorithm can be compared to the student’s true cognitive state. A total of 180 simulated students with different knowledge levels were generated. Then, the diagnostic algorithm estimated the set of known concepts, and the fitness of this estimation was analyzed. The use of the Bayesian student model with random question selection criteria produced up to 90.27% correctly diagnosed concepts. These results can be improved by using the proposed adaptive criterion, going up to 94.53% correctly diagnosed concepts. Moreover, the number of questions needed to obtain these estimations using the adaptive criterion proposed was smaller, so the gain is not only in accuracy but also in efficiency.

This paper is structured as follows: in the next section we briefly describe the theoretical background underlying our integrated student model. Sections 3 and 4 are devoted to the definition of the BN (nodes, links, and parameters) that supports
the student model and to the description of the Adaptive Testing Algorithm, respectively. An in-depth evaluation of the proposed integrated student model and diagnostic algorithm is presented in Section 5. Finally, we present a comparative review of some related work and outline some conclusions and future lines of research.

2. Theoretical Background

As already explained, our work is based on the use of Bayesian Networks and Adaptive Testing Theory. In this section we briefly present the basics of both theories.

2.1. BAYESIAN NETWORKS

A Bayesian Network (BN) (Pearl, 1988) is a directed acyclic graph in which nodes represent variables and arcs represent probabilistic dependence among variables\(^1\). The parameters used to represent the uncertainty are the conditional probabilities of each node given each combination of states of its parents; that is, if \(\{X_i, i = 1, \ldots, n\}\) are the variables of the network and \(pa(X_i)\) represents the set of the parents of \(X_i\), for each \(i = 1, \ldots, n\), then the parameters of the network are \(\{P(X_i|pa(X_i))\}, i = 1, \ldots, n\), that is, the set of discrete conditional probability distributions of each variable given its parents. This set of probabilities defines the joint probability distribution for the entire network as,

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))
\]

Thus, to define a BN, we have to specify:

- The set of variables, \(X_1, X_2, \ldots, X_n\).
- The set of links (arcs) between those variables. These arcs represent a causal influence between the variables. The network formed with these variables and arcs must be a Directed Acyclic Graph (DAG).
- For each variable \(X_i\), its probability conditioned to its parents, that is, \(P(X_i|pa(X_i))\), \(i = 1, \ldots, n\).

If we are using BNs to define a student model, the variables can represent different things depending on the domain. The variables can be rules, concepts, problems, abilities, skills, etc. These variables are linked by relationships between them, such as part-of, prerequisite-of, etc. Once the links and the variables have been defined, the conditional probabilities must be specified. Our integrated student model will be defined in line with this description in Section 3.

\(^1\)For an easy introduction to BNs see (Charniak, 1991) and for a complete presentation (Castillo et al., 1997).
2.2. ADAPTIVE TESTING

A Computer Adaptive Test (CAT) is a test administered by a computer where the selection of the next question to ask and the decision to stop the test are performed dynamically based on a student profile, which is created and updated during the interaction with the system. The main difference between CATs and traditional Paper and Pencil Tests (PPTs) is the same difference that exists between traditional Training Systems and ITSs, that is, the capability to adapt to each individual student. The advantages of CATs have been widely discussed in the literature (Kingsbury and Weiss, 1983), and more recently reported in (Wainer, 1990). The main advantage is a significant decrease in test length, with equal or better estimations of the student’s knowledge level. This advantage is a direct consequence of using adaptive question selection algorithms, that is, algorithms that choose the best (most informative) question to ask next, given the current estimation of the student’s knowledge. Some other advantages come from using a computer to perform the tests: larger databases of questions can be stored, selection algorithms can be used efficiently, and a great number of students can take the tests at the same time, even if they are in different geographical locations.

In more precise terms, a CAT is an iterative algorithm that starts with an initial estimation of the examinee’s proficiency level and consists of the following steps:

1. All the questions in the database (that have not been administered yet) are examined to determine which will be the best to ask next according to the current estimation of the examinee’s level.
2. The question is asked, and the examinee responds.
3. According to the answer, a new estimation of the proficiency level is computed.
4. Steps 1 to 3 are repeated until the stopping criterion defined is met.

This procedure is illustrated in Figure 1.

![Flow diagram of an adaptive test](image-url)
In (Weiss and Kingsbury, 1984), the basic elements in the development of a CAT are defined. These basic elements are:

- **Item Response model.** This model describes how examinees answer the item depending on their level of ability. When measuring proficiency, the result obtained should be independent of the tool used, that is, this measurement should be invariant with respect to the type of test and to the individual that takes the test.
- **Scoring method,** that is, a method to compute the student’s ability level according to his/her answers.
- **Item pool.** This is one of the most important elements in a CAT. A good item pool must contain a large number of correctly calibrated items at each ability level (Flaugher, 1990). Obviously, the better the quality of the item pool, the better the job that the CAT can do.
- **Initial level.** Suitably choosing the difficulty level of the first question in a test can considerably reduce the length of the test. Different criteria can be used, e.g. taking the average level of knowledge of the examinees that have taken the test previously or creating an examinee profile and using the average level of examinees with a similar profile, as proposed in (Thissen and Mislevy, 1990).
- **Question selection method.** Adaptive tests select the next item to be posed depending on the estimated proficiency level of the examinee (obtained from the answers to items previously administered). Selecting the best item to ask given the estimated proficiency level can improve accuracy and reduce test length.
- **Termination criterion.** Different criteria can be used to decide when the test should finish, depending on the purpose of the test. An adaptive test can finish when a target measurement precision has been achieved, when a fixed number of items has been presented, when the time has finished, etc.

The psychometric theory underlying most CAT implementations is *Item Response Theory* (IRT) (Birnbaum, 1968; Hambleton, 1989). All IRT-based models have some common features: (1) they assume the existence of latent traits or aptitudes that allow us to predict or explain the examinee’s behavior; and (2) the relation between the trait $\theta$ and the answers that a person gives to a test item $Q_i$ can be described with an increasing monotonous function called the *Item Characteristic Curve* (ICC). The most commonly used model to describe the ICC is the three-parameter model (Birnbaum, 1968), which states that the ICC associated with a question $Q_i$ is given by the following function:

$$ P_i(\theta) = P(\text{Correct answer to } Q_i|\theta) = c_i(1 - c_i) \frac{1}{1 + e^{-a_i(\theta - b_i)}} $$

Thus, the probability of correctly answering $Q_i$ given a certain knowledge level $\theta$ is
given by the three-parameter function $P_A(\theta)$. This function is plotted in Figure 2 with $a_i = 1.2$, $b_i = 5$, and $c_i = 0.25$.

Let us examine the meaning of these parameters:

- $a_i$ is called the *discrimination index*, and defines the slope of the curve at its inflection point. Therefore $a_i$ denotes how well the question is able to discriminate between students of slightly different abilities.
- $b_i$ is called the *difficulty degree*, and defines the location of the curve’s inflection point. The higher the value of $b_i$, the more difficult the question.
- $c_i$ is called the *guessing factor* and represents the left asymptote of the curve.

Therefore the probability of a correct answer to question $Q_i$ for students of very low ability is close to $c_i$.

According to (Olea and Ponsoda, 1996), if the three-parameter logistic model is used a good item pool should have the following characteristics:

- Discrimination indexes should be big (most of them bigger than 1.2), so precise estimations can be made with few items.
- There must be approximately the same number of items in each difficulty level.
- The guessing factor should be close to $1/n$, where $n$ is the number of possible answers.

An excellent primer to CATs and IRT can be found in (Rudner, 1998), where it is possible to try an actual CAT online. For more detailed descriptions, (Wainer, 1990) and (Van der Linden and Hambleton, 1997).

Having described the basics of BNs and CATs, the next two sections describe how we use them in the student modeling problem.

### 3. An Integrated Approach to Bayesian Student Modeling

In this section, we describe the structural model used in our approach to Bayesian student modeling. The student model defined is an *overlay student model* (as described in (Van Lehn, 1988)), that is, the student’s knowledge is considered as

![Figure 2. ICC Graphic.](image)

---

*In the interactive tutorial described in (Rudner, 1998) it is possible to play with these parameters to obtain a better understanding of their meaning.*
a subset of the expert’s knowledge. Part of this model has already been described in (Millán et al., 2000). The section is structured as follows: in Section 3.1 we describe the nodes that are used in the BN. Having defined the nodes, the causal relationships between the variables are discussed in Section 3.2.

3.1. VARIABLES

In this section we describe the different types of variables that compose our Bayesian student model: variables to measure a student’s knowledge and variables to collect evidence.

3.1.1. Variables to Measure the Student’s Knowledge

To measure the student’s knowledge, we use variables at different levels of granularity. In order to keep terminology simple, we use the names concept, topic, and subject, while bearing in mind that they could represent declarative knowledge, skills, abilities, etc.

- A concept is an elementary piece of knowledge, in the sense that it cannot be decomposed into smaller parts. Elementary concepts are considered the basic units of knowledge.

To represent an elementary concept C we use a random variable $C$ with a Bernoulli distribution, that is, $C$ takes two different values: 1 if the student knows the concept, or 0 otherwise. The probability law of $C$ will then be:

$$P(C = x) = p^x(1 - p)^{1-x},$$

where $p$ is the probability that the student knows concept C, and $x$ can take values 0 and 1.

- A topic is a pair $(C, w)$, where:
  - $C$ is a set of elementary concepts $C = \{C_1, \ldots, C_n\}$, which are mutually independent.
  - $w = (w_1, \ldots, w_n)$ is a weight vector that measures the relative importance of each concept in the topic it belongs to. Without loss of generality, we assume that $\sum_{i=1}^{n} w_i = 1$.

To measure the student’s knowledge about a topic, we use a random variable $T$ defined by:

$$T = \sum_{j=1}^{n} w_j C_j$$

- A subject is a pair $(T, a)$, where:
  - $T$ is a set of mutually independent topics, $T = \{T_1, \ldots, T_s\}$. 
\[ x = (x_1, \ldots, x_s) \text{ is a weight vector that measures the relative importance of each topic in the subject it belongs to. We also assume that } \sum_{i=1}^{s} x_i = 1. \]

By definition, we know that each topic is composed of a set of mutually independent concepts with their respective weights, that is, for each \( i = 1, \ldots, s \) the topic \( T_i \) is composed of a set of concepts \( \{ C_{ij}, j + 1, \ldots, n_i \} \) and a weight vector \( w = (w_1, \ldots, w_{n_0}) \), defined by the expression:

\[
T_i = \sum_{j=1}^{n_i} w_j C_{ij}
\]

To represent the student’s knowledge about a certain subject \( A \), we use a random variable \( A \) defined by:

\[
A = \sum_{i=1}^{s} x_i T_i
\]

Let us consider the following example in order to illustrate the use of such variables.

**EXAMPLE 1.** Let us suppose that a teacher is designing a Mathematics course, whose contents and structure are given in Table 1.

Although it is not always the case, in this example we consider that the time devoted to each topic and subtopic is a measure of its importance. Therefore, the course specification can be easily translated to the representation defined above. The weight of a topic (subtopic) can be computed as the number of days associated with it over the number of days associated with the subject (topic) it belongs to. Thus, for example, the weight for the subtopic *Functions* of the topic *Calculus* is 18/90. Table 2 shows the granularity hierarchy associated with this example.

Note, however, that other criteria could be used to set the relative importance of each topic (concept) in the subject (topic), such as the desired proportion of questions in an exam or any other subjective estimation of the teacher.
3.1.2. Nodes to Collect Evidence

These nodes are used to collect the information relevant to the student’s knowledge state. In our model, the source of evidence is the set of test items related to knowledge nodes. To represent an evidence node, we use a random variable $P$ with a Bernoulli distribution, that is, it takes the value 1 when the student chooses the right answer, and 0 otherwise. The probability law of $P$ is then given by:

$$P(P = x) = p^x(1 - p)^{1-x},$$

where $p$ represents the probability that the student chooses the right answer, and $x$ takes values 0 or 1.

Although we will be considering only test items, note that other sources of evidence (such as exercises, tasks, problems, etc.) can also be considered in the model and represented by a binary random variable, provided that the ITS has the capability to diagnose whether the student’s solution is right or wrong.

3.2. Modeling Causal Relationships: Links and Parameters

Having defined the nodes, we determine the causal relationships among them, as follows: aggregation relationships between knowledge variables at different levels of granularity, and relationships between evidential and knowledge nodes.

3.2.1. Modeling Aggregation Relationships

In order to discuss these relationships, we use the general expression Knowledge Item (KI) to refer either to a subject, a topic, concept, skill, etc. Aggregation or part-of relationships are established between a KI and the KIs it is composed of. For example, the relationship is established between a subject and its topics or between a skill and the more specific subskills it can be divided into.

Ideally, if such sources of evidence are to be included in the model, their solutions should be evaluated in terms of a discrete or even continuous random variable, that is, there should be different degrees of correctness for the answer. However, the use of such variables increases both the computational complexity and the knowledge engineering effort (number of parameters required) to define the BN.
Let us suppose that $I$ is a KI that can be divided into more specific KIs that will be denoted by $I_1, \ldots, I_n$. Each KI will be represented by a binary random variable with two values: mastered or not mastered. To model the causal relationships between them we have two alternatives:

- **Alternative 1**: We consider that knowing the more specific items has a causal influence on knowing the more general item.
- **Alternative 2**: We consider that knowing the more specific item has a causal influence on dominating each of the more specific items it is composed of.

These two alternatives are graphically depicted in Figure 3.

Next, we analyze the independence structures implied by each alternative, and the parameters that need to be specified.

(a) In Alternative 1, the parameters needed are: the prior probabilities of knowing each $I_i$, that is, $\{P(I_i), i = 1, \ldots, n\}$ and the conditional probability distribution of $I$ given its parents, that is, $P(I|I_1, \ldots, I_n)$. This makes a total of $n + 2^n - 1^4$ values. Regarding independence, this structure implies that the $I_i$’s (for $i = 1, \ldots, n$) are mutually independent.

(b) In Alternative 2, the parameters needed are: the prior probability of $I$, $P(I)$, and the conditional probabilities $\{P(I_i|I), i = 1, \ldots, n\}$. This makes a total of $2n + 1$ values. Regarding independence, the structure implies the conditional independence of the $I_i$’s given $I$ (for each $i = 1, \ldots, n$).

It is also interesting to analyze the evolution of the probabilities of the network as new evidence is acquired:

(a) In Alternative 1, evidence about mastering an item $I_j$ changes the probability of mastering its child $I$. Evidence about mastering $I$ changes the probability of its parents $I_i$, $I = 1, \ldots, n$, and opens communication among them (further evidence about $I_i$ will affect the certainty of $I_j$).

(b) In Alternative 2, evidence about mastering an item $I_j$ changes the probability of its parent $I$, which in turn changes the probabilities of the other children $I_i$ ($i \neq j$). Evidence about mastering $I$ changes the probabilities of its children $I_n$, for

---

The number of parameters is $n + 2^n$, but one of the parameters does not need to be specified as it can be computed providing that the probabilities must add up to 1.
Thus, the main differences are that in Alternative 2, evidence about mastering an item $I_j$ affects the probability of mastering the rest of the items of the same level, $I_i$ (with $i \neq j$) and that the evidence about $I$ opens (Alternative 1) or blocks (Alternative 2) the communication among the $I_i$s. It is not clear which of the two alternatives models aggregation relationships better. Perhaps this is the reason why examples of both of them can be found in the literature. For example, Alternative 1 was chosen by Van Lehn and his team for the ANDES system (Conati et al., 1997; Van Lehn, 1996) and also by the ARIES team (Collins et al., 1996) in their studies about Adaptive Testing, whereas Alternative 2 was chosen by Mislevy and Gitomer (Mislevy and Gitomer, 1996) in HYDRIVE, and also by Murray in his Desktop Associates (Murray, 1998). Nevertheless, none of them compare both alternatives or justify their decision.

In our model we have chosen Alternative 1. The main reasons for this choice are:

(a) From the point of view of knowledge representation, Alternative 1 considers that the student’s learning occurs in a gradual and incremental way. That is, when a student learns a topic, the usual procedure is to study each of the parts that compose the topic (usually in the order suggested by the teacher). In the same way, if a student is acquiring an ability (for example, learning how to use certain instruments), this ability is acquired by learning each of the necessary subskills (learning how to use each instrument).

(b) From the point of view of evidence propagation, we have discarded Alternative 2 because, in our opinion, evidence that a certain item $I_i$ is mastered should not increase our belief that other items $I_j$ are mastered (unless there is independent confirmation that item $I_j$ is mastered), since this would mean that when we study a concept our probability of knowing another concept belonging to the same topic increases.

(c) As for parameter specification, Alternative 1 could seem more complex, because it requires an exponential number of parameters instead of the polynomial number required by Alternative 2. However, in Sections 3.2.2 and 3.2.3 we show how the definition of the knowledge nodes allows us to use an equivalent network whose parameters can be easily computed from the set of weights defined.

3.2.2. Relationships Between Concepts and Topics

As just discussed, we consider that knowing each of the concepts in a topic has a causal influence on knowing the topic, and therefore the BN corresponding to these relationships has the structure depicted in Figure 4.
The parameters of this network are:

- Prior probabilities of each concept, \( \{p_i, i = 1, \ldots, n\} \)
- The conditional probability \( P(T|[C_i]_{i=1,\ldots,n}) \), given by the expression:

\[
P(T = x|[\{C_i = 1\}_{i \in S}, \{C_j = 0\}_{j \notin S}]) = \begin{cases} 
1 & \text{if } x = \sum_{i \in S} w_i \\
0 & \text{otherwise}.
\end{cases}
\]

where \( S = \{j \in \{1, \ldots, n\} \text{ such that } C_j = 1\} \).

Then, when this network is initialized, we obtain the probability law of the random variable \( T \).

The values taken by the random variable \( T \) can be easily interpreted: if \( T \) takes a certain value \( x \in [0,1] \), this means that if the student is asked to apply his/her knowledge about topic \( T \) in \( n \) situations, he/she will demonstrate mastering the topic in \( x n \) situations, where the set of possible situations is content balanced, that is, if the total number of situations is \( n \), \( w_i \) of them are relevant to the elementary concept \( C_i \), for each \( i = 1, \ldots, n \).

The behavior of the BN depicted in Figure 4 can be emulated with an equivalent BN, which we define next and show in Figure 5.

In this network all the variables are binary (i.e., the set of possible values is \( \{0, 1\} \)), and the parameters are defined by:

- Prior probabilities for the \( C_i \), \( P(C_i = 1) = p_i \), for each \( i = 1, \ldots, n \).
- Conditional distribution of \( T' \) given the values of the \( C_i \) that is:

\[
P(T'|[C_1, \ldots, C_n]) = \sum_{i \in S} w_i.
\]

This binary random variable \( T' \) does not have clear semantics. The motivation for introducing it is that, as the next proposition shows, it allows us to determine...
the value that the continuous random variable $T$ takes. In this way, we can use the BN represented in Figure 5 instead of the BN represented in Figure 4, with the advantage that all of its nodes are binary (therefore making its specification and handling easier).

**PROPOSITION 1.** Let us assume that the random variables $C_1, \ldots, C_n$ take a certain set of values, that is, for a certain $S$ subset of $\{1, \ldots, n\}$ we have that $C_i = 1$ for each $i \in S$ and $C_i = 0$ for each $i \notin S$. Then, the random variable $T$ takes a certain value $x$ if and only if the a posteriori probability that the random variable $T'$ takes the value 1 is $x$, that is,

$$P'(T = x) = 1 \leftrightarrow P'(T' = 1) = x$$

**Proof.** First we show the necessary condition. Let $S$ be the subset of $\{1, \ldots, n\}$ such that $C_i = 1$ for each $i \notin S$ and $C_i = 0$ for each $i \notin S$. Then, if $x$ represents the value that the random variable $T$ takes, that is, $x = \sum_{i \in S} w_i$, the a posteriori probability that the random variable $T'$ takes the value 1 is:

$$P'(T' = 1) = \sum_{i \in S} w_i = x$$

To show that the condition is also sufficient, let $x$ be the a posteriori probability that $T'$ takes the value 1. Then, necessarily, $x = \sum_{i \in S} w_i$, and therefore $P'(T = x) = 1$. $\square$

We recall that the importance of this proposition is that it allows us to use the BN shown in Figure 5 to obtain an estimation of the student’s knowledge level in topic $T$, with the advantage that we are dealing with a binary variable $T'$ instead of with a discrete variable $T$.

### 3.2.3. Relationship Between Topics and Subject

As discussed in Section 3.2.1, we also consider that knowing each of the topics in a subject has a causal influence on knowing such a subject, and therefore, adding these relationships to the BN that represents the concepts and topics we obtain the BN shown in Figure 6.

The parameters of this network are:

- The prior probabilities of knowing each concept, $\{p_{ij}, i = 1, \ldots, r; j = 1, \ldots, n_i\}$.
- For each $i = 1, \ldots, r$, the conditional distribution $P(T_i | \{C_{ij} \}_{j=1}^{n_i})$, given by the expression:

$$P(T_i = x | \{C_{ij} = 1\}_{j \in S_i}, \{C_{ik} = 0\}_{k \notin S_i}) = \begin{cases} 1 & \text{if } x = \sum_{j \in S_i} w_j \\ 0 & \text{otherwise} \end{cases}$$

where $S_i = \{j \in \{1, \ldots, n_i\} \text{ such that } C_{ij} = 1\}$ for each $i = 1, \ldots, r$.
The conditional distribution $P(A|\{T_i, i = 1, \ldots, r\})$, given by

$$P(A = x|\{T_i = \sum_{j \in S_i} w_j\}_{i=1,\ldots,r}) = \begin{cases} 1 & \text{if } x = \sum_{i=1}^r \sum_{j \in S_i} w_j \\ 0 & \text{otherwise} \end{cases}$$

If we initialize this network, we obtain the probability law of the random variable $A$.

The interpretation of the values that the random variable $A$ takes is similar to the interpretation of the values of the topics: the random variable $A$ takes a certain value $k \in [0,1]$ when the student shows mastery of the subject in $kn$ out of $n$ situations relevant to subject $A$, where the set of $n$ situations is content balanced, that is, it takes into account the relative importance of each concept in a topic and of each topic in the subject. In this way, if the total number of situations is $n$, some of them are relevant to the elementary concept $C_{ij}$, for each $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$.

Next, we show that, as before, the behavior of the BN shown in Figure 6 can be emulated with the equivalent BN depicted in Figure 7.

In this BN all the variables are binary, and its parameters are:

- Prior probabilities for the $C_{ij}$, $P(C_{ij} = 1) = p_{ij}$ for $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$.
- Conditional distribution of $T_i'$ given the $C_{ij}$, defined as:

$$P(T_i'|C_{ij}, \ldots, C_{im}) = \sum_{j \in S_i} w_{ij}.$$
Conditional distribution of $A'$ given the $T_i$, defined as:

$$P(A'|T_1, \ldots, T_s) = \sum_{i \in V} x_i,$$

where $V = \{i|1, \ldots, r\}$ such that $T_i = 1$.

The following proposition shows that the BNs depicted in Figures 6 and 7 have an equivalent behavior.

**Proposition 2.** Let us assume that the random variables $C_1, \ldots, C_n$ take a certain set of values, that is, for a certain $S$ subset of $\{1, \ldots, n\}$ we have that $C_i = 1$ for each $i \in S$ and $C - i = 0$ for each $i \notin S$.

- For each $i = 1, \ldots, s$, the random variable $T_i$ takes a certain value $x$ if and only if the a posteriori probability that the random variable $T_i$ takes the value 1 is $x$.
- The random variable $A$ takes a certain value $x$ if and only if the a posteriori probability that the random variable $A'$ takes the value 1 is $x$.

**Proof.** The first part has already been shown in Proposition 1. For the second part, we only need to apply the same proposition to the part of the network that contains the binary random variables $T_i$ and the random variable $A$.

In order to illustrate these results let us consider a simple example.

**Example 2.** Let us assume that a student is learning how to identify a certain vegetable species, in such a way that knowing the subject consists of being able to correctly identify vegetables belonging to three different species, that we call species 1, 2, and 3.

The relative importance of the topics is measured in terms of a set of weights that are specified by the teacher. Let us assume that these weights are $w_1 = 0.2$, $w_2 = 0.5$, and $w_3 = 0.3$, meaning that a balanced exam for this subject should contain 20% of questions relevant to species 1, 50% relevant to species 2, and 30% relevant to species 3.

Let us also assume that there is a student whose probabilities of correctly identifying species 1, 2, and 3 are 0.8, 0.6, and 0.7, respectively. What is the knowledge level reached by this particular student in the subject?

The traditional way of measuring this knowledge level is to calculate the percentage of correct answers in the exam. This value can be computed using the total probability law. Let $A$ be the event ‘the student gives the correct answer to a question about the subject’, and, for each $i = 1, 2, 3$ let $B_i$ be the event ‘the question is relevant to species $i$’.
Then, by the total probability law we have that:

\[
P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)
\]

\[
= 0.8 \cdot 0.2 + 0.6 \cdot 0.5 + 0.7 \cdot 0.3 = 0.73
\]

Meaning that, if this student is presented with a balanced set of \( n \) questions, he/she will give the correct answer to 0.73 \( n \) of them.

Let us now show how the BN defined can emulate this behavior. The nodes in the network are: \( I = \) knowledge about the subject, and \( E_i = \) knowledge about species \( i \), for \( i = 1, 2, 3 \). Then, the random variable \( I \) is defined as \( I = 0.2 \cdot E_1 + 0.5 \cdot E_2 + 0.3 \cdot E_3 \), and the equivalent BN (with \( I \) defined as in Section 3.2.2.) is depicted in Figure 8.

The parameters of this network are the prior probabilities of each \( E_i \) (for this particular student are \( P(E_1 = 1) = 0.8 \), \( P(E_2 = 1) = 0.6 \), and \( P(E_3 = 1) = 0.7 \)) and the conditional distribution \( P(I'|E_1E_2E_3) \) (which is computed by adding up the weights associated with the \( E_i \)'s that take the value 1). This conditional distribution is given in Table 3:

Then, when we initialize the network we obtain that \( P(I' = 1) = 0.73 \), meaning that the knowledge variable \( I \) takes the value 0.73, i.e. the expected percentage of correct answers that a student will give in a balanced exam is 73%.

3.2.4. Modeling Relationships Between Knowledge and Evidential Nodes

In this section we discuss how to model the relationships between knowledge and evidential nodes. Two different models will be presented: a static model, with a traditional BN, and a dynamic model, in which a dynamic BN is used.

3.2.4.1. Static Model. Once again, we have two alternatives: to consider that the knowledge nodes \( K_1, \ldots, K_n \) can have an influence on the evidential nodes, or, conversely, that evidential nodes have a causal influence on knowledge nodes. Both

![Figure 8. BN for the identification of vegetable species.](image)
alternatives are graphically represented in Figure 9.

The first alternative is directly based on the notion of causality: knowledge has a causal influence on being able to solve situations related to these concepts. The second alternative corresponds to representing knowledge in terms of rules: if a situation is solved correctly, then this provides evidence about knowing the items involved. Then, we have:

(a) In Alternative 1, the parameters to specify are the prior probabilities of mastering each $K_i$ ($P(K_i)$, $i = 1, \ldots, n$), and the conditional distribution of the evidential nodes, $P(E_j|\{K_i\text{ such that } K_i \in \text{pa}(E_j)\})$, $j = 1, \ldots, s$. The independence structures implied by this alternative are:

- $K_i$, $i = 1, \ldots, n$, are mutually independent a priori;
- $K_i$ is independent of $E_j$ for each $E_j$ which is not a child of $K_i$, $i = 1, \ldots, n$;
- $E_j$ is independent of every $E_i$ (with $i \neq j$) given $\text{pa}(E_j)$, $j = 1, \ldots, s$;
- $E_j$ is independent of $K_i$ for each $i$ such that $K_i \in \text{pa}(E_j)$, $j = 1, \ldots, s$.

(b) In Alternative 2, the parameters needed are: prior probabilities for the $E_j$, $P(E_j)$, $j = 1, \ldots, s$, and the conditional distribution of $K_i$ given its parents, that is, $P(K_i|\text{pa}(K_i))$, $i = 1, \ldots, n$. This structure implies the following independences:

- $E_j$, $j = 1, \ldots, s$, are mutually independent a priori.
- $E_j$ is independent of $K_i$ for each $K_i$ which is not a child of $E_j$, $j = 1, \ldots, s$;
- $K_i$ is independent of each $K_j$ (with $i \neq j$) given $\text{pa}(K_i)$, $i = 1, \ldots, n$;
- $K_i$ is independent of $E_j$ for each $j$ such that $E_j \notin \text{pa}(K_i)$, $i = 1, \ldots, n$.

Therefore, the second alternative would imply the independence of $K_i$ given the evidence, which simply is not true as already discussed in (VanLehn et al., 1998).

Let us see a simple counterexample: suppose that being able to correctly answer a certain question $P$ requires mastering two KIs $K_1$ and $K_2$, and that question $P$ has been answered incorrectly. Then, knowing that the student knows $K_1$ should decrease the probability that the student knows $K_2$. However, since $K_1$ and $K_2$ are conditionally independent given $P$, evidence about $K_1$ will not affect the probability of $K_2$ in the way it should.

Thus, in this case, the alternative chosen is Alternative 1, that is, the one that better describes the behavior we want the network to have.

**Figure 9.** Alternatives to model relationships between knowledge and evidential nodes.
3.2.4.2. Dynamic Model. In contrast to other domains in which traditional BNs are used, the student modeling problem has the particularity that the state of the knowledge nodes can change over time. This is especially clear in the case of the evidential nodes. The fact that we pose a student a question with several related concepts and the student gives the correct answer does not mean that if we ask another question of the same type (involving the same concepts) the student will also solve it correctly. However, if we use a traditional BN, once a question has been posed the evidential node is blocked with the answer obtained, and therefore it cannot be used again. This behavior does not adequately describe the real situation, in which a teacher can ask the same type of question twice or more to be sure that the student is able/unable to solve it. For this reason, we think that the use of a dynamic model is especially suitable for these relationships.

Let us briefly describe the proposal presented in (Reye, 1998) regarding the application of dynamic BNs to the student modeling problem. In this proposal, for each \( j = 1, \ldots, k, \ldots \), the following nodes are defined:

\[
\begin{align*}
L_j & = \text{student’s state of knowledge after the } j\text{th interaction with the system.} \\
O_j & = \text{result of the } j\text{th interaction.}
\end{align*}
\]

The relationships between these nodes are depicted in Figure 10.

In our case, we define the following variables:

\[
\begin{align*}
K_j^i & = \text{student’s state of knowledge about item } K_i \text{ after } j \text{ interactions with the system, for } i = 1, \ldots, n \text{ and } j = 0, \ldots, k, \ldots \\
E_j^i & = \text{result of the } j\text{th interaction with the system (where, in this case, the interaction consists of evidence gathering), for } i = 1, \ldots, n \text{ and } j = 1, \ldots, k, \ldots
\end{align*}
\]

In this way, nodes \( K_j \) play the role of nodes \( L \), and nodes \( E_j \) play the role of nodes \( O \). The only difference is that, in this case, the interaction with the system is reduced to evidence gathering, so it is not necessary to introduce the links between nodes \( E_j^{i-1} \) and \( K_j^i \). As the discussion about the appropriate direction of the links between knowledge and evidential nodes presented in Section 3.2.4.1 is also applicable to this case, the dynamic BN is constructed from the BN depicted in Figure 10. The relationship between two successive interactions \((j-1)\)th and \(j\)th of the dynamic BN is shown in Figure 11.

The parameters for this BN are:

- The prior probabilities of nodes \( K_0^i \), that is, \( P(K_0^i) \) for \( i = 1, \ldots, n \).

![Figure 10. BN for student modeling.](image-url)
The conditional distribution of $E_j^i$ given its parents, that is, $P(E_j^i | \text{pa}(E_j^i))$ for $i = 1, \ldots, n$ and $j = 0, \ldots, k, \ldots$.

The conditional distribution of $K_j^i$ given $K_j^{i-1}$, that is, $P(K_j^i | K_j^{i-1})$ for $i = 1, \ldots, n$ and $j = 0, \ldots, k, \ldots$.

The relationship between these parameters with the parameters of the BN depicted in Figure 9 is:

- $P(K_j^i) = P(K_j)$, for $i = 1, \ldots, n$.
- $P(E_j^i | \text{pa}(E_j^i)) = P(E_j | \text{pa}(E_j))$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k, \ldots$

The only new parameters are $\{P(K_j^i | K_j^{i-1})\}$ for $i = 1, \ldots, n$ and $j = 0, \ldots, k, \ldots$. As we assume that an interaction consisting in evidence gathering does not change the student’s knowledge state, such probabilities are easy to specify and are given by the following expression:

$$P(K_j^i = x | K_j^{i-1} = y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

In this way, for each $j = 1, \ldots, k, \ldots$ and for each $i = 1, \ldots, n$ the probability distributions of $K_j^{i-1}$ and $K_j^i$ are the same.

3.2.4.3. Relationships Between Concepts and Test Items. As discussed above, the relationships between concepts and test items are modeled with networks like the one depicted in Figure 12.

Figure 11. Dynamic BN to model relationships between knowledge and evidential nodes.
Therefore, the parameters that we need to specify in this part of the network are the prior probabilities of the concepts and the conditional probabilities of the test items given the concepts. In order to simplify as much as possible the specification of the conditional probabilities, we have modified the approach described in (Van Lehn et al., 1998) that basically consists of considering that:

- The probability that a test item is correctly answered, given that the student knows every related concept, is $1 - s$, where $s$ is a slip factor.
- The probability that a test item is correctly answered, given that one or some of the concepts related to the item are not known, is $k/n$, where $n$ is the number of possible answers and $k$ is a factor that represents the probability that a student will try to guess the correct answer.

The main drawback of this approach is that it assumes that it is as equally likely that a student gives a correct answer when only one of the related concepts is not known as when he/she does not know any of them. We consider that this probability should depend on the number of concepts that are mastered and on the importance of these concepts, that is, the more knowledge the student has, the more likely it is that he/she will guess the correct answer. This is especially true in the case of test items, where the student can choose the answer by discarding the incorrect ones.

Our approach is as follows: let $F(x)$ be the 3-parameter logistic function in IRT theory, that is:

$$F(x) = c + \frac{1 - c}{1 + \exp(-1.7a(x - b))} x \in IR$$

where $c = 1/n$, $a$ is the discrimination index and $b$ is the difficulty level (see Section 2.2). A new function $G$ is defined by:

$$G(x) = 1 - \frac{(1 - c)(1 + \exp(-1.7ab))}{1 + \exp(1.7a(x - b))} x \geq 0$$

We show in Figure 13 how the function $F$ has been transformed:

We can see that $G(0) = c$. This function $G$ will be used to compute the probabilities of giving the correct answer to the test item depending on the number of concepts.

---

**Function $G$ has been defined as a linear transform of function $F$, i.e. $G(x) = a + bF(x)$, where $a$ and $b$ have been computed to satisfy $G(0) = c$ and $\lim_{x \to \infty} G(x) = 1$.**
known by the student in the following way: if the student does not know any of the related concepts, the probability of choosing the right answer is set to be $c = 1/n$. If all the concepts are mastered, it will be $1 - s$. The rest of the values are interpolated between $c$ and $1 - s$, using function $G$, as illustrated in Figure 14.

The way in which function $G$ is used is described as follows. Let $x^*$ be such that $G(x^*) = 1 - s$, and let us assume that the test item has $p$ related concepts. Then, the values that will be used to compute the $2^p$ probabilities needed are:

$$\left\{ G(0), G\left(\frac{x^*}{p-1}\right), G\left(\frac{2x^*}{p-1}\right), \ldots, G\left(\frac{(p-2)x^*}{p-1}\right), G(x^*) \right\}$$

In order to assign such values, we also take into account the importance of the concepts. In this way, $G(0)$ (which is $1/n$) will be assigned to the probability of giving the correct answer when none of the related concepts are known, $G(x^*/(p-1))$ will be assigned to the probability of choosing the correct answer when only the least important concept is known, and so on$^6$. In this way, the teacher will only need to provide a discrimination index and a difficulty parameter, and the conditional probabilities needed can be automatically computed using the method described. This procedure is illustrated in Example 3.

$^6$Ties are broken using the binary order.
In this way, our approach takes into account that the probability of choosing the correct answer increases as knowledge is more complete, and therefore it is obvious that it will produce a more accurate diagnosis than the approach described in (Van Lehne et al., 1998).

EXAMPLE 3. To illustrate the procedure described above, we present an easy example in which four concepts $C_1$, $C_2$, $C_3$, and $C_4$ are needed to answer a question $P$. Let us suppose that the concepts are ordered according to their importance, $C_1$ being the most important one and $C_4$ the least important one. In this example, the values for the parameters are set to be $c = 0.25$, $s = 0.01$, $b = 5$, and $a = 0.3$.

First, we need to compute $x^*$ such that $G(x^*) = 0.99$. We obtain that $x^* = 13.716$. The probabilities assigned to each one of the sixteen different combinations of known concepts are given in the last column of Table 4, where the values $kx^*/15$, for $k = 0, \ldots, 15$, are given in the column labeled $x$.

As we can see, the more complete the student’s knowledge is, the bigger the probability of correctly answering the question that the procedure assigns. There are, however, some cases not covered by this rule, such as the case in which we have a student who knows concepts $C_1$ and $C_4$ and a student who knows concepts $C_2$ and $C_3$. The approach we have taken to solve this situation is to assign probabilities according to the binary order, so the bigger probability is assigned to the student that knows concepts $C_1$ and $C_4$.

4. Bayesian Adaptive Tests

In this section, we present a new algorithm for Adaptive Testing based on BNs, that allows diagnosing several abilities at the same time. This algorithm is a crucial part of the evaluation process, since it will perform the diagnostic process.

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
<th>$x$</th>
<th>$G(x) = P(P=1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.914</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.914</td>
<td>0.233</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2.743</td>
<td>0.345</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.658</td>
<td>0.427</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.572</td>
<td>0.522</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5.487</td>
<td>0.622</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.401</td>
<td>0.717</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7.316</td>
<td>0.797</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.230</td>
<td>0.861</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.145</td>
<td>0.907</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10.059</td>
<td>0.939</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10.973</td>
<td>0.961</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11.888</td>
<td>0.975</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>12.802</td>
<td>0.984</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13.717</td>
<td>0.99</td>
</tr>
</tbody>
</table>
4.1. STRUCTURE OF THE NETWORK

Adapted Bayesian tests take place on the network structure presented in Section 3, that is, the knowledge nodes are concepts, topics, and subjects, and the evidential nodes can be test items or general questions (provided that the ITS has the ability to diagnose the correctness of the solution). In this work we have considered three levels of granularity. The extension to allow an arbitrary number of levels of granularity $k$ is immediate.

Thus, the structure of the BN that is used in the adaptive tests is depicted in Figure 15.

The evaluation process consists of two phases:

- **Diagnostic phase**, which is performed in the part of the network that contains the concepts and the relationships between them. The goal of this phase is to determine the set of concepts that the student knows/does not know from the answers given to related test items.
- **Evaluation phase**, where, from the results of the previous phase, the probabilities will be propagated to determine the knowledge level reached by the student at the different levels of granularity, that is, the knowledge level reached in each of the topics and in the subject.

Thus, the adaptive test is responsible for the diagnostic process, in which only the lower part of the network is used (concepts and questions). Once the test has finished, the evaluation process takes care of estimating the degree of knowledge reached by the student in each of the topics and in the subject. The BN is therefore divided in two parts, as illustrated in Figure 16.

Having presented the whole process of evaluating a student, we describe the diagnostic process in detail in the following section.

4.2. BASIC ELEMENTS OF THE BAYESIAN ADAPTIVE TESTING ALGORITHM

As described in Section 2, the basic elements of a CAT are:
We now use this description to present the elements of the algorithm that we propose as a basis to carry out Bayesian adaptive tests. In the following section, we describe each of these basic elements.

4.2.1. Item Response Model
Once the network is defined, the item response model is given by the conditional probability of the item given its parents. To this end, in Section 3.2.4.3 we have proposed the use of a modification of the 3-parameter logistic function to measure the relationship between knowing a set of concepts and correctly answering a question related to them, that is, to compute the conditional distribution needed.

4.2.2. Scoring Method
The scoring method is given by the use of the Bayesian model, since the algorithm of probability propagation provides a sound method to evaluate the answers, i.e. to estimate the knowledge level of the concepts involved according to the answers given by the students to the test items.

To carry out this probability propagation, a goal-oriented algorithm as described in (Castillo et al., 1997) is used to determine the set of relevant nodes with the objective of reducing computational complexity. Thus, each time the student answers a question the goal-oriented algorithm is used to compute the reduced
subgraph where the propagation will take place. In this way, the efficiency of the propagation process is increased.

4.2.3. Item Pool
Regarding the item pool, the use of the 3-parameter logistic function provides a simple way of specifying the required parameters – and therefore of calibrating the questions – which takes into account not only unintentional slips and lucky guesses, but also the fact that the probability of giving the right answer increases as the set of related concepts known by the student is bigger. Moreover, it makes possible the use of the traditional IRT parameters: guessing factor, difficulty level, and discrimination index.

4.2.4. Initial Level
To set the initial level, ideally we should use the available information about the particular student that is going to take the test. However, in many practical cases this information might not be available. A simpler option is to divide the student population into stereotypes with different initial levels (certain types of students are more likely to know certain types of concepts). In the absence of any other information, it seems reasonable to use a uniform distribution, that is, to consider that it is equally likely that the student knows/does not know each of the elementary concepts.

4.2.5. Item Selection Criteria
Regarding the item selection method, several criteria are presented and are described next. These criteria are used to select the best question to ask given the current estimation of the student’s knowledge level. The final goal of using such criteria is to achieve more precise estimations of the student’s knowledge level with shorter tests. In Sections 4.2.5.1 and 4.2.5.2 we describe the criteria proposed.

4.2.5.1. Random Criterion. The easiest criterion is the random criterion, which we denote by $C_R$. With this criterion, questions are selected randomly, with each question in the database having the same probability of being selected. The diagnostic and evaluation methods are based on the BN model described. It is obvious that this criterion is not adaptive, but we have used it so we can test the performance of the Bayesian diagnostic algorithm and compare it with the performance obtained by using adaptive criteria.

4.2.5.2. Adaptive Criteria. Adaptive criteria choose the best question to ask next according to the performance shown by the student in the previous items, or, more precisely, to the estimation of the knowledge level reached by the student that has been obtained from the answers given to previous items. We have defined two different types of adaptive criteria: criteria based on the information gain that
a question provides, and conditioned criteria, which are based on the idea of favoring the behavior shown by the student so far.

4.2.5.2. Criteria Based on the Information Gain. We first define the concept of the utility of a question \( P \) for a knowledge node \( C \).

**DEFINITION 1.** Given an evidential node \( P \) and a knowledge node \( C \), we define the utility \( U_1 \) of node \( P \) for node \( C \) as:

\[
U_1(P, C) = \frac{|P(C = 1|P = 1) - P(C = 1)|P(P = 1)}{P(C = 0)P(P = 0)}
\]

The interpretation of this utility measure is simple: the utility of an evidential node is defined as the expected gain of information. Note that what we do is to calculate the change in the probability of \( C \) according to the result of the evidential node \( P \), and then weighting this change with the probability of each possible result. Therefore, the most informative evidential node for a given item will be the one with the maximum utility.

Due to the type of relationships defined in our network, we know that when an evidential node is answered correctly the probability of the knowledge node increases, and when it is answered incorrectly, the probability of the knowledge node decreases. This means that we can disregard the absolute values in the definition of the utility \( U_1 \) measure, and use the following expression:

\[
U_1(P, C) = \frac{P(C = 1|P = 1) - P(C = 1)}{P(C = 0)P(P = 0)}
\]

Thus, in the context of adaptive tests, the most informative question will be the one with the greatest utility. We can see that the utility of a question is affected by the student’s knowledge level, since the probabilities of correctly/incorrectly answering the question are used as weights, and of course these probabilities depend on the current estimation of the student’s knowledge level.

In (Collins et al., 1996), the concept of utility is defined as

\[
U_C(P, C) = |P(C = 1|P = 1) - P(C = 0|P = 0)|
\]

Although the authors have obtained satisfactory results in the simulations, in our opinion this measure is not appropriate, because ideally both \( P(C = 1|P = 1) \) and \( P(C = 0|P = 0) \) should be maximized and therefore it does not make much sense to maximize their difference.

The utility measure that we propose has a drawback. In an adaptive test, calculating the utility of the questions in the item pool means instantiating the

\footnote{The probability of knowing \( C \) should increase (decrease) as much as possible when the student displays knowledge (no knowledge) when answering question \( P \).}
network twice for each question (considering right and wrong answers). Given that
the number of questions in a good item pool should be large, this process can
be computationally intensive, which cannot be afforded since students’ waiting times
should be short.

Fortunately, this problem can be easily solved. We just need to apply Bayes
Theorem in the expression that defines the utility to obtain:

\[
U_1(P, C) = (P(P = 1|C = 1) - P(P = 1))P(C = 1)
+ (P(P = 0|C = 0) - P(P = 0))P(C = 0)
\]

The advantage of this new expression is that to compute the utilities we need to
instantiate the concepts instead of the questions. This results in a large com-
putational saving, since the number of concepts is typically much smaller than
the number of questions. Thus, for example, in a very simple network with only
one concept C and k related questions P_1, P_2, \ldots, P_k, the computation of the utilities
of each question P_i (i = 1, \ldots, k) for the concept C requires instantiating the network
2k times if we use the first expression and only twice if we use the second one. At
the same time, the instantiations required in the calculation of the utility of a question
take place in the subgraph of relevant nodes generated by the use of the goal-oriented
algorithm. In this way, we have achieved affordable waiting times (less than a
second) in the simulations carried out \(^8\).

We now give an alternative definition to the concept of utility.

DEFINITION 2. Given an evidential node P and a knowledge node C, the utility of
node P for node C is defined as

\[
U_2(P, C) = P(P = 1 | C = 1)P(C = 1) + P(P = 0 | C = 0) = P(C = 0)
\]

This utility measure also has a simple interpretation: we give priority to those ques-
tions with a greater degree of sensitivity and specificity \(^9\), or, equivalently, with a
smaller rate of false positives (students who answer correctly without knowing
the concept) and false negatives (students who answer incorrectly even though they
know the concept) \(^10\).

\(^8\)The trial network has fourteen concepts and one hundred questions.
\(^9\)In medicine, the sensitivity of a test T for an illness I is defined as P(T = 1|I = 1) (proportion of
positive results in the test among people that have the illness), and the specificity of a test T for an
illness I is defined as P(T = 0|I = 0) (proportion of negative results in the test among people that
do not have the illness). Obviously, tests with higher sensitivity and specificity are preferred. These
concepts can be easily extended to the field of student modeling, with questions playing the role of
tests and knowledge items playing the role of illnesses.

\(^10\)False positives are the complement of sensitivity and false negatives are the complement of specific-
ity. Therefore, minimizing false positives (false negatives) is equivalent to maximizing sensitivity
(specificity).
Another interpretation for this utility measure comes from simplifying its expression:

\[
U_2(P, C) = P(P = 1 \land C = 1) + P(P = 0 \land C = 0) = P(P = C)
\]

That is, this utility gives the probability that the variables \( P \) and \( C \) take the same value.

We therefore have two different definitions for the concept of utility: \( U_1 \), based on the \textit{expected gain of information}, and \( U_2 \), based on the concepts of sensitivity and specificity of a question.

Having defined the utility of a question for each of the concepts involved, we have to define the \textit{global utility} of a question. We propose two different criteria, each of which is based on a different definition of global utility:

- \textbf{Criterion of the sum}, in which the global utility of a question is defined as the sum of the utilities of the question for each of the related concepts, i.e.:

\[
U(P) = \sum_{C \in \text{pa}(P)} (P, C)
\]

However, this criterion could penalize those questions related to a small number of concepts, since their definition of global utility would have fewer adding terms. To avoid this, we introduce a second way of defining global utility:

- \textbf{Criterion of the maximum}, in which the global utility of a question is defined as the maximum of the utilities of the question for each of the related concepts, i.e.:

\[
U(P) = \max_{C \in \text{pa}(P)} U(P, C)
\]

Combining the two definitions of utility of a question for a concept and the two definitions of the global utility of a question, we have four adaptive criteria based on the concept of utility:

- Criterion of the \textit{sum} of the utilities, where the utility is defined as the \textit{expected gain of information}, that we denote as \( C_{SG} \).
- Criterion of the \textit{maximum} of the utilities, where the utility is defined as the \textit{expected gain of information}, that we denote as \( C_{MG} \).
- Criterion of the \textit{sum} of the utilities, where the utility of a question for a concept is defined in terms of the concepts of \textit{specificity} and \textit{sensitivity}, that we denote as \( C_{SS} \).
- Criterion of the \textit{maximum} of the utilities, where the utility of a question for a concept is defined according to the concepts of \textit{specificity} and \textit{sensitivity}, that we denote as \( C_{MS} \).

4.2.5.2.2. \textit{Conditional Criteria}. The conditional criteria are based on the idea of taking into account the tendencies shown by the student in previous questions. The utility of the question is then defined as the sensitivity or the specificity...
themselves, depending on the knowledge that the student is showing so far. We propose two different criteria:

- **Criterion conditioned by the probability of the concept.** The utility of a question is calculated by the expression:

\[ U(P) = \max_{C(p|P)} U'(P, C) \]

where \( U'(P, C) \) is defined as:

\[ U'(P, C) = \begin{cases} P(P = 1/C = 1) & \text{if } P(C = 1) > P(C = 0) \\ P(P = 0/C = 0) & \text{otherwise.} \end{cases} \]

The idea of this criterion consists of choosing the most specific or the more sensitive question according to whether the student shows/does not show knowledge about the concept. We denote this criterion by \( C_{CC}. \)

- **Criterion conditioned by the probability of the question.** The utility of a question is computed by the expression:

\[ U'(P) = \begin{cases} \max_{C(p|P)} P(P = 1/C = 1) & \text{if } P(C = 1) > P(C = 0) \\ \max_{C(p|P)} P(P = 0/C = 0) & \text{otherwise.} \end{cases} \]

This criterion is similar to the previous one, but instead of choosing the sensitivity or the specificity depending on the probability of the concept, we take one or the other depending on the probability of the question being answered correctly or not. We denote this criterion by \( C_{CQ}. \)

To summarize, the seven criteria that are analyzed and compared are\(^{11}\):

- Random Criterion, \( C_R. \)
- Criterion of the Sum of the utilities where the utility is defined as the expected Gain of information, \( C_{SG}. \)
- Criterion of the Maximum of the utilities where the utility is defined as the expected Gain of information, \( C_{MG}. \)
- Criterion of the Sum of the utilities where the utility of a question for a concept is defined in terms of the concepts of Specificity and Sensitivity, \( C_{SS}. \)
- Criterion of the Maximum of the utilities where the utility of a question for a concept is defined according to the concepts of Specificity and Sensitivity, \( C_{MS}. \)
- Criterion Conditioned by the probability of the Concept, \( C_{CC}. \)
- Criterion Conditioned by the probability of the Question, \( C_{CQ}. \)

We will discuss the results of this study in Section 5.

\(^{11}\) Some other criteria were also considered and evaluated, such as averaging the sum with the number of concepts involved in the definition of the global utility and a criterion based on the traditional definition of information gain in Information Theory. However, the results obtained by using such criteria were very poor compared to the results obtained with the seven criteria analyzed in this paper.
4.2.6. Termination Criteria

As a termination criterion we have used a combination of two criteria: the test finishes when a previously fixed maximum number of questions is reached, or when all the concepts have been evaluated\footnote{The termination criterion used when testing the random question selection criterion was different. In this case we considered a fixed value for the length of the test.}. To determine whether a concept has been evaluated, we establish a certain level $l \in [0, 0.5)$. If the probability of knowing a concept is greater than or equal to $1 - l$ then the concept is diagnosed as known, whereas if it is smaller than $l$, the concept is diagnosed as unknown. All those concepts whose probability is between $l$ and $1 - l$ will be considered as non-diagnosed. Therefore, a test can finish even though some concepts have not been diagnosed if the fixed maximum number of questions is reached. This mechanism avoids tests which are too long. Note that depending on the regularity of the student’s answers there could be concepts that are never diagnosed.

5. Evaluation of the Algorithm Using Simulated Students

We used simulated students for the evaluation of the algorithm. Simulated students have also been used for this purpose in (Collins et al., 1996; Van Lehn et al., 1998; Van Lehn et al., 1995). This allowed us to evaluate the algorithm without defining a test for a particular subject or having a test group of real students. The main reasons for using simulated students are:

- **Pre-evaluation of the validity of the method.** It does not seem appropriate to test an evaluation method with real people without proving its validity beforehand. Of course, a method that has not been tested previously should never be used to grade students. We could have asked the students to volunteer in the evaluation of the algorithm, but in this case the students’ motivation to answer the test cannot be compared with their motivation to answer a test that will be used to actually evaluate them.

- **Subjectivity of teachers’ estimations.** Even in the case of having a set of sufficiently motivated students, the estimations of the knowledge level that we obtain with our testing algorithm would have to be compared with human estimations, which would be obtained either from direct knowledge about the student’s performance or by the use of traditional evaluation methods, such as exercises, exams, etc. In any case, these estimations are always subject to some degree of error, and therefore can never be considered as completely accurate. The impossibility of comparing the estimations obtained with our method to the student’s real state of knowledge makes the evaluation of our method more complicated, since we could never be sure whether they are worse or better than the estimations performed by the human tutor.
On the other hand, the drawbacks of this technique are well known (Van Lehn et al., 1995). At least two issues must be mentioned:

- **Limitations in AI technology.** In fact, we cannot adequately simulate the way real students interact by means of natural language, non-verbal communication, etc.

- **Limitations of the model.** Many features of real students are not represented in the model (for example, motivation, self-confidence, etc.) Nevertheless, simulated students are instantiations of the model, so these features are not considered in the experiment. Therefore, an empirical validation of the proposed model should be carried out in order to assert the applicability of the experimental results to real students.

Next, we describe how a simulated student is generated. Let \( \{C_1, \ldots, C_n\} \) be the concepts in the diagnostic network. Given a value \( k \in [0, 1] \), the simulated student of type \( k \) is defined as a student that knows \( 100k\% \) of the concepts \( \{C_1, \ldots, C_n\} \). The set of known concepts is generated randomly, so that we can generate simulated students with the same level of knowledge but with different sets of known concepts\(^{13}\). Once a simulated student has been generated, the network is used in order to calculate the probabilities of correctly answering each question. Such probabilities are used to simulate the behavior of the student as follows: let us suppose that the probability of correctly answering question \( P \) is \( p \). If the test poses question \( P \), then a random number \( n \) in the interval \([0, 1]\) is generated. If \( p \geq n \) then it is considered that the student has correctly answered the question, and if \( p < n \) that he/she has answered incorrectly. After obtaining the answer, the diagnostic algorithm uses it to update the probabilities of the concepts and chooses the next question to ask the student, using any of the criteria defined. It is easy to see that this simple mechanism allows us to compare the results obtained with the real state of knowledge of the simulated student.

In the simulations, we have used a trial network consisting of a subject \( A \), four topics \( T_1, T_2, T_3, \) and \( T_4 \), fourteen concepts \( C_1, \ldots, C_{14} \), and one hundred questions \( P_1, \ldots, P_{100} \). The prior probability of knowing each concept is 0.5. Each question is related to one, two or three concepts. Note that, in order to be able to answer a question, we consider it necessary to make use of all the concepts related to it. Note also that each concept in the network is related to several questions. For illustration purposes, in Figure 17 we show the relationships between the fourteen concepts and the first twenty questions.

Each question has six possible answers, and therefore a common guessing factor of \( 1/6 \). There are ten difficulty levels, ranging from 1 to 10, and ten questions in each difficulty level. There are four different groups of 25 questions each. The slip factor \( s \) and the discrimination index \( a \) of each group are shown in Table 5.

---

\(^{13}\)The idea of randomly generating the set of known concepts is motivated by the need to ensure that the performance of the algorithm was good in any kind of situation. This hypothesis will be relaxed in Section 5.1.2.4 to introduce student stereotypes.
As we can see, groups are numbered according to their psychometric quality (the smaller the number the higher the psychometric quality). For example, items in group 1 have the highest psychometric quality (smaller slip factor and higher discrimination index).

In the simulation, 30 students of each of the following six different types were generated: 0.0 students (do not know any concept), 0.2 students (know 20% of concepts), 0.4 students (know 40% of concepts), 0.6 students (know 60% of concepts), 0.8 students (know 80% of concepts), and 1.0 students (know all concepts). Therefore, we have used a total of 180 simulated students.

5.1. RESULTS

We begin by analyzing the results obtained at the end of the test for each of the criteria, and then evaluate in more detail the results for those criteria that have shown a better performance.

5.1.1. Final Results

In order to evaluate the performance of the criteria presented in Section 4.2.5, we proceed by calculating the number of concepts that have been correctly diagnosed, incorrectly diagnosed, and non-diagnosed. A concept has been correctly diagnosed if the simulated student knew the concept and it has been diagnosed as known, or if the simulated student did not know the concept and it has been diagnosed as unknown. A concept has not been diagnosed if its probability is between the

---

Table 5. Slip factor and discrimination index

<table>
<thead>
<tr>
<th>Group</th>
<th>Slip factor $s$</th>
<th>Discrimination index $a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0.001</td>
<td>2</td>
</tr>
<tr>
<td>Group 2</td>
<td>0.01</td>
<td>1.2</td>
</tr>
<tr>
<td>Group 3</td>
<td>0.01</td>
<td>0.3</td>
</tr>
<tr>
<td>Group 4</td>
<td>0.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

---

As we can see, groups are numbered according to their psychometric quality (the smaller the number the higher the psychometric quality). For example, items in group 1 have the highest psychometric quality (smaller slip factor and higher discrimination index).

In the simulation, 30 students of each of the following six different types were generated: 0.0 students (do not know any concept), 0.2 students (know 20% of concepts), 0.4 students (know 40% of concepts), 0.6 students (know 60% of concepts), 0.8 students (know 80% of concepts), and 1.0 students (know all concepts). Therefore, we have used a total of 180 simulated students.

---

When determining the number of concepts known we considered the nearest smaller integer. For example, a 0.6 student must know 8.4 concepts out of the 14 concepts in the trial network, so it is considered that the number of concepts known is 8.
minimum and maximum levels previously fixed by the teacher (in this simulation 0.3 and 0.7). The results are given in Table 6.

We give in Table 7 the same results of Table 6 expressed as the percentages of non-diagnosed concepts, correctly evaluated concepts, and incorrectly evaluated concepts.

The first thing that attracts our attention in this table is the good behavior shown by the random criterion, which correctly diagnoses 90.27% of concepts, 3.06% incorrectly, and has just 6.67% non-diagnosed concepts. Taking into account that the test consists of sixty questions, and that there are fourteen concepts to evaluate, the results obtained are very good. Without any doubt it is due to the theoretical consistency of the model used, since, as we have pointed out in previous sections, BNs constitute a sound theoretical model that shows excellent performance in classification and diagnosis problems.

It was quite surprising to see that only one of the proposed adaptive criteria shows clearly better behavior than the random criterion. We believe that this is due to the fact that the model allows for anomalous situations, that is, lucky guesses and unintentional slips. Let us analyze the performance of each criterion:

- If we look at the criteria based on the utility defined as the gain of information, when an anomalous situation (lucky guess or unintentional slip) occurs the gain

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Based on information gain</th>
<th>Conditioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>C_R</td>
<td>C_SG</td>
</tr>
<tr>
<td>Correct</td>
<td>2275</td>
<td>2304</td>
</tr>
<tr>
<td>Incorrect</td>
<td>77</td>
<td>209</td>
</tr>
<tr>
<td>Non-diagnosed</td>
<td>168</td>
<td>7</td>
</tr>
<tr>
<td>Average number of questions</td>
<td>60</td>
<td>16.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Based on information gain</th>
<th>Conditioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>90.27%</td>
<td>91.4%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>3.06%</td>
<td>8.29%</td>
</tr>
<tr>
<td>Non-diagnosed</td>
<td>6.67%</td>
<td>0.28%</td>
</tr>
<tr>
<td>Average number of questions</td>
<td>60</td>
<td>16.88</td>
</tr>
</tbody>
</table>

15Although we use the term anomalous to refer to the situations in which students guess the right answer or fail a question whose answer they know, in practice these situations are very frequent, especially in test exams, the former being more probable than the latter.
of information is in the direction opposite to that desired. In this way, since we are selecting those questions that produce a maximum gain, this non-desired gain is also maximum and therefore the diagnostic process is being distorted, resulting in a greater number of incorrectly evaluated concepts. However, the average number of questions required is really small (only around 15/16 questions to evaluate 14 concepts).

- Regarding the criteria based on the concept of utility defined in terms of the concepts of sensitivity and specificity, it is worth pointing out that for students whose behavior is more predictable, that is, for 0.0 and 1.0 students, both criteria give better results than the random criterion. However, the results are worse for students whose behavior is less predictable (0.2, 0.4, 0.6, and 0.8 students), which makes the global results worse.

- The criterion conditioned by the probability of the concept is the one that has presented the worst performance. This might be due to the fact that the utility of a question for a concept \( U(P, C) \) can be defined as its sensitivity for those concepts whose probability \( P(C) \) is greater than 0.5, and as its specificity for those whose \( P(C) \) is smaller than 0.5. It does not make much sense then to take the utility of the question \( U(P) \) as the maximum of these utilities \( U \), as it is sometimes given by a sensitivity and sometimes by a specificity.

- Finally, the best behavior has been presented by the criterion in which the definition of the utility is conditioned by the probability of the question, for which we have obtained the most precise diagnosis. The distribution of the number of questions required is shown in the graph in Figure 18, where we represent the number of students (vertical axis) that required each number of questions (horizontal axis). Except for the case of requiring all the sixty questions, the number of questions has been grouped in intervals of five questions.

The average number of questions required for the evaluation of all the concepts with the adaptive test is 51.98, with a standard deviation of 10.53. It is true that the reduction in the number of questions required is not significant, which might be due to the good performance of the Bayesian model as a diagnostic algorithm, but together with the greater precision achieved and with the simplicity of the
criterion proposed, we consider that its application is worthwhile. In the next section, we make an in-depth comparative analysis of the performance of the two criteria that have shown better performance, i.e. the random criterion and the criterion conditioned by the probability of the question (that we will call the adaptive criterion from now on).

The criteria based on the information gain have been discarded because of the high percentage of incorrectly diagnosed concepts (around 10%). However, the great reduction achieved in the test time can make the application of these criteria worthwhile in some cases. As an example, we show in Table 8 the performance (in percentages) of the criteria based on information gain and the random and adaptive ones when the number of questions is fixed at 15.

As is shown in Table 8, after 15 questions the two criteria based on the information gain significantly outperform the other two policies. Therefore, they should be considered if the goal of keeping the number of questions low is preferred to the goal of achieving maximum accuracy.

5.1.2. Comparison Between the Random and Adaptive Criteria
In order to carry out this in-depth analysis, we study the evolution of the test by analyzing the results after 15, 30, 40, and 50 questions, and when the test finishes. We also analyze the results for each student type and the errors in the evaluation part, in which – once the diagnostic process has finished – the knowledge level reached by the student in each topic and in the subject is estimated. Finally, we will consider the case in which the initial level is not set to be uniformly distributed and student stereotypes are used in the simulation. Let us first study the evolution of the test.

5.1.2.1. Test Evolution. In order to compare the evolution of the random and adaptive\textsuperscript{16} tests, in Table 9 we display the number of concepts that are not diagnosed, that are correctly diagnosed, and that are incorrectly diagnosed after a fixed number of questions have been presented (15, 30, 40, 50) and also when the test finishes. The same data are depicted in Figures 19 to 21.

Note that the scale in these three graphics is different, and, in particular, that the range in Figure 20 is much smaller. The three graphics show that the performance of the test with the adaptive criterion is always better than the performance of

\textbf{Table 8. Results after 15 questions}

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Cr</th>
<th>CSG</th>
<th>CMG</th>
<th>CCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>34.01%</td>
<td>83.06%</td>
<td>83.49%</td>
<td>36.59%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>6.07%</td>
<td>8.33%</td>
<td>8.45%</td>
<td>3.53%</td>
</tr>
<tr>
<td>Non-diagnosed</td>
<td>59.92%</td>
<td>8.61%</td>
<td>8.06%</td>
<td>59.88%</td>
</tr>
</tbody>
</table>

\textsuperscript{16}An adaptive test is a test in which questions are selected according to the adaptive criterion which we have defined to be the criterion conditioned by the probability of the question.
Table 9. Test results evolution

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th></th>
<th>Incorrect</th>
<th></th>
<th>Non-diagnosed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random</td>
<td>Adaptive</td>
<td>Random</td>
<td>Adaptive</td>
<td>Random</td>
<td>Adaptive</td>
</tr>
<tr>
<td>15</td>
<td>857</td>
<td>922</td>
<td>153</td>
<td>89</td>
<td>1510</td>
<td>1509</td>
</tr>
<tr>
<td>30</td>
<td>1514</td>
<td>1648</td>
<td>134</td>
<td>73</td>
<td>872</td>
<td>799</td>
</tr>
<tr>
<td>40</td>
<td>1878</td>
<td>1971</td>
<td>117</td>
<td>69</td>
<td>525</td>
<td>480</td>
</tr>
<tr>
<td>50</td>
<td>2100</td>
<td>2247</td>
<td>97</td>
<td>60</td>
<td>323</td>
<td>213</td>
</tr>
<tr>
<td>End</td>
<td>2275</td>
<td>2382</td>
<td>77</td>
<td>58</td>
<td>168</td>
<td>80</td>
</tr>
</tbody>
</table>

Figure 19. Number of questions/Correctly diagnosed concepts.

Figure 20. Number of questions/Incorrectly diagnosed concepts.

Figure 21. Number of questions/Non-diagnosed concepts.
the test with the random criterion, and therefore it will always generate shorter and more precise tests.

In order to study the statistical significance of these results, we performed a significance test. Let \( N_c \) be the number of concepts correctly diagnosed by the system using criterion \( c \), where \( c = r \) for the random criterion and \( c = a \) for the adaptive one. For the sample of 180 simulated students, the average of \( N_r \) is \( \overline{N_r} = 12.639 \) in the case of the random criterion and \( \overline{N_a} = 13.233 \) in the case of the adaptive one. Performing the Wilcoxon matched-pairs signed-rank test, we find 145 students with a non-zero difference and a smaller sum of ranks \( R = 2776 \). This implies that the distributions of \( N_r \) and \( N_a \) are significantly different (at the confidence level 99.999%).

Next, we analyze the diagnostic algorithm’s tendencies, that is, we analyze whether it tends to overestimate or underestimate the student’s knowledge. To this end, we again show the final results obtained with both criteria, that are presented in Table 6 and represented (in percentages) in Figure 22.

Let us now split the concepts that have been incorrectly evaluated into two categories: concepts that have been overestimated and concepts that have been underestimated. The results are shown in Table 10, and represented in percentages in Figure 23.

Note that both methods tend to overestimate. However, we believe that this is not due to the Bayesian diagnostic method, but to the item pool used. A student that does not know the concepts required for a question has a probability of 0.16667 of guessing it, whilst a student that knows all the concepts required for a question has an average probability of 0.05715 of slipping\(^\text{17}\). Thus, the test’s tendencies are determined by the item pool (in our case, the tendency is to overestimate students, since it is much more likely to guess than to slip).

![Random](image1.png) ![Adaptive](image2.png)

**Figure 22.** Final results.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Random</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overestimated</td>
<td>53</td>
<td>39</td>
</tr>
<tr>
<td>Underestimated</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>58</td>
</tr>
</tbody>
</table>

\(^\text{17}\)As mentioned in footnote 15, this is a common situation in test exams.
It is also interesting to analyze how many times each question has been used. Such data is depicted in Figure 24.

We see that the random criterion tends to uniformly distribute the items chosen, where each item has been used a minimum of 91 times and a maximum of 134 times. By contrast, the adaptive criterion uses items in a different way, since there are items that are hardly used and other items that are used 179 times (such as P₁, P₂, and P₇₀). The average number of times that items in each group were used is shown in Table 11, where we can see that items with higher psychometric quality have been used more frequently.

5.1.2.2. Results by Student Type. Next we analyze the results by student type. As already explained, we considered six different types. First, we show in Table 12 the mean number of questions needed to evaluate each student type in the adaptive test.

The results by student type are shown in Table 13.

We see that, for every student type (except for Student 1.0) the results obtained with the adaptive criterion are significantly better than those obtained with the

<table>
<thead>
<tr>
<th>Table 11. Average usage of items by group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Group 1 (σ = 0.001, α = 2)</td>
</tr>
<tr>
<td>Group 2 (σ = 0.01, α = 2)</td>
</tr>
<tr>
<td>Group 3 (σ = 0.01, α = 0.3)</td>
</tr>
<tr>
<td>Group 4 (σ = 0.02, α = 1.2)</td>
</tr>
</tbody>
</table>
random criterion, since more concepts are correctly diagnosed and fewer concepts are incorrectly diagnosed/non-diagnosed. The most significant improvement is achieved for type 0.4, with the shortest tests (an average of 41.73 questions) and 11% more correctly diagnosed concepts. The only case in which the random criterion seems to have a better performance is in the case of type 1.0, but the improvement is not significant, given that almost every concept is diagnosed correctly in both cases.

Next, we analyze the results of the evaluation process, in which the knowledge level reached by the student in each topic and in the subject is determined.

### 5.1.2.3. Results of the Evaluation Process

The procedure to analyze the evaluation process is as follows: at the end of the test, those concepts whose probability belongs to the interval [0, 0.5) are considered as not known and therefore instantiated to 0, and those concepts whose probability belongs to the interval [0.5, 1)

| Table 12. Mean number of questions by student type |
|-----------------------------------|--------------------------|
| Student 0.0                       | 54.23                    |
| Student 0.2                       | 52.67                    |
| Student 0.4                       | 41.73                    |
| Student 0.6                       | 51.8                     |
| Student 0.8                       | 54.76                    |
| Student 1.0                       | 56.73                    |

| Table 13. Results by student type |
|-----------------------------------|--------------------------|
| Diagnosis                        | Random                   | Adaptive                |
| Student 0.0                       | Correct                  | 371                     | 395                     |
|                                   | Incorrect                | 17                      | 4                       |
|                                   | Not diagnosed            | 32                      | 21                      |
| Student 0.2                       | Correct                  | 366                     | 385                     |
|                                   | Incorrect                | 17                      | 14                      |
|                                   | Not diagnosed            | 37                      | 21                      |
| Student 0.4                       | Correct                  | 357                     | 387                     |
|                                   | Incorrect                | 24                      | 20                      |
|                                   | Not diagnosed            | 59                      | 13                      |
| Student 0.6                       | Correct                  | 376                     | 400                     |
|                                   | Incorrect                | 9                       | 10                      |
|                                   | Not diagnosed            | 35                      | 10                      |
| Student 0.8                       | Correct                  | 390                     | 402                     |
|                                   | Incorrect                | 10                      | 8                       |
|                                   | Not diagnosed            | 20                      | 10                      |
| Student 1.0                       | Correct                  | 415                     | 413                     |
|                                   | Incorrect                | 0                       | 2                       |
|                                   | Not diagnosed            | 5                       | 5                       |
are considered as known and therefore instantiated to $1^{18}$. This evidence is propagated through the BN and the probabilities that the subject and each of the related topics take the value $1$ are obtained. As already shown in Section 3, these probabilities can be interpreted as the knowledge level reached by the student, and can be compared with the real knowledge level obtained from real data in an analogous way.

Next, we analyze the distribution of the errors in the evaluation for each type of test (random and adaptive). The error is defined as the difference of the real evaluation and the evaluations obtained with the adaptive and random criterion, respectively. The distribution of the errors in the evaluation of each topic and of the subject (number of students for which the error in the evaluation of the topic belongs to the interval shown in the abscise) is represented in Figures 25 to 29.

We see that the estimations of the knowledge level obtained with the adaptive criterion are closer to the real values than those obtained with the random method, due to the greater precision of the diagnostic process. The number of students whose estimation of the knowledge level obtained with the adaptive criterion is coincident with the real knowledge level is 166 for Topic 1, 164 for Topic 2, 134 for Topic 3, and 119 for Topic 4 (out of a total of 180 students). The different results for the four topics are explained by the different numbers of concepts involved. In Figure 29 we can see that at a higher level of granularity (subject) the errors in the lower level (over the four topics) are accumulated, so only 80 of 180 students have obtained their real grade exactly.

Next, we analyze the errors. To this end, Table 14 shows the average and standard deviation of the absolute values of the errors.

In Table 14 we can see that the average absolute error with the adaptive criterion is between a minimum of 0.0142 and a maximum of 0.0399 (with very small standard deviations), which seems an acceptable error given that the model allows students

![Figure 25](image)

*Figure 25* Distribution of errors in the evaluation of Topic 1.

$^{18}$Proposition 2 needs all concepts to be instantiated, therefore at this level we do not consider any concept as *non-diagnosed*. 
Figure 26. Distribution of errors in the evaluation of Topic 2.

Figure 27. Distribution of errors in the evaluation of Topic 3.

Figure 28. Distribution of errors in the evaluation of Topic 4.
that do not master any of the related concepts to guess and students that master all the related concepts to fail.

5.1.2.4. Using Concept Categories and Student Stereotypes. Two of the assumptions on which the previous study relies are as follows:

- The prior probability of knowing each concept is 0.5.
- The set of known concepts for each simulated student is randomly generated.

Under these two assumptions the testing algorithm considers that concepts are indistinguishable. However, these hypotheses can be considered unrealistic since concepts are usually ordered according to some criteria, such as their difficulty level, the teacher’s preferences, the student’s interests, etc. In order to take into account these kinds of assumptions, we have divided concepts into different groups or categories, \( G_1, \ldots, G_n \) and both prior probabilities and student stereotypes are defined in terms of these groups. The criterion used in our example is the difficulty level, but notice again that other criteria like the ones mentioned above can also be considered under the same schema.

In this new experiment, the trial network presented in Section 5 was modified in the following way: concepts \( G_1, \ldots, C_{14} \) (which are assumed to be ordered according to their difficulty level, i.e., from easier to more difficult) were divided into three groups:

- **Group 1** (easy concepts). Concepts \( G_1, \ldots, C_5 \). Their prior probability is 0.75.
- **Group 2** (medium difficulty concepts). Concepts \( G_6, \ldots, C_{11} \). Their prior probability is 0.5.
- **Group 3** (advanced concepts). Concepts \( G_{11}, \ldots, C_{14} \). Their prior probability is 0.25.
### Table 14: Mean and standard deviation of absolute error

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Adaptive</td>
<td>Random</td>
<td>Adaptive</td>
<td>Random</td>
</tr>
<tr>
<td>Average</td>
<td>0.0409</td>
<td>0.0270</td>
<td>0.0264</td>
<td>0.0142</td>
</tr>
<tr>
<td>Deviation</td>
<td>0.0857</td>
<td>0.0757</td>
<td>0.0635</td>
<td>0.0395</td>
</tr>
</tbody>
</table>
Four different types of simulated students were generated: *novice, intermediate, good, and expert* students. *Novice students* know some easy concepts. *Intermediate and good students* know all easy concepts and some of medium difficulty (*good students* know more concepts of medium difficulty than *intermediate students*). *Expert students* know all easy and medium difficulty concepts and some advanced ones. The procedure for generating these random students is the following: a random number \( n \) between \( i \) and \( j \) is generated, and then it is assumed that the simulated student knows concepts \( C_1, \ldots, C_n \), where the values of \( i \) and \( j \) for each type of student are given in Table 15.

In total, 180 students were generated (45 of each type). The number of concepts correctly/incorrectly/non-diagnosed for each student type are shown in Table 16, where the last two columns show the results in percentages and the last row shows the global results (independently of the type of student).

As we can see, the introduction of student stereotypes slightly improves the results, both for the random and the adaptive criteria. This improvement in performance might be explained by the testing algorithm having more complete information. The performance of the adaptive criterion is still better than the random one, attaining more accurate results with fewer questions.

### Table 15. Values of \( i \) and \( j \) for the different student types

<table>
<thead>
<tr>
<th>Type</th>
<th>( i )</th>
<th>( j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Intermediate</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Good</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Expert</td>
<td>12</td>
<td>14</td>
</tr>
</tbody>
</table>

### Table 16. Results using student stereotypes

<table>
<thead>
<tr>
<th>Type</th>
<th>Mean number of questions</th>
<th>Diagnosis</th>
<th>Random</th>
<th>Adaptive</th>
<th>Random</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>55.78</td>
<td>Correct</td>
<td>508</td>
<td>589</td>
<td>80.63%</td>
<td>93.49%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect</td>
<td>38</td>
<td>15</td>
<td>6.03%</td>
<td>2.38%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not diagnosed</td>
<td>84</td>
<td>26</td>
<td>13.33%</td>
<td>4.13%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>589</td>
<td>606</td>
<td>93.49%</td>
<td>96.19%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect</td>
<td>19</td>
<td>7</td>
<td>3.02%</td>
<td>1.11%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not diagnosed</td>
<td>22</td>
<td>17</td>
<td>3.49%</td>
<td>2.70%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>589</td>
<td>610</td>
<td>93.49%</td>
<td>96.83%</td>
</tr>
<tr>
<td>Intermediate</td>
<td>55.29</td>
<td>Incorrect</td>
<td>13</td>
<td>4</td>
<td>2.06%</td>
<td>0.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not diagnosed</td>
<td>28</td>
<td>16</td>
<td>4.44%</td>
<td>2.54%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>610</td>
<td>605</td>
<td>96.83%</td>
<td>96.03%</td>
</tr>
<tr>
<td>Good</td>
<td>52.11</td>
<td>Incorrect</td>
<td>10</td>
<td>8</td>
<td>1.59%</td>
<td>1.27%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not diagnosed</td>
<td>10</td>
<td>17</td>
<td>1.59%</td>
<td>2.70%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>2296</td>
<td>2410</td>
<td>91.11%</td>
<td>95.63%</td>
</tr>
<tr>
<td>Expert</td>
<td>52.04</td>
<td>Incorrect</td>
<td>80</td>
<td>34</td>
<td>3.17%</td>
<td>1.35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not diagnosed</td>
<td>144</td>
<td>76</td>
<td>5.71%</td>
<td>3.02%</td>
</tr>
<tr>
<td>Global results</td>
<td>53.81</td>
<td>Incorrect</td>
<td>80</td>
<td>34</td>
<td>3.17%</td>
<td>1.35%</td>
</tr>
</tbody>
</table>
6. Related Work

BNs have been successfully applied to build student models in several systems. In contrast to BNs, CATs have not often been used in student modeling, despite the great improvement in accuracy and efficiency that can be achieved by using adaptive question selection algorithms. In this section, we briefly review those works more directly related to our research (many of them have already been discussed). An excellent discussion about the use of approximate reasoning techniques in user and student modeling can be found in (Jameson, 1996).

- HYDRIVE (Mislevy and Gitomer, 1996) models a student’s competence at troubleshooting an aircraft hydraulics system. The student’s knowledge is characterized in terms of general constructs (dimensional variables), and a BN is used to update these student model dimensional variables, using the student’s actions as evidence. As already pointed out in Section 3.2.1, our work differs from Mislevy and Gitomer’s in the definition of the aggregation relationships.

- ANDES (Conati et al., 1997) is an ITS that teaches Newtonian Physics via coached problem solving. This system evolved from OLAE (Martin and Van Lehn, 1995b) and POLA (Conati and Van Lehn, 1996a), and uses BNs to carry out long-term knowledge assessment, plan recognition, and prediction of the student’s action during problem solving. In (Van Lehn et al., 1998), diagnostic testing was used to find the prior probabilities needed for the ANDES system. This work has already been compared to our approach in Section 3.2.4.3.

- Our use of a dynamic BN was inspired by Reye’s work (Reye, 1996), described in Section 3.2.4.2.

- In (Collins et al., 1996), BNs are applied together with granularity hierarchies. Test items (questions) are used as evidence to determine if the student masters the learning objectives defined. Three different structures for the BN are compared in terms of the knowledge engineering effort required, test length, and test coverage. However, we have already shown the inadequacy of the adaptive question selection method presented. It is also interesting to note that the performance of the diagnostic algorithm was only evaluated in terms of test length and coverage, but not in terms of the accuracy of the student model obtained. Moreover, the evaluation was performed only with good (0.8 and 1) or bad (0 and 0.2) artificial students, and not with intermediate students (in our studies 0.4 and 0.6 simulated students have also been used) that are obviously the most difficult to evaluate due to their unpredictable behavior.

- SQL-Tutor (Mitrovic, 1998) is an ITS for the SQL database language. SQL-Tutor is based on Constraint-Based Modeling, a student modeling approach proposed in (Ohlsson, 1994). A probabilistic student model is used
to select problems of appropriate difficulty (Mayo and Mitrovic, 2000). The
student model consists of a set of binary random variables representing the
constraints. When the student solves a problem, the probabilities are updated
using heuristics. The reasons that the authors give for using such heuristics
are: (a) the size of the network (more than 500 constraints) and the com-
putational complexity of Bayesian propagation algorithms make the online
selection of problems impracticable; and (b) the nature of the domain
(non-independent variables, the difficulty of defining a granularity hierarchy)
makes the use of other approaches like those proposed in (Reye, 1998) and
(Collins et al., 1996) infeasible. In (Mitrovic et al., 2002), the performance
of this probabilistic student model is evaluated, showing promising results.
However, we think that this performance could be improved by avoiding as
much as possible the use of such ad hoc heuristics, which do not have the firm
theoretical foundation of BNs. Instead, other approaches to reduce com-
putational complexity, such as the ones proposed in (Jameson, 1996) or the
use of goal-oriented algorithms (Castillo et al., 1997) should be considered.

7. Conclusions and Future Work

In this work we have presented a new integrated approach to Bayesian student
modeling. In our new integrated student model, nodes have a well-defined semantics
and links accurately describe the relationships between them. The students’ state of
knowledge is represented in terms of more than one variable and is described at
the level of granularity required. Moreover, the student model allows substantial
simplifications when defining the conditional probabilities needed for the BN, that
can be automatically computed from a set of weights (that measure the relative
importance of each subitem in the aggregated item) or from certain data associated
with each question (concepts which are necessary to know along with their
importance, and parameters such as slip and guessing factors, difficulty level,
and discrimination index).

The validity of the structural model proposed has been tested by using simulated
students. The results obtained are very promising, as they show that the Bayesian
integrated model so defined produces highly accurate estimations of the student’s
cognitive state at all levels of granularity. However, the simulated students are just
instantiations of the model presented here, whereas real student behavior is influ-
enced by many other factors that are not explicitly represented. Therefore, in order
to assert the validity of the proposed model in the real world, a formal evaluation
with real students should be performed. In particular, one of the greatest limitations
of the model is that it assumes that the student’s knowledge does not change, which is
a valid assumption for this experiment but might be considered unrealistic in real
settings.
Even when the results obtained are very satisfactory (90.27% correctly diagnosed concepts), it has been possible to improve them by combining the structural model with adaptive testing technologies, that is, by applying adaptive question selection methods (going up to 94.53% correctly diagnosed concepts with a model that allows lucky guesses and random slips). To this end, several adaptive criteria have been defined, and their performance tested using *simulated students*. Once the best adaptive criterion has been chosen (the criterion conditioned by the probability of the question), we have shown that its behavior is better at all possible levels: the adaptive criterion requires a smaller number of questions and yields more accurate results independently of the number of questions that have been asked so far and independently of the student type. However, we must insist that, in spite of the excellent results, this empirical evaluation should be only considered as a first step towards a formal evaluation with real students.

Regarding future work, there are several directions to be explored, which we group into two categories: improvements in the integrated structural model, and applications of the model developed. Regarding improvements in the integrated student model, we plan to investigate: (a) the introduction of *prerequisite relationships* in the model, as this could contribute to improving the precision and efficiency of the diagnosis process. However, the way of introducing such relations in the model has to be studied carefully, because these would change the independence relationships implicit in the model; and (b) the use of new sources of evidence about the student’s cognitive state, such as the instruction sessions he/she has gone through, teachers’ opinions, etc. Again, a detailed analysis of the exact meaning of such nodes and of the relationships with existing nodes needs to be performed. Once the whole model has been defined, evaluations with simulated students and then with real students should be carried out to test its performance. Regarding applications of such a model, our final goal is the development and implementation of a Bayesian evaluation system (*SIBET, Sistema Inteligente Bayesiano de Evaluación mediante Tests*). *SIBET* will be accessible through the Web, and will allow people without knowledge of programming and BNs to implement their own adaptive tests based on BNs. To this end, *SIBET* will have two different modules: (a) a *test editor*, that is, a module to define the curriculum structure and to edit the tests, that will be used by the designer; and (b) a *virtual classroom* for the evaluation process where the students will be able to take the tests previously defined online, and their answers will be used to diagnose the set of concepts that the student masters and to compute measures of the knowledge achieved at the different levels of granularity defined. In this way, *SIBET* could be used as a stand-alone system for student assessment or as a diagnostic module in a more complex architecture (ITS) that would enable curriculum adjustment and remediation. This system is inspired by the SIETTE system (Rios et al., 1999), (http://alcor.lcc.uma.es/siette), which basically has the same characteristics but only enables diagnosing one ability at a time, since it is based on the unidimensional IRT.
Acknowledgements

The authors would like to express their deepest gratitude to Francisco Durán for his invaluable help in this work. They would also like to thank the anonymous reviewers and the editors for their professional and careful reading of the paper and for their many detailed and valuable comments.

References


**Author’s vitae**

**Dr. Eva Millán** is Associate Professor in the Department of Computer Sciences at Málaga University (Spain), where she lectures on Operations Research and Expert Systems. From April to September 1998 and 1999 she worked as International Fellow at SRI International, California (formerly Stanford Research Institute). She received her master degree in Mathematics from the University of Málaga in 1991, and her Ph.D. degree in Computer Science from the same University in 2000. Her research interests lie in the areas of Intelligent Tutoring Systems, Computerized Adaptive Testing and Bayesian Networks.

**Dr. José Luis Pérez-de-la-Cruz** is Associate Professor in the Department of Computer Sciences at Málaga University (Spain), where he lectures on Artificial Intelligence. He received his Ph.D. degree in Civil Engineering from the Technical University of Madrid in 1990. His research is aimed to the application of Artificial Intelligence techniques to Intelligent Tutoring Systems and to Engineering problems.