Graduate Management Admission Council®

Partial Order Knowledge Structures for CAT Applications

Michel C. Desmarais and Xiaoming Pu,

École Polytechnique de Montréal

and

Jean-Guy Blais

Université de Montreal

Presented at the CAT and Cognitive Structure Paper Session, June 7, 2007

Abstract

Bayesian and graph models of student knowledge assessment have made significant progress in the last decadeand are challenging the more traditional IRT approach for CAT applications. We review some of the most prominent frameworks in Bayesian knowledge assessment and how they compare to IRT and introduce one such framework in the family of Bayesian models, the POKS (Partial Order Knowlege Structure). A comparison of the POKS approach to IRT and a Bayesian Network approach showed that it can perform detailed knowledge assessment at a computational cost of orders of magnitude less than a Bayesian Network and IRT. The assessment accuracy results of experiments show that it is at least as good as a one-dimensional IRT model and generally outperforms a Bayesian Network with small data sets. However, a number of challenges remain for the POKS approach as well as for other Bayesian frameworks in CAT applications. One of the most important issue is how scalable the approaches are over a large number of items. Another issue is the estimation of reliability and error margins, which are currently almost ignored by these approaches. We review these challenges and the work ahead.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Desmarais, M. C., Pu, X, & Blais, J.-G. (2007). Partial Order Knowledge Structures for CAT Applications. In D. J. Weiss (Ed.), ***Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*** **Retrieved [date] from** www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Michel C. Desmarais or Xiaoming Pu, Polytechnique Montreal, C.P. 6079, succ. Centre-Ville, Montréal, Québec, Canada, H3C 3A7.
Email : michel.desmarais@polymtl.ca, xiaoming.pu@polymtl.ca, jean-guy.blais@umontreal.ca

Partial Order Knowledge Structures for CAT Applications

Adaptive testing can be considered one of the first applications of what is currently a very active research topic, that of adaptive and personalized interfaces. The principle of adaptive interfaces is based on constructing a user model of an application and then adapting the performance of the application to the model. This is exactly what adaptive testing does. It constructs a model of the respondent's knowledge and adapts the administered items as a function of the model, generally with the specific goal of making a knowledge assessment with a minimum number of items. Although the area of adaptive interfaces includes a wide range of adaptation, from user preferences to user intentions (McTear 1993), it remains that computerized adaptive testing (CAT) might be the very first large-scale application of the basic principle of these applications: to construct a user model and adapt the performance of the application as a function of the model.

The domains of CAT and adaptive interfaces evolved separately from each other, largely unaware of each other's developments. The area of adaptive interfaces and, in particular, that of adaptive learning environments gave rise to several models of learner knowledge and several techniques for its evaluation (Self, 1988). While item response theory (IRT) was rapidly developing in the area of adaptive testing, the area of intelligent tutorials was developing its own approaches for representing and assessing competences for adaptive learning environments (Carr & Goldstein, 1977). Most of these efforts were based on rule-based systems and aimed at providing a very detailed assessment of learner knowledge. The main feature of these models was to arrive at an accurate assessment that referred, not only to the precise concepts mastered, but also to incorrect concepts, or *mal-rules* (Payne & Squibb, 1990). These models had the advantage of providing a high level of *granularity* in that they could provide a very precise assessment of acquired or missing knowledge/competences; however, they did not integrate any notion of uncertainty, which is inherent to the modeling of knowledge.

Inversely, work in the area of psychometrics and IRT models incorporated, from the outset, the notion of uncertainty and were devoted mainly to estimating the reliability of the models and the confidence intervals used to make an assessment with a known degree of certainty. However, the granularity of IRT-based models still remains low and generally limited to one dimension, or, in the case of more recent work on multidimensional IRT, to a few dimensions simultaneously, well below the level of granularity that can be attained with the rule-based models of intelligent tutorial environments.

These divergences are easily explained considering that, in the case of psychometrics, the most frequent requirements originate from the context of summative evaluation and consist in determining whether the respondent will pass or fail a test. The requirements of intelligent tutorial environments are aimed, instead, at determining the learning problems of the learner in order to select very specific capsules of pedagogical content aimed at remedying incorrect concepts or guiding the learner toward a more advanced content. The respective requirements of the two domains are, therefore, very different, which explains in large part the little influence they have had on each other.

The link between the two domains emerged from work on graphical models of knowledge and we mention, among others, that of Almond and Mislevy (1999), Mislevy and Gitomer

(1995), and the overview of Jameson (1995). In addition to demonstrating that the IRT model could be considered as a graphical model, this work emphasized the importance, for any learner assessment, to incorporate measures of uncertainty on one hand, and, on the other, to arrive at an accurate assessment as required by advanced computerized learning environments.

The present paper provides a comparison of various approaches specific to each domain. The approaches are described in the first sections; we then report on a series of experiments used for a quantitative comparison. Finally, a discussion about the observed results and the qualitative criteria that will clarify the advantages and inconveniences of each approach is presented.

Graphical Models and Knowledge Assessment

Graphical models have long been used for representing a domain of knowledge (see, for example, Findler, 1979). They offer a visualization that is intuitive and, also, a formalism as a starting point for constructing rigorous and even mathematical models. They are now commonly used in the domain of representation and assessment of acquired knowledge and competences.

Concepts, Items, and Hidden Nodes

The nodes of a graphical model typically belong to two categories: concepts and items. The item nodes represent questions, exercises, or any other observable manifestation of the individual's competence. The concept nodes represent latent competences. These are not directly observable. Thus, the nodes of a graphical structure are classified according to whether they are observable or unobservable (or "hidden") nodes.

The distinction between an observable node and a hidden node is fundamental, not only from an epistemological point of view, but also with regard to the techniques used to automatically construct these structures from data and make the knowledge assessment.

Figure 1 illustrates a graphical model containing hidden nodes, concepts, and observable nodes, items. The concepts are in fact divided into two groups:

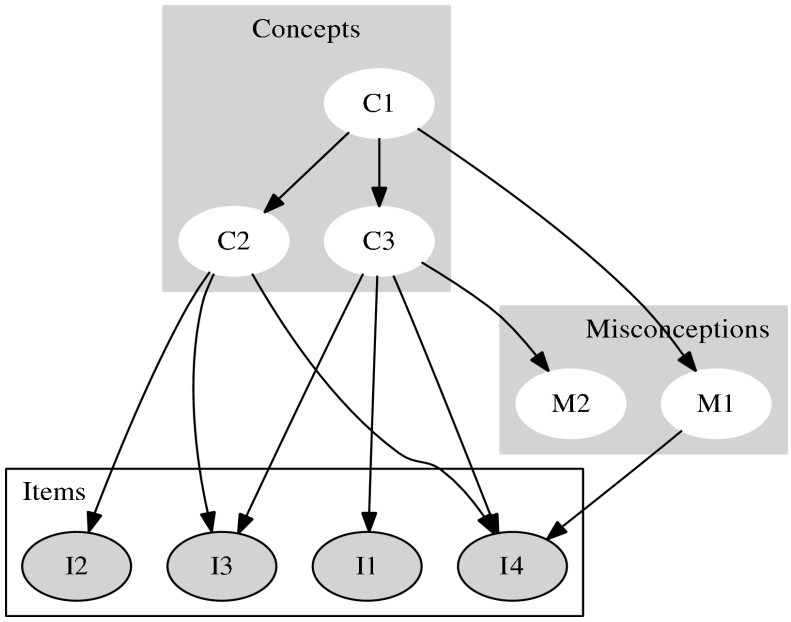
- Competences, called *concepts* in the figure;
- False competences, called *incorrect concepts*.

While the competences determine success of the items, the false competences, or incorrect concepts, are rather the source of failures. These are typically errors frequently found in a domain of knowledge.

For example, in the learning of operations on fractions, the addition of two fractions will often contain the error which consists in adding the numerators and the denominators, and shown here by rule :

$$\frac{a}{c}+\frac{c}{d}\overset{e_1}{\rightarrow}\frac{a+c}{b+d} \quad . \tag{1}$$

Figure 1. Example of a Graphical Model



Obviously, the correct way to proceed consists of transforming the fractions in order to obtain a common denominator. This competence could be broken down, corresponding to rule :

$$\frac{a}{c}+\frac{c}{d}\xrightarrow{c_1}\frac{ad+cb}{bd} \quad , \tag{2}$$

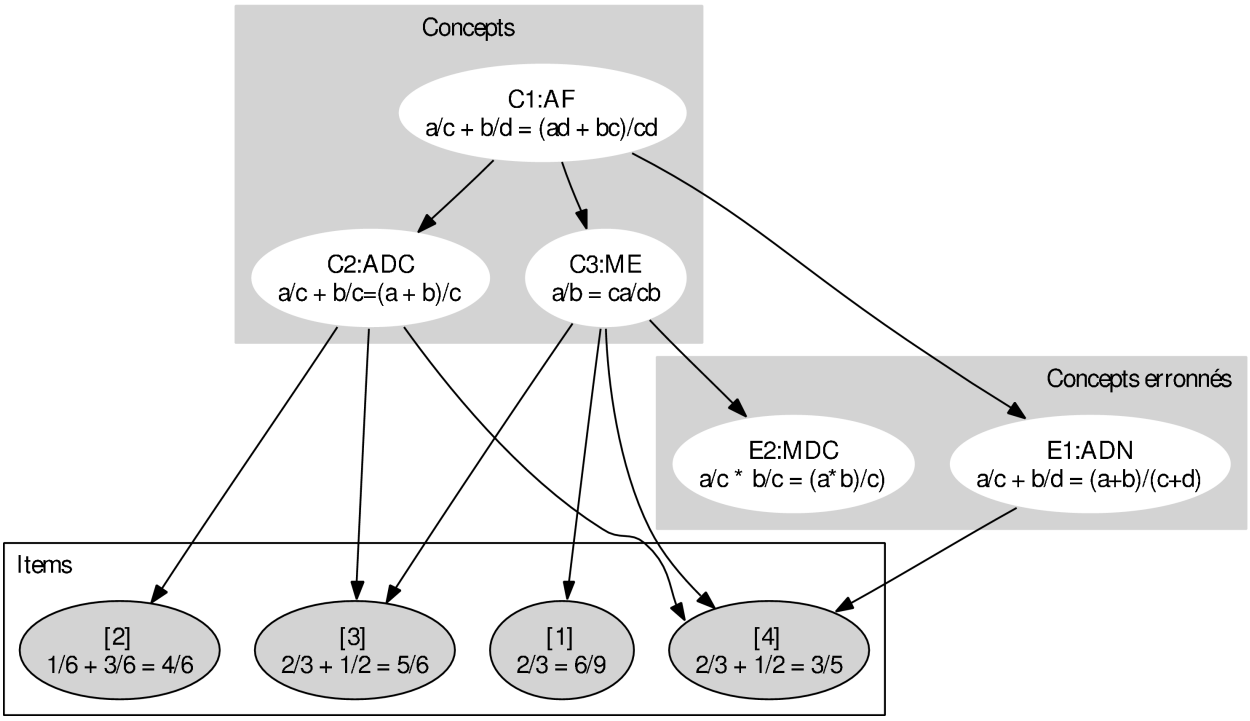
into two sub-competences, corresponding to rules and :

$$\frac{a}{c}+\frac{c}{d}\xrightarrow{c_2}\frac{ad}{cd}+\frac{cb}{cd}\xrightarrow{c_3}\frac{ad+cb}{bd} \quad . \tag{3}$$

The notion “to add with a common denominator” will thus be considered a concept, or a precise competence that the student will have to master for the addition of fractions, and it will be linked to two other more precise concepts.

Figure 2 shows a graphical model that illustrates these interactions among the items, the concepts and the incorrect concepts in the domain of the arithmetic of fractions. It contains four items with number [4] being a wrong answer. It is, therefore, linked to the incorrect concept E1:ADN, that represents the error of rule E1. The concepts, or competences C2:ADC (addition with common denominator) and C2:ME (multiplication by an integer) respectively represent notions of algebraic transformations that are involved in the mastery of the items to which they are linked in the figure. The highest-level concept, C1:AF (addition of fractions) is, for its part, linked to the two lower-level concepts C2:ADC and C3:ME

Figure 2. Example of a Graphical Model With Concepts and Items From the Domain of the Arithmetic of Fractions



As can be seen, graphical models have a high level of expressiveness due to their representation of complex links, precise competences and higher-level competences, and incorrect concepts that explain particular errors that the student can make faced with very precise items.

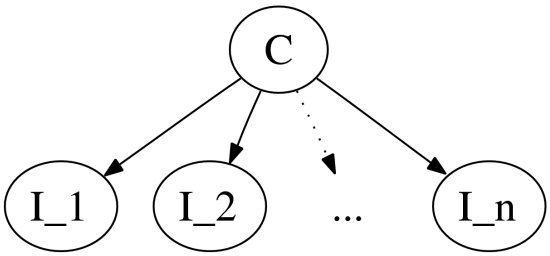
We note, however, that the cognitive value of a graphical model depends essentially on the choice of nodes, relations, and the semantics that are given to a relation among the nodes. Thus, the model in Figure 2 is based only on an intuition of the latent competences involved in the resolution of the four items and not on a demonstrated cognitive model. The meaning of the relations is not clearly defined and it too remains intuitive. It could in fact mean “is related to.”

We will see, however, that the semantics of links are completely different from one graphical model to another, especially when a probabilistic graphical model is used with a view to predict or to assess. This is notably the case with Bayesian networks to be seen later and whose semantics are very specific. It is important to remember that all graphical models are not necessarily founded on a solid cognitive base and that the semantics of their links vary from one model to another.

The IRT Approach in the Graphical Perspective

IRT models can actually be represented by a graphical model whose semantics are those of Bayesian networks. Figure 3 illustrates the principle.

Figure 3. The IRT Model Represented Graphically



In the basic IRT model, a single competence is represented by concept C in the graph and it is referred as θ in the IRT framework. The θ competence essentially represents the fact that a test measures one dimension and that success on all the items depends on this dimension. In addition, by stipulating that Figure 3 conforms to a representation of a Bayesian network, we then conclude that, given a determined level of competence, θ, success on an item is independent of success on another item. In Bayesian terms, we can therefore conclude that

$$P(I_i, I_j|\theta) = P(I_i|\theta)P(I_j|\theta) \quad . \tag{4}$$

This is to say that $P(I_i|\theta)$ and $P(I_j|\theta)$ are independent of each other. This conclusion in fact corresponds to one of the fundamental IRT hypotheses: success on an item given a level of competence, θ, does not influence success on another item.

This definition corresponds to what is called a "naïve" Bayesian network. Given the above-mentioned independence hypothesis, it follows that

$$P(I_1, I_2, \ldots, I_n|\theta) = P(I_1|\theta)P(I_2|\theta)\ldots P(I_n) \quad . \tag{5}$$

This equation corresponds to the simplifying hypothesis of a naive Bayesian network. A direct link between IRT and graphical representation of naive Bayesian networks therefore exists (Almond & Mislevy 1999). The fundamental difference between a naive Bayesian network and the three-parameter logistic IRT model, for example, resides in the fact that the link between the probability of success on an item and a level of competence, θ, is determined by a sigmoid function with three parameters, a_i, b_i, and c_i, that corresponds to

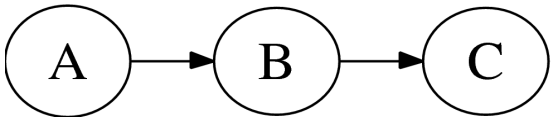
$$P(I_i|\theta) = c_i + \frac{1}{1+e^{-a_i(\theta - b_i)}} \quad , \tag{6}$$

Where a_i is the discrimination capacity of an item i; b_i the level of difficulty of the item i and c_i the chance factor (also called pseudo-chance) that determines a minimum probability of success.

Formalism of Bayesian Networks

Bayesian networks (BN) are accurate representations of dependence and independence relationships among the nodes of the network. Without entering into the details of the semantics of formalism, we will state the fundamental principles and readers can refer to Neapolitan (2004) for a more complete description.

BN formalism is based on the notion of a *Markov blanket* that stipulates independence relationships. In the following figure, for example:



the *B* node represents a Markov blanket, since, if it is observed, then nodes *A* and *C* will be independent of each other, while, if it is not, *A* and *C* are dependent through *B*.

Bayesian networks have the quality of possessing both this formal definition of their semantics and also of being an intuitive representation of relationships among events. For example, if:

(A) John is a smoker;

(B) John has cancer;

(C) John's lung X-ray shows a "spot",

we understand from the outset that the above graph represents relationships of causality. In addition, if we know that *A* is a true statement, we will then tend to believe that events *B* and *C* are more probable. Conversely, if we observe that *C* is true, then we will also tend to believe that *A* is more probable. However, if we know that *B* is true, then the observation of *A* as true does not increase the probability of *C* since we know that it is cancer, *B*, that is the cause of *C*. Finally, it is also the case in the opposite direction of the arrows: knowing that *B* is true, the observation of *C* does not provide additional information concerning *A* and no longer influences its probability. (It must be presumed, however, that cigarettes cannot cause a spot on the lung other than through cancer. Otherwise, there would then be an $A \rightarrow C$ link, indicating that cancer is not the only mechanism through which cigarettes can bring about such an X-ray.) This simple example shows one of the types of independence that is defined with a Markov Blanket and that serves as the basis of BN construction.

The POKS Approach and the Knowledge Spaces Theory

One graphical approach that has recently emerged in the area of adaptive testing is Partial Order Knowledge Structures, or POKS. It is inspired by both the work of Doignon et Falmagne (1999) on knowledge spaces for the representation of knowledge, and on naive Bayesian networks for the inference of knowledge. We first describe the knowledge spaces theory and then the POKS approach.

Knowledge spaces theory

Knowledge spaces theory represents a domain by a set of knowledge items. The items are manifestations of competences such as success on a particular exercise or to a knowledge question. We do not refer to notions of concepts or latent competences in this theory, but exclusively to observable manifestations of these competences or concepts.

The knowledge state of a student is represented by a subset of this domain, the items that a student masters. It is, therefore, a very simple way of representing the state of an individual's knowledge. At first, only those items whose mastery by the respondent is observable are involved in order to model the competences.

Next, the mastery of a latent competence can be indicated by defining which items are manifestations of which competences. We can base ourselves on a summation, potentially weighted, of the expected successes/failures on these items as a measure of the latent competence. Actually, this is what teachers do when they divide an exam into sections corresponding to various topics of the subject matter. This approach contrasts with that of IRT that explicitly represents the respondent's competence in the model by a latent dimension, θ .

The knowledge spaces theory also stipulates that the items of a knowledge domain are mastered in a constrained order. Taking the example in Figure 4 and presuming that the student must find the value of the term on the right side of each equation, we see at the outset that item *a* is more difficult than the other three items. In this sense, success on *a* involves success on the other three items. Inversely, item *d* is, itself, more elementary than the others, and a failure on this item implies a failure on the other items. It is more difficult, however, to draw any conclusions regarding a given order with items *b* and *c*.

Figure 4. Partial Order of Success on Four Knowledge Items {*a,b,c,d*}

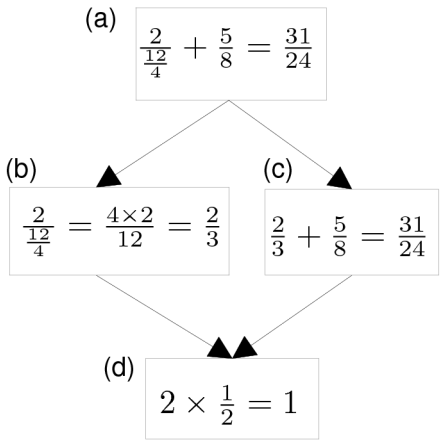


Figure 4 thus shows a partial order that reflects implication relationships among the other items. It constrains the order in which the items can be mastered and, consequently, the various knowledge states to a specific set of subsets:

$$\{\varnothing,\{d\},\{c,d\},\{b,d\},\{b,c,d\},\{a,b,c,d\}\}$$

This representation of ordering constraints by a partial order, in fact corresponds to the representation used in the POKS approach. However, the knowledge spaces theory involves an important nuance that is not integrated in the directed graph of Figure 4 and that concerns alternative orders for acquiring certain competences.

Taking an example of alternative order, we refer to the three concepts of Figure 2, C1:AF, C2:ADC, C3:ME. Even if we presume that transformations C2:ADC and C3:ME are involved in the calculation for C1:AF, as the following equation expresses it, it is just as plausible that a student in fact learn the more specific rule :

$$\frac{a}{c}+\frac{c}{d}\xrightarrow{c_1}\frac{ad+cb}{bd} \qquad (7)$$

and that he actually has never acquired the more elementary notions corresponding to C2:ADC and C3:ME (rules and). This situation corresponds to learning based on memorization of rules to solve always increasingly specific problems, rather than learning based on problem solving based on more general rules that are combined to solve more complex problems.

The knowledge spaces theory is based, therefore, not on partial order, as in Figure 4, but on a graph that represents alternatives. We will not go any further into the details of the knowledge spaces theory, but instead refer to Falmagne et al. (1990) and Doignon and Falmagne (1999) for a complete description of the theory. In addition, it is interesting to note that this theory has given rise to an application of a commercial study guide called ALEKS (www.aleks.com) and which is also described in Falmagne et al. (2006).

Induction of the structure

Graphical structures are used to represent and infer the individual's competences as shown in the example of Figure 4. Such a structure can be constructed by a domain specialist who can make a judgement regarding ordering. This can be done when the number of items remains low. When the number increases, however, the task quickly becomes overwhelming given the number of links to consider. It is, therefore, vital to develop an approach for automatically constructing this structure using data.

Construction of the structure is done using responses from a sample of examinees for a test including knowledge items that represent the nodes of the structure. The process involves comparing each pair of items in order to determine whether an interaction exists between the two.

Let us take, for example, the distribution of respondents to a test on two items, X_a and X_b , as could be deduced from Table 1. If we have a relationship $X_a \rightarrow X_b$, as seen in Figure 4, we can then expect to find respondents in each of the first three conditions $\left(\left\{x_{ab}, x_{\neg a \neg b}, x_{a \neg b}\right\}\right)$ but not the fourth $X_{\neg a} \rightarrow X_b$. Given the relationship $X_a \rightarrow X_b$, we do not expect to find a respondent who succeeds on X_a , but fails on the normally easier item, X_b .

Table 1. Contingency Table of Combination Possibilities of Items X_a and X_b

	Condition	x_a	x_b	Respondents
(1)	x_{ab}	Success	Success	Yes
(2)	$x_{a\neg b}$	Success	Failure	No
(3)	$x_{\neg a b}$	Failure	Success	Yes
(4)	$x_{\neg a \neg b}$	Failure	Failure	Yes

There is, however, a probabilistic element that must be taken into account. The relationship $X_a \rightarrow X_b$ can exist, but it can be a weak link; or again, there might be noise in the data, such as success items through chance or failed items through inattention. Thus, to determine whether a link $X_a \rightarrow X_b$ exists between the two items X_a and X_b, the following three conditions must be true:

$$P\left(\left[P(X_b|X_a) \geq p_c\right]|D\right) > (1-\alpha_c) \quad (8)$$

$$P\left(\left[P(\neg X_a|\neg X_b) \geq p_c\right]\right) > (1-\alpha_c) \quad (9)$$

$$P(X_b|X_a) \neq P(X_b)(p < \alpha_i) \quad (10)$$

where p_c is the minimal conditional probability for $P\left(X_b|X_a\right)$ and $P(\neg X_a \vee \neg X_b)$; only one value is retained for the test of all the relationships of the network (0.5 in general),

α_c is the alpha error of conditional probability tests; this error determines the tolerated proportion of relationships whose conditional probability of the population is below the p_c threshold; the usual values are between 0.2 and 0.5,

$p < a_i$ corresponds to the alpha tolerance error for the interaction test, and

D is the joint frequency distribution of X_a and X_b in a sample of test data. This distribution is a 2×2 contingency table, as can be seen in Table 1.

The first condition (inequality 1) stipulates that the conditional probability of a success on item X_b given a success on X_a must be greater than a threshold p_c, and that we can come to this conclusion using a response sample on items *D*, with a rate of error less than a_c.

The second condition (inequality 2) is analogous to the first and stipulates that the probability of a failure on item X_a given a failure on X_b must be greater than p_c, with a maximum rate of error of a_c given the distribution *D* of responses.

These two conditions are calculated using a cumulative binomial distribution. In inequality 8, the value of $P([P(X_b|X_a)]|D)$ is obtained by the summation of the binomial function for all the distributions in which $x_{a\neg b}$ is less than the frequency observed in *D*, that is:

$$\begin{aligned} P([P(X_b|X_a)]|D) &= P(x \leq x_{a\neg b} \,|\, X_a) \\ &= \sum_{i=0}^{x_{a\neg b}} \mathrm{Bp}(i, x_a, p_c) \\ &= \sum_{i=0}^{x_{a\neg b}} \binom{x_a}{i} p_c^{[x_a - i]}(1 - p_c^i) \end{aligned} \qquad (11)$$

where: $x_a = x_{ab} + x_{a\neg b}$.

The conditional probability of the second condition is based on the same function, but uses $Bp(i, x_{\neg b}, p_c)$ rather than $Bp(i, x_a, p_c)$.

The third condition (inequality 10) represents an independence test and is verified by a χ^2 test with distribution *D*, for the 2×2 contingency table:

$P(\chi^2) < \alpha_c$

For small samples, this test can also be replaced by the Fisher exact test.

The choice of the value for p_c determines the strength of the implication relationship between two items. For example, if we have $X_a \rightarrow X_b$ and the order in which these two items is mastered is highly constrained, then the value of *P*(*B*|*A*), as determined by the distribution of frequencies *D*, will be very close to 1. The value of p_c thus represents the lower limit at which we accept to retain a relationship. The choice is relatively arbitrary, but it must not logically be less than p_c = 0.5.

The two values α_c and α_i represent the alpha errors that we are prepared to tolerate to conclude that the condition is satisfied. For very small samples, these values can be as high as 0.5 in order to keep the greatest possible number of relationships. In the experiments described below, these values varied between 0.2 and 0.1.

The inference of knowledge.

Once the structure is constructed, it is used to choose the items to be administered to the respondent of a test, and then update the probability of success on the other items following success or failure. As was seen in the preceding sections, POKS uses a model of a subset of mastered or non-mastered items to model the status of the respondent's knowledge, following the example of knowledge spaces theory. It is, therefore, on the basis of the subset of items deemed successful, therefore having a probability of 0.5 or more of being successful, that the acquired competences can be evaluated. The calculation of the updated probabilities, following the observation of a success or a failure, is explained in this section.

Updating of probabilities, or knowledge inference, can be done according to two versions. In the first, information about the observation of a success or failure is propagated in a transitive fashion throughout the network. For example, if we have $X_a \rightarrow X_b \rightarrow X_c$ and success is

observed at X_a, then the probabilities of X_b and X_c will be affected. In the second version of this algorithm, we propagate only for X_b and the value of X_c is not affected, unless, at the time of the construction of the network, the transitive link $X_a \rightarrow X_c$ was also derived. We describe, here, only this last version; however, details of the first version are described in Desmarais, Maluf and Lui (1996) and Desmarais and Pu (2005).

Probability updating in the POKS framework is based on the calculation of posterior probabilities with the local independence assumption: i.e. for two items, X_1 and X_2, their joint and conditional probability, given a third, X_3, is independent and, by definition,

$$P(X_1, X_2 | X_3) = P(X_1 | X_3)\, P(X_2 | X_3). \tag{12}$$

This hypothesis is necessary to simplify calculations not only for the construction of the structure and inference, but especially because, otherwise, a much larger quantity of data would be needed. This hypothesis is used to presume that only binary relationships can be evaluated, while in other cases, relationships with three or more variables must be validated. We then have to gather data in a table with 2^n entries, in which *n* corresponds to the number of variables that can be interacting.

According to the local independence hypothesis, the calculation of the probability of an item following observation of a series of items corresponds to a chain of posterior probability calculations as described in the following.

Let us first suppose a relationship $X_a \rightarrow X_b$, the posterior probability of X_b, given observation of a success or failure on X_a, is calculated on the basis of Bayes' theorem in its odds ratio version:

$$O(X_b \mid X_a) = O(X_b) \frac{P(X_a \mid X_b)}{P(X_a \mid \overline{X}_b)} \tag{13}$$

$$O(X_b \mid \overline{X}_a) = O(X_b) \frac{P(\overline{X}_a \mid X_b)}{P(\overline{X}_a \mid \overline{X}_b)} \tag{14}$$

in which $O(X_b)$ is the initial odds and $O(X_b | X_A)$ represents the odds of X_b given the observation of X_a. Conditional odds is here defined in its usual form:

$$O(X_b|X_a) = \frac{P(X_b|X_a)}{1 - P(X_b|X_a)} \ . \tag{15}$$

In order to combine observations from several items, the inference process is based on the local independence hypothesis to simplify the calculation. Thus, the updating following observation of several items X_i, X_j, … , X_n of item X_b, is the simple product of the likelihood ratios. For example, supposing that we have *n* relationships of the form $X_i \rightarrow X_b$, then

$$P(X_1, X_2, \ldots, X_n | X_b) = \prod_i^n P(X_i | X_b) \ . \tag{16}$$

From this equation, the new probability of X_b, given $X_1, \ldots, X_n$, can be rewritten in the odds ratio form as

$$O(X_b|X_1,X_2,\ldots,X_n)=O(X_b)\prod_i^n \frac{P(X_i|X_b)}{P(X_i|\overline{X}_b)} \quad . \tag{17}$$

If an observation corresponds to a fail on , then the ratio

$$\frac{P(\overline{X}_i|X_b)}{P(\overline{X}_i|\overline{X}_b)} \tag{18}$$

is used.

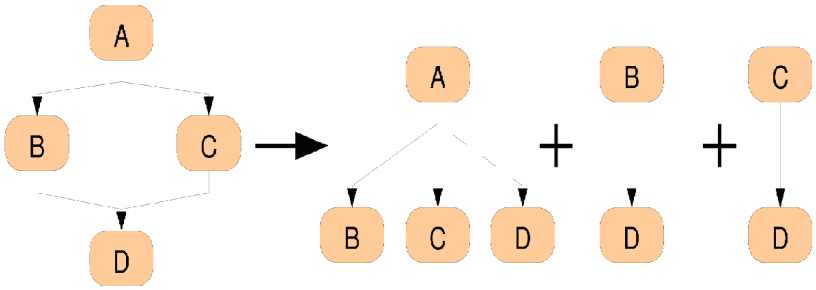
The local independence assumption on which these calculations are based is very strong and characterizes the Bayesian approache called “naive.” It greatly simplifies the calculations and the necessary quantity of calibration data, as mentioned. Although this hypothesis is often incorrect in many situations, it has been shown that, despite this, results obtained are relatively robust (see Domingos & Pazzani 1997; Rish 2001; Friedman, Geiger & Goldszmidt 1997).

From partial order to a set of one-level networks

As mentioned in the preceeding section, the propagation of an observation is not carried over transitive relations, contrary to the studies of Desmarais and Pu (2005), and Desmarais et al. (1996). For example, if we have $A \rightarrow B$ and $B \rightarrow C$, the probability of *C* remains intact following observation of *A*, unless the $A \rightarrow C$ link is explicitely derived from the data. However, if a strong ordering of the type $A \rightarrow B \rightarrow C$ exists, then we expect to also find $A \rightarrow C$.

This principle is shown in Figure 5. A POKS network is here transformed into three simpler networks with only one level. The dotted line of the partial order is normally derived from the data if the network nodes are strongly ordered, i.e. if the three conditions in inequalities 8, 9, and 10 tend toward maximum values.

Figure 5. Correspondence Between a Directed Graph and a Set of One-Level Networks Used for the Inference of a Positive Observation (*P* = 1)



Compared to the algorithm in Desmarais et al. (1996), this version has the advantage of not requiring any framework for the so-called “partial” propagation, throughout transitive relationships, from *A* to *C* in a structure such as $A \rightarrow B \rightarrow C$, for example. Considering that we can expect transitive relationships to be derived by the structure construction algorithm, i.e. that

the algorithm will also infer the $A \rightarrow C$ relationship using the same data, the results should therefore be similar. This result was confirmed in our exploratory experiments for comparing the two versions.

Symmetrical Relationships

If no two items were equivalent in the node structure, then the structure obtained using a data sample would, in fact, be a partial order, or a directed acyclical graph. However, if two items were equivalent in terms of preconditions and difficulty, then these items would be linked to each other by symmetric relationships and the network would contain cycles. We could, therefore, simply group together the symmetric items in a single node to reproduce a partial ordering with no cycle. Success or failure on a single item of the group would then be sufficient to conclude for success or failure on all the other items.

Unfortunately, the reality is not so simple. In practice, several nodes have symmetrical relationships derived from the algorithm above, but they differ in terms of precondition and level of difficulty. The more tolerant the values retained for p_c and a_i are for conditions in inequalities 8, 9, and 10, the greater will be the number of symmetric relationships whose items are not equivalent. This will be reflected by symmetric relationships but with very different respective values of $O(X_a|X_b)$ and $O(X_b|X_a)$. For example, in the symmetric relationship $A \neq B$, if the two items are equivalent then $O(A|B)$ is approximately equal to $O(B|A)$. However, if they are not equivalent, a symmetric relationship could still be derived even though $O(A|B)$ and $O(B|A)$are very different, it all depends on the error tolerance and the minimum strength of the links, both determined respectively by parameters p_c and a_i .

The consequence of having symmetric relationships of this kind is that the structure derived by the induction algorithm is not a partial order. However, with the knowledge inference algorithm used, the cycles introduced by the symmetric relationships have no impact on the propagation of inference insofar as the algorithm does not propagate the observations in a transitive fashion. With the algorithm of Desmarais and Pu (2005) and Desmarais et al. (1996), which is recursive and follows transitive relationships, measures must be taken to avoid entering infinite loops. The simplest principle consists of stopping as soon as a node has already been visited for a same observation and to propagate breadth first, rather than depth first.

It must be noted that symmetry is the only possible source of cycles in a POKS structure, considering that a cycle $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow A$ cannot be produced otherwise than by having symmetric relationships (see Desmarais et al. 2006, for a formal proof).

Performance Comparison of the Various Models

We have seen that IRT models can be considered graphical models akin to naive Bayesian networks. The model is based on a local independence hypothesis and on a relationship among the items and a competence based on a sigmoid function. The POKS model is also akin to naive Bayesian networks and is based on a posterior probability calculation and a structure induced on the basis of statistical tests for inferring the learning order among the items. In addition to these two models, there is also the model of the general Bayesian networks that is also used to model and infer competences. How do these models compare with regard to their capacity to predict the results on a test of competences?

We report, in the following, on a series of experiments aimed at determining how these approaches compare in terms of predictive performance. The first of these experiments is described in the next section and deals with the comparison between the two-parameter logistic (2PL) IRT model and POKS. It is followed by a comparison between the POKS approach and Bayesian networks.

Comparison of POKS with IRT

Desmarais, Fu, and Pu (2005) implemented simulations for evaluating the predictions of the POKS and IRT models. Essentially, these simulations dealt with the respective capacity of each model to predict the classification of a respondent regarding his/her overall success on the test, as well as their capacity to predict success on each item individually.

The experiments were based on post-hoc simulation of a question-response process. For each respondent, the process was simulated with previously gathered complete test data. Thus, the predictions of the IRT and POKS approaches were compared to real data.

Different variables were used to explore the behavior of the models. Thus, two methods for the choice of the next item were used: Fisher information and the reduction of entropy. Fisher information is widely used in IRT applications (see Eggen, 1998, for a comparison of the various item selection strategies with IRT), while entropy reduction is a widely used approach in the area of probabilistic models and has been used with the POKS approach (Desmarais et al. 1996a). Comparisons used two data sets: a 160-item test on the knowledge of French used in the Canadian civil service and a 34-item test on the knowledge of Unix commands.

Prediction of success on the test

The results of Desmarais et al. (2005) showed that the two approaches were very effective for categorizing respondents, with a rate of well-categorized respondents over 90% after only 10% of administered items in approximately half of the simulation cases. However, results differ according to various parameters, in particular the cut score and the item selection procedure. Overall, we note that when choosing items based on Fisher information, both approaches had similar scores (with an overall average score of 0.73 for both and when the POKS' α parameter was optimal—which is to be expected once calibration is done). When choosing items based on information gain, POKS then performed better than the 2PL IRT model, with an average overall score of 0.79 compared to an average score of 0.73 for the IRT[1].

Prediction of success on the items

Other simulations were carried out to determine their capacity to predict success on items individually. This objective is obviously more difficult, but it offers the possibility of providing a very accurate assessment insofar as certain items concern more specific competences and that an assessment can be made with heightened accuracy.

Results varied greatly between the two data sets, those of the Unix test showing a clearly better performance than those of the French test, and for both models. However, the POKS model had better performance in both cases. This result does not seem surprising considering that the IRT model is not designed for this task and cannot deal with several dimensions simultaneously, i.e. distinguishing the various competences involved in a test. On the contrary,

[1]Although the information gain approach could be implemented in the simulation environment with POKS, it could not be implemented for the IRT model.

the POKS model possesses this quality and it comes from the fact that the formalism used is a partial order rather than a linear order (for example, $A \rightarrow B \rightarrow C$) that could adequately model only one skill. It would, therefore, be worthwhile to investigate whether a multidimensional IRT model could perform at the same level as POKS.

Comparison of POKS With a Bayesian Network

The simple IRT model cannot manage several dimensions at once; however, this is certainly not the case with Bayesian networks, as we have seenabove. Desmarais et al. (2006) have, therefore, compared the capacity of a Bayesian network to predict success on items individually. A Bayesian network is constructed using the data following two current methods: the PC algorithm (Spirtes, Glymour & Scheines 2000) and the K2 algorithm (Cooper & Herskovits 1992). Inferences with the networks are also based on current algorithms in the domain. Question-response simulations were done following the example of the methodology used for comparison with the 2PL IRT model. The same data were also used and data of a third test of some twenty items in arithmetic were added.

Results show that the POKS model performed better than the Bayesian networks. Not only were the predictions more accurate, with a reduction in the relative error of approximately 5% to 20% in general, but the standard deviation of the predictions on several simulations was much smaller, pointing to a greater stability than the model based on a Bayesian network. Just as it is with the comparison with the 2PL model, the size of the differences varied from one test to another, but the tendency was systematic.

Combining the Approaches

We, therefore, conclude that in the matter of prediction for individual items, the POKS model appears to have an advantage in relation to models based on IRT and Bayesian networks. This makes it the better choice for the task of making an accurate assessment when using just one test.

However, we can ask whether Bayesian networks are not better suited for exploiting relationships between higher-level competences and concepts, considering that POKS does not represent these competences in its model that is solely made up of items, i.e. observable manifestations of competences. This question has also been investigated in the study by Desmarais et al. (2006).

The authors attempted to improve the assessment of higher-level competences using a technique that involves increasing the initial observations with POKS. For example, if items X_a and X_b are posed to the respondent and POKS estimates that item X_c has a chance of success beyond a certain threshold, then the three items will be considered as observed and provided as an entry in the Bayesian network for it to perform its assessment of competences.

An experiment with the arithmetic test data was conducted using this technique. The higher-level competences in arithmetic of each respondent had been estimated previously independently by experts[2], thus serving to validate the predictions made by the model.

The results of the experiment remain mitigated. It was shown that POKS can improve predictions on individual items and even surpass predictions made with POKS only, though the difference was not statistically significant at $\alpha = 0.05$. However, the prediction of success for

[2]The data came from Vomlel (2004).

higher-level competences was not improved. We can conclude that the POKS inferences are already contained in the relationships of the Bayesian network that express links among competences. Increasing the set of observed items would, therefore, add nothing to the assessment. Nevertheless, the fact that the prediction for the items was improved somewhat contradicts this hypothesis. It is also possible that independent evaluations of competences contain too much noise, or again, that improvement of predictions for the items is only the result of chance. Other studies are needed to elucidate these questions.

Conclusions

This paper presented a graphical and Bayesian approach for the modeling of knowledge, the POKS approach. The approach is inspired both by a widely recognized theory of learning in the field of cognitive psychology, the theory of knowledge spaces (Doignon & Falmagne 1999), and graphical Bayesian networks for managing probability concerns. First, we showed that the approach can classify respondents as a "master" with a rate generally as good or better than IRT's two-parameter logistic model. Moreover, considering that one of the main advantages of the POKS approach is to provide a prediction for the mastery of each item individually, we also compared the predictions at the level of the items themselves. The comparison was made with a model based on Bayesian networks, because these also predict mastery at the level of individual items. The comparison revealed that the POKS approach predicted mastery on items in a more accurate fashion than the model based on Bayesian networks.

The POKS approach, therefore, seems to offer an efficient and effective model for the prediction of results of a competence test, both on the overall level of the test, as the comparison with the 2PL IRT model showed, and at the precise level of items individually, as shown by the comparison with a model based on Bayesian networks. However, in spite of these apparent advantages, several uncertainties and major questions remain.

The POKS approach is shown to perform generally better than the two-parameter logistic IRT model, but questions remain with regard to its comparison with the more recent developments in the area of the IRT, in particular, multidimensional models. If a test was constructed with items that conformed perfectly to a specific number of dimensions, would the POKS approach still show the advantage that we observed? The results reported in this paper deal with tests that cover several dimensions, as do most tests, and are therefore probably not optimal for a multidimensional model. It is, therefore, possible that results would be different for a multidimensional IRT approach.

A second question concerns evaluation of the margin of error and the reliability of POKS predictions. We do not know the distribution of the accuracy rate of POKS predictions, other than by deriving it from simulations such as those carried out for the experiments described here. In certain contexts, however, it is crucial to be able to estimate the margin of error of predictions. The most obvious case is that of a test aimed at accreditation in which a decision must be made to stop or continue presenting items to the respondent. It is, therefore, important to better know the conditions that influence the reliability of the POKS model predictions and, ultimately, to be able to derive a margin of error relative to these predictions.

In addition, it should be noted that the experiments carried until thus far concern tests with from approximately 20 to 160 items. Several contexts for the use of adaptive testing require several hundreds, even thousands, of items. These contexts present an additional difficulty, that

of having to construct a POKS model using partial data. Considering that it is not realistic to administer a test with several hundred of items to a respondent, the model must therefore be constructed using incomplete data for each individual. The knowledge construction and inference model must be adapted, as a consequence. In addition, these contexts also bring up the question of whether the model remains robust with tests of that size. This has yet to be demonstrated.

References

Almond, R. G. & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.

Carr, B. & Goldstein, I. (1977). *Overlays: A theory of modelling for computer aided instruction* (Technical report). AI Memo 406, MIT, Cambridge, MA.

Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309–347.

Desmarais, M. C., Fu, S. & Pu, X. (2005). Tradeoff analysis between knowledge assessment approaches. In C. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.), *Proceedings of the 12th international conference on artificial intelligence in education, AEID'2005* (pp. 209–216). Amsterdam: IOS Press.

Desmarais, M. C., Maluf, A. & Liu, J. (1996). User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-adapted Interaction*, *5*(3-4), 283–315.

Desmarais, M. C., Meshkinfam, P. & Gagnon, M. (2006). Learned student models with item to item knowledge structures. *User Modeling and User-adapted Interaction*, *16*(5), 403–434.

Desmarais, M. C. & Pu, X. (2005). A Bayesian inference adaptive testing framework and its comparison with item response theory. *International Journal of Artificial Intelligence in Education*, *15*, 291–323.

Doignon, J.-P. & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Duda, R. & Hart, P. (1973). *Pattern classification and scene analysis*. Wiley, New York.

Eggen, T. J. (1998). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249–261.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P. & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui, & J. Schmid (Eds.), *ICFCA*, Vol. 3874 of *Lecture Notes in Computer Science* (pp. 61–79). Springer.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P. & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review, 97*, 201–224.

Findler, N. V. (Ed.). (1979). *Associative networks: Representation and use of knowledge by computers*. Orlando: Academic Press.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*(2-3), 131–163.

Jameson, A. (1995). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-adapted Interaction*, *5*(3-4), 193–251.

McTear, M. F. (1993). User modelling for adaptive computer systems: a survey of recent developments. *Artificial Intelligence Review*, *7*, 157–184.

Mislevy, R. J. & Gitomer, D. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-adapted Interaction*, *42*(5), 253–282.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. New Jersey: Prentice Hall.

Payne, S. J. & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science*, *14*(3), 445–481.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (pp. 41–46).

Self, J. (1988). Bypassing the intractable problem of student modelling. *Proceedings of Intelligent Tutoring Systems, ITS'88* (pp. 18–24).

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, Massachusetts: The MIT Press, 2nd edition.

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, *12*(Supplementary Issue 1), 83–100.