Submission type: Full paper

Title: Tradeoff analysis between knowledge assessment approaches

Authors: Michel C. Desmarais (michel.desmarais@polymtl.ca), Shunkai Fu (shukai.fu@polymtl.ca) and Xiaoming Pu (xiaoming.pu@polymtl.ca)

Abstract: The problem of modeling and assessing an individual's ability level is central to learning environments. Numerous approaches exists to this end. Computer Adaptive Testing (CAT) techniques, such as IRT and Bayesian posterior updating, are amongst the early approaches. Bayesian networks and graphs models are more recent approaches to this problem. These frameworks differ on their expressiveness and on their ability to automate model building and calibration with empirical data. We discuss the implication of expressiveness and data-driven properties of different frameworks, and analyze how it affects the applicability and accuracy of the knowledge assessment process. We conjecture that although expressive models such as Bayesian networks provide better cognitive diagnostic ability, their applicability, reliability, and accuracy is strongly affected by the knowledge engineering effort they require. We conclude with a comparative analysis of data driven approaches and provide empirical estimates of their respective performance for two data sets.

Keywords: Student models, Bayesian inference, graphical models, adaptive testing, CAT, IRT, Bayesian networks

Institution: École Polytechnique de Montréal
Computer Engineering
C.P. 6079, succ. Centre-Ville,
Montréal Québec,
Canada, H3C 3A7
telephone 1.514.340.4711x3914

Main contact person: michel.desmarais@polymtl.ca

Tradeoff analysis between knowledge assessment approaches

Michel C. Desmarais, Shunkai Fu Xiaoming Pu
École Polytechnique de Montréal

Abstract. The problem of modeling and assessing an individual's ability level is central to learning environments. Numerous approaches exists to this end. Computer Adaptive Testing (CAT) techniques, such as IRT and Bayesian posterior updating, are amongst the early approaches. Bayesian networks and graphs models are more recent approaches to this problem. These frameworks differ on their expressiveness and on their ability to automate model building and calibration with empirical data. We discuss the implication of expressiveness and data-driven properties of different frameworks, and analyze how it affects the applicability and accuracy of the knowledge assessment process. We conjecture that although expressive models such as Bayesian networks provide better cognitive diagnostic ability, their applicability, reliability, and accuracy is strongly affected by the knowledge engineering effort they require. We conclude with a comparative analysis of data driven approaches and provide empirical estimates of their respective performance for two data sets.

Keywords. Student models, Bayesian inference, graphical models, adaptive testing, CAT, IRT, Bayesian networks

1. INTRODUCTION

Assessing the user's mastery level with respect to one or more abilities is a key issue in learning environments. Any system that aims to provide intelligent help/assistance to a user is bound to model what that person already knows and doesn't know.

The Item Response Theory (IRT) emerged as one of the earliest and most successful approaches to perform such assessment (Birnbaum, 1968). The field of Computer Adaptive Testing, which aims to assess an individual's mastery of a subject domain with the least number of question items administered, has relied on this theory since its conception. Although it was initially limited to assessing a single ability dimension, the IRT framework has since been extended to multidimensional assessment (Reckase, 1997).

IRT has the characteristic of being data driven: knowledge assessment is purely based on model calibration with sample data. Model building is limited to defining which item belongs to which skill dimension. These are important characteristics that IRT shares with other student modeling approaches such as Bayesian posterior updates (Rudner, 2002) and POKS (Desmarais, Maluf, & Liu, 1995). We return to this issue later.

Curiously, until fairly recently, the field of intelligent learning environments did not adopt the IRT approach to modeling the learner's expertise, even though this approach was cognitively and mathematically sound. Instead, techniques known as "overlay models" (Carr & Goldstein, 1977) and "stereotypes" (Rich, 1979) were used to model what the user knows (see Kobsa, 2001). It remains speculative to explain why the research community working on intelligent learning applications has, at least initially, largely ignored the work on IRT and other data driven approaches, but we can evoke some possibilities:

- training data that could prove difficult to collect if large samples are required;
- IRT requires numerical methods (eg. multi-parameters maximum likelihood estimation) that were non trivial to implement and not widely available as software packages until recently;
- the AI community was not familiar with the field from which IRT comes from, psychometric research;
- intelligent learning applications focused on fine grained mastery of specific concepts and student misconceptions in order to allow highly specific help/tutoring content to be delivered; IRT was not designed for such fine grained assessment but focuses instead on the determining the mastery of one, or a few, ability dimensions.

However, in the last five to ten years, this situation has changed considerably. Overlay and stereotype-based models are no longer the standard for performing knowledge assessments in AI-based learning systems. Ap-

proaches that better manage the uncertainty inherent to student assessment, such as probabilistic graphical models and Bayesian networks, are now favored. In fact, researchers from the psychometric and the Student/User Modeling communities are recently working on common approaches. These approaches rely on probabilistic graph models that share many commonalities with IRT-based models, or encompass and extend such models (Almond & Mislevy, 1999; VanLehn & Niu, 2001; Reye, 2004; Mayo & Mitrovic, 2001). Reflecting on these last developments, we can envision that the data driven and the probabilistic/statistical models, of which IRT is an early example, and the fine grained diagnostic approaches, typical of Intelligent Learning Environments, are gradually merging. In doing so, they can yield powerful models and raise the hope of combining the best of both fields, namely cognitively and mathematically sound approaches that are amenable to statistical parameter estimation (cf. full automation), and high modeling flexibility necessary for intelligent learning environments.

We review some of the emerging models and compare their respective advantages from a qualitative perspective, and conclude with a performance analysis of three data driven approaches over two domains of assessment.

2. Qualitative factors

Student modeling approaches differ over a number of dimensions that can determine the choice of a specific technique in a given context of application. These dimensions are summarized in the following list.

Flexibility and expressiveness: As hinted above, AI-based systems often rely on fine-grained assessment of abilities and misconceptions. As hinted above, that factor itself may be sufficient to throw out a single skill dimension IRT model. Although global skill dimensions are appropriate in the context of assessing general mastery of a subject matter, many learning environments will require more fine-grained assessment.

Cost of model definition: Fine-grained models such as those found in Bayesian Networks (see for example Vomlel, 2004; Conati, Gertner, & VanLehn, 2002) require considerable expert modeling effort. On the contrary, data driven approaches such as IRT can completely waive the knowledge engineering effort. Because of the modeling effort, fine-grained models can prove overly costly for many applications.

Scalability: The number of concepts/skills and test items that can be modeled in a single system is another factor that weights into evaluating the appropriateness of an approach. The underlying model in IRT allows good scalability to large tests and for a limited number of ability dimensions. For fine grained student models, this factor is more difficult to assess and must be addressed on a per case basis. For example, in a Bayesian Network where items and concepts are highly interconnected, complexity grows rapidly and can be an significant obstacle to scalability.

Cost of updating: The business of skill assessment is often confronted with frequent updating to avoid over exposure of the same test items. Moreover, in domains where the skills evolve rapidly, such as in technical training, new items and concepts must be introduced regularly. Approaches that reduce the cost of updating the models are at significant advantage here. This issue is closely tied to the knowledge engineering effort required and the ability of the model to be constructed and parametrized with a small data sample.

Accuracy of prediction: Student modeling applications such as Computer Adaptive Testing (CAT) are critically dependent on the ability of the model to provide an accurate assessment with the least number of questions. Models that can yield confidence intervals, or the degree of uncertainty of their inferences/assessment, are thus very important in this field as well as in many context in which measures of accuracy is relevant.

Reliability and sensitivity to external factors: A factor that is often difficult to assess and overlooked is the reliability of a model to environmental factors such as the skills of the knowledge engineer, the robustness to noise in the model, and to noise in the data used to calibrate a model. Extensive research in IRT has been conducted on the issue of reliability and robustness under different conditions, but little has been done in intelligent learning environments[1].

Mathematical foundations: The advantages of formal and mathematical models need not be defended Models that rely on sound and rigorous mathematical foundations are generally considered better candidates over *ad hoc* models without such qualities because they provide better support to assess accuracy and reliability, and they can often be automated using standard numerical modeling techniques and software packages. Both the Bayesian Network and IRT approaches do fairly well on this ground, but they also make a number of assumptions that can temper their applicability.

[1] With some notable exceptions such as VanLehn and Niu (2001).

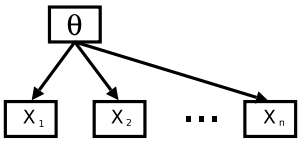

Figure 1. Graphical representation of the links between θ, the examinee's mastery or ability level, and $\{X_1, X_2, ..., X_n\}$, the test items.

Approximations, assumptions, and hypothesis: In the complex field of cognitive and skill modeling, all models must make a number of simplifying assumptions, hypothesis, or approximations in order to be applicable. This is also true of Bayesian modeling in general, and for a number of reasons that range from the lack of sufficient data to calibrate full joint conditional probability tables, to the complexity of constructing and inference in Bayesian networks. Of course, the more assumptions and approximations are made, the less accurate and reliable a model becomes. This issue is closely linked to the reliability and sensitivity one. Some approach may work well in one context and poorly in another because of violated assumptions.

These factors will determine the value of a student modeling approach. For example, a model approach that requires highly skilled Ph.D.'s in Bayesian modeling, combined with expert content knowledge, and that performs poorly if some modeling errors are introduced, will be much less appealing than an approach that can be fully automated, whose reliably is good and measurable with small samples to build and calibrate the model, and yet, that permits fine grained cognitive modeling.

3. Qualitative comparison of approaches

The previous section establishes the qualitative factors by which we compare different approaches to student skill modeling. This section pursues with an analysis of how models fare with respect to the factors mentioned. A more specific quantitative comparison will follow.

The student models we focus on are (1) IRT, (2) a simple Bayesian posterior probability update, (3) a graphical model that links items among themselves and uses a Bayesian update algorithm (POKS), and (4) more complex Bayesian and graphical models that link together concept and misconceptions (hidden variables), and items (evidence nodes) within the same structure.

3.1. Bayesian posterior updates

The simplest approach to assessing mastery of a subject matter is the Bayesian posterior update. It consists in the application of Bayes rule to determine the posterior probability: $P(m|X_1, X_2, \ldots, X_n)$, where m stands for *master* and $X_1, X_2, \ldots, X_n$ is the response sequence after n item responses are given. According to Bayes theorem and under strong independence assumptions, the posterior probability of m given the observation of item X_i is determined by:

$$P(m|X_i) = \frac{P(X_i|m)\ P(m)}{P(X_i|m)\ P(m) + P(X_i|\neg m)\ (1 - P(m))} \tag{1}$$

$P(m|X_i)$ will serve as the next value for $P(m)$ for computing $P(m|X_{i+1})$. The initial and conditional probabilities, $P(m)$ and $P(m|X_i)$, are obtained from sample data. We refer the reader to (Rudner, 2002) for further details.

The approach can be graphically represented by figure 1 and by considering θ as the *mastery* node and $\{X_1, X_2, ..., X_n\}$ as the test items. The interpretation of this graph is that θ, the event that the student masters the subject matter, will influence the probability of correctly answering each test items. Almond and Mislevy (1999) shows that this graph also corresponds to the IRT model, although the probability updating scheme is different. More on this later in section 3.2.

This approach has many advantages that stem from its simplicity. It does not require knowledge engineering and can be fully automated and calibrated with small data sets. It is also computationally and conceptually very simple.

That simplicity comes at the price of low granularity and strong assumptions. In equation 1, the student model is limited two states, *master* or *non-master* with regards to a subject matter[2]. The model also makes the assumption that all test items have similar discrimination power, whereas it is common to find items significantly more discriminant than others.

[2]Mastery is determined by an arbitrary passing score.

Although figure 1 illustrates a single dimension example, multiple dimensions, or multiple concepts, can readily be modeled with this approach. Each concept or subject matter, s, can be represented by their respective θ_s. Moreover, the model can be extended to more than two states, although a larger data set will be necessary to obtain equivalent accuracy as in a two-states model. Some intelligent tutoring systems have used such extensions to the basic principle of Bayesian posterior probability updates to build intelligent learning environments (see Jameson, 1995; Almond & Mislevy, 1999). Some also relied on subjective assessments to derive the conditional probabilities, but that strategy is highly subject to human biases and low agreement amongst experts that can result in poor accuracy and low reliability.

3.2. Item Response theory

IRT can be considered as a graphical network similar to the one in figure 1. However, in contrast to the Bayesian posterior update method, the variable θ represents an ability level on a continuous scale. The probability of succeeding an item X_i is determined by a logistic function named the Item Characteristic Curve (ICC)[3]:

$$P(z_i \mid \theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \tag{2}$$

Note that this particular function is called the "two-parameter logistic model". Other variants exists, dropping parameters a and b, or adding a guessing parameter c. The function defines an 'S' shaped curve where the probability $P(X_i)$ increases as a function of θ, as one would expect. The parameter a determines the slope of increase around a value of θ determined by the second parameter, b.

The value of θ is determined by maximizing the likelihood of the responses provided by the student, generally using a maximum-likelihood numerical method. IRT is a well documented and details can be found in Reckase (1997).

IRT has the advantage of being a fully automated method that can be calibrated with relatively small data set, depending on the desired accuracy of the assessment. Contrary to the Bayesian posterior update approach in section 3.1, the two-parameter IRT model takes into account the discrimination factor of individual test items, and it models ability on a continuous scale as opposed to a dichomotous variable, or a multinomial variable when the model is extended. This last property of the model also means that a greater accuracy can be expected for computing $P(X_i|\theta)$. That information can, in turn, be useful for the purpose of computing the most informative test items or adjusting item difficulty. Finally and as mentioned, the model can be extended for multidimensionality. In short, it is a more sophisticated model than the Bayesian posterior updating model, but it does not allow fine-grained modeling of a large number of dimensions such as found in some intelligent tutoring systems where individual concepts and misconceptions are often modeled.

3.3. Probabilistic graphs models

Figure 1's graph model is limited to a single ability dimension and test items are singly-connected the ability node. However, graph models can also embed specific concepts and misconceptions in addition to general skill dimension and test items. The network structure can be a multilevel tree structure. Test items can be connected together in a directed graph such as figure 2's structure. We refer to such extensions as probabilistic graph models (for a more detailed discussion on the subject, see Almond & Mislevy, 1999)

To model fine-grained skill acquisition, such as individual concepts and misconceptions, probabilistic graphical models are arguably the preferred approach nowadays. Such models represent the domain of skills/misconceptions as the set of nodes in the graph, $\{X_1, X_2, ..., X_n\}$. A student model consists in assigning a probability to each of the node's value. The arcs of the graph represent the interrelationships amongst these nodes. The semantics of the arcs varies according to the approach, but it necessarily has the effect that changes occurring in the probability of a node affects neighboring nodes and, under some conditions according to the inference approach, it can propagate further.

3.4. Item to item graph models

One probabilistic graph model approach is to link test items among themselves. The domain of expertise is thus defined solely by observable nodes. A "latent" ability (i.e. non directly observable) can be defined by a subset of nodes, possibly carrying different weights. This is essentially the principle behind exam tests where questions can carry weights and where the mastery of a topic in the exam is defined as the weighted sum of success items.

[3]The ICC curve can also be defined after what is known as the normal ogive model but the logistic function is nowadays preferred for its greater computational tractability.

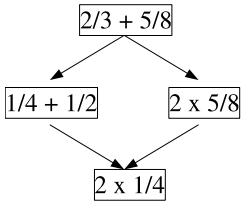

Figure 2. Graphical example of the interrelationships between abilities to solve different arithmetic problems.

Item to item graph models derive probabilities of mastery of an item given the partial order structure of items (as in figure 2), and given the items observed so far. The semantics of links in such structures simply represents the order in which items are mastered. The cognitive basis behind such an approach is the Knowledge space theory (Falmagne, Koppen, Villano, M., & Johannesen, 1990), which states that the order in which people learn to master knowledge items is constrained by an AND/OR graph structure. The example in figure 2 illustrates the order in which we could expect people to learn to solve these simple arithmetic problems. For example, we learn to solve $2 \times 1/4$ before we can solve $1/4 + 1/2$, but the order is not clearly defined between abilities for solving $2 \times 5/8$ and solving $1/4 + 1/2$. Figure 2 is, in fact, a directed acyclic graph (DAG), not an AND/OR graph, but it does capture the partial ordering of mastery amongst knowledge items and allows to make valuable inferences. What it does not capture, are alternative methods of mastery. We refer the reader to Falmagne et al. (1990) for more details on this theory.

Some researchers adopted this type of graph representation to perform knowledge assessment. Kambouri, Koppen, Villano, and Falmagne (1994) used a combination of data driven and knowledge engineering approach to build knowledge structures (AND/OR graphs), whereas Desmarais et al. (1995) used a data driven only, automated approach to building a simplified version of knowledge structures represented as DAG. That approach is named Partial Order Knowledge Structures (POKS). We will compare the POKS performance to the IRT and the Bayesian posterior update approaches in section 4.

The advantage of leaving out the latent abilities from the graph structure is that model construction can be fully automated (at least for the POKS approach). It also involves benefits in terms of reliability and replicability by avoiding expert-based model building and subjective and individual biases. The disadvantages is the loss of explicit links between concepts or misconceptions in the graph structure. However, note that latent abilities (concepts) can later be derived by determining the items that are evidence for given concepts. For example, if concept C_1 has evidence nodes X_1, X_2, X_3, mastery of C_1 can be defined as a weighted sum of the probabilities of its evidence nodes: $C_1 = w_1X_1 + w_2X_2 + w_3X_3$.

3.5. Concept and misconception nodes graph models

Graph structures that also include concept and misconceptions nodes in addition to test items can derive the probability of success in a more sophisticated manner than the item to item graph models described above. Probability of mastery of a concept can be determined by estimated mastery of other concepts and by the presence of misconceptions in the student model. Most research in intelligent learning environments used different variations of this general approach to build graph models and Bayesian networks to perform student expertise assessment (to name only a few: Vomlel, 2004; Conati et al., 2002; Millán & Pérez-de-la-Cruz, 2002; Martin & Vanlehn, 1995).

By modeling the interdependencies between concepts of different level of granularity and abstractions, misconceptions, and test items that represent evidence, it comes as no surprise that a wide variety of modeling approaches are introduced. We will not attempt to further categorize graph models and Bayesian networks here, but try to summarize some general observations that can be made on these.

A first observation is that the student models can comprise fine-grained and highly relevant pedagogical information such as misconceptions. It entails that detailed help or tutoring content can be delivered to the user once the student cognitive assessment is derived.

We also note that many approaches rely on a combination of data driven and knowledge engineering to derive the domain model. However, we know of no example that is completely data driven. This is understandable since detailed modeling of concepts and misconceptions necessarily requires pedagogical expertise. What can be data driven is the calibration of the model, namely the conditional probabilities in Bayesian networks.

The variety of approaches in using Bayesian networks and graph models to build student models that include concepts and misconceptions is much too large for a proper coverage in the space allotted here. Let us only conclude this section by mentioning that, although these approaches are currently more complex to build and to use, they have strong potential because of their flexibility. The effort required is most appropriate for knowledge domains that are stable such as mathematics teaching.

4. Performance comparison

In the previous sections, we attempted to draw a comparative picture of some student modeling approaches over dimensions such as data-driven vs human engineered models, which in turn has impacts on how appropriate is an approach for a given context. Very simple approaches based on Bayesian posterior updates, and slightly more sophisticated ones such as IRT and item to item graph structures, can be entirely data driven and require no knowledge engineering effort. By contrast, more complex structures involving concepts and misconceptions are not currently easily amenable to fully automated model construction, although model calibration is feasible in some cases.

We conducted an empirical comparison of the data driven approaches over two knowledge domains, the Unix shell commands and the French language. The approaches are briefly described and the results reported. First, a short description of the simulation method for the performance comparison is described.

4.1. Simulation method

The performance comparison is based on the simulation of the question answering process. For each examinee, we simulate the adaptive questioning process with the examinees' actual responses[4]. The same process is repeated for every approach. After each item administered, we evaluate the accuracy of the examinee's classification as a *master* or *non master* according to a pre-defined passing score.

The choice of the next question is adaptive. Each approach uses a different method for determining the next question because the optimal method depends on the approach. We use the method for choosing the next question that yields the best result for each approach.

The performance score of each approach corresponds to the number of correctly classified examinees after i items are administered.

The simulations are made on two sets of data: (1) a 34 items test on the knowledge of Unix shell commands administered to 48 examinees, and (2) a 160 items test on French language administered to 41 examinees.

4.2. Bayesian posterior updates, IRT, and POKS comparison

All approaches compared are well documented elsewhere and we limit their descriptions to brief overviews.

4.2.1. Bayesian posterior updates

The Bayesian posterior updates procedure consists in applying Bayes rules according to equation (1).

The choice of the next question to ask is the maximum discrimination measure (Rudner, 2002):

$$M_i = |\log(P(z_i|m)/P(z_i|\neg m))|$$

4.2.2. Item Response Theory (IRT)

The simulation uses the two-parameters logistic model version of IRT which corresponds to equation (2). Values for parameters a and b are calibrated using the sample data sets. Estimation of θ is performed with a maximum likelihood estimation procedure after each item is administered.

Choice of the next question corresponds is based on the Fisher information measure, which is the most widely used for the IRT approach and it was introduced early on in IRT (Birnbaum, 1968). The Fisher information is a function of θ and the parameters a and b of equation (2).

4.2.3. Partial Order Knowledge Structure (POKS)

The POKS method is described in Desmarais et al. (1995) (see also Desmarais & Pu, 2005). It consists in inferring structures such as the one in figure 2 from the data sample. Updating of the conditional probabilities is based on Bayesian posterior probabilities of the parent nodes. Evidence is propagated from observed nodes (items answered) in accordance to Bayes rule for the nodes directly connected with the observed one. Evidence is further propagated to indirectly linked nodes according to an interpolation scheme known as the PROSPECTOR algorithm (see Giarratano & Riley, 1998). For linear structures (eg. $A \rightarrow B \rightarrow C$), test shows that this approximation yields probability values that are within less than 1 percent of those obtained with Bayesian Network software applications such as Microsoft MSBNx (Microsoft Corporation, 2005). For other structures, the values will differ according to how valid are the assumptions of conditional independence of the POKS framework for the data set, and how accurate is the approximation. To a large extent, an empirical answer to this question is provided by the performance evaluation.

[4]Taking care of removing from the calibration data the simulation's current examinee's data case in order to avoid over-calibration.

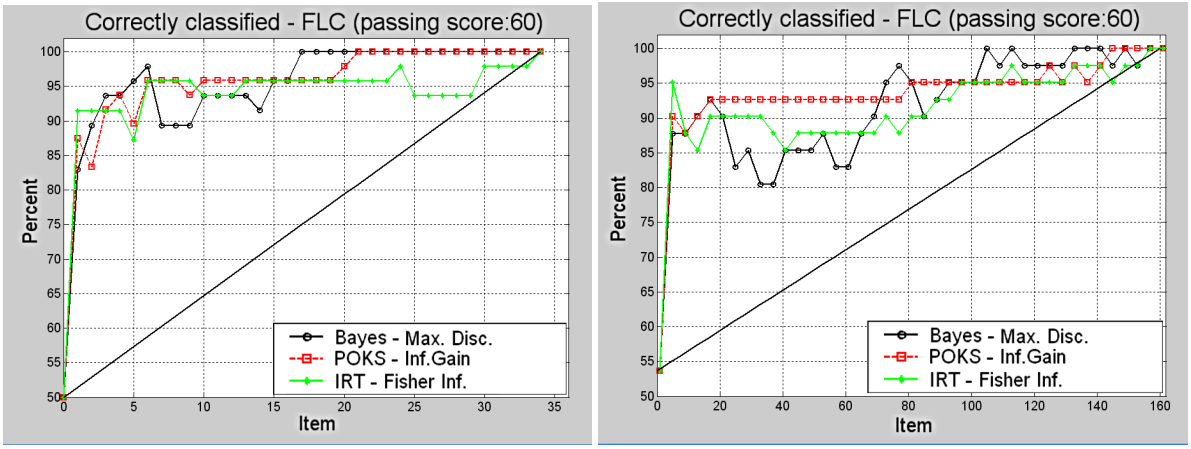

Figure 3. Performance comparison of three knowledge assessment methods.

The choice of the next question is determined by the minimal entropy measure. The item chosen corresponds to the one that is expected to reduces the most the entropy of the test items set. The entropy measure is based on the standard formula $-[p\log(p) + (1-p)\log(1-p)]$ and the test's entropy is the summation over all test item entropies. Test entropy value is highest if all items have a probability of 0.5 (i.e. maximum uncertainty), and it is 0 if all items have a probability of 1 or 0.

4.3. Results

The performance of the three approaches are compared in figure 3. It reports the results of the simulation for the Unix and French language tests comprised respectively of 34 and 160 items. The percentage of correctly classified examinees, averaged over 48 simulation cases for the Unix test and 41 for the French language one, are plotted as a function of the number of item responses. Passing score is 60%. The diagonal line is for baseline comparison.

Both plots start at 0 questions, which corresponds to the number of correctly classified examinees that correctly fall into the most likely state (master or non master) according to the sample. For the Unix test about half were master, thus the starting score is around 50%, whereas for the French test a little more than half were master. The x-axis end at the number of questions in the test and at a 100% correctly classified score, when all items are answered. After about 5 question items, all three approaches correctly classify more than 85% of examinees for both tests but, for the French test and after about 5 items, the POKS approach perform a little better than the Bayes posterior update and the IRT approaches. The Bayes approach also appears to be less reliable as the curve fluctuates more than the other two throughout the simulation.

Other simulations shows that POKS and IRT are in general better than Bayes posterior update at cutting scores varying from 50% to 70%[5], and that POKS is slightly better than IRT but not systematically (further details can be found in Desmarais & Pu, 2005).

5. Conclusion

Student models are gradually converging towards a probabilistic representation of mastery of skill sets. Automated and data driven models such as Bayesian posterior update, IRT, and Partial Order Knowledge Structures (POKS), limit their representation to observable test items. Subsets of items can be used to define higher level skills, but knowledge assessment is not based on them directly. These approaches have the advantages of avoiding the knowledge engineering effort to building the student model. With this come further advantages such as avoidance of human biases and individual differences in modeling, the possibility of full automation and reduced costs for building and updating the models, and a reliability and accuracy that can better be measured and controlled as a function of sample size.

We show that the accuracy of the three data driven approaches for classifying examinees as master and non master is relatively good. Even the simplest method, namely the Bayesian posterior updates, performs relatively well with small data samples below 50 cases, but it is less accurate and reliable than the other two.

Graphical models and Bayesian networks that include concept and misconception nodes provide more flexibility and diagnostic power than the data-driven approaches reviewed. However, they generally require a knowl-

[5] Simulations beyond the 50% to 70% range is unreliable because almost all examinees are already correctly classified before any item is answered.

edge engineering effort that hampers their applicability and can also affect their accuracy. It would be interesting to have a Bayesian network approach to add to the comparison study to better assess their comparative accuracy. This paper aims to nurture some effort in this direction.

6. Acknowledgements

This work has been supported by the National Research Council of Canada.

7. References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 223–237.

Birnbaum, A. (1968). Some latent trait models and their use in infering an examinee's ability. In F. Lord, & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Carr, B., & Goldstein, I. (1977). *Overlays: A theory of modelling for computer aided instruction* (Technical report).

Conati, C., Gertner, A., & VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, *12*(4), 371–417.

Desmarais, M. C., Maluf, A., & Liu, J. (1995). User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction*, *5*(3-4), 283–315.

Desmarais, M. C., & Pu, X. (2005). Computer adaptive testing: Comparison of a probabilistic network approach with item response theory. *Proceedings of the 10th International Conference on User Modeling (UM'2005)* (p. (to appear)). Edinburg.

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.

Giarratano, J., & Riley, G. (1998). *Expert systems: Principles and programming (3rd edition)*. Boston, MA: PWS-KENT Publishing.

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.

Kambouri, M., Koppen, M., Villano, M., & Falmagne, J.-C. (1994). Knowledge assessment: tapping human expertise by the query routine. *International Journal of Human-Computer Studies*, *40*(1), 119–151.

Kobsa, A. (2001). Generic user modeling systems. *User Modeling and User-Adapted Interaction*, *11*(1-2), 49–63.

Martin, J., & Vanlehn, K. (1995). Student assessment using bayesian nets. *International Journal of Human-Computer Studies*, *42*(6), 575–591.

Mayo, M., & Mitrovic, A. (2001). Optimising ITS behaviour with bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, *12*, 124–153.

Microsoft Corporation (2005).

Millán, E., & Pérez-de-la-Cruz, J. L. (2002). A bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*(2–3), 281–330,.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*, 63–96.

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*, 329–354.

Rudner, L. M. (2002). An examination of decision-theory adaptive testing procedures. *Proceedings of American Educational Research Association* (pp. 437–446). New Orleans.

VanLehn, K., & Niu, Z. (2001). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education*, *12*, 154–184.

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