

# Adaptive Testing by Bayesian Networks with Application to Language Assessment

Francesca Mangili<sup>1</sup>, Claudio Bonesana<sup>1</sup>, Alessandro Antonucci<sup>1</sup>,  
Marco Zaffalon<sup>1</sup>, Elisa Rubegni<sup>2</sup>, and Loredana Addimando<sup>2</sup>

<sup>1</sup> Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI-SUPSI  
{francesca,claudio,alessandro,zaffalon}@idsia.ch

<sup>2</sup> Scuola Universitaria Professionale della Svizzera Italiana (SUPSI),  
{elisa.rubegni,loredana.addimando}@supsi.ch

**Abstract.** We present a general procedure for computerized adaptive testing based on probabilistic graphical models, and show on a real-world benchmark how this procedure can increase the internal consistency of the test and reduce the number of questions without affecting accuracy.

**Keywords:** Computerized adaptive testing; Bayesian networks; Entropy.

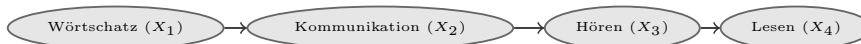
## 1 Introduction

The goal of *Computer Adaptive Testing* (CAT) is to reduce the assessment time and to challenge test takers by adapting the sequence of questions to their ability level. *Item Response Theory* (IRT) is CAT traditional background. *Bayesian networks* (BNs) can offer IRT a powerful language for describing dependencies between skills and modeling richer tasks [1]. Although several researchers have explored BNs in educational assessment, real-world applications and extensive studies of their effectiveness are hardly found in the literature. In this work, we present a general procedure for BNs-based CAT and we test it in a real-world benchmark about German language proficiency assessment.

## 2 Adaptive Testing by Bayesian Networks

Students skills are modeled by a set  $\mathbf{X} := (X_1, X_2, \dots, X_n)$  of categorical variables whose joint *probability*  $P(\mathbf{X})$  is described by a BN through (i) a directed acyclic graph whose nodes represent the variables in  $\mathbf{X}$ ; (ii) conditional probability tables (CPTs)  $P(X_i|\Pi_{X_i})$ ,  $i = 1, \dots, n$ , where  $\Pi_{X_i}$  is the joint variable of the *parents* (i.e., the immediate predecessors) of  $X_i$  (see, e.g., Fig. 1 for the model used in the German language assessment). We point the reader to [2] for the theoretical notions about BNs. To evaluate the informativeness level about  $\mathbf{X}$  provided by  $P$ , we adopt the *entropy*  $H(\mathbf{X}) := -\sum_{\mathbf{x}} P(\mathbf{X}) \cdot \log P(\mathbf{X})$ . Low entropy levels indicate high informativeness. To evaluate the student we formulate a number of *questions*, described as a collection of variables  $\mathbf{Y} := (Y_1, \dots, Y_m)$ . Each question node is represented as a leaf child of the background skills “required”

to answer it. To make our approach adaptive, we chose the  $(k + 1)$ -th question to be asked based on the  $k$ -th previous answers  $y_1, \dots, y_k$ , by minimizing the conditional entropy  $H(\mathbf{X}|y_1, \dots, y_k, Y_{k+1}) := -\sum_{y_{k+1}} H(\mathbf{X}|y_1, \dots, y_k)P(y_{k+1})$ . Finally, we stop the test when the entropy  $H(\mathbf{X}|y_1, y_2, \dots, y_n)$  is sufficiently low.



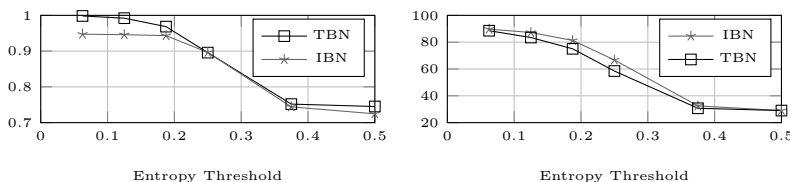
**Fig. 1.** Graph of a BN for German language skills.

*An Application to Language Assessment.* We use the answers of 170 students to 95 questions about German language to reproduce our CAT approach. The *Traditional Evaluation Method* (TEM) assigns to each skill a level A1, A2, B1, B2 by setting thresholds on the fraction of correct answers. We compare TEM with the *independent skills* model (IBN) and the *tree* (TBN) topology in Fig. 1.

Tab. 1 shows in the non-adaptive case the relative agreement between the TEM, IBN and TBN, and the internal consistency of the three tests evaluated using the split-half methodology. Both BN approaches have larger reliability than TEM. Concerning the adaptive case, Fig. 2 shows the relative agreement of the adaptive IBN and TBN with the non-adaptive TBN, and the average number of questions asked. Both models show a strong reduction in the number of questions as the entropy threshold increases. For instance, using the TBN model, we can save 20 questions on average at the price of only a 3% accuracy reduction. This shows that a relevant number of question are little informative and could be avoided by means of an adaptive approach.

Algorithm	Relative agreement					Split-half reliability					
	Wört.	Kom.	Hör.	Les.	All	Algor.	Wört.	Kom.	Hör.	Les.	All
TEM/IBN	.80	.87	.89	.85	.85	TEM	.28	.82	.88	.79	.84
TEM/TBN	.79	.87	.88	.83	.84	IBN	.71	.89	.83	.87	.90
IBN/TBN	.98	.95	.94	.92	.95	TBN	.79	.91	.87	.89	.92

**Table 1.** Relative agreement between models and their split-half reliability.



**Fig. 2.** Agreement with the non-adaptive TBN (left) and average number of questions asked by the adaptive methods (right).

## References

1. R. G. Almond, R. J. Mislevy, Graphical models and computerized adaptive testing, *Applied Psychological Measurement* 23 (3) (1999) 223–237.
2. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.