Question Selection Methods for Adaptive Testing with Bayesian Networks

Martin PLAJNER[1] and Amal MAGAUINA[1] and Jiří VOMLEL[2]

[1] *Faculty of Nuclear Sciences and Physical Engineering,*
Czech Technical University, Prague
Trojanova 13, Prague, 120 00, Czech Republic
martin.plajner@fjfi.cvut.cz

[2] *Institute of Information Theory and Automation,*
Czech Academy of Sciences,
Pod vodárenskou věží 4, Prague 8, 182 08, Czech Republic

Abstract

The performance of Computerized Adaptive Testing systems, which are used for testing of human knowledge, relies heavily on methods selecting correct questions for tested students. In this article we propose three different methods selecting questions with Bayesian networks as students' models. We present the motivation to use these methods and their mathematical description. Two empirical datasets, paper tests of specific topics in mathematics and Czech language for foreigners, were collected for the purpose of methods' testing. All three methods were tested using simulated testing procedure and results are compared for individual methods. The comparison is done also with the sequential selection of questions to provide a relation to the classical way of testing. The proposed methods are behaving much better than the sequential selection which verifies the need to use a better selection method. Individually, our methods behave differently, i.e., select different questions but the success rate of model's predictions is very similar for all of them. This motivates further research in this topic to find an ordering between methods and to find the best method which would provide the best possible selections in computerized adaptive tests.

Keywords: Computerized Adaptive Testing, Question Selection, Bayesian Networks.

1 Introduction

In our research we focus on Computerized Adaptive Testing (CAT). In CAT there is not a single static version of a test distributed to many students but an individual test is dynamically created during the course of testing for each individual participant. The next question is selected with regard to student's previous answers. This leads to several benefits as a better student assessment, a better motivation, etc. [4, 8]

We employ Bayesian networks for our research in this domain. The most recent papers we have published consider the beneficial effect of monotonicity conditions while learning model parameters. In this paper we aim at the testing process itself while the network is already learned. We take a closer look at the question selection procedure. There are many options how to select the next question from a bank of possible questions. This selection process is crucial for a successful adaptive testing procedure because the order in which questions are selected affects the rate in which the model improve its estimations. There is so far no definite answer which objective function produces the best possible results. In this article we discuss several question selection functions and compare them on two real models.

The paper is organized as follows. First, we describe the concept of Computerized Adaptive Testing, our models, and the notation we use. A short overview of two empirical data sets is presented. Both sets contain results of a paper (written) test collected for the purpose of our

research. The first dataset is formed by results of high school tests of mathematical skills in the domain of functions; the second dataset has been collected from test results of foreign students of Czech language. In Section 4, we propose three different types of methods and to compare we also use linear selection process (questions are asked in the same order as they are ordered in the set of possible questions). All methods are tested on two available data sets. Results of experiments are presented in Section 5 of this paper where methods are compared and contrasted. The concluding section summarizes our results and points out possibilities for further improvements in this area.

2 Computerized Adaptive Testing

CAT is a concept of testing which is getting a large scientific attention for about two decades [9, 10, 12]. With CAT we build computer administered and computer controlled tests. The computer system is selecting questions for a student taking the test and evaluating his/her performance.

The process can be divided into two phases: model creation and testing. In the first phase the student model is created while in the second phase the model is used to actually test examinees. There are many different model types usable for adaptive testing as can be found, for example, in [1, 2, 3]. In this work we are working with Bayesian Networks. Regardless of the model the testing part follows the same scheme. With a prepared and calibrated model, CAT repeats following steps:

- The next question to be asked is selected.
- This question is asked and an answer is obtained.
- This answer is inserted into the model.
- The model (which provides estimates of the student's skills) is updated.
- Answers to all questions are estimated given the current estimates of student's skills. (optional)

This procedure is repeated until a termination criterion is reached. Criteria can be of various types, for example, a time restriction, a number of questions, or a confidence interval of the estimated variables (i.e., reliability of the test).

In this article we consider the first step of the testing procedure which is the question selection procedure.

3 Bayesian Network Models

We use Bayesian Networks (BNs) to model students. Details about BNs can be found in, for example, [6, 5]. We restrict ourselves to the following BN structure. Networks have two levels, variables in the parent's level are addressed as skill variables $S \in \mathcal{S}$ where $\mathcal{S}$ is the set of all skills. The children level contains question variables $X \in \mathcal{X}$ where $\mathcal{X}$ is the set of all questions.

- We will use symbol $\boldsymbol{X}$ to denote the multivariable $(X_1, \ldots, X_n)$ taking states $\boldsymbol{x} = (x_1, \ldots, x_n)$. The total number of question variables is n, the set of all indexes of question variables is $\boldsymbol{N} = \{1, \ldots, n\}$. Question variables are binary and they are observable.
- We will use symbol $\boldsymbol{S}$ to denote the multivariable $(S_1, \ldots, S_m)$ taking states $\boldsymbol{s} = (s_1, \ldots, s_m)$. The set of all indexes of skill variables is $\boldsymbol{M} = \{1, \ldots, m\}$. In this article we use only binary skill variables. The set of all possible state configurations of $\boldsymbol{S}$ is $Val(\boldsymbol{S})$. Skill variables are all unobservable.

3.1 Data and specific models

To test our theoretical methods we have collected empirical data. We obtained two different data sets which are described here.

First, we designed a paper test of mathematical knowledge of grammar school students. The test focuses on simple functions (mostly polynomial, trigonometric, and exponential/logarithmic).

Students were asked to solve various mathematical problems (reffered as questions) including graph drawing and reading, calculating points on the graph, root finding, describing function shapes and other function properties. Questions are open (the mathematical problem's solution has to be included) and results are stored as binary correct/wrong values. In total 281 participants took the test. For the purpose of this paper this data set is modeled by two different Bayesian network models. One of them is shown in Figure 1. It consists of 53 questions and 8 skill nodes. These skill nodes represent different student skills connected to questions. This models is further referred to as Mathematical knowledge test Model (MM).

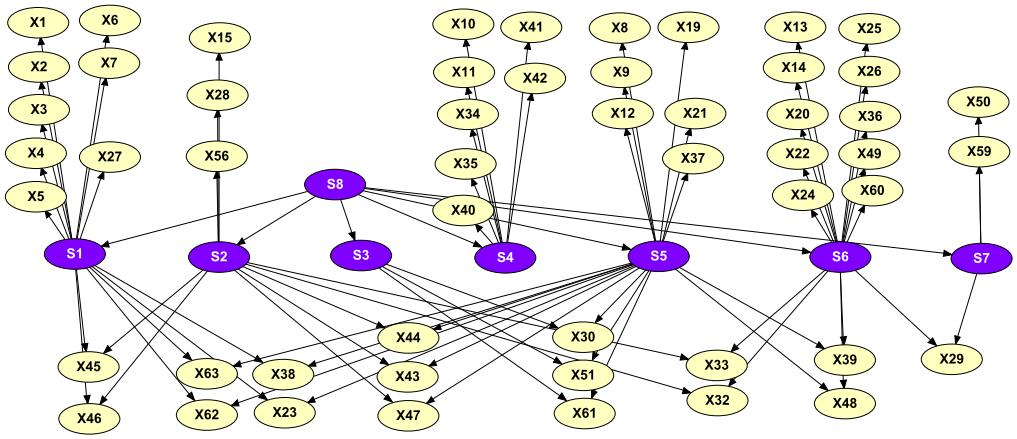

Figure 1: CAT MM structure

The second dataset was collected with a test of Czech language for non-native speaker students. This test contained multiple choice questions with four possible answers. One answer was correct. The test was assessed in a binary way where each question was either correct or incorrect. This test contains 30 questions and 143 students participated in the testing process. The model which was created by a domain expert is shown in Figure 2. Apart from 30 question nodes it has 11 skill nodes. Each skill, again, represents a specific ability a student should have to answer a connected question correctly. The skills include abilities related to morphology, vocabulary, conjugation, etc. This model is referred to as Czech language test Model (CM)[1].

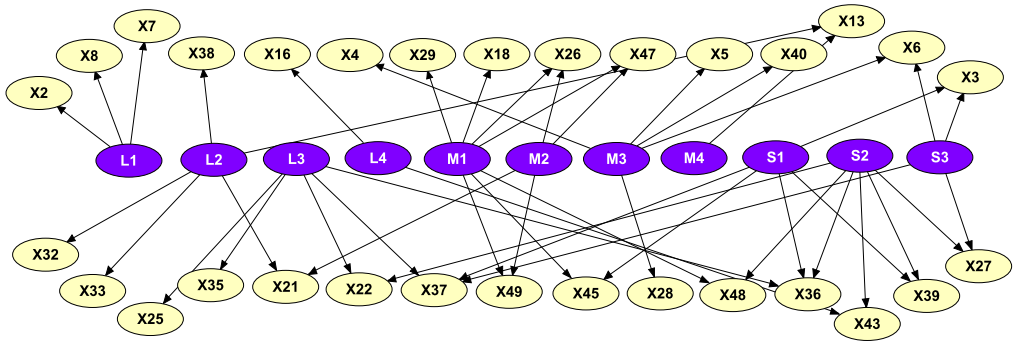

Figure 2: CAT CM structure

4 Question Selection Methods

The task of the question selection is repeated in every step of testing of an individual student. Its process is described in detail below.

[1]More detailed information can be found (in Czech) in the master thesis of Amal Magauina available at https://dspace.cvut.cz/

We define the question evidence e as:

$$e = \{X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n} | i_1, \ldots, i_n \in \boldsymbol{N}\}.$$

where $\{i_1, \ldots, i_n\} = \boldsymbol{I}$ are indexes of already answered questions. Remaining questions are unobserved (unanswered) $\hat{\mathcal{X}} = \{X_i | i \in \boldsymbol{N} \setminus \boldsymbol{I}\}$.

The goal is to select a question from $\hat{\mathcal{X}}$ to be asked next. The selection is dependent on a criterion function which may take different forms. Below, we describe three possible question selection methods. In this paper we also use, as a comparison method, a sequential selection. While using the sequential selection, the question we select is simply chosen in the same order as they are ordered in the question input list. This type of question selection is often used in non-adaptive tests where questions are always asked in the same sequence.

Three methods, we present further, are:

- Maximization of the Expected Entropy Reduction (also called Information Gain)
- Maximization of the Expected Skills Variance
- Maximization of the Expected Question Variance

The motivation for selecting these three possibilities is discussed for each criterion separately.

4.1 Maximization of the Expected Entropy Reduction

The purpose of an adaptive test is to provide the best possible information about a tested student. Each student is modeled by his skills. The criterion described in this section uses the Shannon entropy calculated over all skill values which we define in this section. It is a measure of the certainty of skills estimation. Because of that we want to select a question which provides the largest expected information gain if asked, i.e., a question which reduces uncertainty the most. This method is further referred to as Skills' Entropy.

We compute the cumulative Shannon entropy over all skill variables of S given the evidence e:

$$H(e) = \sum_{j \in \boldsymbol{M}} \sum_{s=0}^{1} -P(S_j = s|e) \cdot \log P(S_j = s|e) \ . \tag{1}$$

Assume we decide to ask a question $\hat{X} \in \hat{\mathcal{X}}$. After inserting the observed outcome the entropy over all skills changes. We can compute the value of new entropy for evidence extended by $\hat{X} = \hat{x}$ as:

$$H(e, \hat{X} = \hat{x}) = \sum_{j \in \boldsymbol{M}} \sum_{s=0}^{1} -P(S_j = s|e, \hat{X} = \hat{x}) \cdot \log P(S_j = s|e, \hat{X} = \hat{x}) \ . \tag{2}$$

This entropy $H(e, \hat{X} = \hat{x})$ is the sum of individual entropies over all skill nodes. Another option would be to compute the entropy of the joint probability distribution of all skill nodes. This would take into account correlations between these nodes. In our task we want to estimate marginal probabilities of all skill nodes. In the case of high correlations between two (or more) skills the latter criterion would assign them a lower significance in the model. This is the behavior we wanted to avoid. The first criterion assigns the same significance to all skill nodes which is a better solution. Moreover, the computational time required for the proposed method is lower.

Now, we can compute the expected entropy after answering question $\hat{X}$:

$$EH(\hat{X}, e) \quad = \quad \sum_{\hat{x}=0}^{1} P(\hat{X} = \hat{x}|e) \cdot H(e, \hat{X} = \hat{x}) \ . \tag{3}$$

Finally, we choose a question X^* that maximizes the information gain $IG(\hat{X}, e)$

$$X^* \quad = \quad \arg\max_{\hat{X} \in \hat{\mathcal{X}}} IG(\hat{X}, e) \ , \text{ where} \tag{4}$$

$$IG(\hat{X}, e) \quad = \quad H(e) - EH(\hat{X}, e) \ . \tag{5}$$

4.2 Maximization of the Expected Skills Variance

With this criterion we want to select a question which leads to the largest variance of state probabilities of skill variables. The rationale behind this selection is very similar to the one discussed in the previous method. The goal is to provide the most accurate estimation of student's skills and also to provide the best separation of students based on their skills. We measure the variance between skill's state probabilities (student having the skill). The variance is measured for two possible answers to one question, i.e., correct and incorrect. The criterion searches for a question which provides the largest variance in these two possibilities. This method is further referred to as Skills' Variance.

We consider unaswered question $\hat{X} \in \hat{\mathcal{X}}$ to be asked. First, we establish following notation:

$$p_0^j = P(S_j = 1|\hat{X} = 0, e) \ ,$$
$$p_1^j = P(S_j = 1|\hat{X} = 1, e) \ ,$$

where $S_j \in \mathcal{S}$. The symbol p_0^j stands for the probability of a student having the examined skill S_j even though the answer to the question $\hat{X}$ was incorrect. p_1^j is the case where the answer was correct and the student has the skill S_j. Naturally, the value of p_1^j should be larger than p_0^j. We compute the average value $\overline{p}^j$:

$$\overline{p}^j = P(\hat{X} = 0|e) \cdot p_0^j + P(\hat{X} = 1|e) \cdot p_1^j \ .$$

Then, the expected variance of states' probabilities of the skill S_j after answering the question $\hat{X}$ can be obtained using the following formula:

$$var_j(S_j|e, \hat{X}) = (\overline{p}^j - p_0^j)^2 \cdot P(\hat{X} = 0|e) + (\overline{p}^j - p_1^j)^2 \cdot P(\hat{X} = 1|e) \ . \tag{6}$$

This value has to be computed for each skill in the model. Afterwards, we compute the average of these values for the question $\hat{X}$:

$$var(\mathcal{S}|e, \hat{X}) = \frac{1}{m} \sum_{j \in \boldsymbol{M}} var_j(S_j|e, \hat{X}) \ . \tag{7}$$

We select a question which has the highest average value computed from (7):

$$X^* \quad = \quad \arg\max_{\hat{X} \in \hat{\mathcal{X}}} var(\mathcal{S}|e, \hat{X}) \ . \tag{8}$$

Maximization of (6) can be viewed as a generalization of the criterion of student separation described in our previous article [7]. The difference is that in this case we consider the probability of $S_j = 1$ after answering $\hat{X}$ instead of the most probable state of S_j.

4.3 Maximization of the Expected Question Variance

Previous two criteria aimed at skills directly. This third one aims at questions instead. From all unanswered questions we want to find a question with the highest expected variance of correct answer probabilities for all possible state combinations. This criterion is motivated as follows: if the question's correct answer probability varies a lot with changing skill states it means that this question is significantly affected when student skills shifts. It follows from the Bayes rule that this question also has a significant influence on the skills. This method is further referred to as Questions' Variance.

The expected variance of the question's $\hat{X}$ correct answer probability is computed given the following formula:

$$var(\hat{X}|e) = \sum_{\boldsymbol{s} \in Val(\boldsymbol{S})} (P(\hat{X} = 1|e) - P(\hat{X} = 1|\boldsymbol{s}, e))^2 \cdot P(\boldsymbol{s}|e) \ . \tag{9}$$

A question with the highest value of expected variance given by (9) is selected to be asked next. The use of this function for computations during testing is impractical because of its computational

complexity as it would take long time to select the next question. We propose an approximation of Formula (9). We compute the variance for a single skill node and then take into account their combined average instead of the full computation over all states' combinations.

We establish following notation:

$$r_0^j = P(\hat{X} = 1|S_j = 0) \ ,$$
$$r_1^j = P(\hat{X} = 1|S_j = 1) \ ,$$

where $S_j \in \mathcal{S}$, $\hat{X} \in \hat{\mathcal{X}}$. r_0^j stands for the probability, that the student answers correctly to the question even though he/she has no skill in question. r_1^j is the same situation while the student has all examined skills. Intuitively, the value r_1^j has to be larger than r_0^j.

With the average value

$$\overline{r}^j = P(S_j = 0|e) \cdot r_0^j + P(S_j = 1|e) \cdot r_1^j$$

we can compute the expected variance of correct answer probability for the question $\hat{X}$ using the next formula:

$$\begin{aligned} var_j(\hat{X}|e) &= (\overline{r}^j - r_0^j)^2 \cdot P(S_j = 0|e) + (\overline{r}^j - r_1^j)^2 \cdot P(S_j = 1|e) \ , \\ var(\hat{X}|e) &= \frac{1}{m} \sum_{j \in \boldsymbol{M}} var_j(\hat{X}|e) \ . \end{aligned} \tag{10}$$

A question X^* we select is maximizing this variance.

$$X^* \quad = \quad \arg\max_{\hat{X} \in \hat{\mathcal{X}}} var(\mathcal{S}|e, \hat{X}) \ . \tag{11}$$

The value $(\overline{r}^j - r_s^i)$ can be viewed as differential of $P(\hat{X} = 1|S_j = s)$ of skill variables S_j that have only two states $s \in \{0, 1\}$. Therefore, if P is the probability density function of the continuous skill variable S_j, we can view it as a finite equivalent of the probability $P(\hat{X} = 1|S_j = s)$ derivative with respect to s. It means that $var(\hat{X}|e)$ is similar to Fisher's information which is a commonly used criterion for IRT (Item Response Theory) [11] – another possible type of model for CAT. More detailed explanation of this criterion in the case of continuous skill variables can be found in [7].

5 Experiments

5.1 Experimental Setup

To evaluate models we have done experiments on both data sets with models MM and CM described above. We have used 10 fold cross-validation method for both data sets. Models were first learned using standard EM algorithm from learning data. Next, we performed a simulation of CAT test for every model and for every student using testing data.

During simulated testing we first estimated the skills of a student based on his/her answers. At the start of each step we compute marginal probability distributions for all skills $\boldsymbol{S}$. This happens before selecting a new question and updating the model with the new answer. We use evidence e obtained in previous steps which is at the start of testing empty. Then, based on estimated skills we predict answers to all questions, where we select the most probable state of each question $X \in \mathcal{X}$:

$$x^* = \arg\max_x P(X = x|\boldsymbol{S}) \ . \tag{12}$$

By comparing this value to the real answer x' of the question X we obtain a success rate of the response estimates for all questions $X \in \mathcal{X}$ of a test result t (particular student's result) in one step

$$\mathrm{SR}^t \quad = \quad \frac{\sum_{X \in \mathcal{X}} I(x^* = x')}{|\mathcal{X}|} \ , \text{ where} \tag{13}$$

$$I(expr) \quad = \quad \begin{cases} 1 & \text{if } expr \text{ is true} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The total success rate of one model in one step for all test data is defined as

$$\mathrm{SR} \quad = \quad \frac{\sum_{t=1}^{D} \mathrm{SR}^t}{D} \ , \tag{15}$$

where D is the dataset size.

5.2 Experimental Results

Average results of simulated tests for both data sets, i.e., both models described above, are displayed in graphs 3 and 4 for each model separately. Graphs show success rates SR in the first 30 steps. Step 0 is the state before asking any questions. At this point the prediction is based only on data itself. There is no evidence and the selection criterion adds no benefit. Therefore the SR is the same over all cases for a single model. For comparison we include also the sequential selection as described in Section 4.

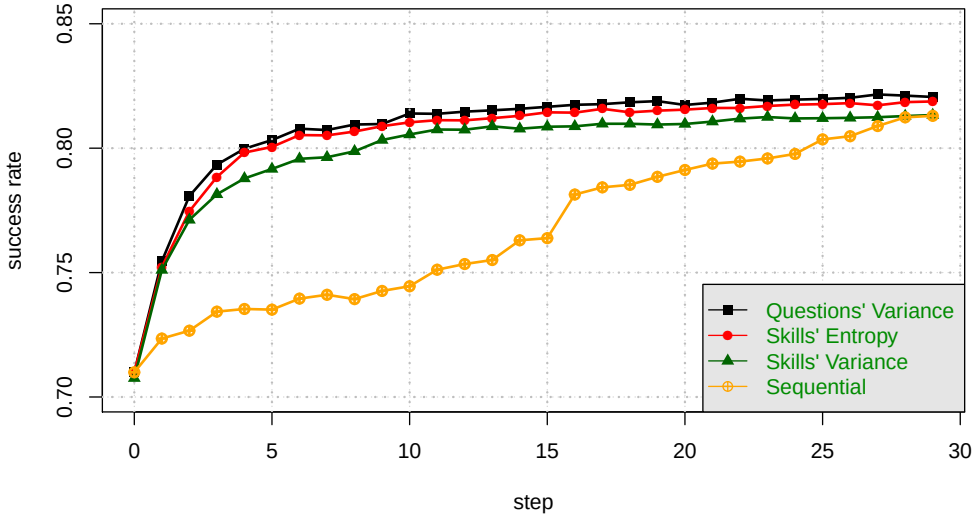

Figure 3: MM success rates for first 30 questions of simulated testing

As we can see in these graphs the worst performing method is the sequential selection. It is apparent that the rate in which this method improves its estimates is lower than the rate of the remaining methods. This is caused by the fact that it is selecting questions which are not most informative in the current test situation for the tested student.

The Skills' Variance method of the selection has the lowest performance (or same) from three proposed methods in both models. As explained below it is the only method which has statistically significantly worse results in one instance. Particular reasons for this behavior has to be explored further as there might be many possible causes.

Questions selected by individual methods are not the same even though the success rate of question estimates is very similar. The selected questions are displayed in the Tables 3 and 4. Numbers displayed correspond to the total number of selections of the particular question in the MM and the CM in the first five steps of simulated testing. Questions which were not selected at all are not included in the tables. By inspecting these tables we can easily see that there are differences in individual methods. Some questions were not selected at all by one method while the other two methods selected them in some cases only. For example, in the MM the question X42 was not selected by Skills' Entropy while Questions' Variance selected it in 112 cases in the first five steps. Nevertheless, we can see a trend of good questions which are selected very often and soon in the process of testing by all methods. For example, in the MM the question X43 was

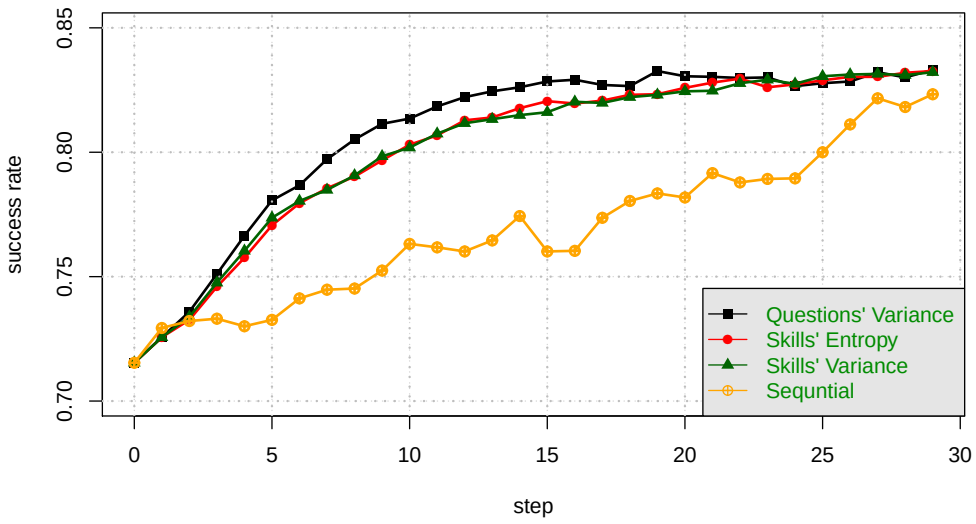

Figure 4: CM success rates for first 30 questions of simulated testing

selected for all students by Skills' Entropy and only for 7/10 of the dataset by the two remaining methods.

5.3 Wilcoxon tests

To confirm our conclusions described above we used the Wilcoxon signed-rank test. Tables 1 and 2 contain p-values obtained from Wilcoxon tests to compare the success rates of two criteria. An alternative hypothesis is that the overall success rate of the i-th criterion (row index) is greater than the overall success rate of the j-th criterion (column index).

Table 1: MM Wilcoxon tests p-values

	sequential	**Skills' Entropy**	**Skills' Variance**	**Questions' Variance**
sequential	-	1	1	1
Skills' Entropy	$1.17 \cdot 10^{-5}$	-	$1.62 \cdot 10^{-1}$	$9.39 \cdot 10^{-1}$
Skills' Variance	$2.20 \cdot 10^{-4}$	$8.40 \cdot 10^{-1}$	-	$9.99 \cdot 10^{-1}$
Questions' Variance	$2.04 \cdot 10^{-8}$	$6.15 \cdot 10^{-2}$	$9.22 \cdot 10^{-3}$	-

Table 2: CM Wilcoxon tests p-values

	sequential	**Skills' Entropy**	**Skills' Variance**	**Questions' Variance**
sequential	-	1	1	1
Skills' Entropy	$1.10 \cdot 10^{-4}$	-	$5.14 \cdot 10^{-1}$	$8.76 \cdot 10^{-1}$
Skills' Variance	$8.77 \cdot 10^{-5}$	$4.92 \cdot 10^{-1}$	-	$8.58 \cdot 10^{-1}$
Questions' Variance	$7.72 \cdot 10^{-6}$	$1.27 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$	-

As we can see in both cases, p-values of the sequential selection compared to all other criteria are much smaller than the borderline of $\alpha = 0.05$. This confirms the fact that the sequential selection provides the worst results. The table of the MM also shows that the success rate of Questions' Variance method is greater than the SR of Skills' Variance method (p-value $= 9.22 \cdot 10^{-3}$). This

shows there is statistically important improvement in success rates of the former over the latter method. All other pairs of different selection criteria show statistically insignificant difference within the selected confidence interval. Therefore, for the remaining pairs we can not establish any statistically sound order.

6 Conclusions and Future Work

This article considered different ways of selecting questions during the procedure of Computerized Adaptive Testing. We presented three different types of methods to select questions during CAT which were afterwards tested. For testing we used two data sets collected for this purpose.

The first important empirical observation is that the question selection method has a significant impact on the quality of predictions during the CAT procedure. In the comparisons all three proposed methods clearly outperformed the sequential selection. The motivation to study these methods is thus valid.

The next observation is that three proposed methods behave differently. In this case the difference in the quality of prediction is not large, it is statistically insignificant, but the methods are distinguishable since they select different questions. The first step in the future research is to provide generalizations of these methods to support multi-state skill variables. It seems that especially for Skills' Variance it may be very beneficial to test a model with skill nodes having more than two states. It is necessary to show if we can improve these methods and establish any ordering between them which would be valid generally over different models.

Acknowledgement
This work was supported by the Czech Science Foundation (project No. 16-12010S) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/198/OHK4/3T/14.

Table 3: The frequency of questions X1-X61 selections during simulated testing using criterion (a) Skills' Entropy, (b) Skills' Variance, (c) Questions' Variance for the MM model. Questions which were never selected are not included.

	X3	X5	X7	X10	X11	X12	X13	X19	X25	X26	X27	X29	X30	X32	X33	X38	X39	X41	X42	X43	X44	X50	X51	X61
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	281	-	-	-	-
2	-	-	-	-	-	-	-	118	-	-	-	-	-	149	-	-	-	14	-	-	-	-	-	-
3	-	62	10	5	-	-	-	70	-	8	-	-	-	-	-	-	-	38	-	-	-	-	68	20
4	-	9	-	11	-	1	8	-	-	23	7	25	18	-	28	-	-	113	-	-	-	16	2	20
5	-	24	-	13	13	14	-	9	3	1	9	14	49	-	19	-	-	33	-	-	5	38	10	27

	X3	X5	X7	X10	X11	X12	X13	X19	X25	X26	X27	X29	X30	X32	X33	X38	X39	X41	X42	X43	X44	X50	X51	X61
1	-	-	-	-	-	-	-	-	-	-	-	-	-	56	-	-	-	-	-	197	28	-	-	-
2	-	-	-	-	-	-	-	49	-	-	-	-	-	105	-	-	-	108	-	10	9	-	-	-
3	10	53	-	-	-	-	-	48	-	-	-	-	-	14	38	-	-	35	9	15	-	-	45	14
4	4	6	5	7	-	1	-	16	-	13	-	-	-	-	11	3	-	80	35	13	9	49	18	11
5	9	8	2	1	-	-	8	32	-	24	1	14	9	-	13	11	-	32	20	2	7	40	20	28

	X3	X5	X7	X10	X11	X12	X13	X19	X25	X26	X27	X29	X30	X32	X33	X38	X39	X41	X42	X43	X44	X50	X51	X61
1	-	-	-	-	-	-	-	-	-	-	-	-	-	84	-	-	-	-	-	197	-	-	-	-
2	-	-	-	-	-	-	-	113	-	-	-	-	-	15	-	-	-	106	-	-	47	-	-	-
3	-	16	3	-	-	-	-	22	-	-	-	-	-	21	4	-	-	61	66	25	-	-	49	14
4	6	59	7	2	-	1	-	-	-	-	-	-	2	5	17	16	-	18	36	12	-	47	39	14
5	12	14	-	10	-	1	-	12	7	17	1	2	15	25	10	16	5	18	10	-	-	70	10	26

Table 4: The frequency of questions X2-X49 selections during simulated testing using criterion (a) Skills' Entropy, (b) Skills' Variance, (c) Xuestions' Variance for the CM model. Xuestions which were never selected are not included.

	X2	X3	X4	X6	X7	X13	X16	X18	X21	X22	X26	X28	X35	X36	X37	X39	X43	X45	X47	X48	X49
1	-	-	-	-	-	-	-	-	-	-	-	-	143	-	-	-	-	-	-	-	-
2	-	14	-	-	-	-	86	-	-	43	-	-	-	-	-	-	-	-	-	-	-
3	19	4	-	13	-	4	25	-	-	15	-	-	-	-	-	-	3	-	-	60	-
4	12	15	8	14	8	12	9	-	-	12	6	-	-	-	1	6	5	-	-	11	24
5	26	22	1	13	2	13	10	-	5	7	10	-	-	4	-	20	1	2	-	7	-

	X2	X3	X4	X6	X7	X13	X16	X18	X21	X22	X26	X28	X35	X36	X37	X39	X43	X45	X47	X48	X49
1	-	-	-	-	-	-	-	-	-	-	-	-	143	-	-	-	-	-	-	-	-
2	-	14	-	-	-	-	77	-	-	52	-	-	-	-	-	-	-	-	-	-	-
3	14	4	-	18	-	-	29	-	-	6	-	-	-	-	-	-	3	-	-	69	-
4	4	15	7	19	-	8	21	-	-	8	6	-	-	-	-	4	6	-	1	17	27
5	15	33	3	31	-	5	9	-	2	14	9	2	-	-	-	-	2	2	1	11	4

	X2	X3	X4	X6	X7	X13	X16	X18	X21	X22	X26	X28	X35	X36	X37	X39	X43	X45	X47	X48	X49
1	-	-	-	-	-	-	-	-	-	-	-	-	143	-	-	-	-	-	-	-	-
2	23	18	-	-	-	-	29	-	-	58	-	-	-	-	-	-	-	-	-	15	-
3	16	25	-	15	-	-	13	-	-	-	6	-	-	-	-	-	3	-	-	65	-
4	20	29	-	18	3	-	13	-	-	33	6	-	-	1	-	4	1	2	-	13	-
5	10	29	2	11	6	-	24	3	-	9	5	-	-	6	-	20	8	1	-	3	6

References

[1] R. G. Almond and R. J. Mislevy. Graphical Models and Computerized Adaptive Testing. *Applied Psychological Measurement*, 23(3):223–237, 1999.

[2] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

[3] M. J. Culbertson. *Graphical Models for Student Knowledge: Networks, Parameters, and Item Selection*. PhD thesis, University of Illinois at Urbana, 2014.

[4] K. C. Moe and M. F. Johnson. Participants' Reactions To Computerized Testing. *Journal of Educational Computing Research*, 4(1):79–86, jan 1988.

[5] T. D. Nielsen and F. V. Jensen. *Bayesian Networks and Decision Graphs (Information Science and Statistics)*. Springer, 2007.

[6] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., dec 1988.

[7] M. Plajner and J. Vomlel. Student Skill Models in Adaptive Testing. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 403–414. JMLR.org, 2016.

[8] S. Tonidandel, M. A. Quiñones, and A. A. Adams. Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *The Journal of applied psychology*, 87(2):320–32, apr 2002.

[9] W. J. van der Linden and C. A. W. Glas. *Computerized Adaptive Testing: Theory and Practice*, volume 13. Kluwer Academic Publishers, 2000.

[10] W. J. van der Linden and C. A. W. Glas, editors. *Elements of Adaptive Testing*. Springer New York, NY, 2010.

[11] W. J. van der Linden and R. K. Hambleton. *Handbook of Modern Item Response Theory*. Springer New York, 2013.

[12] H. Wainer and N. J. Dorans. *Computerized Adaptive Testing: A Primer*. Routledge, 2015.