Multidimensional Computerized Adaptive Testing Based on Bayesian Theory

Abstract - **Effective and efficient assessment of a learner's proficiency has always been a high priority for intelligent e-Learning environments. The fields of psychometrics and Computer Adaptive Testing (CAT) provide a strong theoretical and practical basis for performing skills assessment, of which Item Response Theory (IRT) is the best recognized approach. For assessing multiple skills at once, which is called for in e-learning environments because they rely on fine-grained skill models to know the knowledge state of learners and provide them appropriate study path, multidimensional IRT (MIRT) is a necessity and emerged as a candidate. However, MIRT is computationally expensive. A simpler multidimensional model, based on classical Bayesian theory, is proposed. It is an extension of Rudner's work on unidimensional Bayesian decision theory. The explanation of theory basis is based on binary classification (master/non-master) test. Its evaluation is performed through simulations with pseudo-random data samples comprised of 6 skill dimensions. We firstly show that the model can take advantage of multidimensional test items to accelerate assessment and that it performs better than the uni-dimensional version. Secondly, we compare its classification accuracy rate with its unidimensional version and MIRT approach. The results show that its performance is much better than existing unidimensional model, and at least as good as MIRT, in spite of the fact that its complexity and computational burden is lighter than MIRT.**

Index Terms – Assessment, e-Learning, Bayesian theory, computerized adaptive testing (CAT), multidimensional, item response theory (IRT).

INTRODUCTION

Compared with traditional classroom instruction, e-learning offers a virtual one-on-one tutoring environment by tailoring the learning experience to the characteristics of different learner. This can represent a significant advantage. Bloom reported that one-on-one instruction helps average students to perform as well as the top 2 percent of students receiving classroom instruction [3]. E-learning's emergence provides an opportunity to realize the goal of large scale personalized instruction at reasonable cost. With a personalized study plan designed by an e-learning system, we can actively engage the learner with a teaching strategy and material that appeals to the learner's knowledge, style of learning, etc [4]. Effective and efficient assessment of a learner's proficiency is a necessary requirement to achieve the desired adaptivity.

The fields of psychometrics and Computer Adaptive Testing (CAT) [1,6, 8,13,14,15]provide us with a large body of theory that can be most helpful for this purpose. Since its birth three decades ago, CAT has been implemented for a number of large scale tests, such as GRE, TOEFL, and GMAT. CAT is attracting attention in the e-Learning community because it offers a flexible and efficient assessment technique that can yield high accuracy. It represents an attractive mean to build a precise picture of a user's knowledge state. During the learning process, an automatic, quick and accurate assessment feedback would enable the e-Learning system to adjust its teaching accordingly.

A number of e-Learning systems that rely on an assessment module to tailor user feedback are already in usage, such as Microsoft E-Learning [12], APeLS [2,4,5] ALEKS [11]. The assessment module is at the heart of the ALEKS. With the quick and accurate evaluation of student's knowledge state provided by the assessment module, ALEKS is able to readily determine the most relevant material, and thus efficiently guide the individual's learning path. In ALEKS, learning is powered and optimized by student assessment. In fact, competence-based personalized learning is becoming an important approach to create adaptive learning paths [9,10].

Student assessment is the central focus of CAT. For CAT, IRT [8] has always been the prevalent approach since its beginning. It represents the golden standard in this field, and can assign an individual on a continuous scale of proficiency.

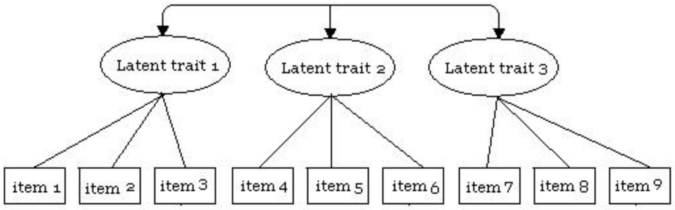
However, it suffers from a relatively high degree of complexity and computational cost. Moreover, in many cases, we are only interested in classifying examinees into a finite number of discrete categories, such as *master/non-master*, or *excellent/good/fair/fail*. For this coarser outcome, a discrete classification scheme can suffice. In fact, what we are often looking for is a discrete classification over a number of individual skills. The Bayesian theory based CAT approach (BT-CAT) is proposed as a candidate [6,13,14].

Rudner showed that a high classification accuracy rate can be obtained with a Decision theoretic framework with only a few test items administered to a student and that a small dataset was required for calibration. However, Rudner's model is limited to a single dimension, which means that only one skill can be measured through a single exam. To meet the typical e-learning requirement of assessing a number of skills, an upgraded multidimensional model based on the decision theory is described in this article, and we call it Bayesian decision theory based multidimensional CAT model (BT-MCAT).

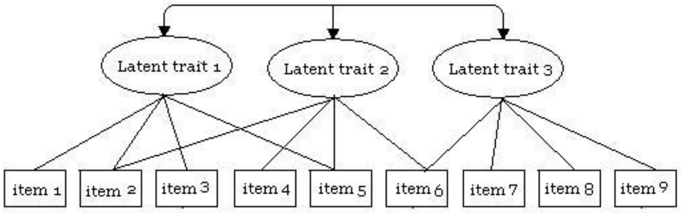
MULTIDIMENSIONAL CAT BASED ON BAYESIAN DECISION THEORY

With the basic introduction of e-learning and assessment in the previous section, we, from this section, will turn directly to the discussion of our work – a multidimensional CAT model based on Bayesian theory. In this section, we focus on introducing the theory basis of this multidimensional model, and for those readers who are interested in its unidimensional version, please refer Rudner`s article [13,14].

Test multidimensionality refers to the fact that the success on a test depends on multiple skills, as opposed to a single skill in a unidimensional test [15]. The concept of dimensionality also applies to items. There are two kinds of item multi-dimensionality: *between-item* and *within-item*, which is determined by the number of dimensions linked to a single item, as shown in Figure 1. A test solely composed of unidimensional items covering many latent traits is called between-item multidimensional. A test is called within-item multidimensional if it contains multidimensional items that can be used to measure several latent trait simultaneously. In our project, we only study the within-item model.



(Between-Item)



(Within-Item)

Figure 1 Two kinds of multidimensionality

For simplicity, only two categories are assumed for each latent trait (skill), *master* and *non-master*. However, the model can readily be expanded to multinomial classification. Note that in the decision theory framework, we could assign different "costs" values for deciding whether a student is *master* or not and to each type of mistake we could make. Again, for simplicity, we assign the same cost to all, but bear in mind that the approach is readily extensible to account for different cost structures.

We describe the basis of the BT-MCAT model below and cover the three standard elements of a CAT model: calibration, assessment updating procedure, and item selection rule. Many CAT system also include a stopping rule but, because we want to observe the trend of performance as items are administered, this is not applied in our case.

Calibration

In our discussion, we assume there are K dimensions of traits (i.e. K skills), so the corresponding ability vector is θ = [θ^1 , θ^2 , ..., θ^k , ..., θ^K]. Considering there are two categories, *master* and *non-master*, in each dimension, we use θ_m^k and $\theta_{\overline{m}}^k$ to refer *master* (m) and *non-master* ($\overline{m}$) ability level of the skill k respectively. Each item is assigned one or more dimensions. Let us use $c(i,k)=1$ to indicate that item i can be used to measure the dimension k ; otherwise $c(i,k)=0$. A correct response of item i is indicated as X_i , otherwise $\overline{X}_i$. Although multivariate analysis and factor analysis can be applied to discover the links between items and dimensions, we assume that the dimensionality of each item is predefined explicitly beforehand.

Given each item's respective dimensions, two parameters need to be derived from data, the *a priori* probability that the learner is a *master* and the conditional probability of a correct response given that the learner is a *master* for category *k*. Estimation procedures for each are described below.

- **Estimating** $P(\theta_m^k)$ **and** $P(\theta_{\overline{m}}^k)$ **, the *a priori* probabilities of the proportions of *master* and *non-master* category in each dimension** $k \in [1..K]$ **in a sample from the population.** With sample response data of a group of examinees available, a passing score, which is calculated simply as the sum of total correct responses, is set first to determine with what raw score one examinee can be classified as *master* or not. The knowledge state of all the examinees can be known by this way; then, the corresponding amount of master and non-master students in the sample population can be calculated. The proportion of master and non-master is easily determined, which is regarded as the estimated proportion of After the passing score is set for each dimension, each examinee's category can be determined based on whether or not his or her raw score is higher than the passing score. In our experiment, the same passing score is set for each dimension to simply the discussion.

- **Estimating** $P(X_i \mid \theta_m^k)$ **and** $P(X_i \mid \theta_{\overline{m}}^k)$ **, the conditional probability of correct response on each item given known category, *master* or *non-master,* in each dimension.** In the previous step, not only the state of each sample student is known, *master* or *non-master*, but the amount of master and non-master students in the sample population. Furthermore, for each item, we can calculate the amount of students who are master and response correctly. The ratio of this outcome to the total master student is the estimate of $P(X_i \mid \theta_m^k)$ **.** $P(X_i \mid \theta_{\overline{m}}^k)$ can be determined in a similar way**.** Because only dichotomous items are used in the test, the probability of incorrect response given specific category $P(\overline{X}_i \mid \theta^k)$ is compensatory to $P(X_i \mid \theta^k)$**.**

Probability update

Given a calibrated item bank, the adaptive testing procedure starts with each dimension set to their initial probabilities; then the first item is selected and administered to the examinee. The response is observed and the probabilities of all dimensions are updated accordingly. This computation is a typical Bayesian inference procedure:

$$P_i(\theta_m^k \mid X_i) = \begin{cases} \dfrac{P(X_i \mid \theta_m^k)P_{i-1}(\theta_m^k)}{P(X)}, if & c(j,k) = 1 \\ P_{i-1}(\theta_m^k), & if \quad c(j,k) = 0 \end{cases}$$

Equation 1

where

$$P(X_i) = P(X_i \mid \theta_m^k)P_{j-1}(\theta_m^k) + P(X_i \mid \theta_{\overline{m}}^k)P_{i-1}(\theta_{\overline{m}}^k)$$

Equation 2

Note that for the first item, $P_{i-1}(\theta_m^k)$ is $P(\theta_m^k)$, the a priori probability calculated in the calibration step**.** In (Equation 1), only the probability of the dimension(s) linked with the currently active item will be updated; the remaining dimensions are kept constant. The outcome of Equation 1, $P_i(\theta_m^k \mid X_i)$, serves as the new value for $P_i(\theta_m^k)$ in the next iteration when another item is chosen. The update given wrong response $\overline{X_i}$ is similar. This process is repeated until a decision is made or, for the purpose our simulations, after a fixed number of items are administered.

Item selection rule

The selection rule plays a critical role in an adaptive testing model. Through the optimal choice of item sequence, test length can be greatly reduced while maintaining a high classification accuracy level.

An optimal selection rule in an MCAT model will take all dimensions into consideration. In this study, the Maximum Information Gain is applied to select the item, resulting in a maximum reduction of entropy as the next one. Entropy is calculated over all dimensions as shown in Equation 3.

$$H_i = \sum_{k=1}^{K} [-P_i(\theta_m^k)\log(P_i(\theta_m^k) - P_i(\theta_{\overline{m}}^k)\log(P_i(\theta_{\overline{m}}^k)]$$

Equation 3

A search made over each of the non administered item to find the most informative one. The expected entropy of each item value is calculated given correct and wrong responses by Equation 4.

$$\begin{aligned} E(H_j) = \\ \sum_{k=1}^{K} \{ & [-P(\theta_m^k \mid X_j)\log(P(\theta_m^k \mid X_j) - P(\theta_{\overline{m}}^k \mid X_j)\log(P(\theta_{\overline{m}}^k \mid X_j)]P(X_i) \\ & + [-P(\theta_m^k \mid \overline{X}_j)\log(P(\theta_m^k \mid \overline{X}_j) - P(\theta_{\overline{m}}^k \mid \overline{X}_j)\log(P(\theta_{\overline{m}}^k \mid \overline{X}_j)]P(\overline{X}_j) \} \end{aligned}$$

Equation 4

The item that brings global maximum entropy reduction will be chosen as the next one

$$\max_j (H_i - E(H_j))$$

Equation 5

Note that multidimensional items are more likely to be chosen than unidimensional ones since they bring more information.

SIMULATION DESIGN AND RESULTS

The validation of the proposed approach is carried through a simulation process. Samples are generated with predefined proficiency level values, θ , using Monte Carlo simulation. This procedure is standard for the validation of multidimensional CAT frameworks.

Three simulations are designed to make a comparison between 1) BT-MCAT with adaptive selection rule and a random selection rule; 2) BT-MCAT vs. BT-UCAT; 3) BT-MCAT vs. IRT-MCAT. With these three groups of comparison, we will have chance to study the performance of this newly developed model relative to traditional exam, unidimensional CAT, and multidimensional CAT based IRT respectively.

Monte Carlo simulation description

Our experiments will replicate Wang & Chen's [15] within-item multidimensional IRT (MIRT) experiments, which will allow us to compare our results based on BT-MCAT to theirs.

Six different latent traits serve as our dimensions, and nine item banks are generated to cover these 6 dimensions. Table 1 contains a description of the item banks, namely the corresponding number of item in each bank and the dimension(s) each bank covers.

Table 1 Design of Wang and Cheng's test item banks

		Dimension (D)					
Test (T)	# of item	1	2	3	4	5	6
1	160	x					
2	160		x				
3	160			x			
4	20	x	x				
5	20	x		x			
6	20		x	x			
7	200				x		
8	200					x	
9	200						x

Among those nine banks, T1/2/3/7/8/9 (we use this shorthand notation to refer here to test item banks {1,2,3,7,8,9}) are unidimensional, and they respectively correspond to a single dimension D1/2/3/4/5/6. However, T4/5/6 (shaded rows) are two dimensional ones with only 20 items in each one. For example, T4 is for D1 and D2. As Wang & Chen explained, the test design such that there are fewer multidimensional items than unidimensinoal ones and is justified on the basis of its similarity to the Basic Competence Test for junior high school students in Taiwan [15]. The total number of items in Table 1 is 1040, with a fixed 200 for each dimension.

To simulate these pseudo-random items, the examinees, and their responses with desired structure, a series of steps are required. Firstly, we generate a number of examinees from multivariate normal distribution with adaptivity levels corresponding to the pre-defined θ's mean and standard deviation in each dimension. Next, items with known parameters, such as discrimination a and difficulty b , are produced. Finally, we simulate the examinees' response to these items by using a three-parameter multidimensional ICC (Item Characteristic Curve) formula [8,15]. The ICC formula provides the probability of correct response on each item for each examinee based on the parameters determined in the first two steps. In our simulation, 1000 examinees are drawn randomly from the multivariate normal distribution with mean $\theta^T = [\ 0.2\ ,\ 0.0\ ,\ -0.2\ ,\ -0.1\ ,\ 0.1\ ,\ 0.0\ ]$, and standard deviation of 1.0. Moreover, skill dimensions are correlated as defined by the Σ matrix. We see that D1 to D3 are moderately to highly correlated with value 0.8 and 0.7; D4 and D5 are moderately correlated; D6 is negatively correlated with other latent traits. Recall that these values are replicated from Wang & Chen [15] and that they reflect realistic conditions. A total of 1000 test response samples are generated. Among them, 600 samples are selected randomly for parameter calibration and the remaining 400 for validation purpose.

$$\Sigma = \begin{bmatrix} 1.0 & .8 & .8 & .3 & .3 & -.4 \\ .8 & 1.0 & .7 & .2 & .2 & -.3 \\ .8 & .7 & 1.0 & .1 & .2 & -.2 \\ .3 & .2 & .1 & 1.0 & .7 & -.2 \\ .3 & .2 & .2 & .7 & 1.0 & -.2 \\ -.4 & -.3 & -.2 & -.2 & -.2 & 1.0 \end{bmatrix}$$

This data is used for the simulations reported in the following three sections.

BT-MCAT with adaptive selection rule vs. RA

In this first experiment, we look at the ability of the BT-MCAT framework to classify examinees according to their predefined skill mastery and over each of the six dimensions. We compare the performance of the dimensions composed of within-item vs. between-item questions. We also compare the approach using the entropy driven item selection and a random item selection.

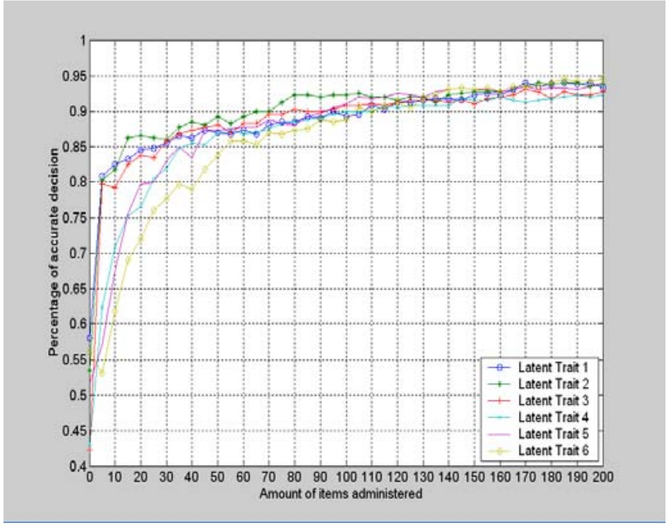


Figure 2 Accuracy rate of decision for all 6 dimensions in BT-MCAT with adaptive selection rule

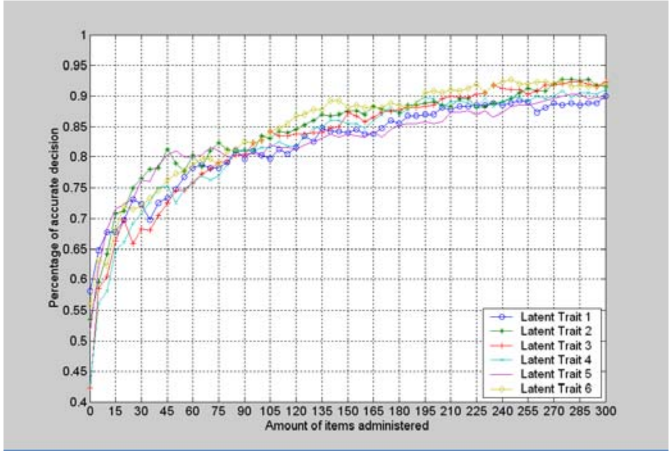


Figure 3 Accuracy rate of decision for all 6 dimensions in BT-MCAT with RA rule.

The results are reported in Figure 2. The X-axis represents the amount of items administered and the Y-axis is the percentage of correct decision. We notice that D1/2/3 have a significantly better performance compared with D4/5/6 at the beginning when less than 40 items are used. This confirms that the multidimensional items, D1/2/3, have a higher potential of entropy reduction.

Figure 3 reports the results for the RA selection rule condition. The accuracy rate evolves at about the same rate for all six dimensions, contrary to the results of Figure 2 where multidimensional items showed greater accuracy than unidimensional items. These results indicate the importance of the selection rule in order to take advantage of the multidimensional items.

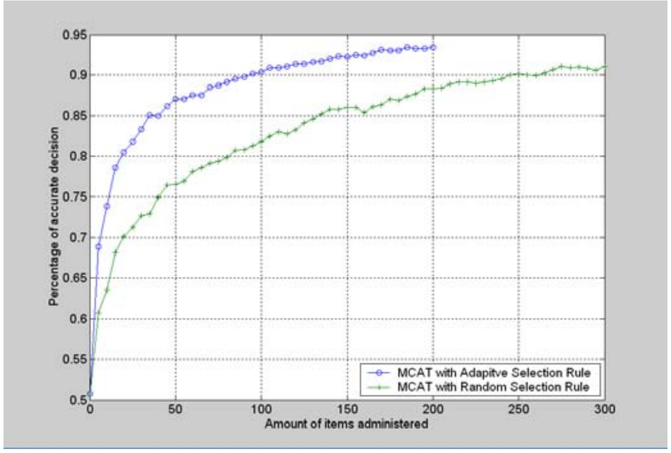


Figure 4 The comparison of the average performance over all 6 dimensions for MCAT with adaptive selection and RA rule.

Figure 4 shows the average performance over all six dimensions for each condition, adaptive vs. random. A clear difference emerges. MCAT with adaptive selection rule reaches to about 94% after 200 items whereas the random condition reaches only 91% after 300 items.

BT-MCAT vs. BT-UCAT

The previous section confirms the expected ability of the item selection strategy to take advantage of multidimensional items. In this section, we show that this advantage is reflected by the ability of the multidimensional approach, BT-MCAT, to assess the learner's skills with fewer items than the unidimensional approach developed by Rudner [13,14].

Since multidimensional items are not appropriate for unidimensional model, we take out T4/5/6 from Table 1 and add 40 additional items to T1/2/3 respectively so that each dimension still maintain equivalent item size, 200. This scheme allows us to have comparable item banks for all dimensions. The new item bank design for BT-UCAT is shown in Table 2:

Table 2 Design of item bank for BT-UCAT simulation

		Dimension (D)					
Test (T)	# of item	1	2	3	4	5	6
1	200	x					
2	200		x				
3	200			x			
7	200				x		
8	200					x	
9	200						x

The performance of each dimension for BT-UCAT is shown in Figure 5, in which we see that the accuracy all of the six curves evolve at a similar rate, unlike BT-MCAT in which dimensions with multidimensional items designed reach higher accuracy level. The average performance comparison is made in Figure 6. In the graph, we notice that it requires BT-MCAT about 100 items to reach 90%, whereas the corresponding percentage is around 180 for BT-UCAT. Obviously, BT-MCAT needs fewer items to reach the same performance level than UCAT.

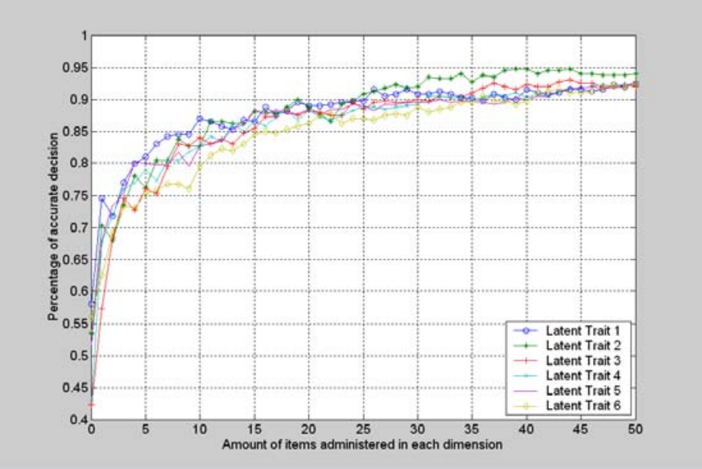


Figure 5 Accuracy rate of decision for all 6 dimensions in BT-UCAT with adaptive selection rule.

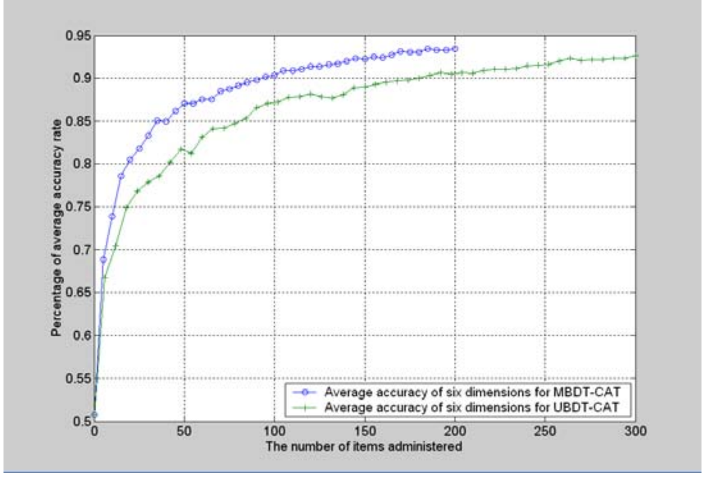


Figure 6 The comparison of the average performance over all 6 dimensions for BT-MCAT and -UCAT.

BT-MCAT vs. IRT-MCAT

We have shown the ability of the BT-MCAT model to take advantage of multidimensional items for assessing a set of six skills. That advantage was shown to be reflected in comparison with the unidimensional BT-UCAT model. We now turn to the comparison of the BT-MCAT model with the *de facto* standard in the CAT field, the multidimensional IRT model, MIRT. As mentioned before, we simulate the design of the experiment by [15], which allows us to compare our results directly with theirs.

The following graph is reproduced based on Wang & Chen's result [15]. The X-axis value is the number of items administered, and the Y-axis indicates the degree of test reliability, r . Test reliability is defined as the square correlation between the true θ and the estimated latent trait $\hat{\theta}$ in their article. A high r value means the observed scores are highly correlated with its true scores, which indicates that the corresponding test is reliable.

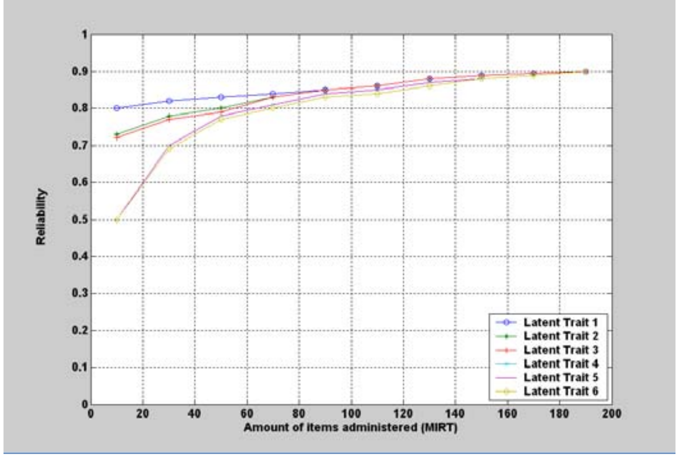


Figure 7 Wang & Chen's MIRT simulation results of MIRT-CAT about reliability performance.

Test reliability relies on a continuous scale metric for θ, which is consistent with the IRT framework. However, BT uses discrete categories instead of a continuous θ and a classification accuracy rate instead of the reliability measure. A transformation from the reliability measure to an accuracy rate is thus required to make the comparison between our results and Wang and Cheng' s results. The transformation is based on a Monte Carlo simulation. We generate samples using the Test reliability parameters and measure the corresponding accuracy rate. The details of the procedure are described below:

1. Generate a group of examinee with predefined ability value $\theta_{jk}(j=1..1000, k=1..6)$, and their pseudo random responses. In fact, this step has been done before;
2. Generate the "estimated skill value" of those examinee samples mentioned in step (1), $\hat{\theta}(j=1..1000, k=1..6)$, based on the reliability value r obtained by Wang and Chen in their MIRT simulation. The following known formula is applied here to generate a new variable $\hat{\theta}$ based on a known variables θ, and their desired correlation value is r:

$$\widehat{\theta} = \theta\sqrt{r} + v\sqrt{1-r}$$

In this formula, v is a parameter with a normal distribution, and its vector size is the same as θ.

3. Generate the pseudo random responses according to the "estimated skill value" $\widehat{\theta}$ as we did in step (1). The outcome of this step will be response vectors of 1000 examinees in our simulation, and the category of each examinee, *master* or *non-master*, can be determined with predefined passing score.
4. An accurate decision is reached if the category of examinee with $\widehat{\theta}$ is the same as that with θ, and classification rate can be determined as well.

By applying this transformation procedure, each r value in Figure 7 can be mapped to a corresponding classification rate value, and the result is shown in Figure 8. Figure 9 It shows the comparative performance of the BT-MCAT and MIRT models. We see that when the 190 items are administered in both models, BT-MCAT reaches to around 94%, but MIRT only 88%.

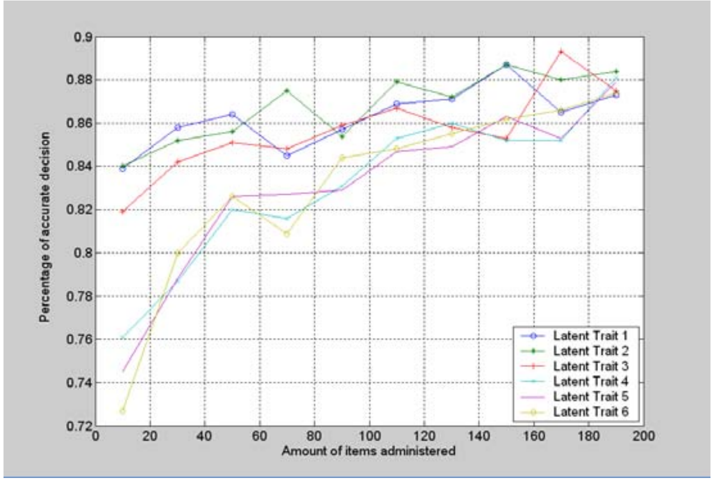


Figure 8 The corresponding accuracy rate for MIRT-CAT after a transformation from the reliability index.

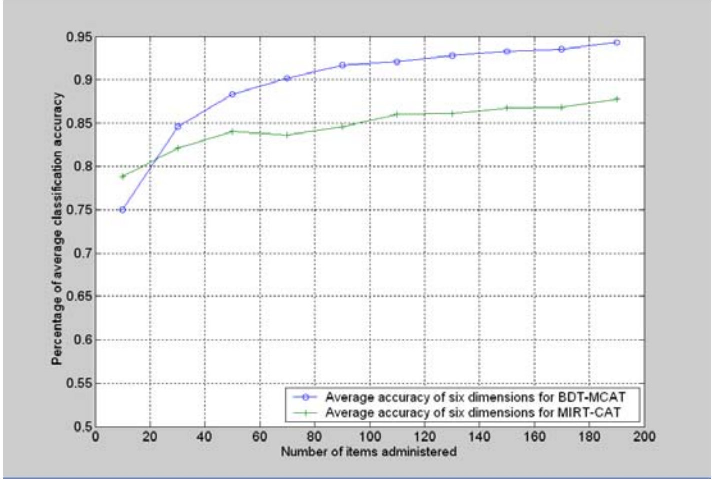


Figure 9 Comparison of average accuracy rate for all the same 6 dimensions involved between BT- MCAT and MIRT under the same experimental conditions.

DISCUSSION AND CONCLUSION

This study extends the unidimensional BT-UCAT model to multidimensional tests. This extensions allows the simultaneous assessment of multiple, potentially correlated skills through a single adaptive testing procedure. This is a step towards bringing the Bayesian theory framework closer to a useful tool in intelligent learning environments that require student skills assessment over a number of different dimensions. As we mentioned before, accurate assessment is highly demanded in many personalized learning to make the right recommendation of learning path.

In this project, we show that BT-MCAT is capable of optimizing the choice of item to take advantage of multidimensional ones and that it brings a significant improvement over the unidimensional framework. A comparison with MIRT also reveals that it performs better than this framework. Given that it is computationally simpler and more efficient, this makes the BT-MCAT approach an attractive alternative when the aim is to classify the student into a discrete set of categories such as master or non-master.

Future work can be to embed this model in someone e-learning system, and study how its assessment result can be used to create a compelling adaptive learning environment.

REFERENCE

1. Allen, M. J. and Yen, W. M., "Introduction to

Measurement Theory." (1979).

2. APeLS, "Http://Css.Uni-Graz.at/Demos/Apels/.".

3. Bloom, B. S., "The 2 Sigma Problem: The Search for Methods of Group Instruction As Effective As One-to-One Tutoring." , Educational Researcher, 13, 4 (1984): 4-16.

4. Conlan, O., Hockemeyer, C., Wade,V.and Albert,D "Metadata Driven Approaches to Faciliate Adaptivity in Personalized ELearning Systems.", Journal of the Japanese Society for Information and Systems in Education (2003).

5. Conlan, O. and Wade,V., "Evaluation of APeLS - An Adaptive ELearning Service Based on Multi-Model, Metadata-Driven Approach.", Proceedings of AH2004, (2004): 291-95.

6. Desmarais, M. C., Fu, Shunkai, and Pu, Xiaoming "Tradeoff Analysis Between Knowledge Assessment Approaches." , Proceedingsof Artificial Intelligence Intelligence in Education (AIED) (2005).

7. Falmagne, J., Cosyn,E., Doignon,J., Thiery,N. "The Assessment of Knowledge in Theory and in Practice.", http://www.highedmath.aleks.com/about/Science_Behind_ALEKS.pdf.

8. Hambleton, R. K., Swaminathan, H., and Rogers, H. J. "Fundamentals of Item Response Theory.", Sage Publications, F.W.Hesse & Y.Tamura (Eds.) (1991).

9. Hockemeyer, C., "Competence Based Adaptive E-Learning in Dynamic Domains." The Joint Workshop of Cognition and Learning through Media-Communication for Advanced E-Learning (JWCL), Berlin (2003): 79-82.

10. Hockemeyer, C., "Extending the Competence-Performance-Approach for Building Adaptive and Dynamic Tutoring Systems.", Talk at the 33rd European Mathematical Psychology Group (EMPG) Meeting,Bremen,Germany (2002).

11. ALEKS, "http://www.aleks.com".

12. Microsoft E-Learning , "http: //www.microsoft.com/learning".

13. Rudner, L. M., "The Classification Accuracy of Measurement Decision Theory.", National Council on Measurement in Education,Chicago(2003).

14. Rudner, L. M., "An Examination of Decision-Theory Adaptive Testing Procedures.", Proceedings of American Educational Research Association (2002): 437-46.

15. Wang, W.-C. and Chen, P. H., "Implementation and Measured Efficiency of Multidimensional Computerized Adaptive Testing." Applied Psychological Measurement, 28, 5 (2004): 295-316.