Titre: Title:	Computerized adaptive testing based on bayesian decision theory : uni- and multidimensional models
Auteur: Author:	Shunkai Fu
Date:	2005
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Fu, S. (2005). Computerized adaptive testing based on bayesian decision theory : uni- and multidimensional models [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7617/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7617/
Directeurs de recherche: Advisors:	Michel C. Desmarais
Programme: Program:	Non spécifié

UNIVERSITÉ DE MONTRÉAL

COMPUTERIZED ADAPTIVE TESTING BASED ON BAYESIAN DECISION THEORY: UNI- AND MULTIDIMENSIONAL MODELS

SHUNKAI FU
DÉPARTEMENT DE GÉNIE INFORMATIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE EN SCIENCES APPLIQUÉES
(GÉNIE INFORMATIQUE)
Décembre 2005

Your file Votre référence
ISBN: 978-0-494-16783-0
Our file Notre référence
ISBN: 978-0-494-16783-0

UNIVERSITÉ DE MONTRÉAL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

COMPUTERIZED ADAPTIVE TESTING BASED ON BAYESIAN DECISION THEORY: UNI- AND MULTIDIMENSIONAL MODELS

Présenté par : SHUNKAI FU
En vue de l'obtention du diplôme de : Maîtrise en sciences appliquée
A été dûment accepté par le jury d'examen constitué de :

M. Michel Gagnon, Ph.D., président
M. Michel C. Desmarais, Ph.D., membre et directeur de recherche
M. Jean-Guy Blais, Ph.D., membre

Acknowledgement

The author would like to express his deep gratitude to his advisor, Dr. Michel Desmarais, for his assistance and inspiration throughout this work. His support, patience, guidance and suggestions during the author's graduate study at the École Polytechnique de Montréal have helped to smooth away the difficulties he met in his thesis research. In addition, the author would like to thank his committee members: Dr. Jean-Guy Blais and Dr. Michel Gagnon for their time and comments.

The author also would like to thank his wife, Shan Huang, his parents, his sister and his Uncle Joe Kee Bo for their love, comfort, continuing encouragement, support and understanding throughout his education.

Thanks goes to all those who offered the author valuable comments and assistance too, Xiaoming Pu, Lei Ma, Dr. Marc Bourdeau, Dr. Peng Zhan and many anonymous ones.

Sincerely,

Abstract

Future learning environments need to be intelligent, adaptive and personalized, which calls for fine-grained assessment of the learner's knowledge level. Computerized adaptive testing (CAT) can play a key role towards this goal, administering items that match in difficulty level to the user's knowledge state. CAT results in the administration of considerably fewer items while keeping measurement precision equal to full-length paper and pencil (P&P) version of the same tests.

Item response theory (IRT) has been the primary framework of CAT model for decades. IRT is a mathematical model that represents the probability of an examinee's successes to test items in terms of underlying latent ability parameters. A large body of research has been devoted to building models based on IRT for assessing knowledge efficiently. Many well known applications, like TOEFL and GRE, are making use of this CAT framework.

IRT is concerned primarily with ranking examinees across an ability continuum. Although this theory provides good predictive accuracy, it requires strict assumptions – local independence and normal distribution – and an intensive computation burden due to its maximum likelihood estimate algorithm. However, very often, we are only interested in classifying examinees into finite discrete categories, like excellent/good/fair/poor/fail. For such a situation, a simpler measurement model would be sufficient [24]. Rudner proposed a Bayesian decision theory (BDT) measurement model that can be applied to classify examinees into discrete category based on their item response patterns. The model has a simple framework that starts with the conditional probabilities of examinees in each category or mastery state responding correctly to each item, and reaches the exact classification of examinees based on prior knowledge and observed response with classical Bayesian inference procedure.

A first part of our project concentrates on comparing Rudner's BDT approach to IRT. Rudner's model is limited to unidimensional exam, in which only one single latent trait can be measured. An overview of this unidimensional model is presented first, and simulation with two real tests follows to give us a comparison about their classification rate: a 34-item test on the knowledge of UNIX shell commands and a 160-item French language test. The result shows that BDT method can provide comparable result to IRT, though somewhat more noisy performance.

As mentioned in Rudner's article [25], the literature on the use of decision theory to analyze item responses is fairly scant, and none exists about the multidimensional decision theory to our knowledge. Therefore, in the second half part of our project, we augment the existing BDT-UCAT (Unidimensional CAT) model to multidimensional version so that more than one latent trait can be assessed through a single exam. How to do prior information calibration and how the adaptive procedure can work within BDT-MCAT (Multidimensional CAT) framework are introduced and validated through a number of simulation experiments. For the experiment, we resort to a Monte Carlo method for generating data samples to check the correctness of the model and replicate the experimental data. This procedure will allow us to compare our results with MIRT (Multidimensional IRT).

In our simulations, data is generated by a Monte Carlo method that has the following structure: (1) test that measures six different dimensions and nine item banks to cover those six dimensions, of which six banks include only unidimensional items and the other three are composed of multidimensional ones, and the amount of item in each of the nine banks are 200/200/200/20/20/20/200/200/200 respectively; (2) 1000 examinees with predefined correlations among those six dimensions of latent traits; (3) pseudo random responses of these 1000 examinees on those items in the nine banks.

Simulations of BDT-MCAT framework are performed using two kinds of item selection rule: (1) the Maximum Information Gain item selection rule and (2) the Random Administer (RA). The result shows that adaptive model works much better than the random selection rule when the same amount of items are administered. The MCAT model is then compared to the UCAT model, and results show that the simultaneous application of MCAT on all six dimensions perform better than the application of UCAT to each individual dimension. These results show that MCAT is able to leverage items shared among dimensions to yield more accurate assessment level when the same number of item candidate is maintained in each dimension. The result also shows that MCAT performs better than UCAT. Next, BDT-UCAT is compared with IRT-MCAT by using the data from Wang and Chen's article [29] on a specific MIRT approach, Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM). The result shows that BDT-MCAT's performance is comparable to IRT-MCAT model.

The conclusion drawn from our simulation not only gives support for the BDT-CAT model, including uni- and multidimensional, but also suggests that BDT offers an interesting lightweight alternative to MIRT for multidimensional assessment that can potentially meet the fine grained requirements imposed by intelligent learning environment. Limitation of the approach and future work are mentioned at the end of this thesis.

Keyword: Computerized adaptive testing, CAT, Bayesian decision theory (BDT), unidimensional, multidimensional, IRT, Monte Carlo simulation, BDT-CAT

Condensé en français

Introduction

Dans un environnement d'apprentissage conventionnel, le contenu pédagogique présenté est le même pour tous les étudiants, peu importe leurs différences de profils, de besoins ou de connaissances acquises. Pour les étudiants possédant de fortes ou de faibles capacités d'apprentissage, cette stratégie peut s'avérer inadéquate. Bien que les ordinateurs soient maintenant très présent dans le cadre de la formation, ils ne fournissent pas un enseignement personnalisé aux besoins de l'utilisateur. En 1984, Bejamin Bloom a défini le « problème des deux écarts types », qui stipule que les étudiants qui recoivent une formation individualisée réussissent mieux que ceux soumis à une formation conventionnelle dans une mesure de deux écarts types. Une telle amélioration signifie que l'étudiant moyen recevant une formation individualisée réussira au même niveau que les 2% les plus performants d'une classe traditionnelle [4].

Les environments d'apprentissage du futur vont être conçus pour tenir compte des besoins et différences individuelles. Ils seront intelligents, adaptatifs et personnalisés. De tels systèmes vont reposer sur un modèle de l'apprenant qui stockera l'information spécifique à chaque individu. Avec cette information, il est possible d'adapter le contenu pédagogique présenté afin d'optimiser le processus d'apprentissage. Il existe plusieurs dimensions à un modèle de l'apprenant; la dimension de l'évaluation des compétences est une des plus fondamentales. Elle vise à fournir un diagnostic précis des connaissances acquises et permettre ainsi de guider l'apprenant dans le matériel pédagogique. Les tests adaptatifs, TA ou CAT (Computer Adaptive Testing), sont un outil fort utile pour modéliser les connaissances acquises de façon efficace.

Les TA constituent une alternative moderne aux tests « papier-crayon » conventionnels. Le principe de ces tests est d'adapter la séquence de questions présentés selon les réponses

des répondants. Comme les items les plus informatifs sont présentés en premier, la longueur du test peut être réduite de façon significative tout en conservant un niveau de précision équivalent à un test papier-crayon complet du même test.

La Théorie des Réponses aux Items (TRI) a été le premier cadre théorique aux TA durant plusieurs décennies maintenant. La TRI est un modèle mathématique qui represente la probabilité de succès d'un répondant aux items d'un test en fonction de « traits » latents, c'est-à-dire d'habiletés non mesurables directement. Une grande quantité de recherche a été consacrée à la construction de modèle de TRI pour évaluer efficacement la connaissance d'un individu. Plusieurs applications de TA comme les tests TOEFL et GRE reposent sur ce cadre théorique.

Cette étude vise à comparer une approche alternative aux TA avec l'approche conventionnelle de la TRI.

Survol de l'approche de la Théorie Bayésienne de la Décision (TBD) appliquée aux TA

La TRI permet de situer sur un axe continue un répondant à un test par rapport à une habileté. Bien que cette théorie fournit une estimation avec une bonne précision, elle repose sur des hypothèses relativement fortes, celles de l'indépendance locale entre les items et leur distribution normale. En outre, elle nécessite une capacité de calcul importante étant donné l'algorithme itératif de maximisation de la vraisemblance utilisé pour l'estimation. Toutefois, il est fréquent que le cadre d'application de TA ne vise qu'à classifier le répondant en un petit nombre de catégories, comme « excellent/bon/passable/échec » ou « réussite/échec ». Dans ces cas, un schème de mesure plus simble suffit, comme l'avance Rudner [24, 25] qui préconise plutôt la Théorie bayésienne de la décision (TBD) pour classifier les répondants dans une catégorie discrète selon leur réponses aux items d'un test. Ce modèle, plus simble, débute avec une

probabilité initiale pour chaque catégorie, obtenues à partir d'un échantillon de la population visée, et met cette probabilité à jour après chaque question répondue selon le principe classique des probabilités postérieures de la théorie bayésienne.

Notre projet porte sur l'application de la TBD aux TA et comporte deux volets : les TA unidimensionnels et multidimensionnels. Nous débutons avec l'étude et l'implantation du modèle TBD de Rudner [24], dans lequel une seule habileté est évaluée durant un test. Puis, nous étendons ce modèle au problème de la mesure de plusieurs habiletés simultanément dans un même test. Le cadre théorique du modèle TBD multidimensionnel est présenté et sa validation est réalisée par différentes simulations.

Noter que pour nos besoins, nous limiterons le cadre de cette étude à seulement deux catégories pour la TBD, réussite ou échec. Il demeure toutefois possible d'étendre les principes à plusieurs catégories.

Modèle TBD unidimensionnel

TBD-U (Théorie Bayésienne de la Décision—pour tests adaptatifs Unidimensionnels) a été proposée par Rudner [24]. Nous y référerons aussi par simplement TBD. Elle possède l'avantage de sa grande simplicité tout en offrant une performance satisfaisante.
Tout comme pour les TA typiques, quatre éléments se retrouvent dans un modèle TBD : (1) un ensemble d'items (questions) calibrées, (2) un algorithme de mise à jour des probabilités, (3) un algorithme pour la sélection du prochain item à administrer et (4) une règle d'arrêt (décision). Ces éléments sont décrits dans ce qui suit.

Banque d'items calibrés. Tous les items doivent être préalablement calibrés. Les paramètres de calibration varient d'un modèle à l'autre. Dans le cas de TBD, il faut estimer la probabilité de succès à l'item selon que le répondant est classé dans la catégorie

« réussite » ou « échec » à l'examen. Une note de passage doit donc être fixée au préalable pour déterminer la catégorie du répondant.

Mise à jour des probabilités. Après chaque item répondu, la probabilité postérieure que le répondant soit classé dans l'une ou l'autre des catégories est estimée en tenant compte de la probabilité antérieure. Conformément au cadre bayésien classique, la probabilité postérieure deviendra la probabilité antérieure au prochain item présenté.

Règle de sélection du prochain item. Le choix du prochain item doit maximiser l'information fournie par la réponse. Différents algorithmes existent à cette fin, notamment la réduction de l'entropie (le gain d'information) qui est retenue dans cette étude et l'information de Fisher très utilisée avec la TRI.

Règle d'arrêt. La règle d'arrêt détermine si une décision doit être fixée. Dans notre étude, nous n'utilisons pas de règle d'arrêt. Nous observons plutôt l'évolution du taux de bonnes décisions tout au long de la simulation du test adaptatif, de 0 questions jusqu'à la fin du test.

Simulations et comparaisons avec la TRI

Une évaluation du modèle TBD est effectuée avec des données réelles provenant de deux tests, un premier portant sur la connaissance des commandes du système UNIX composé de 34 items, et un second portant sur la connaissance de la langue française de 160 items. Les données de quarante-huit répondants pour le text UNIX et de 40 répondants pour le test de français servent de base à l'étude.

La procédure de simulation consiste à émuler le processus de questions-réponses des TA et à comparer les prévisions de l'approche TBD avec celles de la TRI. Dans le cas de la TBD, on compare la classification (réussite ou échec) obtenue par estimation avec la TRI

ou l'approche TBD à la classification basée sur le score réel en fonction d'un seuil de passage donné, en l'occurence 60%. Pour la TRI, l'évaluation du niveau de compétence, qui varie généralement de -3 -a +3, est transformé par la fonction logistique sur une échelle de 0% à 100% pour être ensuite comparée au seuil de passage. Si l'habileté estimée dépasse le seuil, le résultat attendu du répondant est classifié « réussite », autrement il est classifié « échec ». On obtient ainsi pour chaque approche un taux de bonne classification des répondants qui évolue tout au long du processus de questions-réponses.

Les deux modèles sont calibrés à partir des données des répondants. Deux règles de sélection sont utilisées, la méthode de la réduction d'entropie et l'information de Fisher.

Résultats

Les résultats démontrent que les approches ont des performances moyennes comparables. Cependant, on observe une plus grande variabilité pour la méthode TBD dans le cas du test de français. Néanmoins, compte tenu de la plus grande simplicité de l'approche TBD et de l'algorithme beaucoup plus efficace, ces résultats sont encourageants.

Modèle TBD multidimensionnel

Nous avons étendu le modèle TBD à la situation où plusieurs dimensions, ou compétences, sont évaluées simultanément. Nous utiliserons l'acronyme TBD-M pour y référer. À notre connaissance, seul le cas unidimensionnel a été étudié [24] et cette approche n'a pas encore été explorée.

Un test multidimensionnel peut comporter des items de deux types, ceux qui évaluent une seule dimension, les items unidimensionnels, et ceux qui en évaluent plusieurs simultanément, les items multidimensionnels. Si un test multidimensionne ne comporte

que des items unidimensionnels, il est libellé « inter-items », alors que s'il comporte des items multidimensionnels, il sera libellé « intra-items ». Les tests intra-items offre un potentiel plus intéressant pour les tests adaptatifs et c'est donc ce type de test qui est retenu ici.

Comme dans le cas du TBD-U, la banque d'items doit être calibrée préalablement à l'application de l'algorithme de TA. Cependant, chaque item est calibré par rapport à une dimension et les items multidimensionnels comportent ainsi deux paramètres ou plus selon le nombre de dimensions auxquels ils sont rattachés. Le principe de calibration demeure toutefois le même par rapport à chaque dimension.

Quant au mécanisme d'inférence, il demeure identique aux items unidimensionnels: la probabilité postérieure que le répondant réussisse par rapport à une dimension, c.-à-d. que le score estimé soit au delà du seuil, est calculée après chaque que chaque item relié à cette dimension soit observé. Pour les items multidimensionnels, ils entrent ainsi dans le calcul de chaque dimension auxquels ils sont rattachés. Leur poids est donc plus important dans l'évaluation de l'ensemble des dimensions.

Simulations et données

Le processus de simulation est analogue à celui de TBD-M mais certaines différences s'appliquent dans le choix des items et la simulation est basée sur des données générées plutôt que réelles.. Premièrement, seule la technique de la réduction de l'entropie est utilisée. L'entropie est calculée pour l'ensemble des dimensions et non plus seulement pour une seule. L'item qui diminue l'entropie globale sera le prochain choix.
Contrairement à l'étude TBD-U, les données pour le TBD-M ne sont pas réelles mais simulées. Elles sont générées par la méthode Monte Carlo. Cette méthode consiste à définir des paramètres théoriques d'un modèle de données pour ensuite générer des échantillons à partir de ce modèle. Ceci permet d'une part de connaître la « véritable »

valeur des paramètres qui ne sont pas directement observables, comme c'est le cas pour les compétences. Pour générer des tests intra-items, cette possibilité est très avantageuse car il est difficile de créer de tels tests et d'obtenir des données réelles. D'autre part, cette approche nous permettre de comparer la technique TBD-M avec la TRI multidimentionnelle en répliquant les données et résultats d'un article pour cette comparaison [29].

Les données des simulations comportent six dimensions inter-corrélées avec une matrice de covariance prédéfinie et possèdent les caractéristiques suivantes : (1) les données comprennent des items unidimensionnel et multidimensionnels qui se décomposent en 9 segments; (2) les segments 4, 5 et 6 sont multidimensionnels alors que les autres 6 sont unidimensionnels; (3) le nombre d'items pour chacun des 9 tests est respectivement de 160/160/160/20/20/20/200/200/200 pour un total de 1140 items; (4) 1000 répondants sont définis et leur compétences sont interreliées conformément à la matrice de covariance des dimensions; (5) des réponses sont générées pour chacun des 1140 items des 1000 répondants conformément à leur compétences individuelles.

Ces caractéristiques sont identiques à celles de Wang et Chen [29] et elles s'inspirent de batteries de tests utilisées dans des écoles taiwanaises.

La calibration est basée sur les réponses générées. Six cents répondants sont utilisés pour la calibration et 400 pour la simulation.

Résultats et transformation de métrique par simulation

Une première étude consiste à comparer la performance de TBD-M basée sur l'algorithme de la réduction de l'entropie avec un choix aléatoire. Elle confirme que le choix d'items basé sur cet algorithme est près de 10% plus précis qu'un choix aléatoire avec un nombre

équivalent d'items administrés. La simulation révèle aussi que ce sont les dimensions qui comportent les items multidimensionnels qui affichent une meilleure performance que les dimensions qui ne comportent que des items unidimensionnels. Cet avantage ne s'observe que pour la condition avec l'algorithme de réduction d'entropie et confirme ainsi que TBD-M est en mesure de maximiser l'information provenant des items multidimensionnels. D'autre part, une comparaison de TBD-M avec TBD relève que les résultats sont de 3-4% plus précis dans le premier cas, confirmant que l'approche multidimensionelle exploite les items multidimensionels à leur plein potentiel.

La seconde étape consiste à comparer les résultats de la simulation TBD-M avec ceux de l'approche IRT multidimensionnelle. Cette approche étant complexe à implanter, ce sont les résultats de Wang et Chen [29] qui servent de point de comparaison. C'est d'ailleurs pour cette finque nous avons répliqué leur données.

Cependant, Wang et Chen rapportent leurs résultats sur la base de la corrélation entre les habiletés réelles et celles estimées par le modèle TRI. Or, ces mesures sont celles de la de theta qui est une variable continue et correspond à une distribution normale centrée autour de 0 et qui est sur une échelle de -3 à +3 en pratique. Pour rapporter ces les résultats de Wang et Chen en termes de pourcentage de répondants correctement classifiés par rapport avec un score de passage, nous devons effectuer une nouvelle simulation de Monte Carlo. Le principe consiste à générer des répondants dont le theta corrèle conformément au données rapportées par Wang et Chen, puis de générer des réponses à partir de ces répondants. Le ratio de répondants correctement classifiés peut alors être calculé.

Les résultats de la comparaison de TBD-M avec ceux de la TRI multidimensionnelle rapportés dans Wang et Chen et transformés selon la procédure décrite dans le paragraphe précédent nous indiquent que la performance de TBD-M est au moins aussi bonne que celle de la TRI multidimensionnelle.

Discussion

Ce projet vise à comparer une méthode simple d'évaluation des habiletés, la TBD, avec la méthode bien reconnue dans le domaine, la TRI. La TBD classifie un répondant à un test à partir d'un échantillon de réponses. Elle peut en théorie classifier le répondant à partir d'un nombre arbitraire de de catégories mais pour les besoins de cette étude seules les catégories « réussite » et « échec » sont utilisées. Cette approche se démarque de la TRI qui fournit un chiffre sur une échelle continue pour représenter l'estimation de l'habileté et permet donc en principe une évaluation plus fine de la compétence d'un répondant. Cependant, dans le cas où une simple décision est désirée, la TBD peut s'avérer avantageuse. D'une part, elle est plus simple en terme calculatoire puisque les calculs sont faits sur la base d'équations fermées (calcul fonctionnel) alors que la TRI nécessite un algorithme itératif pour estimer la valeur de theta. D'autre part, elle est plus simple à calibrer que la TRI qui peut reposer sur trois paramètres reliés à chaque item (discrimination, difficulté et chance), alors qu'un seul paramètre par item doit être calibré pour la TBD, le ratio de répondants ayant réussi le test pour ceux qui ont réussi l'item en question.

Les résultats des simulations révèlent que l'algorithme TBD réussit à égaler les résultats de la TRI à la fois pour des tests unidimensionnels et pour un test multidimensionnel comportant aussi des items multidimensionnels. Cependant, certaines conditions nous indiquent une mise en garde, notamment la combinaison de TBD avec l'algorithme de choix d'item basé sur l'information Fisher qui affiche une performance bien en deça des autres conditions. On remarque aussi une plus grande variabilité dans la courbe de résultats de la TBD que ceux de la TRI pour les tests unidimensionnels. Par contre, avec les tests volumineux et multidimensionnels, la TBD semblent se comporter au moins aussi bien que la TRI.

Bien que cette étude ne couvre pas l'ensemble des questions quant à la comparaison TBD et TRI, elle démontre un potentiel intéressant pour cette approche dans des situations où une décision de nature discrète est à prendre et où l'on désire utiliser une approche plus simple que la TRI. Dans le cadre des environnements personnalisés d'apprentissage, il peut s'avérer opportun de choisir une approche plus simple à la modélisation et l'estimation des habiletés de l'apprenant pour réduire le « coût d'entrée » à l'utilisation de techniques de personnalisation ainsi et favoriser l'essort de ces environnement intelligents.

Table of Content

Index of Figures

Index of Tables

List of symbols and abbreviations

BDT	Bayesian Decision Theory
CAT	Computerized Adaptive Testing
CBT	Computer-based Training
CAI	Computer-aided Instruction
IRT	Item Response Theory
ITS	Intelligent Tutoring System
MCAT	Multidimensional CAT
MIRT	Multidimensional IRT
MRCMLM	Multidimensional Random Coefficients Multinomial Logit Model
RA	Random Administer
SPRT	Sequential Probability Ratio Test
UCAT	Unidimensional CAT

Chapter 1. Introduction

1.1 Intelligent and Adaptive Learning

In a traditional learning environment, the same content is presented to all students regardless of their background knowledge and abilities. This approach does not take into account learning differences among individual learners, such that for student with high or low studying abilities, the content and the pace of learning can be inappropriate.

The outcome of one study led by Benjamin Bloom revealed the "two-sigma problem", which states that students who receive one-on-one instruction perform two standard deviation better than those receiving classroom instruction. An improvement of two standard deviation means that the average tutored student performed as well as the top 2 percent of those who are educated in classroom [4]. The value of one-on-one tutoring has attracted attentions from government, academia, and commercial organizations; however, it suffers from high cost for large-scale application. We could not, for example, image providing presonalized instruction at École Polytechnique de Montréal to 5000 students.

Practical solutions that can provide such kind of individualized instruction at affordable cost have been explored, such as computer-based tutoring (CBT) and computer-aided instruction (CAI). However, most of them are not truly individualized to the learners' needs. More recently, intelligent learning environment (ILE) is emerging as an attractive solution that offers a technology-based one-on-one tutoring environment. The goal of ILE is to provide intelligent, adaptive and personalized instruction to learners. Notice that, here, "one-on-one" instruction refers to machine-to-person, which requires far less human resource as it provides an automatic or semi-automatic system solution.

ILE encompasses a large family of systems that aim to be adaptive and responsive to the learner's individual needs. Student model is the foundation of ILE that stores information specific to individual learner, which can be used to tailor the instruction to student's different learning needs. Among those components of a typical student model, assessment is a critical component since accurate information about learner's knowledge level can help to build a learner's skill profile and allow the whole system to customize the learning route or strategy for that specific profile. Furthermore, future ILE require quick, accurate, as well as customizable test that can be accessed by students with convenience, so that learners can truly control the learning pace by themselves without waiting for the lecturer to administer and mark an exam.

1.2 Student model quality factors

Student modeling approaches differ over a number of dimensions and qualities. These differences can be drastically different whether the approach is based on automated learning or on handcrafted model building and calibration. These dimensions are discussed below:

- Flexibility and expressiveness: AI-based learning system often relies on fine-grained assessment of abilities and misconceptions. Graphical probabilistic models, which are a marriage between probability theory and graph theory, are highly suitable for fine-grained cognitive diagnostic, but they do not readily lend themselves to automated learning. On the other hand, a large body of psychometric theory that dwells on classifying an examinee as *master* or *non-master* offers a sound statistical framework for performing global assessment. Item Response Theory (IRT) is the prevalent approach in this field, but it lacks the fine grained diagnostic quality required in adaptive learning systems. Nevertheless, many of the building blocks of IRT can, and have been used for finer grained assessment [7]. For example, the developments over multidimensional IRT (MIRT) and

Tatsuoka's Rule Space using IRT data fitting for knowledge assessment offer an avenue towards merging finer grained diagnostic with IRT.

- Cost of model definition: Fine-grained models such as those found in Bayesian networks (see, for example [6, 28]) require considerable expert modeling effort. Because of this modeling effort, fine-grained models can prove overly costly for many applications. On the contrary, data driven approaches such as IRT can completely waive the knowledge engineering effort.
- 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 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 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 parameterized with a small data sample.
- Accuracy and reliability of prediction: Student modeling applications such as Computerized Adaptive Testing (CAT) [2, 16, 17], which is known as providing an adaptive testing by using computer technology, is 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 environment factors such as skills of the knowledge engineer, the robustness to noise in the model, and to noise in the data used to calibrate a model. Handcrafted models, in particular, are subject individual differences and human biases. They cannot readily offer means to predict their reliability. Whereas extensive research in IRT has been conducted to investigate reliability and robustness under different conditions, little has been done n intelligent learning environments.
- 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.
- 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 and IRT. Of course, the more assumptions and approximations are made, the less accurate and reliable a model becomes. Some approach may work well in one context and poorly in another because of violated assumptions.

These factors will determine the practical value of a student modeling approach. Ideally, we would like a fully automated student model learning approach that requires little data to build and calibrate and, yet, that yields detailed and accurate knowledge assessment.

Such an approach would limit the effort of model building to that of data gathering. It would limit the effort of model update, such as adding new test items and new concepts, to re-running the learning algorithm, which could in principle be done in real time as new data is gathered. Finally, given an algorithmic approach to model building, reliability and accuracy are not dependent on environment factors other than the parameters that characterize the learning data and, if they can be defined and measured, reliability and accuracy can thus be predicted through these parameters.

The aim of this project is focused on providing an efficient and effective assessing module for typical student models. We start with the discussion of non-fine-grained unidimensional CAT model, including theory basis and simulation results, and then refine the model towards multidimensional one that is more fine-grained, providing more detailed and comprehensive information about examinees. Our framework is based in the CAT field, but it is not limited to that application and can be considered as a general learning model framework. We will return to the discussion of these model quality factors mentioned above in the discussion part, concluding that how well our model fit these standards.

1.3 Computerized Adaptive Testing

CAT is an ideal candidate for the assessment requirement of student models since it can provide a quick but accurate feedback of learners' knowledge state, and such information is necessary for the other modules to decide the next study plan.

Before CAT was born, traditional paper-and-pencil (P&P) was the only method of measurement, in which fixed-length exam was presented to all examines. Fixed-item tests waste students' time because they give students a large number of items that are either too easy or too difficult. As a result, the tests give little information about the particular level of ability of each student. Furthermore, presenting too easy or too difficult items increases

the possibility of bringing more error of measurement, so when a group of examinees with a variety of trait values are tested with the same test, it is impossible for the test to be maximally efficient for all examinees simultaneously [1]. An adaptive testing solves this issue. As shown in [16], a test provides the most precise measurement of an examinee's ability when the difficulty of the test is matched to the ability level of the examinee, so adaptive testing is ideal for a group of examinees constructed with different ability levels. For an ITS to track individual subject's learning process accurately and effectively, adaptive testing will be a helpful module since personalized sequence of item can be administered, and a high level of accuracy for ALL students can be maintained by CAT.

IRT has been the prevalent theory since the introduction of CAT, but more recently, two kinds of Bayesian approaches are suggested: Bayesian Networks and Bayesian Decision Theory. The first is a combination of probability theory and graph theory, and has been used in a number of learning environments projects [2, 9, 10, 28]. It is regarded as a fine-grained model, but it also suffers a significant shortcoming, namely that it depends largely on expert modeling, which limits its application. Bayesian decision theory (BDT) is another possible solution, with classical Bayesian theory at its basis. It was proposed by Rudner [24, 25], and this is what we adopted in our project. BDT is a data driven model like IRT, but it is designed for discrete classification, where examinees only need to be classified into a limited number of categories, like *master* and *non-master*.

1.3.1 Basic Components of CAT System

What is CAT constructed with? The following scenario description of a typical CAT session will give an overview.

"You booked a CAT on the day you prefer in advance according to your review plan and personal schedule. You arrive at the testing center, a place like normal office or lab, 20 minutes earlier on the exam day. When your turn comes, you will be led to a computer assigned to take your CAT. After providing the necessary information, the test begins.

Your first item will be presented on the screen, and normally it will not be difficult. You make a response by mouse operation, and click "next". Re-confirm operation will be required by the CAT system to make sure that no mistake will happen. Then, the computer system will proceed, choosing next item according to your previous response: if your answer is right, the new item will be more difficult; else, it will be easier. If you find it be easier, unfortunately, you cannot go back to previous item. You can only go forward until your time expires or other stopping rules be met. So different examinee will meet different item sequence during CAT even their final scores are the same."

CAT is currently in use in many applications and by organizations such as ETS (Educational Testing Service) for well known tests such as GMAT, TOEFL and GRE. A CAT system is composed of the following components:

- **Item bank**. It is a database of all calibrated items available. For items used in CAT, they must be measured in advance to calibrate the related parameters, and an accurate calibration plays as a necessary precondition for the high performance of CAT considering that CAT is kind of item-based, not like traditional exam, which is test-based.

- **Selection rule of first item**. The item to begin the exam. Although the choice of the first question is not critical to measurement, it may be critical to the psychological state of the examinee. If it is much too hard, the candidate may immediately feel despair, and not even attempt to do well. This is particularly the case if the candidate already suffers anxiety about the test. Else if it is much too easy, the candidate may not take the test seriously and so make careless mistakes. Gershon [15] suggests that the first item, and perhaps all items, should be a little on the easy side, giving the candidate a feeling of accomplishment, but in a situation of challenge.

- **Selection rule of next item**. Its role is to select optimal next item to be administered. Commonly used rules are maximum information gain, minimum risk cost, and Fisher information. However, practical concerns of operational testing programs may override the selection rule, such as ensuring proper content balance and limiting the number of examinees permitted to view each item, a problem known as "item exposure." So, normally, in a practical application, a compound rule will be designed for item selection. For example, a group of items with similar maximum information is chosen first and then items are, furthermore, selected from this set based on pre-defined exposure and content balance controlling rules. In our research, we will focus on selection rules that optimize the accuracy of skill assessment, leaving aside practical consideration.
- **Stopping rule.** When to stop the adaptive testing is determined by a stopping rule. There are some potential candidates available: time expiry, fixed number of item, or a degree of confidence in deciding the outcomes of the test (mastery or not) for which SPRT (Sequential Probability Ratio Test) is generally used [27]. In our simulation, no stopping rule is employed since we are only interested in observing the accuracy as a function of the items administered one after the other.
- **Scoring mechanism**. It is optional. In some standard exam, like GRE, the final score will be transformed by some rules so that it can be more explainable or comparable when a long-term data is put together.

In a practical setting, all the modules above should be taken into consideration. In our discussion of BDT-CAT, we will focus on item calibration and selection, ignoring the other details required by application.

1.3.2 Categories of CAT

CAT can incorporate two types of items: dichotomous and polytomous. Dichotomous items have only one correct answer, and they are more frequently found in current CAT;

polytomous items have one or more correct choices with different scoring schema. Unless explicitly mentioned, item type in this article is dichotomous.

CAT can also be classified into uni- and multidimensional. In UCAT (unidimensional CAT) model, a single primary skill is measured through one exam. TOEFL is such an example of UCAT. Although there are several sections in the current TOEFL, each addressing a different dimension, and the adaptive assessment category is based on the assessment of a simple skill at once. Considerable research on various aspects of UCAT has been conducted in the past years, and virtually all existing applications are limited in unidimensionality now. They assume that there is only one primary skill to determine examinees' response in the specific exam. However, unidimensional models are not always appropriate for real tests. A review of many new forms of assessment and the associated scoring protocols, such as those represented in cognitive tasks, portfolio assessments, performance tasks, standardized patient methodology, clinical skills assessments, writing assessments, oral presentations, and projects, suggests that multidimensional model are needed to adequately and accurately account for examinees' test performances [27]. Such need of more fine-grained assessment drives more recent research work towards MCAT (Multidimensional CAT), where more than one latent trait can be measured through a single exam. The following Table 1-1 lists the commonly found kinds of CAT based on those two standards of classification mentioned here.

Table 1-1 Category of CAT

	Unidimensional	**Multidimensional**
Dichotomous	UCAT with dichotomous items	MCAT with dichotomous items
Polytomous	UCAT with polytomous items	MCAT with polytomous items

In our project, only exam with dichotomous items will be studied in both uni- and multidimensional model.

1.4 Overview of Bayesian Decision Theory

A brief introduction to BDT, the foundation of our CAT models, is presented here, and we will return to this topic with more details on BDT-UCAT and BDT-MCAT in Section 4 and 5 respectively.

BDT is a branch of probability theory that allows one to model uncertainty about the world and outcomes of interest by combining prior knowledge and observational evidence. Since most problems involve some levels of uncertainty and prior knowledge, Bayesian approach has been playing as an increasing role in knowledge learning and discovery because of its ability to infer posterior information based on a priori gained experience and observed fact, which lets it differ from other statistical techniques of prediction. There are many known stories on the application of Bayesian theory in different fields, among which some typical ones are NASA's astrophysics research assisting tool-AutoClass [21], weather forecasting [5], health care policy [13,14], marketing [23], and more fields requiring risk analysis, prediction and decision making.

Studies in [8,11,24] show that the performance of CAT model based on BDT can be comparable to IRT approach for classification decisions, while much less computation is required. The model can be used to build independent adaptive testing system or be embedded into other comprehensive system, such as study guide or intelligent tutorial system (ITS) [19]. There are other Bayesian modeling techniques such as Bayesian graph models that are being used to build adaptive testing model [2,6,28], but this study only compares the performance of BDT and IRT. For more comparison, please refer to [8].

1.5 Overview of the Project

Our project focuses on building simple but effective uni- and multidimensional CAT model with BDT. Therefore, we will start by repeating Rudner's unidimensional model of

CAT based on BDT in Section 4. Accuracy performance is studied with simulations of real data samples. As we will show, the outcome of our simulation is satisfactory, comparable to the IRT approach, but the computational complexity is far smaller. We will then pursue our investigation towards testing a multidimensional BDT model. This is the key part of our project, and Section 5 contains a discussion of BDT-MCAT.

In section 3, we will present a brief introduction to the traditional CAT framework, namely IRT. It can be considered as a baseline from which to compare our work.

Content about BDT-UCAT and -MCAT will follow in section 4 and 5, in which both theory and simulation validation will be included for both kinds of model.

Note that in this thesis, the terms dimension, trait, ability and skill will be used alternative, as found in most articles on this field due to authors' preference, to refer the same concept: a non-directly observed characteristic of an individual that determines his ability to perform in a test.

Chapter 2. Simulation Methods Overview

In this section, we discuss the simulation methods used for this study.

2.1 Simulation methods

There are different approaches to validate CAT models. Commonly, they are classified into two primary kinds: live-test studies and simulations, of which the later one can be further divided into simulation with real data (called post hoc simulation) and pseudo random data (Monte Carlo simulation).

Live CAT testing involves the administration of real tests to real (or live) examinees. A real item bank is set up, and items are applied to a group of real examinees in an adaptive way. It is an expensive and time-consuming job considering an iterative procedure is un-avoidable to revise the experiment; besides, only small scale of items and examinees are possibly allowed in such environment especially during the experiment step, which makes it difficult to have a comprehensive study of the underlying CAT model. Therefore, it is not a good choice for the starting point of a CAT research project, but can be used at a later time when the model is more mature to see if it could still work well in real application, where noise exists and personal test preference factor will be considered.

As mentioned above, there are two kinds of simulation available and widely applied in CAT research depending on the research target - post hoc and Monte Carlo simulations. In post hoc simulation, the "item bank" used consists of the actual answers of examinees to a full-length test. The objective of such CAT simulation procedure is to determine how much reduction in test length can be achieved by "re-administering" the items in an adaptive way. It is applied widely in [7, 9-11, 22]. By using this method, we assume that what is observed in the real test is just the "true" ability of examinees. The greatest advantages is that the data is from real world and the results are expected to be more valid.

In contrast to post hoc method, a Monte Carlo simulation defines a theoretical examinee population and item banks, and generates their responses over items by using a pseudo random data generation model. Model parameters can be simulated in similar way, or sometimes, they are estimated from real test data sample. By resorting to Monte Carlo simulation, 1) we are allowed to observe the performance of CAT model given known "true" ability level because they are generated first; 2) we can define examinee and item population parameters according to our research needs. Moreover, the scale of sample can be very large, contrary to the previous two experimentation methods. We thus have the opportunity to observe and study the corresponding performance of target CAT models under various pre-designed conditions through adjusting the distribution of items and examinees in a flexible and easy way. This method is preferred at the early stage of CAT research, providing a quick as well as economical experimental framework.

2.2 Choice of simulation methods in our project

Simulations are chosen for our CAT research project. Post hoc simulation with real data is applied for the study of unidimensional model. Our data is from two tests, a French language and a UNIX shell command test. However, a Monte Carlo method is chosen for the study of multidimensional CAT model due to the requirement of large scale samples, and adjustable parameters of items and examinees. Monte Carlo simulation also permits us to compare our results to one multidimensional IRT model [29] by simulating the same experimental conditions, which not only frees us from the burden to implement another complex model, the golden standard in this field, namely MIRT.

2.3 Overview of the simulation implementation

In our project, all programs are implemented with Matlab considering its repaid development cycle and large function base available for complex computation. The version we use is 6.0. One disadvantage of using Matlab is that a program written in Matlab's own programming language runs in interpretation mode, which makes the simulation running speed much slower than programs written in C or C++.

Basic settings of the simulation include the following:

- Amount of items and examinees to be created for Monte Carlo simulation;
- Percentage of samples to be used for calibration and verification purpose respectively, including the post hoc and Monte Carlo simulation;
- Passing score to determine whether an examinee is a master or non-master based on their raw score in full-length conventional test;
- When the testing will stop for each examinee. In our research, we support to specify certain amount of item need to be administer.

Chapter 3. Brief Introduction to IRT-based Computer Adaptive Testing

Item response theory, or IRT, is the theoretical foundation of most Computer Adaptive Testing frameworks. It represents a golden standard in the field. The BDT model thus has to be compared with this leading theory. Therefore, a brief introduction of IRT is included here.

Item response models are particularly suitable for adaptive testing because it is possible to obtain ability estimates that are independent of test items administered. Even if each examinee receives a different set of items, IRT provides a framework for comparing the ability estimate of different examinees. However, this kind of model is based on strong assumptions (e.g., unidimensionality and item independence given the ability level). In practice, these assumptions are often violated, which may seriously compromise the quality of examinee's test scores and trait estimates. Model assumption violations could lead to a response pattern that may not fit the underlying test model, and result in an ability estimate that does not accurately reflect the latent trait of the examinee [30].

A large body of research has been devoted to IRT in the past three decades. However, the classical IRT models assume that a single ability accounts for the examinee's performance, which is called UIRT (Unidimensional IRT). More recently, a multidimensional version emerged, MIRT (Multidimensional IRT), and it allows a more fine grained measurement [18]. We will focus on UIRT in this section, and turn to MIRT late in section . Note that, for MIRT in our project, we refer to one specific version named MRCMLM (Multidimensional Random Coefficients Multinomial Logit Model). Since it is the unique MIRT model we consider in this thesis, we just use the term MIRT instead of MRCMLM to refer this specific MCAT model.

IRT is a well-documented theory, so we only briefly review its theoretical basis here. IRT models the relationship between examinee's ability and his probability of correct response into a theoretical function named item characteristic curve (ICC). The ICC functions normally have an "S" shaped curve which implies that the higher the ability level one examinee poses, the higher chance he will succeed in the response to that question. Each test item has its own ICC. There are three different ICC functions according to the different number of parameters included: one-, two- and three-parameter logistic model (1PL, 2PL, and 3PL respectively) :

$$P(X_i = 1 \mid \theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \qquad \textbf{Equation 3-1}$$

$$P(X_i = 1 \mid \theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ii})}} \qquad \textbf{Equation 3-2}$$

$$P(X_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}} \qquad \textbf{Equation 3-3}$$

Note: Xi is the response variable, and θ refers to examinee's ability level

The one-parameter model is the simplest (Equation 3-1), and it is often called the Rasch model [20], in honor of its developer. For 1PL and 2PL model, the b_i parameter for an item is the point on the ability level scale where the probability of a correct response is 0.5. The greater the value of b_i, the greater the ability that is required for an examinee to have a 50% chance of answering the item rightly; hence, the harder the item. The item characteristic curve (ICC) relates the probability of success on an item to the ability measured by the test and the characteristic of the item, and Figure 3-1 is such an example.

The logistic model is more mathematically tractable than the normal ogive model because the later involves integration, whereas the former is an explicit function of item and ability parameters and also has important statistical properties [16]. Lord [18] was the first to

develop a two-parameter item response model, based on cumulative normal distribution (normal ogive). Birnbaum [3] substituted the two-parameter normal ogive function with logistic function as the form of the item characteristic function. Logistic functions have the important advantage of being more convenient to work with than normal ogive functions. The 2PL model resembles the one-parameter model except one more parameter, a_i, is added. It is the item's discrimination factor, and it is proportional to the slope of the ICC at the inflection point b_i on the ability scale. Items with steeper slopes are more useful for separating examinees into different ability levels than are items with less steep slopes. Though item discrimination parameter is defined, theoretically, on the scale $(-\infty, +\infty)$, the usual limited range is (0, 2). The two-parameter model can be regarded as a generalization of the one-parameter model that allows for different discriminating items.

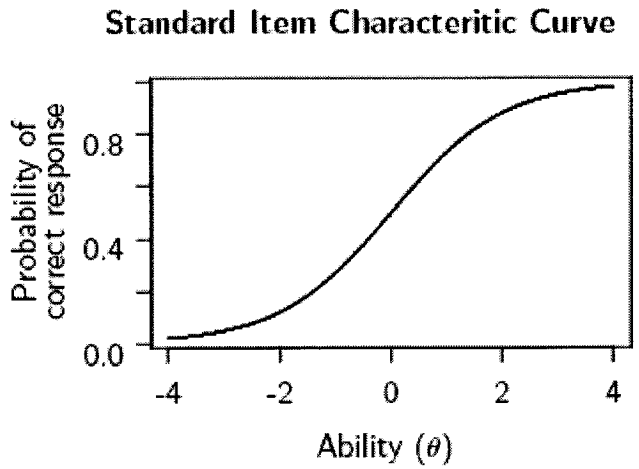

Figure 3-1 Item response curve (a = 1, b = 0, c=0)

For the 3PL model, there is one additional parameter called pseudo-chance-level parameter, c_i. This parameter provides a possibly nonzero lower asymptote for the ICC and represents the probability of examinees with low ability answering (or guessing) the item correctly. Regardless of which model is used, calibration of each item is required.

During the adaptive testing procedure, once one item is answered by the examinee, the value of θ is reassessed. The new estimate of θ is generally based on maximum likelihood estimation (MLE) procedure, the value of θ that maximizes:

$$P(\theta \mid X) \propto P(X_1 \mid \theta)P(X_2 \mid \theta)...P(X_i \mid \theta)$$

where X is the examinee's response vector, $X_1, X_2, ..., X_i$, and where value of $P(X_i \mid \theta)$ is given by Equation 3-1, 3-2 or 3-3 according to the different model choice. Via the MLE technique, it is possible to infer examinee's ability and item parameters from observed response data. The MLE is generally implemented as a root(s) finding process. Newton-Raphson based approaches are usually the applicable solutions in such situations.

The essential goal of IRT is to obtain a measure of an examinee's ability level based on the observation of his response to the n items administered. As mentioned above, the ability estimates is obtained by finding an ability value that maximizes the likelihood function which is just the distribution of item responses for one examinee. The essence of this case is that the joint distribution of item responses for one examinee is treated as a single variable function of ability level θ, and the remaining parameters (a, b, and/or c) are taken as known parameters. To calibrate the corresponding parameters a/b/c of each item as appear in Equation 3-1, 3-2 and 3-3, the IRT joint maximum likelihood estimation (JMLE) is a typical choice. It is an approach that estimates item parameters by finding a, b, c and θ values that maximizes the likelihood function. MLE is a widely accepted approach for precise estimation of parameters; however, it is a time-consuming procedure.

We briefly introduce IRT here since it is not our emphasis to talk about the theory basis of IRT in this project. Comparison with IRT is limited on the classification accuracy performance.

Chapter 4. BDT-UCAT Model

From this section on, our discussion will turn to CAT model based on BDT. In section , our topic is BDT-UCAT, and what will be covered includes two aspects: theory basis and simulation evaluation experiments. Section is the second half of our project on more general multidimensional adaptive model, BDT-MCAT. Although we limit our models to binary category, *master* and *non-master,* note that, it could be expanded to multiple categories or levels, such as low, medium and high.

4.1 Theory basis of BDT-UCAT

In this section, we will talk about the BDT-UCAT model, for which we aim to assess only a single trait at once.

The procedure of adaptive testing based on BDT starts with the proportion of different examinee categories (or the probability of being specific category), which are *master* or *non-master* in our example, in the population. Those proportions are estimates for $P(\theta_m)$ and $P(\theta_{\overline{m}})$. The likeliness for examinees to correctly respond to item i correctly given a specific ability level, or category (*master* and *non-master*) is given by $P(X_i = 1 | \theta_m)$ and $P(X_i = 1 | \theta_{\overline{m}})$. Here, we use θ to indicate ability parameter, and the m and $\overline{m}$ for categories *master* and *non-master* respectively. Given an observed response, the posterior value of $P(\theta_m)$ and $P(\theta_{\overline{m}})$ are recalculated based on the standard Bayesian inference rule, and using the prior information of $P(X | \theta_m)$ and $P(X | \theta_{\overline{m}})$. The posterior probabilities are later used to replace the original $P(\theta_m)$ and $P(\theta_{\overline{m}})$ in the next computation triggered by a new observed response. Recall that an adaptive selection rule (see 1.3.1) is applied to select optimal next item based on the updated value of $P(\theta_m)$ and $P(\theta_{\overline{m}})$. This iteration

will go on till someone stopping rule is met. Local independence is a critical assumption for this referring procedure as required in IRT.

The following sub-sections are devoted to explain the related procedure involved in BDT-CAT with more detailed mathematical formulas.

4.1.1 Bayesian posterior update and the adaptive testing procedure

Within the framework of BDT-CAT, after the j^{th} item is administered, $P(\theta_m)$ and $P(\theta_{\overline{m}})$ can be updated according to Bayesian rule given the response is right ($X_j = 1$):

$$P_j(\theta_m \mid X_j = 1) = \frac{P(X_j = 1 \mid \theta_m) P_{j-1}(\theta_m)}{P(X_j = 1)} \qquad \textbf{Equation 4-1}$$

$$P_j(X_j = 1) = P(X_j = 1 \mid \theta_m) P_{j-1}(\theta_m) + P(X_j = 1 \mid \theta_{\overline{m}}) P_{j-1}(\theta_{\overline{m}}) \qquad \textbf{Equation 4-2}$$

The calculated $P(\theta_m \mid X_j = 1)$ will be used as the new value of $P_j(\theta_m)$, and $P_j(\theta_{\overline{m}}) = 1 - P_j(\theta_m)$. The inference given incorrect response ($X_j = 0$) can be done in similar way. This process continues as more items are administered.

From (Equations 4-1 & 4-2), we know that in BDT-CAT two kinds of prior information are required: the probability of population category $P(\theta)$ and $P(X_i \mid \theta)$ for each item i. The following section will introduce how they can be determined from sample data. Item selection rule will be covered in 4.1.3.

4.1.2 Parameter calibration

Like IRT, calibrations of the examinee and item populations are required with the BDT method. Every item needs to be tested first before they are deposited into item bank with

known parameters and used in the adaptive testing. Calibrated information serves as the prior knowledge, and they assist in inference alone with observed responses.

Two groups of information, as mentioned above, need to be calibrated:

(1) Information about the proportion of population category, *master* and *non-master*, to estimate, $P(\theta_m)$ and $P(\theta_{\bar{m}})$. To do so, passing score need to be determined in advance for target examinee population, and those with higher score than passing score will be regarded as *master*, or *non-master* otherwise. After that, we can count the amount of *master* and *non-master* in the samples, and the corresponding proportions can be determined, which is regarded as estimate of $P(\theta)$. Since there are only two disjoint sets, we have $P(\theta_m) + P(\theta_{\bar{m}}) = 1$.

(2) Information on the probability of correct response to item i given the *master* or *non-master* state: $P(X_i = 1 \mid \theta_m)$ and $P(X_i = 1 \mid \theta_{\bar{m}})$. Because we only study dichotomous items, the probability of incorrect response $P(X_i = 0 \mid \theta)$ is compensatory to $P(X_i = 1 \mid \theta)$ for examinees in the same category. This step still requires a passing score to be determined first so that we can count the statistics of master and non-master categories respectively. For each item, we can sum the amount of master or non-master examinees who answer this item correctly, and the corresponding ratios are estimates for $P(X_i = 1 \mid \theta)$ and $P(X_i = 0 \mid \theta)$.

Both steps above involve only simple computation requiring much less computing burden than that in IRT model.

4.1.3 Item selection rule

Item selection rule determines which item to administer next based on observed responses and prior information. There are several choices available, but we only talk about some typical ones here.

Maximum Information Gain

Maximum entropy reduction, or information gain, is one of the most frequently applied item rule for graph models [9,28]. By applying this selection rule, the item resulting in maximum expected entropy reduction will be selected next. Shannon entropy is applied here to measure the information capacity of current state after j^{th} item is administered:

$$H_j = -P_j(\theta_m)\log(P_j(\theta_m)) - P_j(\theta_{\bar{m}})\log(P_j(\theta_{\bar{m}})) \qquad \textbf{Equation 4-3}$$

For each of the remaining item i, correct and wrong response (considering that only binary item is used) will be assumed respectively so that the corresponding state information entropy can be known by applying Equation 4-1,4-2 and 4-3 above. The expected entropy value H_i then can be determined with the following formula given the known entropy when correct and wrong response are reached:

$$E(H_i) = (-P_i(\theta_m \mid X_i = 1)\log(P_i(\theta_m \mid X_i = 1)) - P_i(\theta_{\bar{m}} \mid X_i = 1)\log(P_i(\theta_{\bar{m}} \mid X_i = 1)))P_i(X_i = 1) + (-P_i(\theta_m \mid X_i = 0)\log(P_i(\theta_m \mid X_i = 0)) - P_i(\theta_{\bar{m}} \mid X_i = 0)\log(P_i(\theta_{\bar{m}} \mid X_i = 0)))P_i(X_i = 0)$$

Equation 4-4

Finally, the item with maximum reduction of information entropy will be selected, and this procedure will guarantee optimal sequence of items be administered, resulting possibly minimum length of test length:

$$\max_i(H_j - E(H_i)) \qquad \textbf{Equation 4-5}$$

This selection rule is used in our BDT-based CAT model, including uni- and multidimensional since it is simply implemented and only light computation required.

Maximum Fisher Information

The Maximum Fisher Information is a popular selection rule used in most IRT-CAT model now. By applying this rule, items are selected to maximize the item Fisher information, which means the item will be chosen that minimizes the expected contribution of an item to the standard error of the ability estimate of an examinee.

Birnbaum [3] has defined the test information function as the sum of information included in each item:

$$I(\theta)=\sum_{i=1}^{n}\frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad \textbf{Equation 4-6}$$

where $P_i(\theta)$ is obtained by observing the item response characteristic curve model at θ, $P_i'(\theta)=\partial P_i(\theta)/\partial\theta$, and $Q_i(\theta)=1-P_i(\theta)$. Birnhaum has shown that Equatino 4-6 is the upper bound on the amount of information that can be yieded by any possible test scoring formula. Since the right-hand side of the equation above is a sum, it can be decomposed into contribution of individual item to the amount of test information.

The item information function is the amount of information contributed by each item, which is given by

$$I_i(\theta)=\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad \textbf{Equation 4-7}$$

The item function is also called Fisher information function due to their equivalence in IRT context. The equations of Fisher information for logistic ICC models are

$$I_i(\theta)=P_i(\theta)Q_i(\theta) \qquad \textbf{1PL or Rasch} \qquad \textbf{Equation 4-8}$$

$$I_i(\theta)=a_i^2P_i(\theta)Q_i(\theta) \qquad \textbf{2PL} \qquad \textbf{Equation 4-9}$$

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} [\frac{P_i(\theta) - c_i}{1 - c_i}] \qquad \textbf{3PL} \qquad \textbf{Equation 4-10}$$

For a test consisting of n items, the test information is the sum of Fisher information of every item. Selecting the item with maximum Fisher information maximizes the contribution to the test information.

Fisher information is used in our unidimensional IRT simulation. To make a fair comparison between IRT and BDT unidimensional models, Fisher information is used for BDT too, but the related parameters, a_i, b_i, and c_i are estimated with the IRT model.

Random administer (RA)

RA is not an adaptive selection rule, but it is regarded as a base line performance for adaptive testing. It simply consider on selecting the next item randomly from the remaining items. RA is only used in our multidimensional simulation.

4.2 Simulation and Evaluation of BDT-UCAT

As mentioned in , our simulation of BDT-UCAT employs real data samples. The simulations are made on two sets of data: (1) a 34 items test on the knowledge of UNIX shell commands and administered to 48 examinees, and (2) a 160 items test on French language administered to 41 examinees. All items are dichotomous: they are either succeeded or failed. Two different models are compared: BDT and 2PL IRT models. In BDT, Maximum Information Gain and Maximum Fisher Information are chosen as selection rules, of which the later one is employed by IRT model too.

Accuracy at the individual level is defined as the agreement between the classification predicted by the model and the actual classifications determined by traditional test with passing score be set.

To prevent over-calibration, the examinee to be studied with adaptive test procedure is not included for calibration.

The performance comparison is based on the simulation of the question answering process. For each examinee, we simulate the adaptive questioning process by feeding the CAT simulation process the examinees' actual responses. After each item administered, the CAT algorithm classifies the examinee as a *master* or *non-master* based on previous answers and according to a pre-defined passing score θ_{ps}, 60% here.

Statistics on data set

Recall that how to do calibration on data samples has been introduced in 4.1.2. In our project, passing score θ_{ps} is set as 60% in all our experiments, although different values were tried during unreported experiments. It means that examinees who answer 60% or more of the total items correctly will be regarded *master*, or *non-master* otherwise. After the passing score is set, we can determine the related statistics on those sample data collected.

The simulations are performed on two sets of data:

- UNIX test: a 34 items test of the knowledge of UNIX shell commands administered to 48 examinees. It is taken from Desmarais [9] and it assesses a wide range of knowledge of the UNIX commands, from the simple knowledge of 'cd' to change directory, to the knowledge of specialized maintenance commands and data processing (e.g. 'awk', 'sed').

- French language exam: a 160 items of French language administered to 41 examinees. This test is from *Formation Linguistique Canada* (FLC), and it is designed by linguistic professionals, covering a wide range of language skills.

Mean scores for the UNIX and French language tests are respectively 53% and 57%, and mean standard deviation per examinee for both test is about 0.16 and 0.07 respectively. Figure 4-1 illustrates the dispersion of scores for each test.

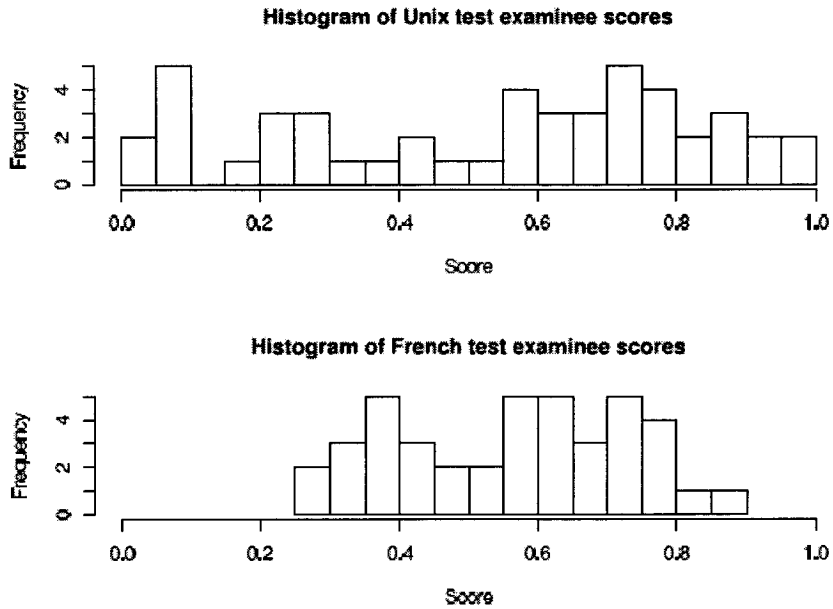

Figure 4-1 Histogram of examinee scores for each test

The following Table 4-1 lists the estimated probabilities of *master* and *non-master* categories of the corresponding examinees in those two real tests: UNIX shell exam and French language exam.

Table 4-1 Proportion of examinee population category (Passing score is 60%)

	Proportion of Master	Proportion of Non-master
UNIX Shell Exam	50.0%	50.0%

French Language Exam	53.66%	46.34%

The conditional probability of right response on the item i given *master* and *non-master* will not be listed here to save space.

Performance measurement

The performance score of each approach corresponds to the number of correctly classified examinees after i items are administered. This score is composed of two parts: the true ability score for the part of the test that has been given so far, and the estimated score of the remaining items. So, the overall ability estimate is a weighted sum of the success to already *given* items and the *estimated* probability of success to the remaining items. That is, if r is the number of items responded, the examinee's estimated score, S, is:

$$S = \frac{\sum_{i=1}^{r} p(X_i = 1) + \sum_{j=r+1}^{n} \hat{P}(X_j = 1)}{n} \qquad \textbf{Equation 4-11}$$

where n is the total number of test item, $X_1 \ldots X_i \ldots X_r$ are known item responses and $X_{r+1} \ldots X_j \ldots X_n$ the remaining items. The probability of a given item, $P(X_i)$, is 1 if the corresponding response to item i is a success, and 0 otherwise. For the remaining items, the respective estimated probability of success, $\hat{P}(X_j = 1)$, can be determined by the following formula:

$$P(X_j = 1) = P(X_j = 1 \mid \theta_m) P(\theta_m) + P(X_j = 1 \mid \theta_{\bar{m}}) P(\theta_{\bar{m}}) \qquad \textbf{Equation 4-12}$$

The category of the examinee is determined by comparing S to the passing score θ_{ps}:

$$\theta = \begin{cases} master, & \text{if } S \geq \theta_{ps} \\ non-master, & \text{otherwise} \end{cases}$$

This procedure results in 100% correctly classified examinees after all test items are administered.

Simulation results

Measuring the performance of each approach is based on a simple metric: the proportion of correctly classified examinee after a specified amount of items is answered by the target examinee population.

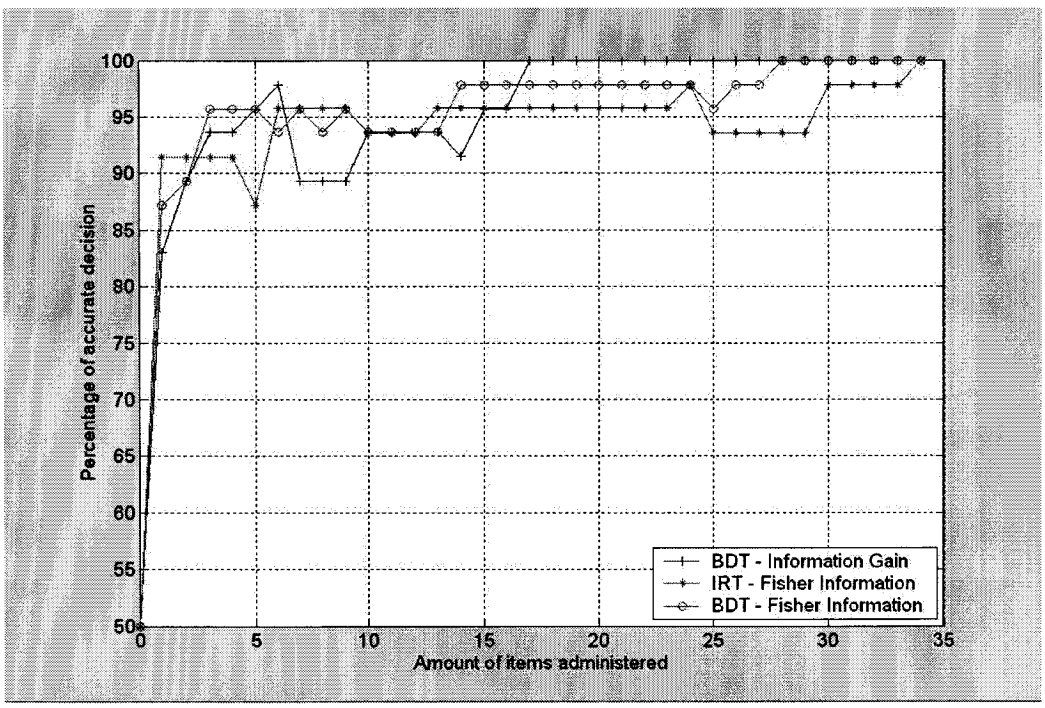

Figure 4-2 Comparison of performance between unidimensional CAT based on BDT and IRT for UNIX tests comprised of 34 items (Passing score is 60%, N=48).

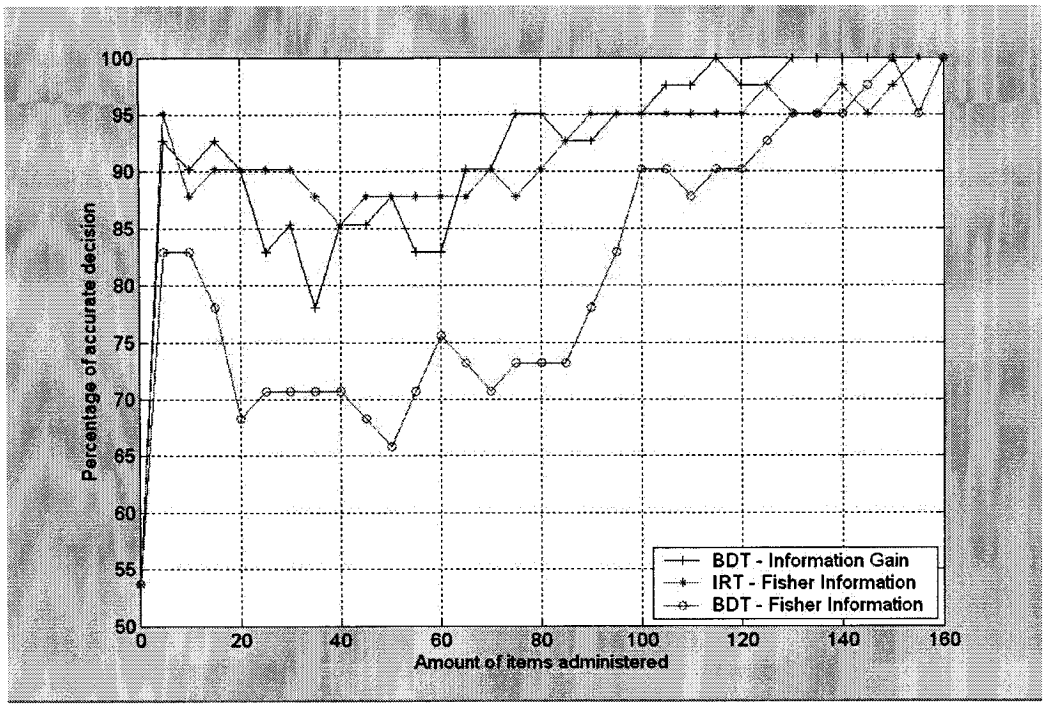

Figure 4-3 Comparison of performance between unidimensional CAT based on BDT and IRT for French language tests comprised of 160 items (Passing score is 60%, N=41).

From these two simulations with real data, we conclude that BDT-UCAT with Information Gain as selection rule can perform with a similar accuracy rate the as the IRT approach with Maximum Fisher Information selection rule, although the IRT solution seems more stable (with flatter curve) than BDT, which can be seen from its flatter curve. From Figure 4-3, we observe that BDT with Fisher Information results especially lower accuracy rate than the other two approaches, which, probably, is the cause of two category, *master* and non-*master*, involved only. Because Fisher information selection rule will put all entropy of the remaining items into consideration, a little change of $P(\theta)$ (or $Q(\theta)$) obtained by the Bayesian inference will bring great influence on the Fisher entropy, which furthermore influence the decision of the next item selection.

4.3 Conclusion

In section 4, BDT-UCAT model is introduced, and its performance is compared to the IRT 2PL model. Simulations with real data show that for the Information Gain selection rule, BDT-based adaptive testing can perform as efficient as classical IRT model in spite of simpler computational complexity. In the next section, we further extend the BDT-CAT model from unidimensional to multidimensional.

Chapter 5. BDT-MCAT Model

In the previous part, we limit our discussion only to UCAT, which models and measures a single latent skill. We can extend the model to multiple skills, which constitute the second part of this project. Readers will be introduced to a more general multidimensional model based on BDT which provides the possibility to simultaneously assess multiple skills using fewer items than that would be required by UCAT one dimension after another. A series of experiments with several dimensions involved are designed to verify the robustness, correctness, and effectiveness of this new model. These simulations are conducted on simulated data sets created with a Monte Carlo procedure. This procedure has been mentioned in section .

5.1 Categories of MCAT

As mentioned before, 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. Moreover, the concept of uni- and multidimensionality can also be applied to items. In this context, there are two kinds of multidimensionality: between-item and within-item, which depends on the items constructed in the exam. An item is called unidimensional if it is intended to assess one latent trait; otherwise, it is a multidimensional item when intended to reflect multiple latent traits. A test constructed with different test batteries, while each battery is composed of purely unidimensional items and doesn't measure the same latent trait is called between-item multidimensional, where multidimensionality exists between items. A test is classified into within-item multidimensional if it contains some items that can measure more than one latent trait simultaneously. Our research work about MCAT focuses on the later since (1) it offers a better basis for MCAT; (2) BDT-MCAT will have no performance gain comparing with BDT-UCAT if only between-item test is used.

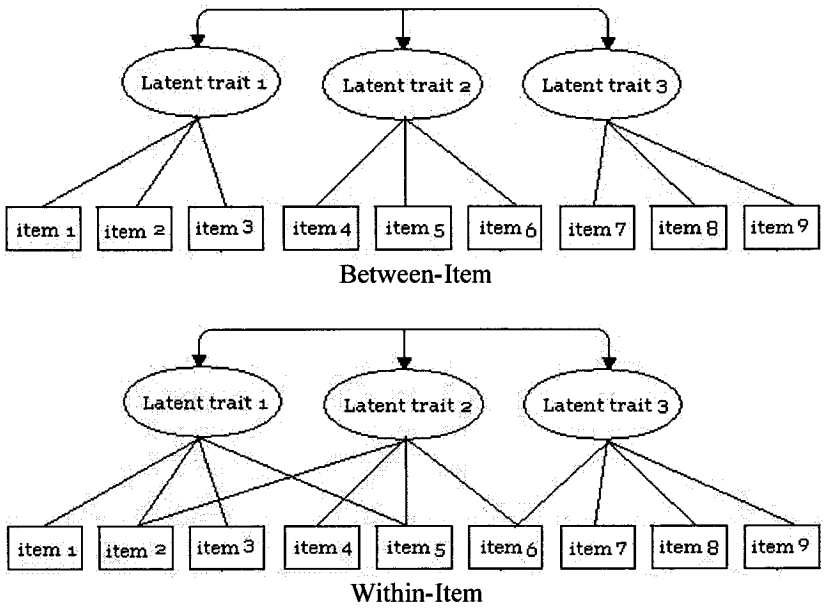

Figure 5-1 Two kinds of multidimensionality

5.2 Theory basis of BDT-MCAT

BDT-MCAT, like its peer BDT-UCAT, requires calibration of prior information about items and examinee population. Therefore, parameter calibration, alone with how the multidimensional adaptive framework can work, will be covered in the coming sections.

5.2.1 Pilot testing and parameter calibration

The calibration in MCAT model involves several dimensions to be measured. Therefore, we assume that there are K dimensions of traits first, and our following discussion will be about these K skills. For BDT-MCAT model, besides estimates of population proportion in each dimension k and the conditional probability of correct response on item i given specific category required in unidimensional model, we also need to know that for each item i, which dimension(s) it measures. That information needs to be determined first before the other two groups can be estimated:

(1) **The correspondence between items and dimensions**. For UCAT, only one dimension is studied at one time, so, by default, all items are related with the same latent trait variable in one test. However, in MCAT, there are multidimensional items with relation to different dimensions. The indication of such correlation existing between item and trait dimension, like that in Figure 5-1, is important because it will let us know to which trait(s) the item contributes. Factor analysis [26] can solve this problem given an observed response vector data when we apply live test or real-data simulation. A data structure like Table 5-1 should be maintained in the model, and we use $c(i,k)=1$ to indicate that item i can be used to measure the dimension k, $k \in [1..K]$; otherwise $c(i,k)=0$. After this information is known, we can do the following step (2) and (3).

Table 5-1 Data structure indicating relation between item and latent trait

	Dimension					
Item#	1	2	...	k	...	K
1	x				x	
2		x				x
...						
i	x			x		
...						

(2) **The proportion of population category in each dimension** $k \in [1..K]$, for example *master* and *non-master* in our discussion here, $P(\theta_m^k)$ and $P(\theta_{\overline{m}}^k)$ after passing score is set for each dimension. In our experiment, passing score is the same for each dimension, which simplify the discussion by ignoring the possible difference existing among dimensions. The outcome of step (1) lets us to classify items into different group they measure, and then the raw score of each examinee in each dimension can be calculated. With known passing score, categories of

examinees can further be determined by a comparison of their raw scores to passing score. Finally, the proportion of different category in each dimension k, $P(\theta_m^k)$ and $P(\theta_{\overline{m}}^k)$, can be estimated.

(3) **Information on the probability of correct response on item i given the *master* or *non-master* state in each dimension**: $P(X_i = 1 \mid \theta_m^k)$ and $P(X_i = 1 \mid \theta_{\overline{m}}^k)$. Because we only study dichotomous items, the probability of incorrect response $P(X_i = 0 \mid \theta^k)$ is compensatory to $P(X_i = 1 \mid \theta^k)$. Based on the outcomes of step (1) and (2), the probability of right response on item i by *master* and *non-master* in each dimension this item i does measure can be estimated.

5.2.2 Adaptive testing procedure

Bayesian probability update

Once an item is administered in MCAT, posterior inference can proceed like in UCAT based on observed response $X_j = 1$ or $X_j = 0$. For example, if the response is right $X_j = 1$, and $c(j,k) = 1$, j^{th} item can be used to measure k^{th} dimension, we have:

Equation 5-1

$$P_j(\theta_m^k \mid X_j = 1) = \begin{cases} \dfrac{P(X_j = 1 \mid \theta_m^k) P_{j-1}(\theta_m^k)}{P(X_j = 1)} & \text{if c(j, k)=1} \\ P_{j-1}(\theta_m^k) & \text{otherwise} \end{cases}$$

where:

$$P(X_j = 1) = P(X_j = 1 \mid \theta_m^k) P(\theta_m^k) + P(X_j = 1 \mid \theta_{\overline{m}}^k) P(\theta_{\overline{m}}^k) \qquad \textbf{Equation5-2}$$

Notice that we only update the related dimensions as indicated by *c(j,k)=1*, ignoring those dimensions that can not be measured by currently administered item i.

Item selection rule

A good selection rule in a multidimensional model will take all dimensions into consideration at once. Here, we choose the Maximum Information Gain rule again since it is easy to be upgraded from its unidimensional version to the multidimensional framework.

$$H_j = \sum_{k=1}^{K} [-P_j(\theta_m^k)\log(P_j(\theta_m^k) - P_j(\theta_{\overline{m}}^k)\log(P_j(\theta_{\overline{m}}^k)] \qquad \textbf{Equation 5-3}$$

The current entropy formula after j^{th} item is administered and responded in Equation 5-3 is like that of UCAT, but aggregates all K dimensions. For each of the remaining item i, the corresponding expected entropy value is calculated given correct and wrong responses, and the expected entropy will be

$$E(H_i) = \sum_{k=1}^{K} \{ [-P(\theta_m^k \mid X_i = 1)\log(P(\theta_m^k \mid X_i = 1) - P(\theta_{\overline{m}}^k \mid X_i = 1)\log(P(\theta_{\overline{m}}^k \mid X_i = 1)]P(X_i = 1) + [-P(\theta_m^k \mid X_i = 0)\log(P(\theta_m^k \mid X_i = 0) - P(\theta_{\overline{m}}^k \mid X_i = 0)\log(P(\theta_{\overline{m}}^k \mid X_i = 0)]P(X_i = 0) \}$$

Equation 5-4

The item that results in maximum entropy reduction globally will be chosen as the next item.

$$\max_i (\mathrm{H_j} - E(H_i)) \qquad \textbf{Equation 5-5}$$

By using this selection rule, we will reach the globally maximum because all the dimensions involved are considered. Besides, multidimensional items are more likely to be chosen since more information can be exposed from the observation of response to items with multiple dimensions, especially at the beginning of adaptive testing procedure.

5.3 Monte Carlo Simulation and Evaluation of MCAT Based on BDT

We compare the performance of BDT-MCAT with BDT-UCAT and IRT-MCAT in this section about their performance.

5.3.1 Overview of simulation design

Our experiments will simulate Wang & Chen's within-item multidimensional simulations [29] based on the following reasons:

- The simulation design and outcome are clearly specified, so repeating the procedure and comparing the results between BDT-MCAT and MIRT-CAT are possible;
- Its design is reasonable, satisfying what are observed from one Taiwan's national exam;
- The work of Wang is well recognized in the MCAT field.

Like Wang & Chen, six different latent traits serve as our dimensions. Nine item banks will be generated, and the following table includes brief information about those banks, including the corresponding amount of item in each bank and which dimension(s) that bank covers.

Table 5-2 Design of item banks for simulation (T refers to Test, and D refers to Dimension)

		Dimension (D)					
Test (T)	# of item	1	2	3	4	5	6
1	200	x					
2	200		x				
3	200			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 are unidimensional, and they are designed for a single dimension D1/2/3/4/5/6 respectively. There are 200 items in each bank. T4/5/6 (with gray shade) are two dimensional item banks, every item being linked to two dimensions. For example, item bank 4 is for D1 and D2. There are 20 items in each bank, much fewer than unidimensional bank.

First we generate examinees with given θ distributions, and next we will generate response to these items from this population of examinee. One thousand examinees are generated in accordance to the specific correlation of those six dimensions of ability, as in Table 5-2. Six hundred samples are selected randomly for calibration and the remaining 400 for validation purpose.

Multi- and unidimensional adaptive tests will be simulated with generated responses. For multidimensional simulation, both uni- and multidimensional item banks T1-9 will be included, with 200 items available for each of the six dimensions. For D1/2/3, 40 items are taken from the within-item multidimensional tests, T4/5/6, and another 160 are randomly chosen from the unidimensional test, T1/2/3 (see Table 5-2 for which test bank correspond to which dimension). For the three remaining dimensions D4/5/6, all the 200 items are

taken from the corresponding unidimensional test, T7/8/9, according to Table 5-2. The final bank structure for the MCAT simulation can be found in Table 5-3 (in page 39).

For the unidimensional simulation, because no multidimensional items can be applied to UCAT model, items from T4/5/6 are not included. So all 200 items from the T1/2/3 banks have to be selected for the D1/2/3, which guarantees that there are the same amount of items, 200, for D1/2/3 in both UCAT and MCAT simulations. The same rule applies to D4/5/6 in UCAT, and the design of item bank is shown in Table 5-4 (in page 52). Those assignment rules above ensure a fair comparison between UCAT and MCAT considering that there are same amount of items available for selection in the measurement of each dimension.

With items and examinees data design defined, we can generate the pseudo-random response of these one thousand examinees on those items available in the banks. The outcome will be one thousand vectors with 1s (right) and 0s (wrong), which can be used for calibration and adaptive simulation in the same way when we process data from real data.

Multidimensional simulations include BDT-CAT with adaptive selection rule and random administer (RA) rule. The result will be presented for comparison. Though no implementation of MIRT has been done in our project, we refer to Wang & Chen's result for a comparison since our data design is identical to theirs.

The procedure for sample data preparation will be introduced first before real simulation results are shown and discussed.

5.3.2 Pseudo-random Examinees, Item Banks, and Response Vectors Preparation

We repeat the simulation of Wang & Chen's within-item multidimensional experiment with dichotomous items in six dimensions [29]. The same data design as Wang & Chen's is shown in Table 5-3, which is constructed based on that in Table 5-2 as explained before . T1 to T3 with 160 items and T7 to T9 with 200 items are unidimensional. T4 to T6 are two dimensional and relatively small in test lengths, with 20 in each one. Therefore, each dimension has 200 items for potential selection, and in total we have 1140 items in the banks. As Wang & Chen explained, the test design in which there are fewer multidimensional items than unidimensional ones is justified on the basis of its similarity to the Basic Competence Test for junior high school students in Taiwan.

Table 5-3 Design of item banks for MCAT simulation (Total: 1140). It is similar to Table 5-2 except for the different number of items included in T1/2/3.

		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

Wang & Chen drew one thousand examinees from the multivariate normal distribution with ability level for each of the six dimensions corresponding to: $\theta^T = [0.2, 0.0, -0.2, -0.1, 0.1, 0.0]$ and standard deviation 1.0, and their correlations are:

$$\Sigma = \begin{bmatrix} 1.0 & .8 & .8 & .3 & .3 & -.4 \\ .8 & 1.0 & .7 & .2 & .2 & -.3 \\ .8 & .7 & 1.0 & .1 & .2 & -.2 \\ .3 & .2 & .1 & 1.0 & .7 & -.2 \\ .3 & .2 & .2 & .7 & 1.0 & -.2 \\ -.4 & -.3 & -.2 & -.2 & -.2 & 1.0 \end{bmatrix} \quad \textbf{Equation 5-6}$$

The correlation matrix indicates that D1 to D3 are moderately to high correlated, and D4 and D5 are moderately correlated. D6 is negatively correlated with other latent traits, which can be regarded as noise and be used to assess the robustness of our model. The close correlations among D1, D2 and D3 are reflected in Table 5-3 with two dimensional T4, T5 and T6. How to create 1000 examinees that meet the given skill levels θ^T and the corresponding correlation matrix Σ will be introduced in the coming sections.

Item parameters a and b

The multidimensional IRT 3PL formula is used to calculate the probability of right response to items in MCAT model given known ability level θ, so for each item we need to design the corresponding a, b and c.

$$P(X_{ni}=1 \mid \theta_n) = c_i + (1-c_i) \times \frac{\exp[a_i^T(\theta_n - b_i 1)]}{1+\exp[a_i^T(\theta_n - b_i 1]} \qquad \textbf{Equation 5-7}$$

There are uni- and multidimensional items in the item banks used by MCAT. For all unidimensional items, we create their parameters a, b and $c(=0.2)$ using the same procedure we have taken in the UCAT simulation. For multidimensional items, b, which is drawn standard normal distribution, and c are set equal across their dimensions; however, discriminating parameter a are drawn from uniform distribution range [0.5, 1.5], and can be different across the item's dimensions.

Examinees with six dimensions of latent trait θ

The following steps illustrate how we can generate the 1000 examinees that satisfy pre-defined multidimensional normal distribution with specified mean and standard deviation, and given correlations among dimensions:

(1) Generate the required number of unidimensional normally distributed examinees; in our simulation, they are $\theta_{jk}(j = 1..1000, k = 1..6)$, of which the index j refers to j^{th} examinee and k refers to k^{th} dimension;

(2) Calculate the Cholesky factor T for the target covariance matrix Σ. The Cholesky factor is an upper triangular matrix which is the "square root" of the covariance matrix.

$$T^T = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.8 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.8 & 0.1 & 0.5916 & 0.0 & 0.0 & 0.0 \\ 0.3 & -0.0667 & -0.2254 & 0.9245 & 0.0 & 0.0 \\ 0.3 & -0.0667 & -0.0563 & 0.6413 & 0.7008 & 0.0 \\ -0.4 & 0.0333 & 0.1972 & -0.0361 & -0.0621 & 0.8915 \end{bmatrix}$$

(3) Post-multiply the independent standard normal variables obtained in step (1), $\theta_{jk}(j = 1..1000, k = 1..6)$, by the Cholesky factor calculated in step (2), T, to give a new data matrix θ'_{jk} with the same size of θ_{jk}. Note that this is a dot product.

$$\begin{bmatrix} \theta'_{1,1} & \theta'_{1,2} & \cdots & \theta'_{1,6} \\ \theta'_{2,1} & \theta'_{2,2} & \cdots & \theta'_{2,6} \\ \theta'_{3,1} & \theta'_{3,2} & \cdots & \theta'_{3,6} \\ \vdots & \vdots & \ddots & \vdots \\ \theta'_{1000,1} & \theta'_{1000,2} & \cdots & \theta'_{1000,6} \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,6} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,6} \\ \theta_{3,1} & \theta'_{3,2} & \cdots & \theta_{3,6} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1000,1} & \theta_{1000,2} & \cdots & \theta_{1000,6} \end{bmatrix} \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,6} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,6} \\ t_{3,1} & t_{3,2} & \cdots & t_{3,6} \\ \vdots & \vdots & \ddots & \vdots \\ t_{6,1} & t_{6,2} & \cdots & t_{6,6} \end{bmatrix}$$

And,

$$\theta'_{jk} = \sum_{p=1}^{6} \theta_{jp} \bullet t_{pk} \qquad \textbf{Equation 5-8}$$

(4) Multiply the desired SD_k (Standard Deviation) and add desired mean μ_k if you want the standard deviation and mean other than 1 and 0.

$$\theta'_{jk} = \theta'_{jk} \bullet SD_k + \mu_k \qquad \textbf{Equation 5-9}$$

θ'_{jk} is the assigned ability level of j^{th} examinee about k^{th} dimension, and $\theta'_{j\bullet}$ is target multidimensional (six dimensions) normal variable we want. We can re-calculate the mean and covariance matrix of these six dimensional variables to verify the correctness of our generating method: $\mu' = [0.1930, -0.0052, -0.2159, -0.0919, 0.1373, 0.0121]$, and the correlation matrix

$$\Sigma' = \begin{bmatrix} 1.000 & .8015 & .7988 & .2828 & .2546 & -.3730 \\ .8015 & 1.000 & .6999 & .2125 & .1826 & -.2715 \\ .7988 & .6999 & 1.000 & .0930 & .1730 & -.1998 \\ .2828 & .2125 & .0930 & 1.000 & .6893 & -.2052 \\ .2546 & .1826 & .1730 & .6893 & 1.000 & -.1864 \\ -.3730 & -.2715 & -.1998 & -.2052 & -.1864 & 1.000 \end{bmatrix}$$

Ignoring the round errors, the results corresponds to the original covariance matrix Σ.

Response vectors of examinees

After we obtain the data sample of examinees' skills θ'_{jk}, we can generate their pseudo random response on items. To do that, we first apply the uni- and multidimensional 3PL IRT formulas (Eq. 3-3 and Eq. 5-7) and to calculate the probability of correct response on uni- and multidimensional items respectively. Then, the following rule is used again to determine the success outcome:

$$X_i = \begin{cases} 1 \text{ if } (P(X_i = 1 \mid \theta) \geq \tau) \\ 0 \text{ if } (P(X_i = 1 \mid \theta) < \tau) \end{cases} \quad \tau \text{ is drawn randomly from uniform distribution}[0,1]$$

The resulting matrix with binary value, for our data is of size 1000×1140 which is the response vectors for examinees. Aggregating the answers by the dimensions, the raw score of each examinee per dimension, of which the size is 1000×6, can be known. Once

again, we can determine the correlation of those 6 latent traits of the examinee population through this obtained raw score matrix. In our data set, the resulting covariance matrix is:

$$\Sigma'' = \begin{bmatrix} 1.000 & .8229 & .8238 & .2523 & .2295 & -.3343 \\ .8229 & 1.000 & .7609 & .1888 & .1681 & -.2473 \\ .8238 & .7609 & 1.000 & .0890 & .1629 & -.1917 \\ .2523 & .1888 & .0890 & 1.000 & .6624 & -.1917 \\ .2295 & .1681 & .1629 & .6624 & 1.000 & -.1790 \\ -.3343 & -.2473 & -.1917 & -.1917 & -.1790 & 1.000 \end{bmatrix}$$

We can see that Σ'' is comparable to Σ and Σ' as expected.

5.3.3 Adaptive Administration vs. Random Administration

We will use the generated data in different simulation studies to evaluate the performance of our model under various conditions. Our first comparison focuses on the item selection rule effect. In this section, BDT-MCAT with Information Gain selection rule will be compared with a random administration (RA) rule.

Recall that for the simulations, 1000 examinees and their corresponding response vectors are generated, among which 600 examinee samples are used for parameter calibration usage, and the remaining 400 for verification purpose. By administering items in an adaptive or random way to the 400 examinees, our accuracy at the individual level was defined as agreement between the classifications predicted by BDT model and the actual classifications determined by the responses generated by Monte Carlo simulation method. The corresponding accuracy measurement is the same as that used in the previous unidimensional model, that is the rate of correctly classified examinees to the total number of examinees for verification. So, unless explicitly mentioned, all the following simulations will employ this standard. However, the number of items will differ between BDT-MCAT with adaptive and RA selection rules, normally 200 for MCAT and 300 for RA considering that adaptive testing requires less items than RA to reach the same performance level.

In our simulation about MCAT with adaptive selection rule, totally 200 items are administered to 400 examinees; therefore there are on average of 33 items for each dimension. However, for MCAT with RA rule, 300 items are administered, on average of 50 items per skill.

Figure 5-2 is the result of our simulation with BDT-MCAT and the Information Gain selection rule, including 200 items and 400 examinees. The x-axis represents the amount of items already administered, and Y-axis is the percentage of correct decision ϕ. Curves of ϕ for all six dimensions are drawn in one graph with different tag. We notice that D1/2/3 have a much better performance compared with D4/5/6 at the beginning, when less than 40 items are administered. It is owed to the fact that there are 40 within-item multidimensional T4/5/6 available for D1/2/3, and they are selected at the beginning since multidimensional items will result in more information reduction. We can observe from Figure 5-2 that after 90 items, about 15 items per dimension on average, all six dimensions reach the 90% accuracy rate level. At then end of 200 items, the average percentage of accurate decision of all six dimensions is about 94%. Considering that there is noise (D6 is negatively correlated with other five dimensions), as discussed in Wang and Chen [29], it is a very good outcome for BDT-MCAT.

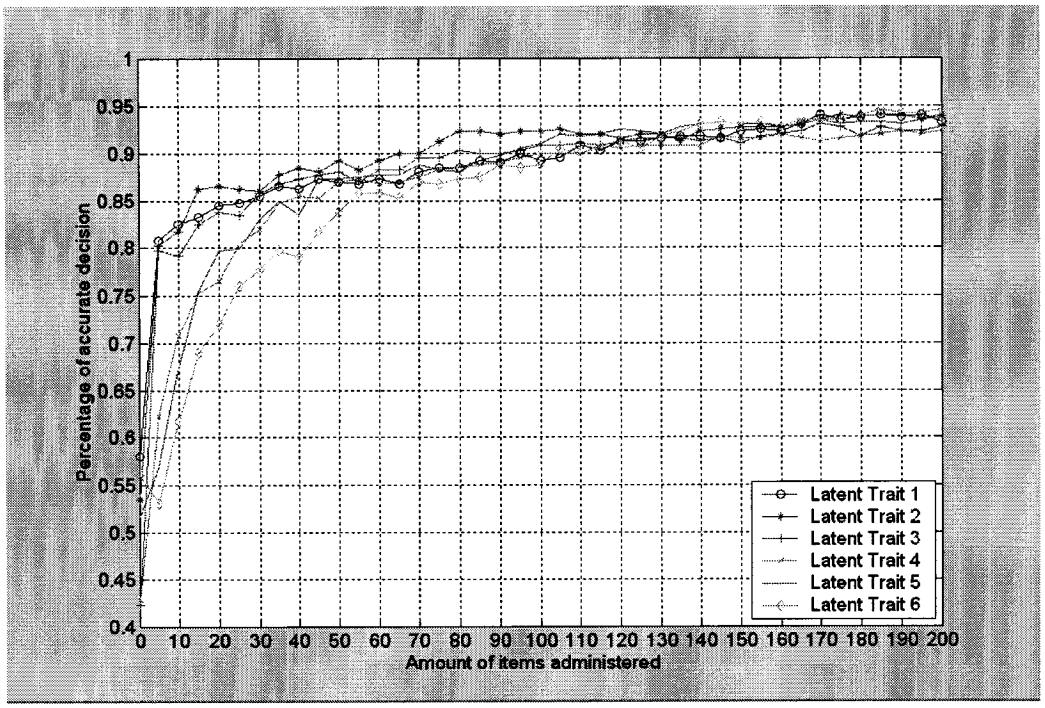

Figure 5-2 Percentage of accurate decisions with items are administered for all six dimensions of latent trait in BDT-MCAT model with Information Gain as selection rule. (N=400)

Next, we turn to the simulation with the RA selection rule over data set. The results for all six dimensions are shown in (Figure 5-3). The following observations can be made over these results:

- All six dimensions develop with a similar speed, unlike in MCAT in which D1/2/3 has a much higher accuracy rate at the beginning. It is the result of random selection, in which all items share the same priority.
- The steepness of curves in Figure 5-2 is bigger than in Figure 5-3, which means that MCAT model reaches higher accuracy and stable state more quickly.
- The curves in Figure 5-3 are less smooth than in Figure 5-2, which indicates that the performance of MCAT with adaptive selection rule is more stable.
- The average accuracy rate for all six dimensions in RA after 300 items administered is about 91.5%, lower than 93.5% of MCAT when 200 items are given. If we observe the Figure 5-2 carefully, we will find that the accuracy rate

increase only 3% when another 100 items are used after the previous 100 items. So, 2% gain while applying adaptive selection rule means that RA rule will cost many more items to reach the same performance level .

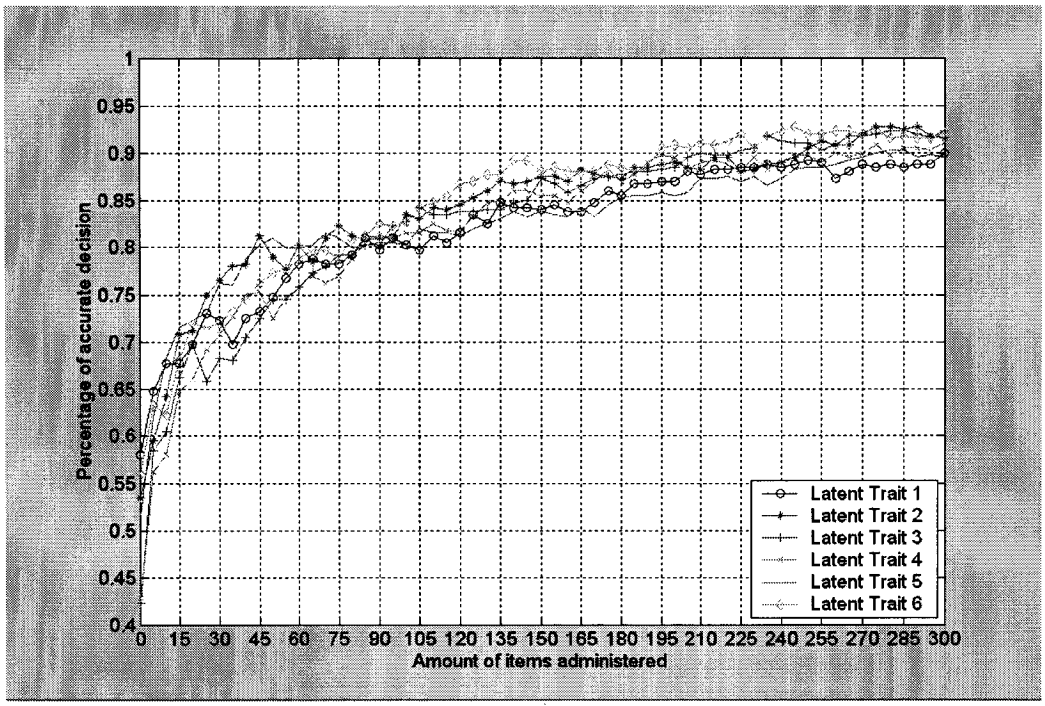

Figure 5-3 Percentage of accurate decisions with items be administered for all six dimensions of latent trait in BDT-MCAT with Random Administration as selection rule (N=400).

To build a clear view of the item selection rule with BDT-MCAT, we average the accuracy rates over all six dimensions and repeat the difference. In Figure 5-4, the average rate between MCAT with Information Gain and RA is compared. Note that the prior curve stops after x=200, however the later one stops after x=300. The first impression with one view is that MCAT with Information Gain performs much better than that with RA. Only after 90 items, MCAT with Information Gain reaches 90%; however, RA needs 245 items, more than two times more items, to get this level. The curve of MCAT with Information Gain has a steep slope at the beginning so that it converges to a stable high performance state more quickly.

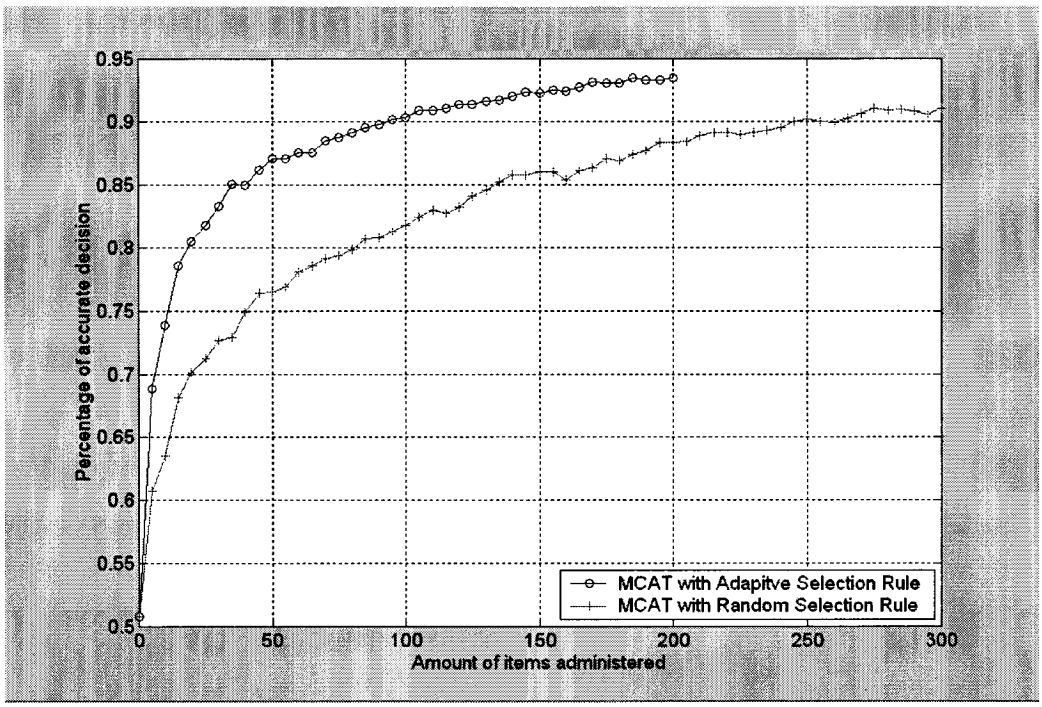

Figure 5-4 Comparison of average accuracy rate for all the six dimensions between MCAT with Information Gain and Random Administer as selection rule. (N=400). For the prior one, only data with 200 items administered is displayed; however, for the later one, 300-item related performance is shown.

In summary: MCAT with Information Gain can benefit us more comparing with that with RA. 90% accuracy rate can be achieved by MCAT with Information Gain with the cost of 90 items, about 60% reduction in item amount required by RA. Furthermore, MCAT with Information Gain can provide a stable and robust solution even when noise is introduced.

This simulation will be regarded as S0 for later reference.

5.3.4 MIRT vs. BDT-MCAT

BDT-MCAT is now compared with the dominating MIRT-CAT model on performance. Since we didn't implement MIRT model in our project, and our simulation is based on the

same data samples as Wang and Chen's [29], naturally we choose to compare our result with theirs.

Wang and Chen's result on MIRT

Wang and Chen's result about the MIRT simulation is based on the same data samples as ours (see Figure 5-5), which is the comparison between multidimensional CAT with adaptive and RA selection rules, is reproduced here so that we can roughly compare our outcome (see Figure 5-4) with theirs. However, they measured the degree of test reliability, r, which is a different measure to ours (classification accuracy). Test reliability is defined as the squared 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.

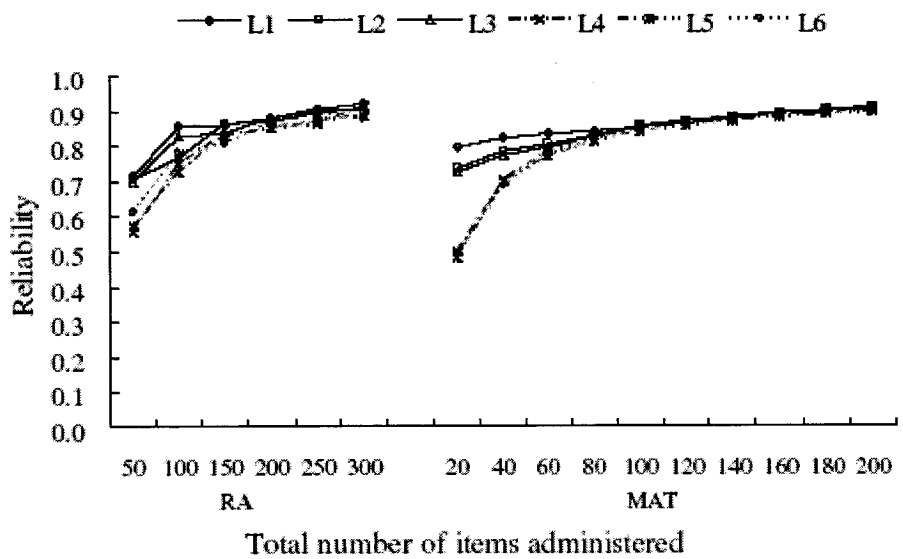

Figure 5-5 This figure is from ([29]), and it is the comparison between MIRT-CAT(MAT shown in the graph) with adaptive and RA selection rules. Its simulation conditions are the same as ours, including the item and examinee population setting.

Through observing the figure above, some conclusions can be made:

- Performance of MIRT-CAT with adaptive selection rule is much better than that with RA selection rule, which is in agreement with our results though different assessment standards are used;
- Curves of MCAT models, whether based on MIRT or BDT, are smooth, compared to those with RA;
- The dimensions (D1/2/3) that contain multidimensional items yield greater accuracy than the other dimensions;

Transformation of Wang and Chen's result for comparison purpose

Since reliability, instead of classification rate, was used in Wang and Chen's project, we cannot compare our outcomes with their MIRT simulation directly. To make them comparable, it is necessary for us to transform Wang and Chen's results based on reliability into the corresponding values about classification accuracy first. How the transformation works is described as below:

- Generate the "real skill value" of examinee samples, $\theta_{jk}(j=1..1000, k=1..6)$, and their pseudo random responses. In fact, this step can be ignored since it has been done in our previous simulation when we prepare data samples like Wang and Chen's (please refer 5.3.2).
- Generate the "estimated skill value" of those examinee samples mentioned in the previous step (1), $\dot{\theta}_{jk}(j=1..1000, k=1..6)$, based on the reliability value r obtained by Wang and Chen in their MIRT simulation (see Figure 5-5). The following known formula is applied here to generate a new variable $\dot{\theta}$ based on a known variables θ, and their desired correlation value is r:

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

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

- Generate the pseudo random responses according to the "estimated skill value" $\dot{\theta}$ 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.
- An accurate decision is reached if the category of examinee with $\dot{\theta}$ 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 5-5 can be mapped to a corresponding classification rate value, and the result is shown in Figure 5-6. Note that we only transform the curves of MIRT here.

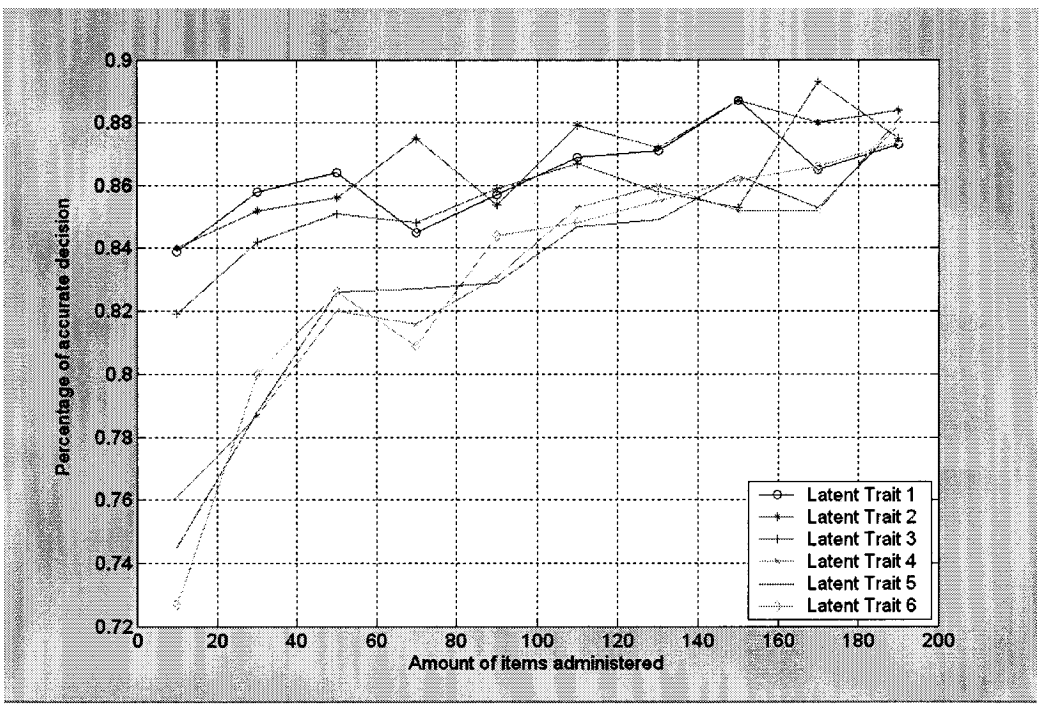

Figure 5-6 Percentage of accurate decision with items be administered for all six dimensions of latent trait in MIRT-CAT as selection rule

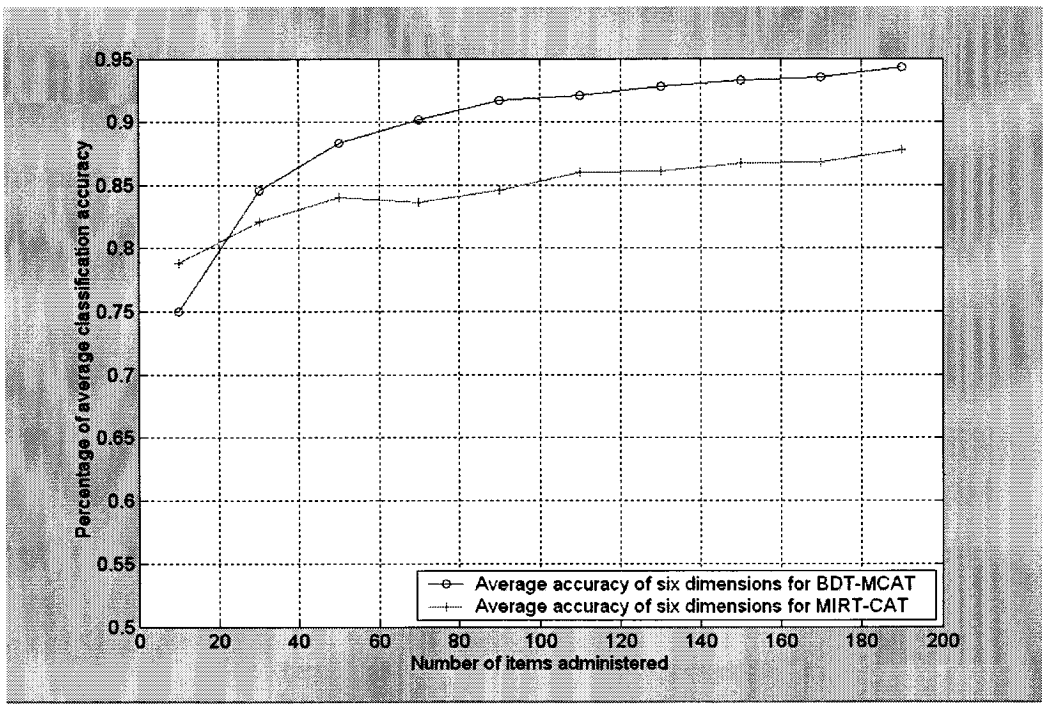

Figure 5-7 Comparison of average accuracy rate for all the six dimensions between BDT-MCAT and MIRT under the same experimental conditions.

Comparison conclusion

With the transformed outcome, we can compare the performance of BDT-MCAT and MIRT directly and fairly. By comparing the two sets of results in Figure 5-2 and Figure 5-6, we find that they share the same trend: those dimensions with multidimensional items D1/2/3 reach higher level more quickly than D4/5/6. At the end of simulation when 190 items are administered, MIRT reaches around 88% on average for all six dimensions, but it is about 94% for BDT-MCAT, seen in Figure 5-7. If we put into consideration the implementation and computation burden, BDT-MCAT would be preferred over MIRT.

5.3.5 UCAT vs. MCAT (all based on BDT)

BDT-UCAT has been discussed in section 4. In this part, we will compare it with the MCAT version to see if MCAT really works better than UCAT. To do that, we will apply

UCAT and MCAT with the same selection rule, Information Gain, to measure the same examinee population with the same dimensions of latent traits. Our comparison will be based on the efficiency of both models – the amount of items required to reach the same average accuracy level of all dimensions.

As explained in 5.3.1, no multidimensional items are included in the UCAT simulation, so only bank T1/2/3/7/8/9 are available as shown in Table 5-4. For each dimension of trait to be measured, there are still 200 items maintained, and in total there are 1200 items for UCAT simulation, 60 items more than 1140 in MCAT example mentioned before.

Table 5-4 Design of item banks for UCAT simulation. Only tests with unidimensional items are used here, and T4/5/6 appearing in Table 5-2 are removed.

		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 data of UCAT on each skill can only be obtained with unidimensional adaptive testing be administered separately in each dimension. Therefore, six dimensions of UCAT simulation run independently given that no correlation among those six dimensions of traits is known. Each simulation will only use the items assigned to them. Like in MCAT, 600 examinees are reserved for calibration, and the remaining 400 for validation. The simulation result of UCAT model is presented in Figure 5-8, in which curves of six dimensions are displayed with different color. After 40 items are given, about 20% of the bank size, all six dimensions reach 90% accuracy rate. We can compare it with MCAT (Figure 5-2) and RA (Figure 5-3). Unlike MCAT, UCAT curves develop with similar speed, and the relative distance is smaller than that in MCAT. It is the result

of independent measurement for each dimension, and no priority exists among the six dimensions. However, in MCAT, since multidimensional items are preferred, the related dimensions will reach higher performance more quickly.

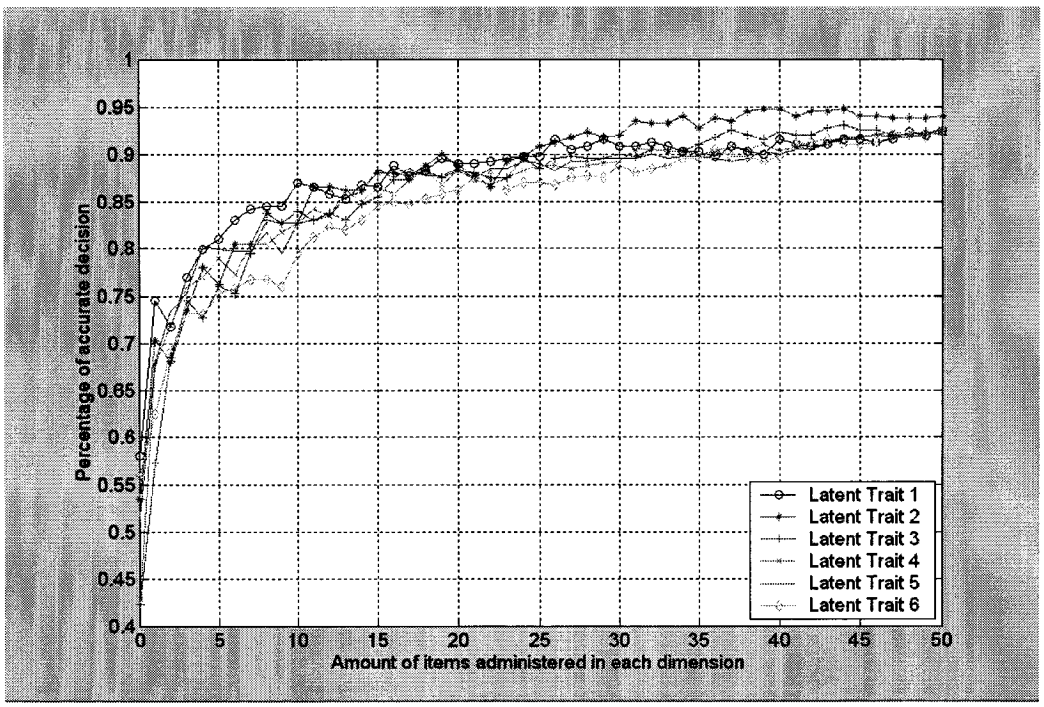

Figure 5-8 Percentage of accurate decisions for each dimension of UCAT model with Information Gain as selection rule. The x-axis indicates the amount of items administered for each dimension.

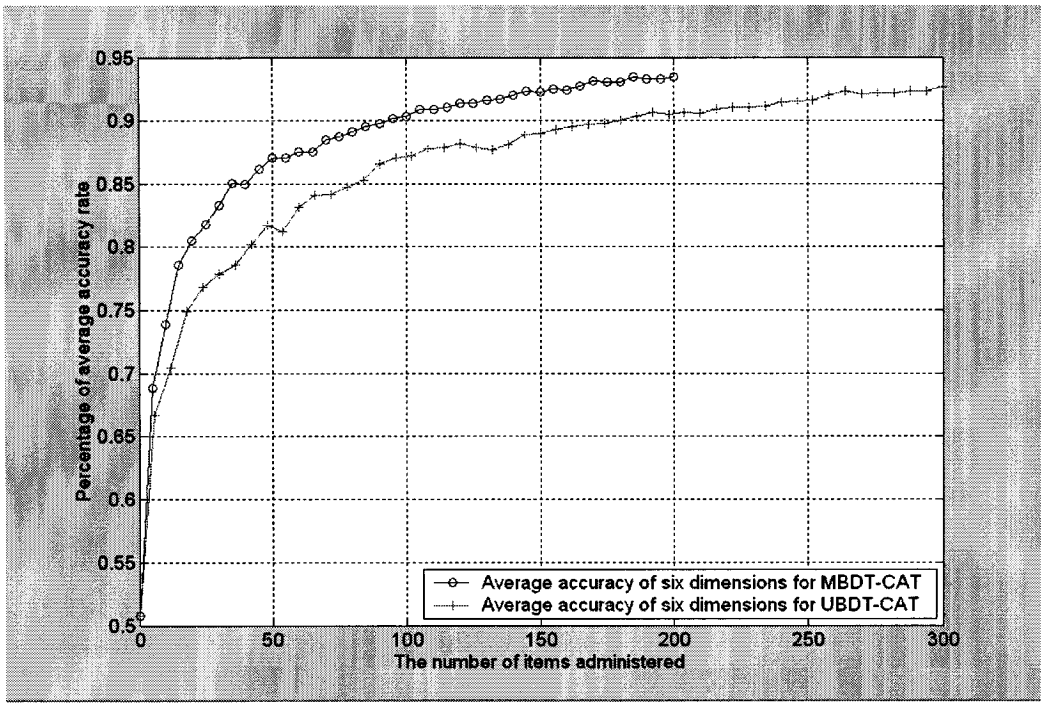

Figure 5-9 Comparison of accuracy rate between BDT-UCAT and –MCAT with the same selection rule, Information Gain.

To compare the overall performance between MCAT and UCAT, we calculate their average accuracy rate over all six dimensions, and the result is shown in Figure 5-9. Note that the steps (or x-axis value) between the adjacent points on the curves of MCAT and UCAT are different: 5 for MCAT, and 6 for UCAT. It is the result of different sampling methods applied in both models. In the MCAT simulation, we observe the performance with every 5 items be administered. However, in the UCAT experiment, the performance data in each dimension is collected when every one more item be given. So, the average of performance for all 6 dimensions of UCAT model is the effect of the total amount of items required by 6 dimensions. So, for every y-axis value of the UCAT curve, it corresponds to multiple of 6 on the x-axis, 6, 12, 18, 24.... From Figure 5-9, we can conclude that MCAT performs better than UCAT model when the same amount of items is consumed. Fox example, in our simulation, MCAT needs about 90 items to reach 90% accuracy rate, but

UCAT needs nearly 180 items. Therefore, about 60% of items are saved in our MCAT experiment.

For a specific object examinee population registering the same exam, correlation exists among the various dimensions of latent traits considering their similar knowledge background. If we can detect this relation, and design the corresponding multidimensional items, we will achieve higher efficiency on test but with no loss of measurement accuracy. Therefore, MCAT is indeed preferred for UCAT.

5.3.6 Conclusion on MCAT based on BDT

Our discussion on MCAT based on BDT above shows that this model is really potential from the following aspect:

- Solid foundation – Bayesian theorem.
- Clear model structure, and easy to understand.
- Easy implementation and application.
- Less computing burden comparing to MIRT approach.
- Comparable performance to its multidimensional competitive MIRT.
- Much better performance than RA and UCAT approaches.

5.3.7 Efficiency of Monte Carlo simulation for MCAT model

The advantageous points of Monte Carlo have been listed above. Though our research focuses on the reasonability and correctness of model, and we can afford tolerable time and resource consumption, we still need to pay enough attention on the efficiency of Monte Carlo approach. The large scale of data generation and decision computation required by our model verification determines that it will be time-consuming simulation. In our experiment, generating one thousand examinees and their response vectors on about

one thousand and two hundreds items costs about 30 seconds. And for adaptive administration, 60 seconds is the cost of simulating one examinee's response to 200 items selected adaptively from the item pools with about nearly one thousand and two hundreds items. We conclude from our simulations that the size of item banks will greatly influence the whole performance.

Chapter 6. Discussion and Conclusion

We started by listing the requirements of an intelligent learning environment and narrowed our discussion to the need for effective assessment model, and CAT is regarded as a very appropriate candidate. Unidimensional CAT based on BDT is the starting point of our research work. We introduced the theory and means to parameterize the model, and assessed its performance. Simulation with real data and Monte Carlo simulation are performed. BDT-CAT with adaptive selection rule can achieve much better performance than that with RA, reaching high accuracy rate with less items. The comparison to IRT-CAT model shows that BDT solution results in similar accuracy rate in decision, but requires less computation due to the saving of iterative algorithms included in the maximum likelihood estimate. Then, this one-dimension model is upgraded to multidimensional version, which can be used to measure more than one latent trait through single exam. Calibration and adaptive procedure are introduced under the umbrella of MCAT like what we do in UCAT part. Intensive simulation follows the introduction of MCAT model to verify that this model's performance is accurate, robust, and trustable. Not only our BDT-MCAT model with Information Gain performs much better than BDT-MCAT with RA, BDT-UCAT, UIRT, but also, at least, as well as MIRT approach. While noise is introduced to sample data, our MCAT model still works well, which is needed for practical application. Therefore, Bayesian decision rule is proven to be effective and efficient for building CAT model, including uni- and multidimensional.

Based on the list of dimension or quality mentioned in section 1 about the evaluation of student modeling, we make a concise conclusion on the merit of our model through our study:

- Flexibility and expressiveness: Single or more than one latent trait can be measured by our proposed model. Relatively fine-grained assessment is achievable with the application of our multidimensional CAT system, to the extent that

several factors can be put into consideration simultaneously and it is critical for AI-based learning system. The result is easily explainable, and profile of examinees' ability level can be reached directly.

- Cost of model definition: Being a data driven approach, CAT based on BDT can waive the knowledge engineering effort. This feature enables BDT based model to be built and updated without much cost, and the final model is more general for application.
- Scalability: The number of concepts/skills and test items that can be modeled is unlimited theoretically. The underlying multidimensional model of BDT allows good scalability to large scale of tests with several ability dimensions, which is an important feature to allow for flexible customization in intelligent learning environment.
- Cost of updating: Since the calibration of items is light, new items can be imported into item banks with low cost. Additional new skills can be introduced within our BDT-MCAT model rapidly.
- Accuracy and reliability of prediction: High accuracy and reliability are provided by BDT-CAT model, including uni- and multidimensional one. Fewer items comparing with traditional test are required to obtain the same performance level, and the model works as well when noise is introduced.
- Reliability and sensitivity to external factors: Since no real sample data are involved in our simulation, we have no opportunity to check the performance of MCAT model under non-ideal application environment. However, considering that our pseudo data are generated randomly and our model still works with satisfactory results, we have confidence on the running accuracy rate of our MCAT model while some so-called noise does exist in the target application.
- Mathematical foundations: Sound theories play as the foundation as our BDT-CAT model. There, it is considered, at least, comparable candidates over ad hoc dominating models, such as IRT. Besides, BDT approach requires much less

computation than IRT, which is much more attractive feature relative to IRT model.

- Approximations, assumptions, and hypothesis: Like other models in the complex field of cognitive and skill modeling, our BDT-based model requires one strong assumption like IRT, local independence.

BDT-CAT is a simple model comparing with currently widely applied IRT, including its theory foundation, parameter calibration, adaptive testing procedure, and implementation. However, it has been proved a potential candidate to build up an adaptive testing system to provide quick and fine-grained measurement. It can run independently as a testing tool as requested in traditional education environment, or can be embedded into modern ITS or Study Guide System as a measurement module, cooperating with other parts to provide an advanced learning environment.

Even though important conclusions were achieved through this project, many problems still remain open for further research:

- For multidimensional model, the amount of multidimensional items occupies a small percent in the item banks, so item exposure will be non-ignorable problem especially when multidimensional items are preferred by the selection rule. Although RA can dampen the problem, we have to make a balance between efficiency and exposure problem.
- An efficient way to determine the correlation between each item and latent ability dimension(s) though there are such software available currently.
- Multidimensional items need work together with unidimensional items to provide finer grained assessment, but how to balance their relative percentage in the item bank to achieve optimal outcome.
- What kinds of internal and external factors will greatly influence the performance of BDT-CAT model?

- Can the BDT-MCAT model works well for test with polytomous items?
- What kind of stopping rule is suitable for BDT-CAT model if it is required?
- More study on the reliability of model is necessary in the future.

Other than known measurement field, BDT framework can also be applied in other situations where binary or discrete classification based on a sequence of observation is requested.

Through this project, we experience the classical Bayesian inference procedure - calibrate a-priori information, observe what happen, and combine them to do inference. Besides, in this project, we learn to verify our assumption and proposal through designed simulations, not only with real data but large-scale pseudo-random sample data. Furthermore, with the outcome of simulations, we have chance to compare, explain it and reach conclusion. More importantly, new guess comes out to guide our further discovery, which directly expose ourselves to more attractive findings.

References

[1] Allen, M. J. and Yen, W. M., Introduction to Measurement Theory *Waveland Press, Inc., IL*, 1979.

[2] Almond, R. G. and Mislevy, R. J., Graphical models and computerized adaptive testing *Applied Psychological Measurement*, vol. 23, no. 3, pp. 223-237, 1999.

[3] Birnbaum, A., Some latent trait models and their use in infering an examinee's ability *Statistical theories of mental test scores*, vol. pp. 397-472, 1968.

[4] Bloom, B.S. , The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, *Educational Researcher*, vol. 13, no.4, pp.4-16, 1984

[5] C.A.S.Coelho , S.Pezzulli, M.Balmaseda, F.J.Doblas-Reyes, and D.B.Stephenson, Forest calibration and combination: A simple Bayesian approach for ENSO *Journal of Climate*, vol. 17, 2004.

[6] Conati, C., Gertner, A., and VanLehn, K., Using bayesian networks to manage uncertainty in student modeling *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 371-417, 2002.

[7] Conejo, R., Guzman, E., Millan, E., Trella, M., Perez-de-la Cruz, J.L., and Rios, A., SIETE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 29-61, 2004.

[8] Desmarais, M. C., Fu, S., and Pu, X., Tradeoff analysis between knowledge assessment approaches *Artificial Intelligence in Education (AIED 2005)*, 2005.

[9] Desmarais, M. C., Maluf, A., and Liu, J., User-experience modeling with empirically derived probabilistic implication networks *User Modeling and User-Adapted Interaction*, vol. 5, no. 3-4, pp. 283-315, 1995.

[10] Desmarais, M. C. and Pu, X. i., Computer adaptive testing: Comparison of probabilistic network approach with item response theory *Proceedings of the 10th International Conference on User Modeling (UM'2005)*, 2005.

[11] Fu, S. and Desmarais, M. C., Computerized adaptive testing: A comparison of item response theory, Bayesian Networks and Bayesian decision theory *Research report of Ecole Polytehnique de Montreal*, 2004.

[12] Fu, S. and Desmarais, M. C., Multidimensional computerized adaptive testing based on Bayesian decision theory *Research report of Ecole Polytechnique de Montreal*, 2005.

[13] Gardner William, Katherine Shear, Kelly J. Kelleher, Pajer Kathleen A., Mammen Oommen, Buysse Daniel, and Frank Ellen, Computerized adaptive measurement of depression: A simulation study *BMC Psychiatry*, vol. 4, no. 13, 2004.

[14] Gardner William, Kelleher Kelly J, and Pajer Katheleen A., Multidimensional adaptive testing for mental health problems in primary care *Medical Care*, vol. 40, no. 9, pp. 812-823, 2002.

[15] Gershon, R. C., Test Anxiety and Item Order: New Concerns for Item Response Theory *Chapter 11 in M Wilson (Ed.) Objectitve Measurement: Theory into Practice.* vol. 1, no. Ablex, Norwood NJ, 1992.

[16] Hambleton, R. K., Swaminathan, H., and Rogers, H. J., Fundamentals of Item Response Theory *Sage Publications*, 1991.

[17] Kreitzberg, C., Stocking, M.L. & Swanson, L. Computerized adaptive testing: Principles and directions *Computers and Education*, vol.2, no.4, pp.319-329

[18] Lord, F. M. , A theory of test scores *Psychometric Monograph (IA: Psychometric Society)*, vol. 7, 1952.

[19] Mayo, M. and Mitrovic, A., Optimising ITS behaviour with bayesian networks and decision theory *International Journal of Artificial Intelligence in Education*, vol. 12, no. 124-153, 2001.

[20] McDonald, R. P., A basis for multidimensional item response theory *Applied Psychological Measurement*, vol. 24, no. 2, pp. 99-114, 2000.

[21] R.Hanson and J.Stutz, P. C., Bayesian classification theory *Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch*, 1991.

[22] Rasch, G., Probabilistic models for some intelligence and attainment tests. *Copenhagen: Danish Institute for Educational Research*, 1960.

[23] Rossi, P. E. and Allenby, G. M., Bayesian statistics and marketing *Marketing Science*, vol. 22, no. 3, pp. 304-328, 2003.

[24] Rudner, L. M., An examination of decision-theory adaptive testing procedures *Proceedings of American Educational Research Association*, vol. pp. 437-446, 2002.

[25] Rudner, L. M., The classification accuracy of measurement decision theory *National Council on Measurement in Education, Chicago*, vol. 2003.

[26] Sharma, S., Factor analysis *Applied multivariate techniques, John Wiley & Sons Publisher*, vol. pp. 90-185, 1996.

[27] van der Linden, W. J. and Hambleton, R. K., Handbook of modern item response theory *New York: Springer-Verlag*, 1997.

[28] Vomlel, J., Bayesian networks in educational testing *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, vol. no. 12(Supplementary Issue 1), pp. 83-100, 2004.

[29] Wang, W.-C. and Chen, P.-H., Implementation and measurement efficiency of multidimensional computerized adaptive testing *Applied Psychological Measurement*, vol. 28, no. 5, pp. 295-316, 2004.

[30] Yi, Q. and MIchael, L., Simulating Nonmodel-Fitting Responses in a CAT Environment *ACT Research Report Series*, Oct, 1998.