This is to certify that the
dissertation entitled

THE EFFECT OF FITTING A UNIDIMENSIONAL IRT MODEL TO MULTIDIMENSIONAL DATA IN CONTENT-BALANCED COMPUTERIZED ADAPTIVE TESTING

presented by

Tian Song

has been accepted towards fulfillment
of the requirements for the

Doctoral degree in Measurement and Quantitative Methods

Mark D. Reckase
Major Professor's Signature

July 5, 2010
Date

MSU is an Affirmative Action/Equal Opportunity Employer

THE EFFECT OF FITTING A UNIDIMENSIONAL IRT MODEL TO MULTIDIMENSIONAL DATA IN CONTENT-BALANCED COMPUTERIZED ADAPTIVE TESTING

By

Tian Song

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2010

ABSTRACT

THE EFFECT OF FITTING A UNIDIMENSIONAL IRT MODEL TO MULTIDIMENSIONAL DATA IN CONTENT-BALANCED COMPUTERIZED ADAPTIVE TESTING

By

Tian Song

This study investigates the effect of fitting a unidimensional IRT model to multidimensional data in content-balanced computerized adaptive testing (CAT). Unconstrained CAT with the maximum information item selection method is chosen as the baseline, and the performances of three content balancing procedures, the constrained CAT (CCAT), the modified multinomial model (MMM), and the modified constrained CAT (MCCAT), are evaluated in terms of measurement precision, item pool utilization and item exposure control. Three simulation factors are considered: (1) multidimensional structure; (2) ability distribution; and (3) difficulty level of content areas. Simulation results show that overall the content balancing methods are similar to or even better than the maximum information method in terms of measurement precision, especially when the content areas have uneven difficulty levels. However, there is no significant difference in item pool usage and item exposure control. Finally, overall the three content balancing methods perform very similarly, but MMM has the most efficient item pool usage.

Dedicated to my parents, and my husband Chen

ACKNOWLEDGEMENTS

There are many people who have helped me tremendously in my dissertation. I would like to express the deepest appreciation to my advisor, Professor Mark Reckase, for his kindness, support, and exceptional guidance. I also thank other members of my dissertation committee, Professor Tenko Raykov, Professor Sharif Shakrani, and Professor Amelia Wenk Gotwals, for their valuable constructive comments and suggestions. Without their help, this work would not have been possible.

I also thank Dr. Steven Pierce, Dr. Connie Page, Dr. Brian Silver and the Center of Statistical Training and Consulting at Michigan State University who have financially supported me for my last two years of doctoral study. I have benefited greatly from the conversations with them about my consulting work, my research, and my career development.

Finally, I would like to thank my parents for making this all possible. They taught me the value of knowledge and hard working. They have been encouraging and supporting me to pursue my dreams over the years. I also thank my husband for his love and unconditional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Over the last few decades, interest in computerized adaptive testing (CAT) has grown considerably. As an alternative to a conventional paper-and-pencil test, it uses a computer to present test items and score responses. In CAT, each examinee is presented with an individually tailored test. Generally, an adaptive test begins with an item with medium difficulty. If the examinee answers it correctly, then he gets a more challenging item; otherwise he gets an easier item. After each response, the examinee's ability is estimated, and the next item that is the most appropriate for the examinee is selected based on the current ability estimate. This process continues until there is enough information to place the person on the ability scale with a specified accuracy, or until a fixed number of items have been administrated (Green, Bock, Humphreys, Linn & Reckase, 1984). A major advantage of CAT is that it provides more efficient and precise ability or latent trait (θ) estimates (Weiss, 1982).

In the current CAT applications, however, there are also a number of challenging issues, such as dimensionality (Green et al., 1984; Liu, 2007; Weiss & Suhadolnik, 1982), content balancing (Kinsbury & Zara, 1991), and item overexposure (Chang & Ying, 1999; Wainer, 2000). This study attempts to address the first two issues.

1.1 Dimensionality

Item response theory (IRT) is a family of mathematical models in which the interactions of a person with test items can be adequately represented by a probabilistic

expression. It plays a central role in almost every aspect of CAT, such as item pool calibration, item selection, and proficiency estimation.

Most IRT models assume that examinees' responses to the items on a test can be accounted for by a single latent trait (Lord, 1980). However, this assumption may rarely hold since most sets of items are not strictly unidimensional and require multiple abilities to obtain a correct response (Reckase, 1985). For example, a mathematical story problem requires reading skills to transform word problem to equations as well as mathematical knowledge to find a solution to the equations.

Due to their popularity and simplicity, most computerized adaptive testing programs use unidimensional IRT models. However, when the unidimensionality assumption is violated, the application of CAT could be seriously affected. If an item pool is composed of items that require a complex of abilities to answer correctly, examinees may be administrated different sets of items that measure completely different combinations of skills. Weiss and Suhadolnik (1982) examined the robustness of adaptive testing to the violation of the unidimensionality assumption. The authors used a factor analysis model to generate multidimensional data and then performed unidimensional adaptive tests with the maximum information item selection strategy. The results showed that as multidimensionality increased, the estimated ability parameters deviated more from their true (first-factor) θ values.

1.2 Content Balancing

Content balancing is a practical consideration in CAT. Unlike traditional paper-and-pencil tests which are built on test blueprint or content specifications, adaptive tests do

not follow content specifications during item selection. Therefore, examinees may be administrated different distribution of items by content area. For example, in a math test, one examinee might receive a test consisting entirely of arithmetic items, and another might receive a test entirely of geometry items. This lack of content comparability could pose a threat to the validity of scores, and may not be acceptable to test takers and test score users. In addition, in licensure and certification testing, if the items administered do not cover all the content areas the test plan requires, it may bring legal challenges to the test (Kinsbury and Zara, 1991).

Since Green et al. (1984) first noted the need to content balance adaptive tests, a number of procedures have been proposed to control the content specifications. Wainer and Kiely (1987) suggested using testlets that are content balanced beforehand instead of items; Kingsbury and Zara (1989) proposed a constrained CAT (CCAT) which selects the most informative item from the content area farthest below its ideal administration percentage; Chen and Ankenmann (2004) developed a modified multinomial model (MMM) to satisfy the practical constraint of content balancing; and Leung, Chang and Hau (2000) proposed a modified constrained CAT (MCCAT) based on Kingsbury and Zara's method.

1.3 Purpose of the Study

The purpose of this study is to investigate, with control of content specifications, the effect of fitting a unidimensional IRT model to multidimensional data in CAT. Unconstrained CAT with the maximum information item selection method is considered

as the baseline, and the performances of three content balancing procedures, CCAT, MMM, and MCCAT, are evaluated. Specifically, the research questions are:

(1) What is the effect of fitting a unidimensional IRT model to multidimensional data in a content-balanced CAT? Does the estimation of ability become more or less accurate when content balancing procedures are applied?

(2) Which content balancing procedure performs the best in terms of ability recovery, item pool usage, and item exposure control?

The present study contributes to the literature in three important ways. First, the robustness of CAT to the violation of unidimensionality assumption and content balancing are jointly considered. Previous study by Ackerman (1991) investigated the effects of fitting a unidimensional IRT model to two-dimensional data in an unconstrained CAT. The results suggested that the estimated unidimensional discrimination values increased when an item's θ_1, θ_2 composite became similar to the composite of the unidimensional calibrated θ scale. These items thus had a greater chance of being selected and administrated in an adaptive testing using the maximum-information item selection strategy. The study also found that if a CAT item pool consisted of items from several content areas measuring dissimilar θ_1, θ_2 composites, examinees at different ability levels might receive different proportion of items from the content areas. Since a balance across content areas is a requirement in practical CAT programs, it is interesting to see how examinees' proficiency would be recovered under this practical constraint. Moreover, in previous literature, after we impose the content constraint in the unidimensional CAT, the measurement precisions are found to be

comparable to the unconstrained maximum information method, with mean squared errors (MSE) of θ slightly higher (Leung, Chang & Hau, 2000; Cheng, Chang & Yi, 2007). Now given the assumption of unidimensionality is violated, applying content balancing procedure might improve measurement efficiency by insuring adequate representation of each dimension. Therefore, it is interesting to investigate the joint effects of those two issues.

Second, three content balancing methods, CCAT, MMM, and MCCAT, are compared in a different context from previous studies. Most existing studies (e.g., Leung, Chang and Hau, 2003a; Leung, Chang and Hau, 2003b) focus on unidimensional data. The results generally showed that the three methods had similar effects on measurement efficiency and item pool utilization. The present study extends the comparison to multidimensional data, where items are assumed to require multiple abilities to answer correctly.

Third, most of the studies on multidimensionality focus on the simple two-dimensional case. In our study, we start with the two-dimensional case, and then turn to the more complicated three-dimensional case. In this way, the results might be generalized to higher dimensional spaces.

Chapter 2

Computerized Adaptive Testing, Content Balancing Procedures, and Multidimensional IRT Models

This chapter introduces the background knowledge and concepts involved in the current project. The mechanism of computerized adaptive testing is described in great detail in section 2.1. Three commonly used content balancing procedures are discussed in section 2.2. Multidimensional IRT models and item characteristics are described in section 2.3. In this section, special attention is also given to the orientation of the unidimensional θ-scale in a multidimensional space.

2.1 Computerized Adaptive Testing

2.1.1 Overview

Computerized adaptive testing (CAT) is a method for administrating tests to match the examinee's ability level. Several large-scale testing programs now use CAT as alternatives to paper-and-pencil tests, for example, the Graduate Records Examination (GRE; Eignor, Stocking, Way & Steffen, 1993), the Test of English as a Foreign Language (TOEFL; Educational Testing Services, 2007), and the Armed Service Vocational Aptitude Battery (ASVAB; U.S. Department of Defense, 1982).

The idea of adapting the difficulty of a test to each individual examinee first appeared in Alfred Binet's (1905) intelligence test in the context of one-on-one administration. From 1970s, with the development of item response theory and the breakthrough in

computer technology, the idea was refined and developed into the current CAT procedures for large-scale testing.

In CAT, items are selected adaptively on the basis of the examinee's responses to the items previously administrated. Figure 2.1 shows the structure of a CAT procedure in a flowchart. It begins with the first item based on an initial estimate of proficiency. After each item response, a new proficiency is estimated and the next optimal item is selected. This process is repeated until it meets certain stopping rules, for instance, the precision of proficiency is adequate, or a fixed number of items have been administrated.

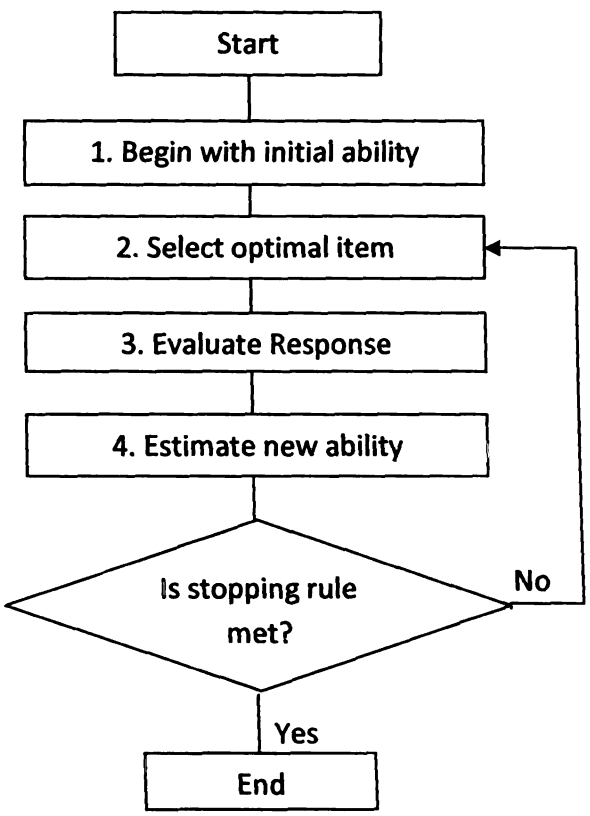


Figure 2.1 A flowchart describing computerized adaptive testing

Compared to a paper-and-pencil test, a CAT offers many advantages. The biggest advantage is that it gives more precise estimates of examinees' ability level with fewer items (Wainer, 1993). This is because the most informative items are selected and administrated, and the items outside of examinees' ability range are excluded during the CAT procedure. In addition, each examinee is presented a test with an appropriate range of difficulty, neither too easy nor too difficult, which reduces the measurement errors induced by confusion, frustration, or boredom. CAT also provides flexible testing schedules and immediate feedback for the examinee after the test.

2.1.2 Components of computerized adaptive testing

A basic CAT application consists of four primary components: item pool, item selection procedure, scoring procedure, and test termination rules (Reckase, 1989).

Item pool

Item pool[1] is a collection of items from which the adaptive test is selected. Items in an item pool are written based on test specifications, and calibrated and linked to a common measurement scale using IRT. To give every examinee precise and efficient measurement, there must be high-quality items with a wide span of difficulty levels. It also needs a sufficient number of items in each content area. For the appropriate size of item pools, six to twelve times the test length is suggested (Luecht, 1998; Patsula & Steffan, 1997;

[1] There are two types of item pools in practice: a master pool and an operational pool. A master pool consists of a collection of items at various stage of development. An operational pool is a pool of items from which individual tests are assembled (Van der Linden, 2005a). This study focuses on the operational item pool.

Stocking, 1998). In practice, more items are needed due to the issues of item exposure, item retirement and etc.

Item selection procedure

The two most widely used item selection procedures are maximum information method (Weiss, 1982) and maximum expected precision method (Owen, 1975).

The maximum information strategy selects the item that provides the maximum amount of item information, $I_i(\hat{\theta}_j)$, at the examinee's current ability estimates $\hat{\theta}_j$,

$$I_i(\hat{\theta}_j) = \frac{[P_i'(\hat{\theta}_j)]^2}{P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)]} \tag{2.1}$$

where $\hat{\theta}_j$ is the ability estimate for examinee j after n preceding responses, $P_i(\hat{\theta}_j)$ is the probability of a correct response to item i given current ability estimate $\hat{\theta}_j$, and $P_i'(\hat{\theta}_j)$ is the first derivative of $P_i(\hat{\theta}_j)$with respect to θ evaluated at $\hat{\theta}_j$. For unidimensional three-parameter logistic IRT model, the equation becomes:

$$I_i(\hat{\theta}_j) = \frac{D^2 a_i^2 (1-c_i)}{\left(c_i + e^{Da_i(\theta_j - b_i)}\right)\left(1 + e^{Da_i(\theta_j - b_i)}\right)^2} \tag{2.2}$$

where a_i is the item discrimination parameter, b_i is the difficulty parameter, c_i is the pseudo-guessing parameter, and D is the scaling constant (typically 1.7). From Equation (2.2), we can see the item information increases as a_i increases, b_i approaches θ, and c_i approaches 0 (Hambleton, Swaminathan, & Rogers, 1991). Therefore, in CAT, items with large discrimination values and difficulty parameter close to the current estimate of

θ are usually desirable. They yield larger information and have a higher probability of being selected when the maximum information method is used.

Owen's maximum expected precision method uses a Bayesian approach. In this procedure, the item that minimizes the expected posterior variance of the θ estimate is selected. Owen developed the mathematical formula for the posterior mean and variance of θ (See Owen (1975) for detailed mathematical formula). Compared to maximum information which is based on iterative numerical methods, the computation burden is smaller for Owen's procedure. However, the θ estimate from Owen's procedure depends on the order of the items administrated. That is, if two examinees are presented the same items and have the same answers, but in different orders, their θ estimates are different. Because of this disadvantage, Owen's procedure is now much less widely used (Wainer, 2000).

In the operational CAT programs, these item selection procedures usually need to be modified for practical considerations, such as item exposure control and content balancing.

Scoring Procedure

In CAT, after each response, the examinee's proficiency is estimated. Based on this estimate, the next item most appropriate for the examinee is selected. Two commonly used estimation procedures are maximum likelihood method and Bayesian method (Bejar & Weiss, 1979).

Maximum likelihood estimation (MLE)

Maximum likelihood estimation is to find an estimate that results in the highest likelihood for the observed string of item responses. Given a response string and a set of items with known parameters, the likelihood function is

$$L(U_j|\theta_j) = \prod_{i=1}^{n} P(u_{ij}|\theta_j) \qquad (2.3)$$

where $L(U_j|\theta_j)$ is the likelihood of response string U_j for a person j located at θ_j;

u_{ij} is the item response on item *i* by person *j* (1 for correct response and 0 for incorrect response);

$P(u_{ij}|\theta_j)$ is the probability of getting response *u* for item *i* by a person *j* located at θ_j;

The maximum likelihood estimate of an examinee's ability, $\hat{\theta}_j$, is the value that maximizes this likelihood function. In practice, we set the derivative of the log-likelihood function (with respect to θ_j) to zero and then solve the equation.

Maximum likelihood method has desirable properties like asymptotical unbiasedness. However, problems can rise at early stage of CAT, since it cannot provide finite estimates for responses to single items or for patterns of responses that are all correct or all incorrect. To solve the problem, we can either constrain θ to a reasonable range (e.g., -4 to 4) or use an alternative estimation method --- Bayesian estimation procedure.

Bayesian ability estimation

In Bayesian estimation, we use the information about the population ability distribution. The initially assumed distribution is called the prior distribution. In CAT, we usually assume that the population ability is normally distributed with a mean of 0 and a standard deviation of 1. Given the prior distribution, after the examinee answers the first item, the posterior distribution of θ is given by the Bayes' theorem,

$$h(\theta|U_j) = \frac{L(U_j|\theta)f(\theta)}{\int_\theta L(U_j|\theta)f(\theta)\,d\theta} \qquad (2.4)$$

where $f(\theta)$ is the prior probability density function for θ,

U_j is the item response string for person j,

$L(U_j|\theta)$ is the probability of the item response string given θ (the likelihood function),

and $h(\theta|U_j)$ is the posterior probability density of θ given the item response string.

This posterior distribution then becomes the new prior distribution for the next item. As the test proceeds, this process continues in a sequential fashion.

There are three common Bayesian-based approaches: Expected a posteriori (EAP), Maximum a posteriori (MAP) and Owen's method.

EAP. The EAP method uses the mean of the posterior distribution, $h(\theta|U_j)$, as the ability estimate:

$$\hat{\theta} = E(\theta|U_j) = \int_{-\infty}^{\infty} \theta h(\theta|U_j)\, d\theta \qquad (2.5)$$

Instead of computing the integral directly, we can approximate it using Gauss-Hermite quadrature points (Stroud & Sechrest, 1966),

$$\hat{\theta} = E(\theta|U_j) = \frac{\sum_{k=1}^{q} X_k L_i(U_j|X_k) W(X_k)}{\sum_{k=1}^{q} L_i(U_j|X_k) W(X_k)} \qquad (2.6)$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with that point, and $L_i(U_j|X_k)$ is the likelihood function after i items evaluated at X_k.

MAP. The MAP method proposed by Samejima (1969) uses the mode of the posterior distribution as the ability estimate, that is, the point that maximizes the posterior probability density. It can be done by setting the derivative of the posterior probability density, $h(\theta|U_j)$, to zero and solving the equation.

Owen's method. Owen (1975) used a normal approximation to the true posterior distribution, which allowed us to derive the mathematical form of the mean and variance of the posterior distribution (See Owen (1975) for detailed mathematical formula). The mean of the posterior distribution is then used as the examinee's ability estimate.

Among these three Bayesian estimation methods, EAP provides the most stable estimates, although the estimates are biased except at the population mean (Bock and Mislevy, 1982). Intuitively, it is better than the Owen's method because it evaluates the posterior distribution directly instead of using a normal approximation. This is also confirmed by Wang and Vispoel (1998), in which the Owen's method yielded the worst performance. In addition, Lord (1986) and Warm (1989) suggested that MAP estimates

could be seriously biased in CAT. Therefore, EAP estimate is adopted at early stage in this study, when there is no finite estimate for the maximum likelihood method.

Stopping rule

An adaptive test can be terminated when a target measurement precision has been achieved, or a fixed number of items have been administrated. Testing each examinee to a prespecified degree of precision insures that the measurements for all individuals are equally precise, but occasionally the test could run out of the items before the target precision is reached or test time could be extremely long for examinees. Segall, Moreno, & Hetter (1997) pointed out that in a variable-length test, examinees with extreme proficiency levels tended to have long tests. It may cause fatigue and raise the chance of careless errors, and each additional item provides little information about the examinee's ability. On the other hand, a fixed-length test is easy to implement and constrains the test time to a reasonable range.

2.2 Content Balancing

Whether to balance the content of items administrated to examinees is one of the first issues that must be addressed in developing a CAT application. By the nature of CAT the examinees receive different items in the same test, and each should get the same number of items from each content area for fairness.

Green et al. (1984) first commented on the need of content balancing in adaptive tests. They noticed that in Bock and Mislevy (1981)'s study, on a test of general science, males performed better than females on natural science items, while females performed better than males on health and nutrition items. In an adaptive test, if a male examinee is administrated all health and nutrition items, he might be disadvantaged and the validity of the score would be threatened. Therefore, content balancing could reduce the impact of subgroup differences. Kingsbury and Zara (1989) also pointed out that administrating a test that covered all the content areas in a test blueprint gave an adequate assessment of the examinee's ability, and reduced the legal challenges to the test (e.g., licensure tests and admission tests).

Previous research on content balancing has developed a number of methods. Kingsbury and Zara (1989) proposed a constrained CAT (CCAT) procedure. In this procedure, the selection of the next optimal item is restricted to the content area that is the farthest below its target percentage. Detailed steps are described as follows:

1. Calculate the target percentages of content areas for the test based on the test blueprint;
2. Estimate the examinee's provisional proficiency after he answers an item;
3. Calculate the percentage of items already administrated in each content area for this examinee;
4. Compare the empirical percentages to the target percentages, and select the content area with the largest discrepancy;
5. Within this selected content area, select and administrate the item with the maximum information at the provisional ability estimate.

In this way, the adaptive test would have any desired content distribution. However, Chen and Ankenmann (2004) argued that this method could lead to high predictability of the content area. Instead, they developed a modified multinomial model to meet the content requirement:

1. Form a cumulative multinomial distribution based on the target percentages of content areas;
2. Generate a random number from the uniform distribution U(0,1) and use it to find the corresponding content area in the cumulative distribution;
3. Within this selected content area, select and administrate the item with the maximum information;
4. This process continues until a content area has reached its target percentage. A new multinomial distribution is formed by adjusting the unfilled percentages of the remaining content areas.

Leung, Chang and Hau (2000) also proposed a modified version of CCAT procedure to eliminate the undesirable order effect. The procedure is similar to CCAT, except that items can be selected from all the content areas for which target percentages are not reached.

Leung, Chang and Hau (2003b) compared these three content balancing methods in a CAT using the maximum information item selection strategy. With simulated unidimensional data they demonstrated that using content balancing methods caused some loss in the measurement efficiency. In addition, the three methods had similar effects on measurement efficiency and item pool utilization, but the MMM method had

the fewest overexposed items. The present study also compares the three methods, but in a different context, where items are assumed to require multiple abilities to answer correctly.

2.3 Multidimensional Item Response Theory (MIRT)

As described in section 2.1, item response theory plays a central role in computerized adaptive testing, from item pool calibration, item selection, to ability estimation. Most computerized adaptive testing programs use unidimensional IRT models, which assumes examinees' responses to test items can be accounted for by a single latent trait. However, the cognitive and psychological processes of responding to test items are very complex, and many researchers believe that multiple skills influence the performance on a test (Ip, 2010; Reckase, 1985; Reckase, Ackerman and Carlson, 1988; Traub, 1983; Walker and Beretvas, 2003). Multidimensional item response theory is a collection of mathematical models that describe the interaction between persons and test items when more than one ability are required to account for test performance.

There are two major types of multidimensional IRT models: compensatory and noncompensatory. The compensatory model is based on a linear combination of ability dimensions, and a high ability on one dimension can compensate for a low ability on another dimension. For example, the compensatory form of the multidimensional three-parameter logistic model is given by (Mckinley and Reckase, 1983),

$$P(u_{ij} = 1 | \boldsymbol{\theta_j}, \boldsymbol{a_i}, d_i, c_i) = c_i + (1 - c_i)\frac{e^{D(\boldsymbol{a_i'}\boldsymbol{\theta_j} + d_i)}}{1 + e^{D(\boldsymbol{a_i'}\boldsymbol{\theta_j} + d_i)}} \qquad (2.7)$$

Where $P(u_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i, c_i)$ is the probability of a correct response to item i by person j;

u_{ij} is the response on item i by person j (1 is correct and 0 is incorrect);

$\boldsymbol{\theta}_j$ is a column vector (m by 1) of person j's abilities in a m-dimensional space;

$\boldsymbol{a}_i$ is a column vector (m by 1) of discrimination parameters for item i;

d_i is a scalar that related to item difficulty;

c_i is the guessing parameter or low asymptote for item i, and

D is the scaling constant (typically 1.7).

In the multidimensional version, $\boldsymbol{\theta}_j$ and $\boldsymbol{a}_i$ are vectors instead of scalars. From equation (2.7), the exponent of e is a linear function of θs plus the intercept term d, $\boldsymbol{a}_i'\boldsymbol{\theta}_j + d_i$. The additivity of the θs implies the compensatory nature of the model.

Sympson (1978) argued that the compensatory model is not realistic for certain types of items. For example, for a mathematics item that requires both arithmetic computation skills and reading skills, if an examinee's reading skills are very low, he might not understand the problem and hence cannot solve the problem even if he has high arithmetic computation skills. For this situation, he developed a noncompensatory model:

$$P(u_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, \boldsymbol{d}_i, c_i) = c_i + (1 - c_i)\prod_{k=1}^{m}\frac{e^{Da_{ik}(\theta_{jk} - d_{ik})}}{1 + e^{Da_{ik}(\theta_{jk} - d_{ik})}} \qquad (2.8)$$

where m is the number of dimensions, a_{ik} and d_{ik} are the discrimination and difficulty parameter respectively, for item i and dimension k, and other parameters are defined as before. In this model, the test item is decomposed into individual components, and the probability of a correct answer is the product of the probabilities of doing each component correctly. Due to the multiplicative nature of the model, the probability of a correct response to an item cannot exceed the lowest probability in the product. The probability does increase with ability increase in one dimension, but up to a limit set by the lowest probability in the product. Therefore, this model is also called partially compensatory model.

Some researchers believe that the partially compensatory model is more theoretically sound, but is less realistic. For example, Ansley (1984) pointed out that data generated with this model did not resemble real test data. Bolt and Lall (2003) also compared the fit of compensatory and partially compensatory models to a common data set from a test of English usage, and found that the compensatory model fit the data better than the partially compensatory model. In addition, estimation difficulty for the partially compensatory model hinders its development and application. As a result, compensatory model is more prevalent in the current literature, and we will not deal with partially compensatory model further in this study.

In a compensatory MIRT model, the parameters sometimes lack intuitive meaning, so Reckase (1985) and Reckase and Mckinley (1991) developed two statistics to interpret the characteristics of the items for compensatory models: multidimensional discrimination (*MDISC*) and multidimensional difficulty (*MDIFF*). They are defined as

$$MDICS_i = \sqrt{\boldsymbol{a}_i'\boldsymbol{a}_i} = \left(\sum_{k=1}^{m} a_{ik}^2\right)^{\frac{1}{2}} \qquad (2.9)$$

$$MDIFF_i = \frac{-d_i}{MDICS_i} \qquad (2.10)$$

where parameters are defined as before. These two statistics are analogous to discrimination and difficulty parameters from the unidimensional IRT models. $MDICS_i$ is the slope of the item response surface at the steepest point, and indicates the discriminating power of the item. $MDIFF_i$ is the distance from the origin to the point of the steepest slope. It represents the multidimensional difficulty of the item: high values indicate difficult items and low values indicate easy items. In addition, the direction of the steepest slope from the origin of the space is given by

$$cos\alpha_{ik} = \frac{a_{ik}}{(\sum_{k=1}^{m} a_{ik}^2)^{\frac{1}{2}}} \qquad (2.11)$$

where α_{ik} is the angel between the kth coordinate axis and the line from the origin to the point that has the greatest slope overall. The cosines above are often called direction cosines.

Using the concept of multidimensional discrimination, multidimensional difficulty, and direction cosines, items can be displayed graphically in the space. Each item is represented by an arrow. The base of the arrow is at the point of maximal slope, the length of the arrow indicates the discrimination of the item, $MDISC_i$. The distance from the origin to the base of the arrow represents the difficulty of the item, $MDIFF_i$, and the

direction of the arrow, α_i, is derived from the direction cosines of the item. Figure 2.2 shows an item vector plot of 45 items in a two-dimensional space.

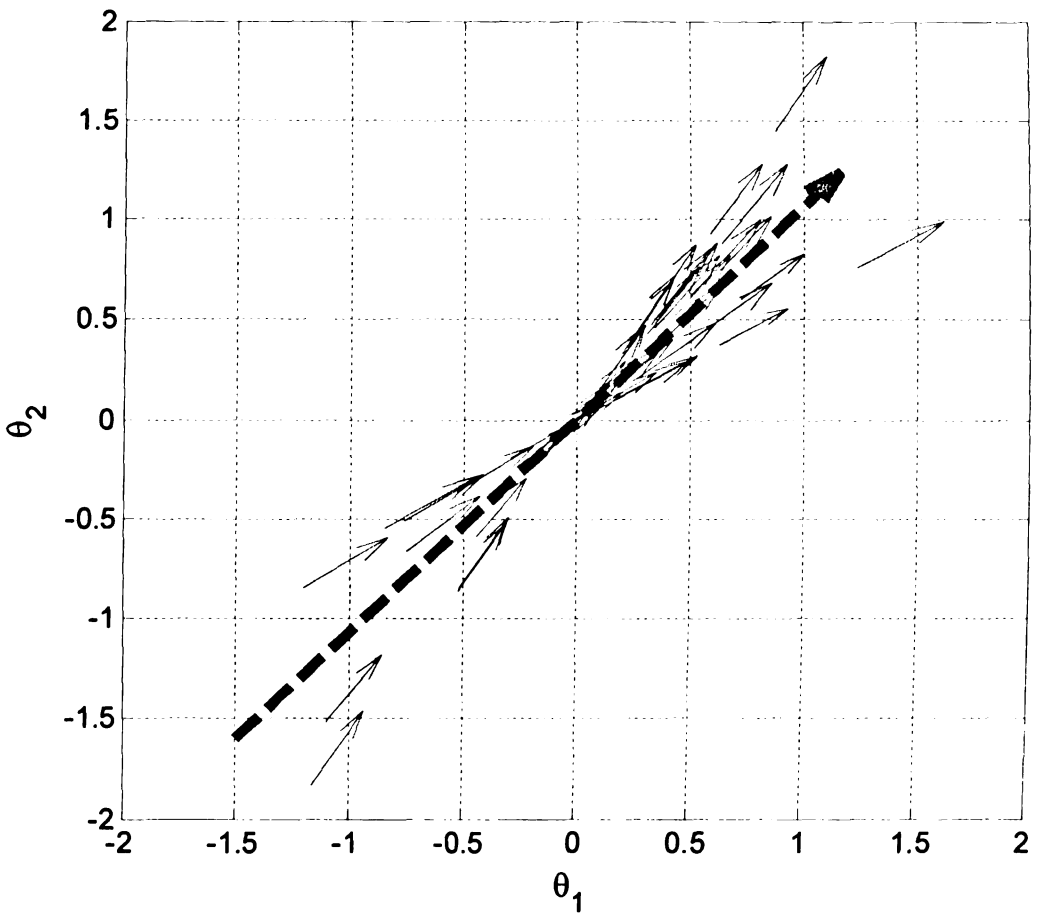


Figure 2.2 Representation of the characteristics of 45 items in a two-dimensional space: item arrows and reference composite.

Orientation of the unidimensional θ-scale in the multidimensional space

Wang (1995, 1996) showed that if we fitted a unidimensional model to a multidimensional test, the orientation of unidimensional θ-scale was related to the matrix of discrimination parameters from the compensatory MIRT model. Specifically, this unidimensional line is defined as the eigenvector of the $\boldsymbol{a}'\boldsymbol{a}$ matrix associated with the largest eigenvalue, and is called the reference composite of the test. In Figure 2.2, the reference composite of the 45 items is represented by the bold dashed arrow.

The projection of the θ-point in the multidimensional space onto the reference composite gives an estimate of the unidimensional θ that would result if the response data from the test items were analyzed using a unidimensional IRT model. Formally, it is given by

$$\theta^* = RC'\boldsymbol{\theta} \qquad (2.12)$$

Where θ* is the projected unidimensional θ; RC is the reference composite vector; and $\boldsymbol{\theta}$ is the multidimensional ability vector.

Chapter 3

Methodology

To examine the effect of fitting a unidimensional IRT model to multidimensional data in a content-balanced CAT, Monte Carlo simulations are conducted. This chapter describes the simulation design and evaluation criteria in details.

3.1 Item Pool

The item pool consists of 400 items, from three content areas. Content 1 has 160 items, and Content 2 and 3 each have 120 items. The unbalanced distribution of items across three content areas resembles a typical item pool in real testing programs. We assume that multiple skills are required to answer the items correctly, and hence assume that the item pool has a multidimensional structure. Two types of representative dimensional structures are adapted from Reckase (2009):

1. Three content areas in a two-dimensional space

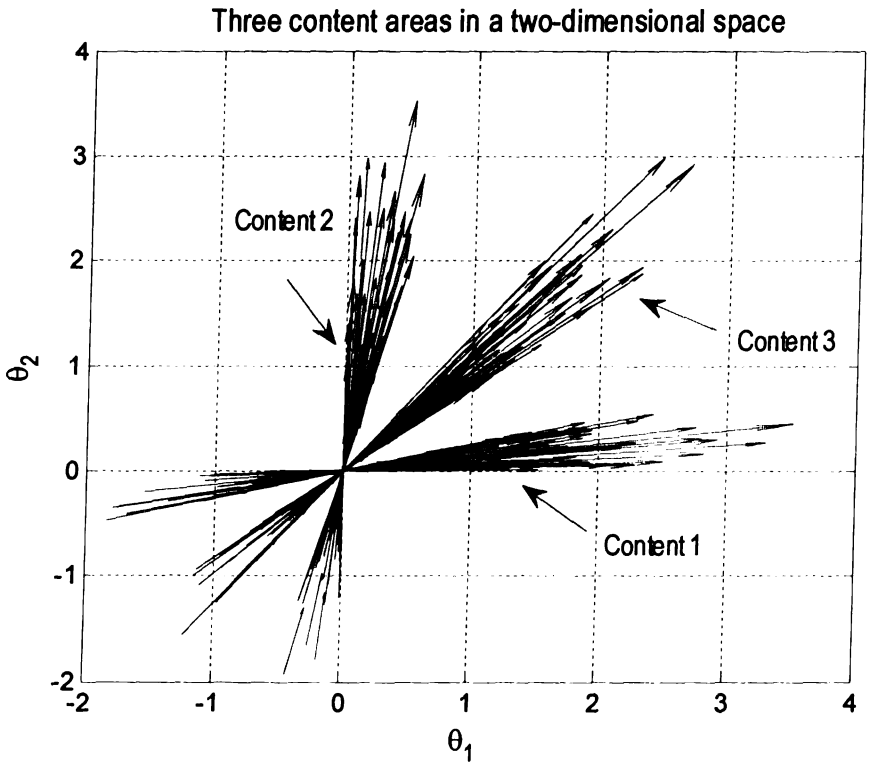
   In this structure, Content 1 and 2 mainly load on either of the two dimensions respectively; Content 3 loads on the composite of the two dimensions. For example, in a mathematics test, there are three content areas: arithmetic, geometry, and algebra. The arithmetic items mainly measure the examinees' computation skills, the geometry items measure the problem solving skills, and the algebra items require both computation and problem solving skills.

2. Three content areas in a three-dimensional space

   In this structure, Content 1 measures Dimension 1, Content 2 measures Dimension 2, and Content 3 measures the composite of all three dimensions. Using the same example above: in the test, arithmetic and geometry items remain unchanged and mainly need skills in either computation or problem solving, but the algebra items are story problems now, which require reading skills in addition to computation and problem solving skills.

To construct the dimensional structure, angles between item vectors and dimensions are specified. In the two-dimensional case, the direction cosines are (1,0), (0,1) and $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ for the three content areas respectively. And in the three-dimensional case, the direction cosines are (1,0,0), (0,1,0) and $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$. Within each content area, the angular variation is 15^o (Roussos, Stout and Marden, 1998). For example, for Content 1 in Structure 1, the angle between the item vector arrow and the first dimension, α_1, is randomly selected from the uniform distribution $U(0^o, 15^o)$; for Content 2 and 3, α_1 is randomly selected from $U(75^o, 90^o)$ and U(37.5^o, 52.5^o) respectively. Then the angle between the item arrow and the second dimension, α_2, is calculated by $90^o - \alpha_1$. Similarly, angles for the three-dimensional case are generated. In Figure 3.1, the two dimensional structures are illustrated.

The item parameters for the compensatory MIRT model are simulated from commonly used distributions. The logs of item discrimination parameters ($MDISC_i$) are randomly drawn from a normal distribution with a mean of 0 and a standard deviation of



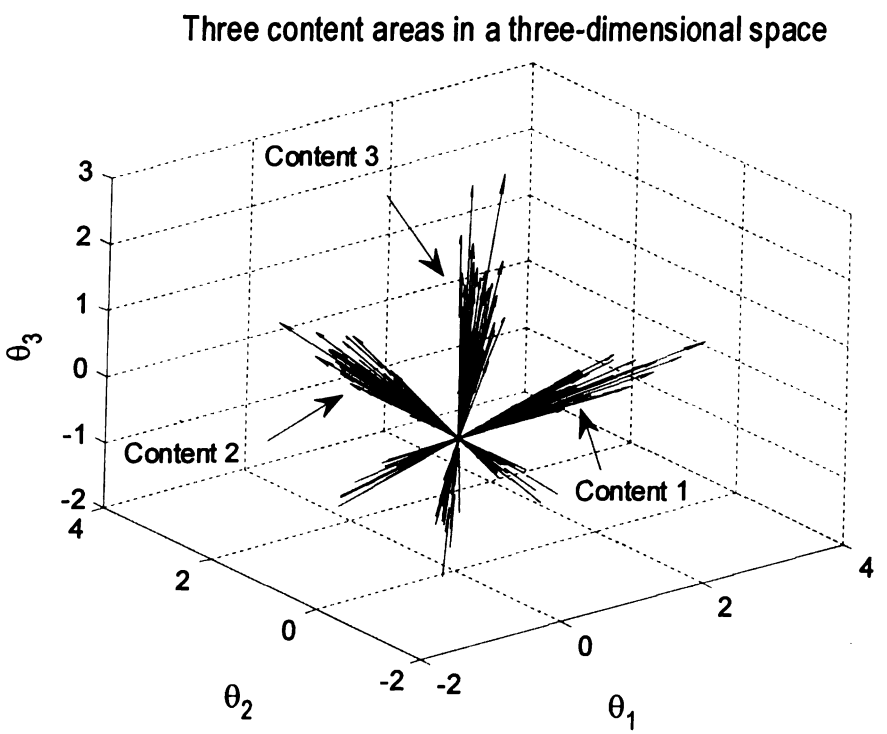


Figure 3.1 Dimensional Structures

0.5, $N(0, 0.5^2)$. Difficulty parameters ($MDIFF_i$) are drawn from $N(0, 0.75^2)$ (Fang, 2008). For simplicity, all items have the same low asymptote value (c-parameters) of 0.2.

Given $MDISC_i$, $MDIFF_i$, and the angles, the parameters a_i and d_i for the compensatory MIRT model are calculated by

$$a_{ik} = MDISC_i * cos\alpha_{ik} \qquad (3.1)$$

$$d_i = -MDIFF_i * MDISC_i \qquad (3.2)$$

3.2 Simulation factors

1. Dimensional Structure

   Two dimensional structures are considered in this study: a) three content areas in a two-dimensional space; b) three content areas in a three-dimensional space. The two-dimensional case is the simplest situation of multidimensionality to start with, and then the more complicated three-dimensional case is examined. By studying these two representative structures, results might be generalized to higher dimensions.

2. Ability Distribution

   Examinees' abilities are simulated from multivariate normal distributions with zero mean vector and three different variance-covariance matrices. Table 3.1 shows mean vector (μ), variance-covariance matrices (Σ), and correlation coefficients between abilities (ρ). Three levels of ability correlations are used: 0, 0.4, and 0.8. A correlation of zero implies that there is no correlation between

multiple abilities; 0.4 indicates a moderate correlation; and 0.8 represents a high correlation. In the three-dimensional case, pairwise correlations are slightly varied to produce a more realistic relationship between multidimensional abilities.

Table 3.1 Three ability distributions used in the simulation study

	Two dimensions	Three dimensions
Distribution 1 $\rho = 0$	$\mu = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}0\\0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
Distribution 2 $\rho = 0.4$	$\mu = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0.4\\0.4 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}0\\0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0.4 & 0.3\\0.4 & 1 & 0.35\\0.3 & 0.35 & 1\end{bmatrix}$
Distribution 3 $\rho = 0.8$	$\mu = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\mu = \begin{bmatrix}0\\0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0.8 & 0.7\\0.8 & 1 & 0.75\\0.7 & 0.75 & 1\end{bmatrix}$

3. Difficulty levels for content areas

   Two cases are examined: a) The average difficulty levels are same for all three content areas: the mean of difficulties (*MDIFFs*) are assumed to be zero; b) Content 1 has less difficult items, and Content 2 has more difficult items: the *MDIFFs* are decreased by 0.6 for items in Content 1, increased by 0.6 for Content 2, and kept unchanged for Content 3. We choose 0.6 to represent a moderate to high change of item difficulty (Swaminathen and Rogers, 1990).

4. Content balancing methods

Three content balancing methods, CCAT, MMM and MCCAT, are compared. Unconstrained CAT with the maximum information item selection strategy is used as the baseline.

In total, the four simulation factors yield 2*3*2*4=48 conditions.

3.3 Simulation procedure

The simulation procedure involves the following steps:

1. We simulate an item pool of 400 items for each combination of dimensional structure and difficulty level condition[2]. And we simulate the ability parameters for 2000 examinees using the distribution described above.

2. Given the item and ability parameters, we generate item responses to all items in the given item pool for 2000 examinees. Using the compensatory three-parameter multidimensional IRT model (Eq. 2.7), the probability of a correct answer for a given item and a given examinee (p) is calculated. The 0/1 response is obtained by comparing p to a random number (x) from a uniform distribution U(0,1). If p>x, then a correct response is obtained, otherwise, an incorrect response is obtained. In this way, a 2000 by 400 item response matrix is generated.

[2] More precisely, we simulate the items for each dimensional structure with even content difficulty levels. Then we alter the difficulty values for the items in Content 1 and 2 to get items for the condition of uneven content difficulty levels. The purpose is to reduce the random noises when comparing the two difficulty-level conditions.

3. Based on all the items in the given item pool, we calculate the reference composite from the $\boldsymbol{a}'\boldsymbol{a}$ matrix and the projections of the θ-points in the multidimensional space onto the reference composite. To be consistent with the scaling of the estimates from the unidimensional IRT program (BILOG-MG), the projected θs are scaled to have a mean of 0 and a standard deviation of 1. It gives a theoretical estimate of the unidimensional θ that would result if the response data from the test items were analyzed using a unidimensional IRT model. The resulting θs are considered as the true θs when the recovery of θ is evaluated.

4. The response data resulted in Step 2 is calibrated using BILOG-MG (Zimowski et al., 2003) to estimate unidimensional item parameters. The three-parameter logistic model[3] with a scaling constant D=1.7 is applied. In BILOG-MG, the convergence criterion is set to be 0.005, and the number of quadrature points for the EM algorithm is set at the default value.

5. Unidimensional CAT is conducted. In this study, the test consists of 30 items. The first item is randomly selected from the 100 items with medium difficulty. The corresponding response is read from the item response matrix generated in Step 2. Based on this response, the provisional ability is estimated. The expected a posterior (EAP) method is adopted at the beginning of the test, assuming N(0,1) is

[3] The unidimensional three-parameter logistic model is $P(u_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{\exp(1.7a_i(\theta_j-b_i))}{1+\exp(1.7a_i(\theta_j-b_i))}$, where a_i, b_i, c_i are discrimination, difficulty, and guessing parameter, respectively, for item i; θ_j is the ability parameter for person j.

the prior distribution. The maximum likelihood estimation (MLE) method is used until at least five items have been administrated and the response pattern contains both 0 and 1 (Cheng, Chang, and Yi, 2007). After each item is administrated, the next most appropriate item for the examinee is selected using the maximum information method. The process continues until a 30-item test has been administrated. During this procedure, the three -parameter logistic model is assumed, and the unidimensional parameters in Step 4 are used for both ability estimation and item selection.

6. Repeat Step 5 with content balancing methods during the item selection procedure.

In addition, to examine the conditional measurement precision, simulation is also conducted for limited points in the ability space, with 50 replications at each ability point. Five equally spaced values of θ from -2 to +2 (θ= -2, -1, 0, 1, 2) are used. Hence, 25 fixed points (5 × 5) are evaluated in the two-dimensional case, and 125 (5 × 5 × 5) points in the three-dimensional case.

3.4 Evaluation Criteria

1. Measurement precision.

The recovery of ability proficiency is assessed by overall bias, mean square error (MSE) of θ, and the correlation between θ and its estimate ($\rho_{\theta,\hat{\theta}}$). Overall bias and MSE are calculated by

$$Bias = \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)/N \qquad (3.3)$$

and

$$MSE = \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)^2/N \qquad (3.4)$$

where $\hat{\theta}_j$ is the estimated ability of the *j*th examinee from the unidimensional CAT, θ_j is the true ability of the *j*th examinee, and N is the number of examinees. Bias and MSE both provide a good indication of the quality of the recovery of examinees' abilities. The smaller the absolute biases and MSEs are, the better the abilities are measured. In addition to overall statistics, conditional measurement precision is calculated at the fixed points.

2. Content balancing

The number of items administrated from each content area is recorded. If the target percentage of certain content area is over or under fulfilled, the test fails to satisfy the content balancing constraint. The percentage of tests violating the content constraint is reported. By design, the CAT with content balancing is expected to have zero percentage of violation.

3. Item pool usage

The three content balancing methods and the maximum information method are also compared in terms of item pool usage. In order to have a maximum item pool usage, a uniform exposure rate distribution is desirable. Chang and Ying (1999) have proposed a scaled chi-square statistics to evaluate the skewness of exposure rate distribution:

$$\chi^2 = \sum_{i=1}^{M} \frac{\left(ER_i - \frac{L}{M}\right)^2}{\frac{L}{M}} \qquad (3.5)$$

where M is the size of item pool, L is the test length, and ER_i is the observed exposure rate for the ith item. This chi-square statistics quantifies the item pool usage efficiency, and indicates the discrepancy between the observed and ideal item exposure rates. Low chi-square statistics is preferred, which implies a more efficient item pool usage.

4. Percentages of underexposed and overexposed items

The exposure rate of an item is defined as the ratio of the number of times the item administrated to the number of examinees. A low exposure rate indicates the item is rarely used. If there are a large proportion of items with low exposure rates in the item pool, the cost-effectiveness of developing the items might not be achieved. On the other hand, if an item is over-selected, it might be known to prospective examinees and test security is threatened. Therefore, an item with either low or high exposure rate is not desirable in CAT programs. In this study, following the literature (Cheng, Chang & Yi, 2007), an item is considered as an underexposed item if its exposure rate is less than 0.02, and an overexposed item if the exposure rate is larger than 0.2. To evaluate the effectiveness of each content balancing method, the percentages of underexposed and overexposed items are reported.

Chapter 4

Results

The simulation results are discussed in this chapter. Section 4.1 summarizes the descriptive statistics of the simulated item parameters. Section 4.2 and 4.3 discuss the estimation of item parameters and person parameters. Section 4.4 to 4.6 evaluate the maximum information method and the three content balancing methods, in terms of the percentage of tests violating the content-balancing requirements, the item pool usage, and the percentages of underexposed and overexposed items.

4.1 Simulated item parameters

The descriptive statistics of the simulated multidimensional item parameters for the two-dimensional case are given in Table 4.1. Along with the standard item parameters (discrimination *a1* and *a2*, difficulty *d*), the generalized discrimination and difficulty indices (*MDISC* and *MDIFF*), and the angles with the coordinate axes (α_1 and α_2) are also shown.

Generally, items are sensitive to differences on a single dimension if they have high discrimination parameters for the dimension and small angles with the corresponding coordinate axis. In Table 4.1, for Content 1, the mean discrimination value for the first dimension (*a1*) is 1.07, and the mean angle with the first dimension (α_1) is 7.8 degree. It is clear that items in Content 1 are mostly sensitive to the first dimension. Similarly, items in Content 2 are mostly sensitive to the second dimension. With roughly equal *a*

parameters and angels with θ_1, θ_2 axes, items in Content 3 measure a combination of the two dimensions. These three distinct sets of items are also shown in Figure 4.1.

Table 4.1 also shows the change of difficulty for the three content areas across conditions. In Panel A, the average *MDIFFs* for the three content areas are -0.09, -0.01 and -0.02, respectively. In Panel B, the difficulty is decreased by 0.6 for Content 1 and increased by 0.6 for Content 2. It results in an average *MDIFF* of -0.69 for Content 1, and 0.59 for Content 2. The change of difficulty is also illustrated in Figure 4.1. In the vector plot of items, the distance from the origin to the base of the arrow indicates the difficulty, *MDIFF*. From Figure 4.1a to 4.1b, the distance changes by -0.6 and 0.6 for Content 1 and 2 respectively.

Similarly, the descriptive statistics of the simulated item parameters and the item vector plots for the three-dimensional case are given in Table 4.2 and Figure 4.2. They demonstrate a clear dimensional structure for the three content areas. Items in Content 1 measure predominantly along θ_1, items in Content 2 measure predominantly along θ_2, and items in Content 3 measure an equally weighted combination of θ_1, θ_2, and θ_3.

Table 4.1 The mean simulated item parameters by content areas in the two-dimensional case

Content	a1	a2	d	MDISC	MDIFF	α_1	α_2
Panel A: Equal difficulty levels across the three content areas: Difficulty =(0, 0, 0)							
1	1.07	0.14	0.15	1.08	-0.09	7.80	82.20
2	0.15	1.08	0.00	1.09	-0.01	81.97	8.03
3	0.78	0.78	0.03	1.10	-0.02	44.97	45.03
Panel B: Unequal difficulty levels across the three content areas: Difficulty =(-0.6,0.6,0)							
1	1.07	0.14	0.80	1.08	-0.69	7.80	82.20
2	0.15	1.08	-0.65	1.09	0.59	81.97	8.03
3	0.78	0.78	0.03	1.10	-0.02	44.97	45.03

Table 4.2 The mean simulated item parameters by content areas in the three-dimensional case

Content	a1	a2	a3	d	MDISC	MDIFF	α_1	α_2	α_3
Panel A: Equal difficulty levels across the three content areas: Difficulty =(0, 0, 0)									
1	1.07	0.08	0.11	0.15	1.08	-0.09	7.80	85.76	84.01
2	0.06	1.08	0.11	0.00	1.09	-0.01	86.85	6.97	84.32
3	0.63	0.64	0.62	0.03	1.10	-0.02	54.70	53.99	56.07
Panel B: Unequal difficulty levels across the three content areas: Difficulty =(-0.6,0.6,0)									
1	1.07	0.08	0.11	0.80	1.08	-0.69	7.80	85.76	84.01
2	0.06	1.08	0.11	-0.65	1.09	0.59	86.85	6.97	84.32
3	0.63	0.64	0.62	0.03	1.10	-0.02	54.70	53.99	56.07

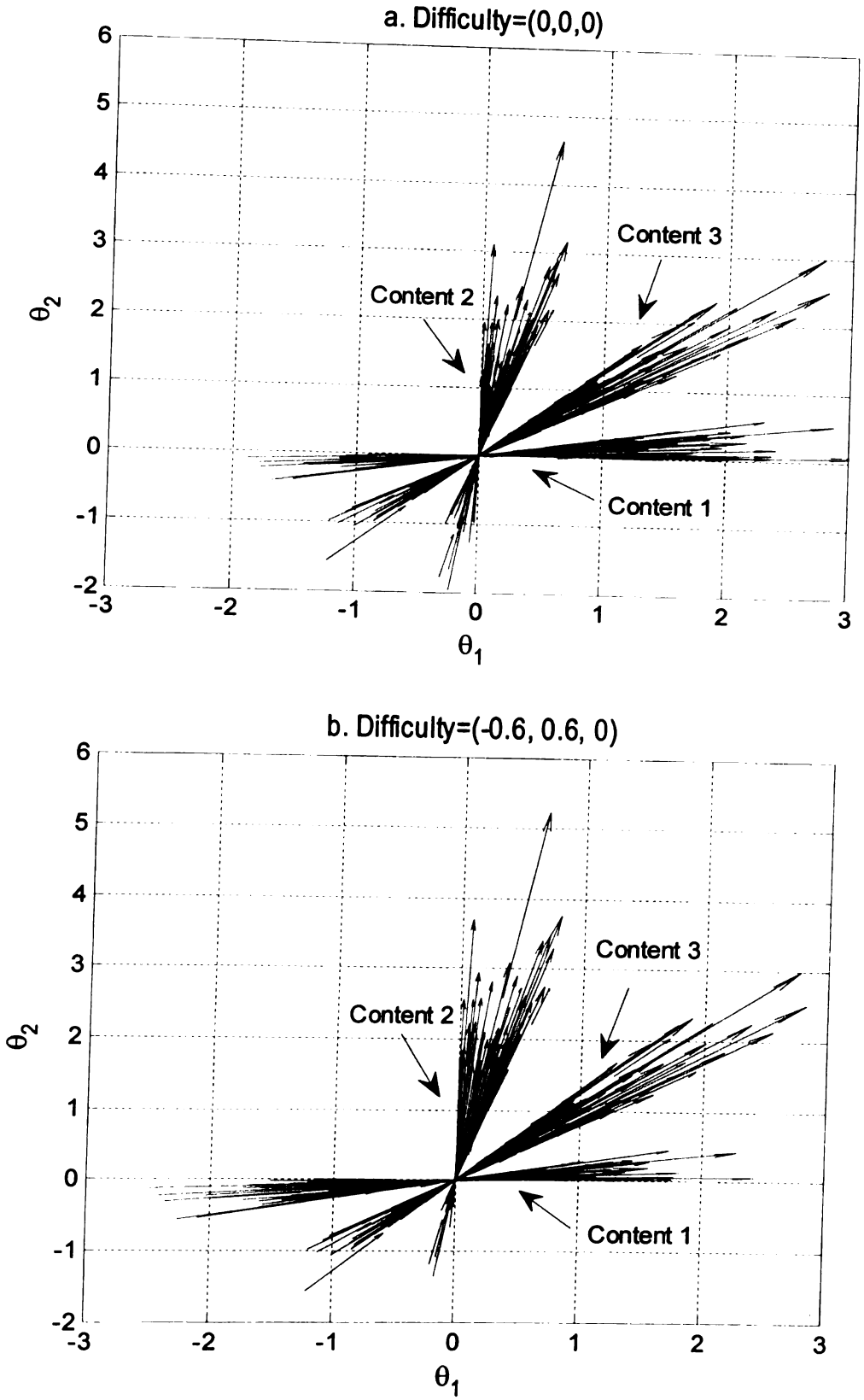


Figure 4.1 Item vector plots in the two-dimensional case

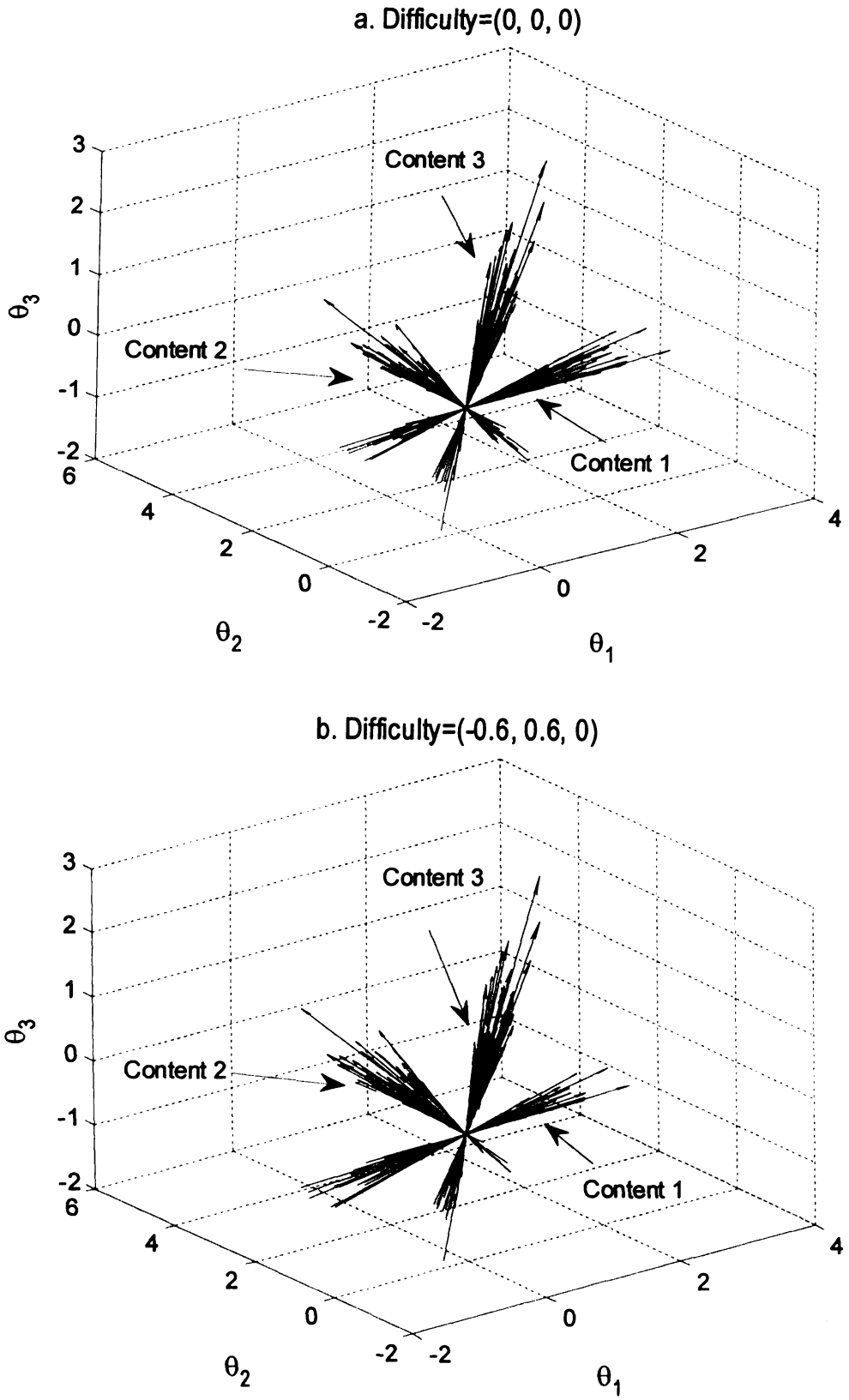


Figure 4.2 Item vector plots in the three-dimensional case

4.2 Estimation of unidimensional item parameters

Following the simulation procedure described in Chapter 3, the item response matrices are generated using the compensatory three-parameter multidimensional IRT model (Eq. 2.7). In order to get the unidimensional item parameters for the CAT procedure, the data is calibrated using the unidimensional three-parameter logistic model (see footnote 3). In BILOG-MG, all the calibration runs converge for both the EM steps and the Newton steps and reach the convergence criterion of 0.005, which indicate that the estimations of item and person parameters reach a good accuracy. Table 4.3 reports the overall chi-square indices of fit from BILOG-MG when the unidimensional model is fitted to our data. Under all simulation conditions, the test shows a good fit of the unidimensional model, with p-values close or equal to 1. It is surprising that unidimensional model fits well when the dimensionality is not one and the correlation is zero. It might be due to the multidimensional structure used in the simulation study. There are two possible reasons. First, by design, items in content 3 measure an equally weighted combination of all constructs, and hence are very close to the estimated unidimensional θ scale. Second, although items in Content 2 and 3 measure a single construct, they are not very far from the estimated unidimensional θ scale. For example, in the two-dimensional case, for Content 1, the angle between the item vector arrow and the first dimension varies from 0^o to 15^o. So it gives us an average angle difference of 37.5^o between Content 1 and the estimated unidimensional θ scale.

Table 4.4 presents the summary statistics of the estimated unidimensional item parameters across content areas for the two-dimensional case. As we expect, Content 3 has a larger estimated discrimination value (a) than Content 1 and 2. This is because the

orientation of the items in this content area is more aligned with the orientation of the reference composite. This founding is also consistent with the results from Ackerman (1991).

Table 4.3 The overall chi-square indices of fit from BILOG-MG calibrations

	Two-dimensional case			Three-dimensional case		
	Chi-square	DF	P-value	Chi-square	DF	P-value
Difficulty=(0,0,0)						
$\rho = 0$	2545.8	3502	1	2726.1	3518	1
$\rho = 0.4$	2486.3	3476	1	2649.3	3480	1
$\rho = 0.8$	2551.4	3447	1	2376.7	3395	1
Difficulty=(-0.6,0.6,0)						
$\rho = 0$	3066.8	3491	1	3453.0	3545	0.8631
$\rho = 0.4$	2586.5	3491	1	2817.5	3463	1
$\rho = 0.8$	2467.9	3398	1	2394.1	3345	1

The estimation of the item difficulty is overall satisfactory. When all the three content areas have an average *MDIFF* of 0, the mean of the unidimensional difficulty values (*b*) is generally close to 0. When *MDIFF* changes by -0.6 and 0.6 for Content 1 and 2, the mean of *b*s also has a similar change in the same direction. For example, with a medium ability correlation of 0.4, the mean unidimensional difficulties are 0.025, 0.070, and 0.023 for the three content areas in Panel A, while they change to -0.637, 0.773, and 0.083 in Panel B. Table 4.4 also shows that as the correlation between true abilities increases, the overall recovery of difficulty improves. For example, in Panel A, with an even difficulty level and an ability correlation of 0.8, the mean unidimensional difficulties are -0.042, 0.016 and 0.006 respectively, which are the closest to the *MDIFF* values.

Finally, most of the guessing parameters are close to 0.2 with small standard deviations.

Similarly, the summary statistics for the estimated unidimensional item parameters in the three-dimensional case are provided in Table 4.5. The observations are largely consistent with those in the two-dimensional case. For example, Content 3 still has the largest mean discrimination value under all conditions.

Table 4.4 Summary statistics for the estimated unidimensional item parameters in the two-dimensional case

Ability		a		b		c	
Correlation	Content	mean	std	mean	std	mean	std
Panel A: Difficulty= (0, 0, 0)							
0	1	0.739	0.227	-0.031	0.826	0.240	0.050
	2	0.738	0.245	0.142	0.868	0.252	0.053
	3	1.131	0.544	0.005	0.689	0.220	0.037
0.4	1	0.940	0.299	0.025	0.736	0.240	0.046
	2	0.886	0.331	0.070	0.758	0.225	0.043
	3	1.326	0.651	0.023	0.584	0.214	0.034
0.8	1	1.107	0.426	-0.042	0.701	0.222	0.039
	2	1.098	0.456	0.016	0.676	0.218	0.037
	3	1.488	0.744	0.006	0.534	0.216	0.033
Panel B: Difficulty= (-0.6, 0.6, 0)							
0	1	0.754	0.276	-0.789	0.810	0.201	0.045
	2	1.222	0.425	0.928	0.537	0.318	0.087
	3	1.216	0.554	0.114	0.658	0.253	0.053
0.4	1	0.896	0.345	-0.637	0.764	0.211	0.041
	2	1.094	0.418	0.773	0.658	0.264	0.059
	3	1.331	0.624	0.083	0.582	0.229	0.044
0.8	1	1.095	0.440	-0.599	0.716	0.225	0.043
	2	1.159	0.519	0.607	0.656	0.221	0.033
	3	1.483	0.743	0.004	0.531	0.211	0.034

Table 4.5 Summary statistics for the estimated unidimensional item parameters in the three-dimensional case

Ability		a		b		c	
Correlation	Content	mean	std	mean	std	mean	std
Panel A: Difficulty= (0, 0, 0)							
0	1	0.622	0.170	-0.053	0.944	0.232	0.049
	2	0.712	0.203	0.257	0.902	0.287	0.053
	3	1.079	0.488	0.030	0.680	0.228	0.047
0.4	1	0.842	0.258	-0.013	0.793	0.228	0.046
	2	0.855	0.299	0.066	0.775	0.229	0.043
	3	1.370	0.639	0.023	0.539	0.219	0.035
0.8	1	1.088	0.411	-0.057	0.660	0.210	0.040
	2	1.154	0.460	0.058	0.665	0.229	0.040
	3	1.696	0.790	0.003	0.456	0.211	0.031
Panel B: Difficulty= (-0.6, 0.6, 0)							
0	1	0.658	0.226	-0.864	0.898	0.197	0.049
	2	1.025	0.321	1.100	0.615	0.336	0.091
	3	1.134	0.503	0.115	0.652	0.253	0.057
0.4	1	0.804	0.287	-0.747	0.773	0.197	0.048
	2	1.142	0.389	0.827	0.614	0.278	0.068
	3	1.444	0.649	0.071	0.512	0.234	0.044
0.8	1	1.106	0.442	-0.614	0.657	0.211	0.044
	2	1.200	0.496	0.628	0.603	0.235	0.042
	3	1.717	0.783	0.018	0.445	0.215	0.029

4.3 Measurement precision

4.3.1 Two-dimensional case:

Table 4.6 presents the estimated bias, mean squared error (MSE) and the correlation between θ and its estimate ($\rho_{\theta,\hat{\theta}}$) for the four methods in the two-dimensional case.

When the three content areas have the same difficulty level, the maximum information method tends to result in lower biases and MSEs than the three content balancing methods, except for a high ability correlation of 0.8. However, the three content balancing methods also yield good measurement precision. The biases and MSEs

are close to those from the maximum information method, and the correlations between θ and its estimate are around 0.96.

On the other hand, when the three content areas have uneven difficulty levels, the maximum information method tends to perform worse. It yields higher MSEs and lower correlations $\rho_{\theta,\hat{\theta}}$ than the three content balancing methods. For example, when there is no correlation between true abilities, the maximum information procedure produces a MSE of 0.173 and a correlation of 0.937. In contrast, the three content balancing methods give MSEs around 0.13 and correlations around 0.95. However, the differences in MSE between the maximum information method and the content balancing methods become smaller as the ability correlation increases. With a correlation of 0.8, the four methods perform comparably. Intuitively, high ability correlation reduces multidimensionality. When the ability correlation approaches 1, multidimensionality diminishes and reduces to a simple unidimensional case. Previous research (Cheng, Chang & Yi, 2007) has suggested that in the unidimensional context, content balancing methods yield the measurement precision close to the maximum information method. Therefore, our finding is consistent with previous research on the unidimensional case.

The difficulty levels for the content areas seem to affect the measurement precision of the four methods, particularly the maximum information method. For example, when the ability correlation is 0, the MSE for the maximum information method increases from 0.131 to 0.173 as the difficulty levels becomes uneven. By design, Content 3 is closer to the orientation of the calibrated unidimensional θ-scale or the reference composite. Therefore, its items have larger unidimensional discriminations and hence have a greater probability of being selected when the maximum information method is used. However,

Table 4.6 Measurement precision for maximum information method and three content balancing methods in two-dimensional case

Ability Correlation	Method	Bias	MSE	$\rho_{\theta,\hat{\theta}}$
Panel A: Difficulty= (0, 0, 0)				
0	Max Information	0.024	0.131	0.968
0	CCAT	0.026	0.146	0.964
0	MMM	0.031	0.149	0.961
0	MCCAT	0.033	0.145	0.962
0.4	Max Information	0.052	0.138	0.967
0.4	CCAT	0.047	0.144	0.968
0.4	MMM	0.050	0.152	0.965
0.4	MCCAT	0.049	0.141	0.967
0.8	Max Information	-0.001	0.189	0.954
0.8	CCAT	0.017	0.196	0.965
0.8	MMM	0.013	0.197	0.965
0.8	MCCAT	0.015	0.178	0.966
Panel B: Difficulty= (-0.6, 0.6, 0)				
0	Max Information	0.038	0.173	0.937
0	CCAT	0.039	0.132	0.954
0	MMM	0.040	0.133	0.952
0	MCCAT	0.040	0.135	0.951
0.4	Max Information	0.067	0.153	0.951
0.4	CCAT	0.063	0.138	0.961
0.4	MMM	0.061	0.143	0.959
0.4	MCCAT	0.064	0.136	0.961
0.8	Max Information	0.029	0.124	0.961
0.8	CCAT	0.038	0.126	0.971
0.8	MMM	0.035	0.124	0.970
0.8	MCCAT	0.039	0.122	0.971

when the difficulty levels for the content areas become uneven, for those who have a very low or high ability, the items with high discrimination from Content 3 may no longer be the most optimal items to be selected, while the items from Content 1 or 2 which have similar difficulty level to their ability level become more informative. As a result, the maximum information method will select a different combination of items in terms of

content areas, which might deviate more from the reference composite and hence affect the measurement precision.

Finally, the three content balancing methods perform comparably. They yield similar biases, MSEs and the correlations between θ and its estimate. However, in terms of MSE, MCCAT tends to have slightly lower values than MMM and CCAT, especially when three content areas have the same difficulty level.

4.3.2 Three-dimensional case:

Table 4.7 summarizes the overall measurement precision for the three-dimensional case. Unlike the two-dimensional case, the maximum information method does not perform the best when three content areas have the same difficulty level. The three content balancing methods yield overall similar values in biases, MSEs and the correlations between θ and its estimate as the maximum information method. Among the three content balancing methods, MCCAT results in the smallest MSEs.

When the three content areas have uneven difficulty levels, the performances of the four methods follow similar patterns as in the two-dimensional case. Clearly, evidenced by larger correlations between θ and its estimate and smaller MSEs, the three content balancing methods give better recovery of person parameter than the maximum information method. In addition, as the correlation between true abilities increases, the bias becomes smaller and the correlation between θ and its estimate becomes larger. For example, for CCAT, the absolute mean bias decreases from 0.052 to 0.031, and the

Table 4.7 Measurement precision for maximum information method and three content balancing methods in three-dimensional case

Ability Correlation	Method	Bias	MSE	$\rho_{\theta,\hat{\theta}}$
Panel A: Difficulty= (0, 0, 0)				
0	Max Information	0.037	0.181	0.950
0	CCAT	0.039	0.161	0.955
0	MMM	0.031	0.178	0.949
0	MCCAT	0.039	0.151	0.955
0.4	Max Information	0.028	0.172	0.962
0.4	CCAT	0.038	0.174	0.963
0.4	MMM	0.029	0.172	0.962
0.4	MCCAT	0.038	0.170	0.962
0.8	Max Information	0.016	0.163	0.964
0.8	CCAT	0.034	0.184	0.967
0.8	MMM	0.029	0.197	0.963
0.8	MCCAT	0.035	0.169	0.968
Panel B: Difficulty= (-0.6, 0.6, 0)				
0	Max Information	0.052	0.238	0.915
0	CCAT	0.059	0.162	0.944
0	MMM	0.050	0.182	0.936
0	MCCAT	0.051	0.193	0.935
0.4	Max Information	0.038	0.167	0.944
0.4	CCAT	0.042	0.135	0.959
0.4	MMM	0.036	0.138	0.956
0.4	MCCAT	0.041	0.155	0.951
0.8	Max Information	0.023	0.151	0.953
0.8	CCAT	0.031	0.147	0.966
0.8	MMM	0.030	0.154	0.962
0.8	MCCAT	0.040	0.136	0.968

correlation between θ and its estimate increases from 0.944 to 0.966, when the ability correlation increases from 0 to 0.8. This is not surprising, since with high correlation between true abilities, the effect of multidimensionality would become smaller and the estimation procedure would become more accurate. Meanwhile, the difference in MSE

between the maximum information method and the three content balancing methods decreases as the correlation between true abilities increases.

4.3.3 Conditional Bias and MSE

In addition, the conditional biases and MSEs for limited points are presented in Figure 4.3 to Figure 4.10. Five equally spaced values of θ from -2 to +2 (θ = -2, -1, 0, 1, 2) are used. Hence, 25 fixed points (5 × 5) are evaluated in the two-dimensional case, and 125 (5 × 5 × 5) points in the three-dimensional case.

Figure 4.3 and 4.4 show the conditional biases for the four methods in the two-dimensional case. Clearly, the maximum information method and the three content balancing methods yield similar conditional biases. The figures demonstrate that the examinees located at the two ends of the θ distribution have more volatile biases, while those who are in the middle have biases close to 0. Generally, with unidimensional data, the estimated θ-values from the three-parameter logistic model have larger measurement errors for high and low ability examinees than for middle ability examinees. In particular, high ability examinees tend to be underestimated and low ability examines tend to be overestimated. However, the shape of conditional biases in this study does not follow the pattern strictly. This is mainly because additional estimation errors are introduced when we fit a unidimensional model to multidimensional data.

Figure 4.5 and 4.6 show the conditional MSEs for the four methods. Again, the four methods perform very similarly. The MSEs are small in the middle while large at the two ends. Also, the MSEs at the lower end are larger than those at the upper end, which is due

to the guessing issue in the three-parameter IRT model. In addition, the difference between the two ends shrinks when the difficulty levels for the three content areas become uneven.

The conditional biases and MSEs for the three-dimensional case generally follow the same patterns as those in the two dimensional case. As shown in figure 4.7 and 4.8, the four methods yield similar conditional biases. Since the estimated θ values are restricted to -4 and 4, the conditional biases are bounded between -4- θ and 4- θ. In addition, in the three-dimensional case, the differences between the maximum information method and the three content balancing methods are more apparent in terms of MSE. In figure 4.9 and 4.10, the MSEs at the lower end of θ distribution for the maximum information method are larger than those for the content balancing methods. This difference is magnified when the difficulty levels for content areas become uneven. The three content balancing methods performed similarly, although MCCAT gives slightly higher MSE than CCAT and MMM at the lower end of θ, especially when the content areas have the same difficult level.

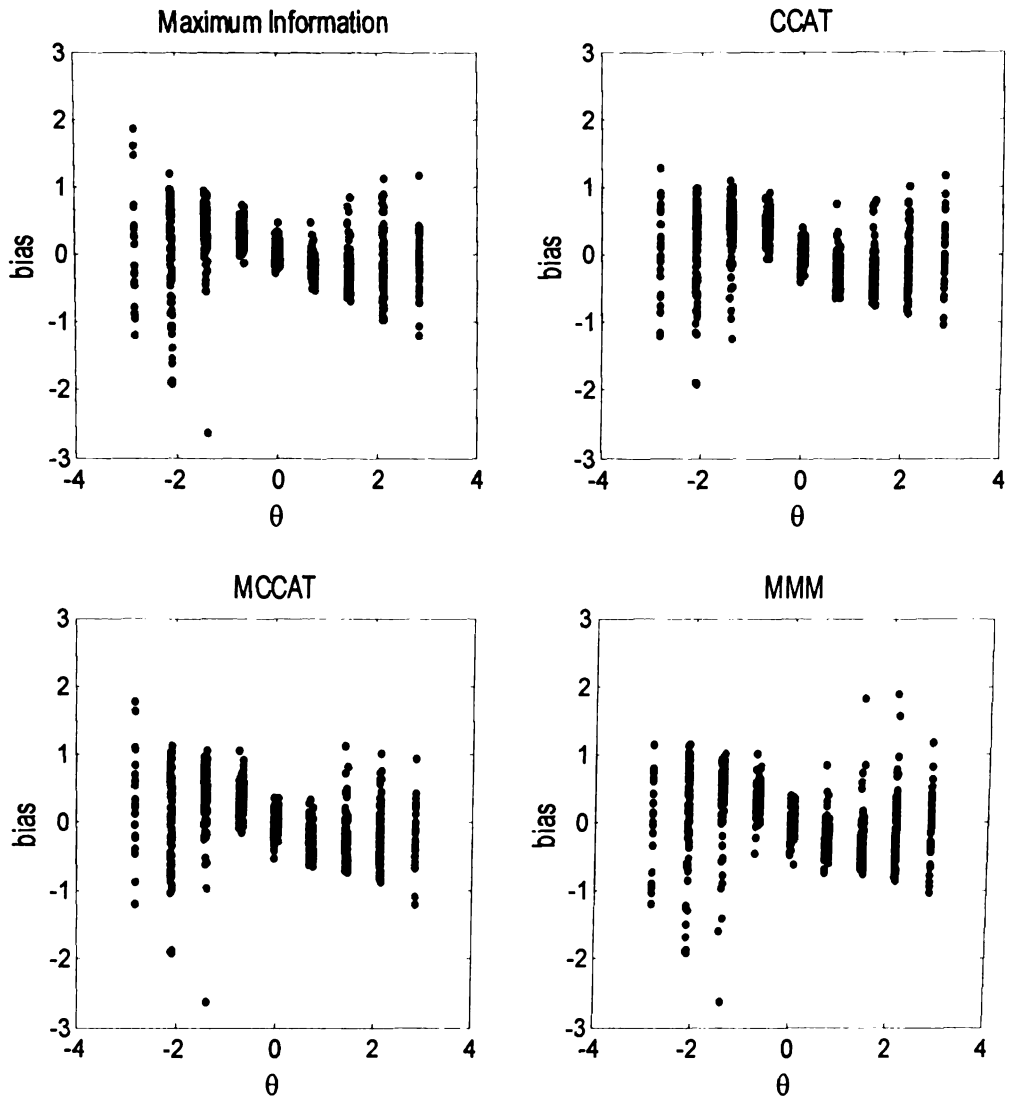


Figure 4.3 Conditional Biases for the four methods, Difficulty=(0, 0, 0), two-dimensional case

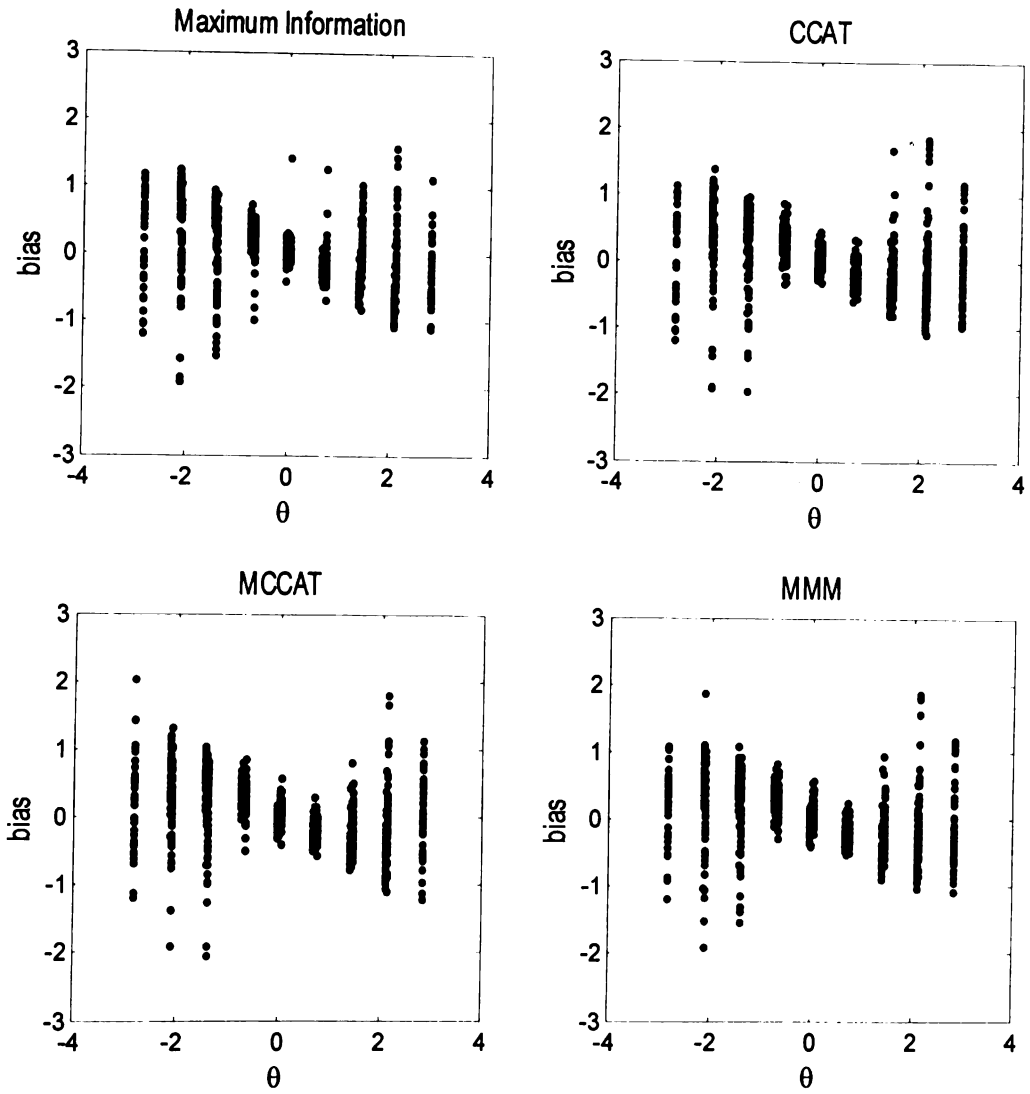


Figure 4.4 Conditional Biases for the four methods, Difficulty=(-0.6, 0.6, 0), two-dimensional case

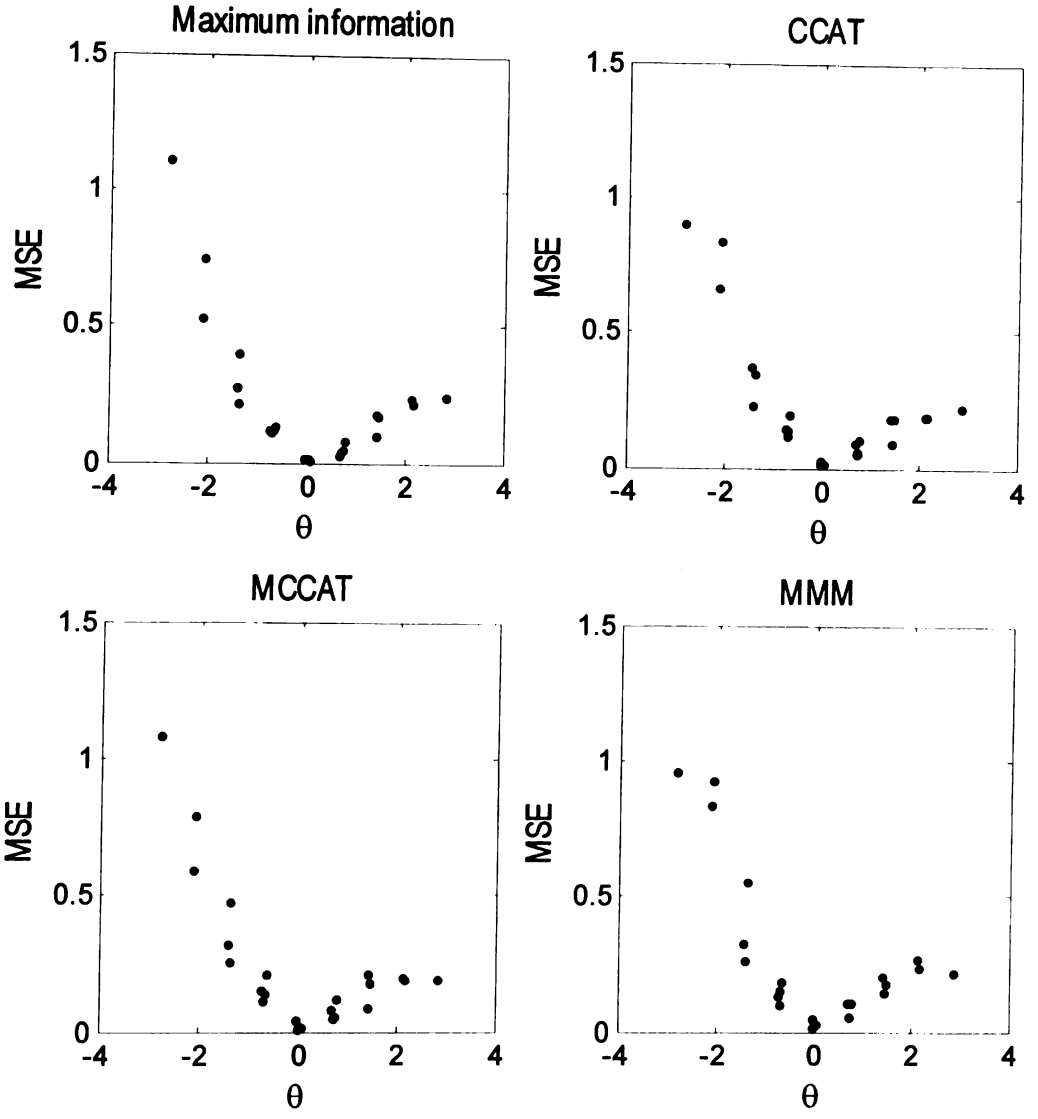


Figure 4.5 Conditional MSEs for the four methods, Difficulty=[0, 0, 0], two-dimensional case

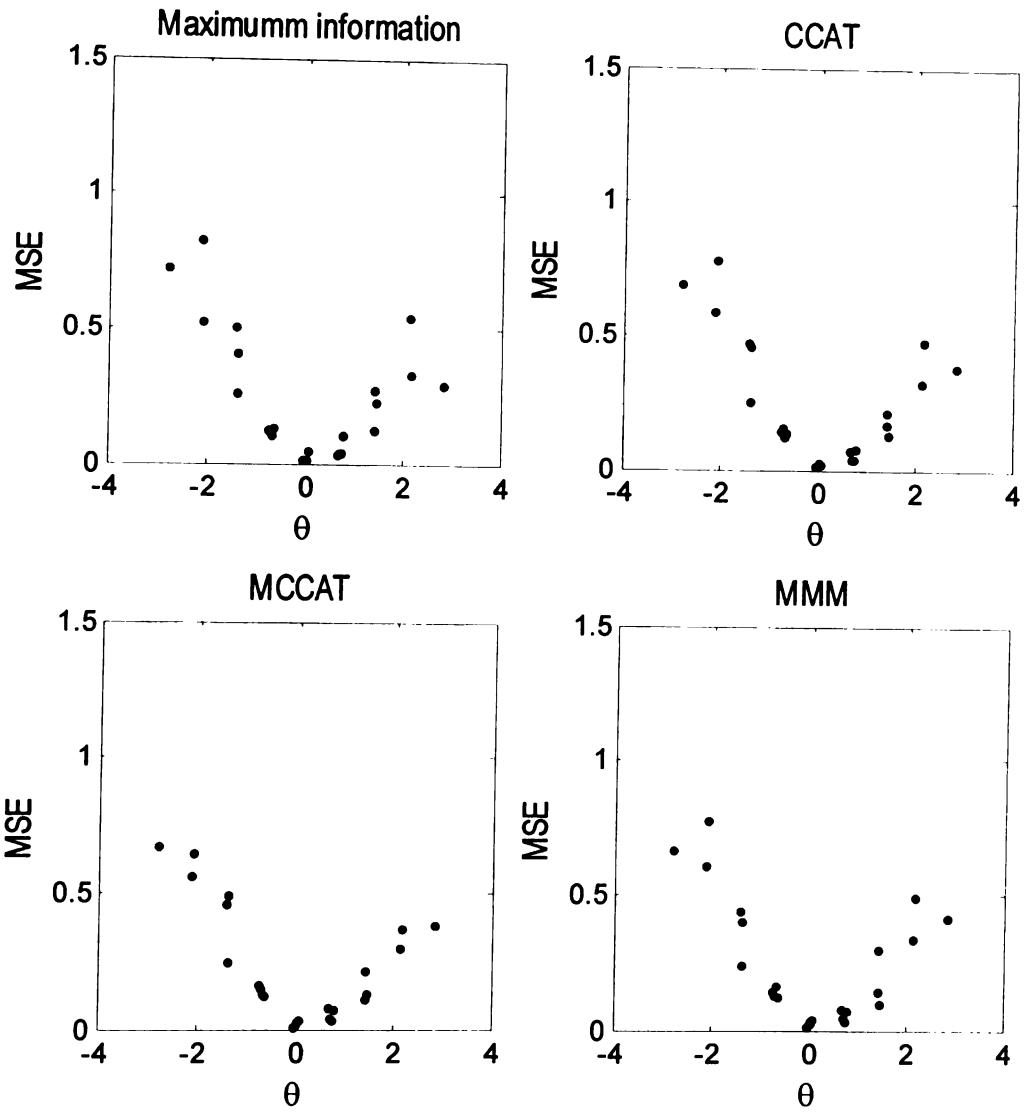


Figure 4.6 Conditional MSEs for the four methods, Difficulty=(-0.6, 0.6, 0), two-dimensional case

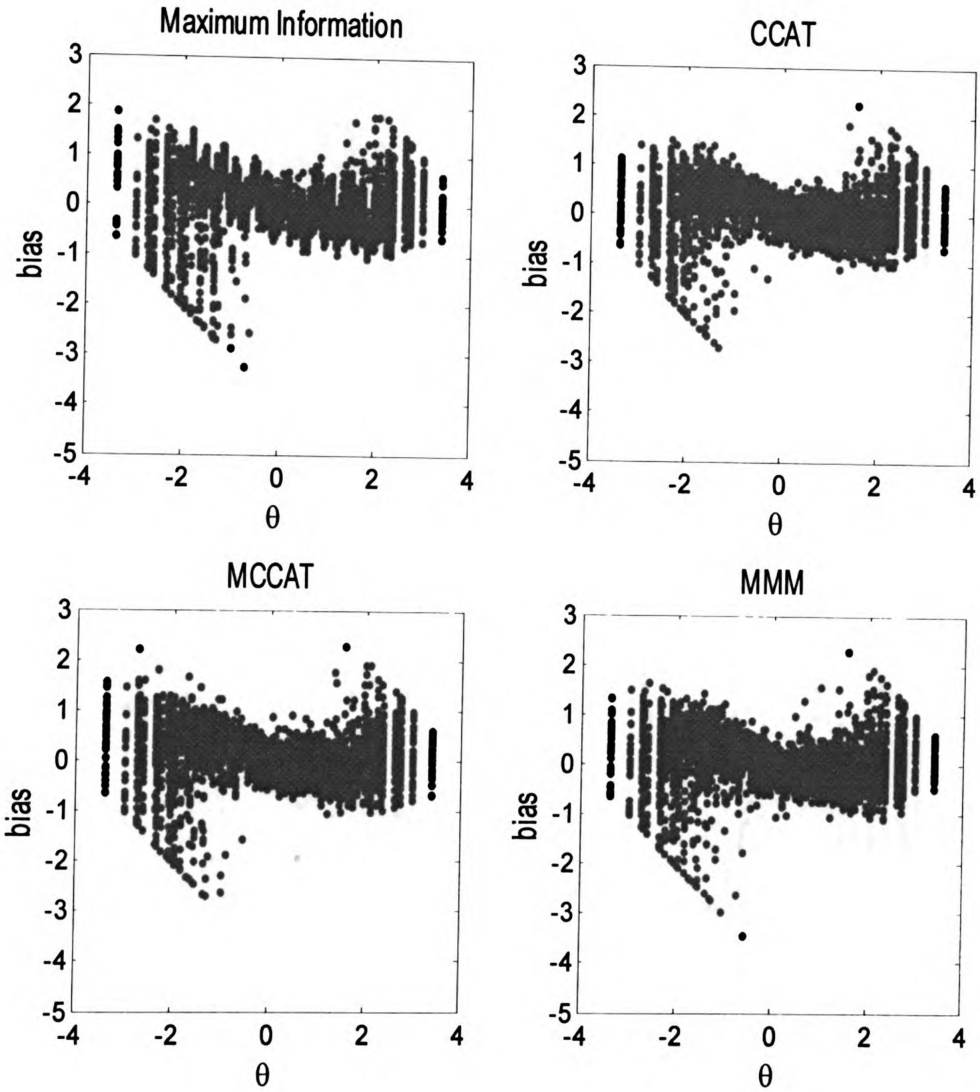


Figure 4.7 Conditional Biases for the four methods, Difficulty=(0, 0, 0), three-dimensional case

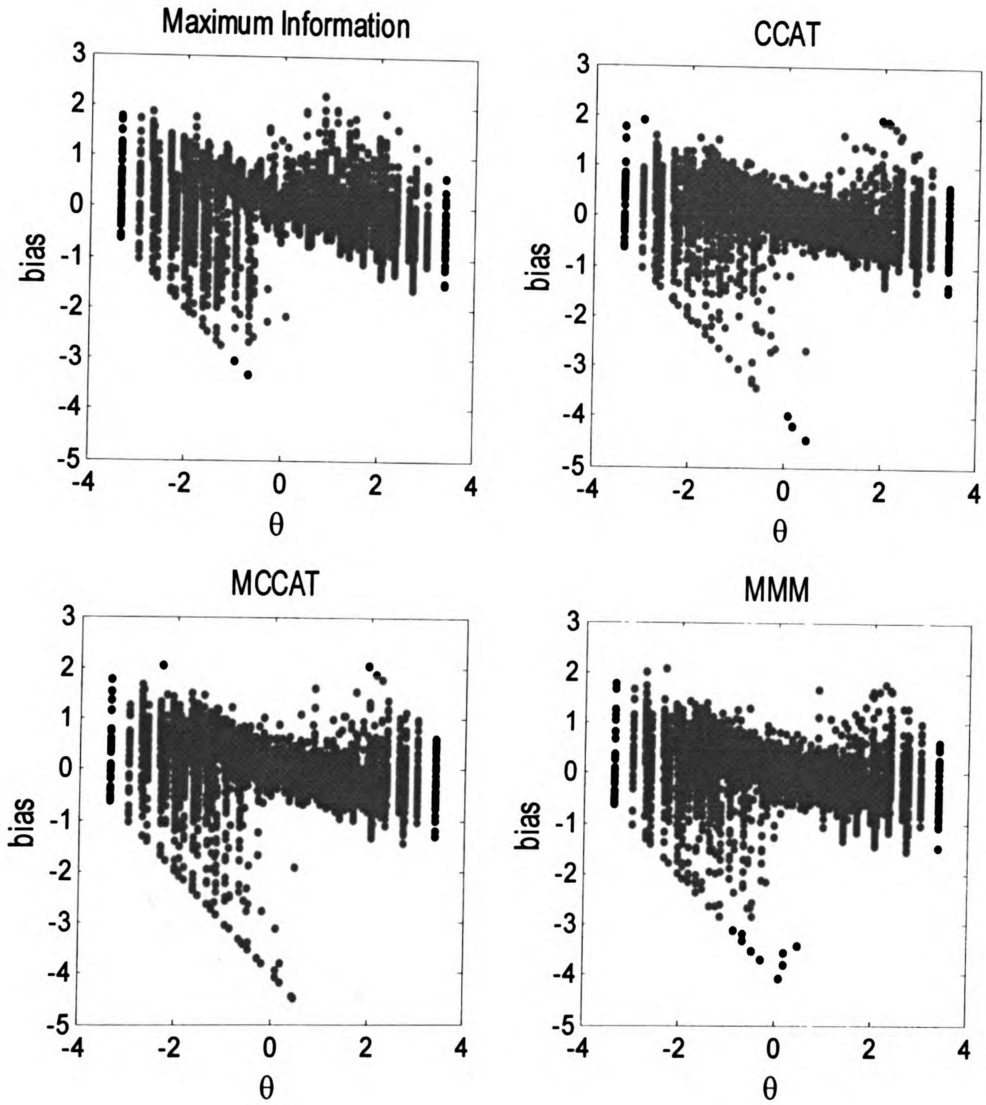


Figure 4.8 Conditional Biases for the four methods, Difficulty=(-0.6, 0.6, 0), three-dimensional case

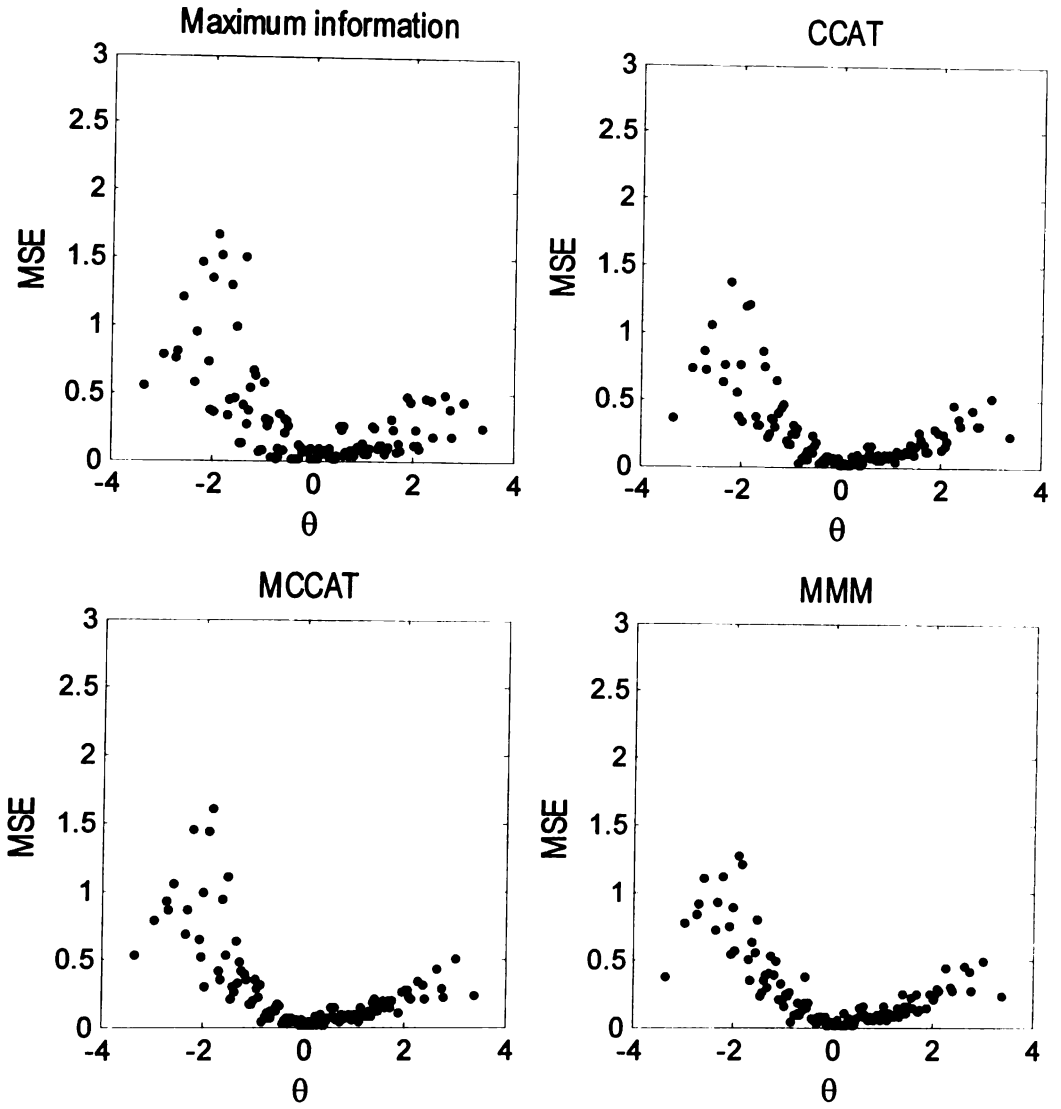


Figure 4.9 Conditional MSEs for the four methods, Difficulty=(0, 0, 0), three-dimensional case

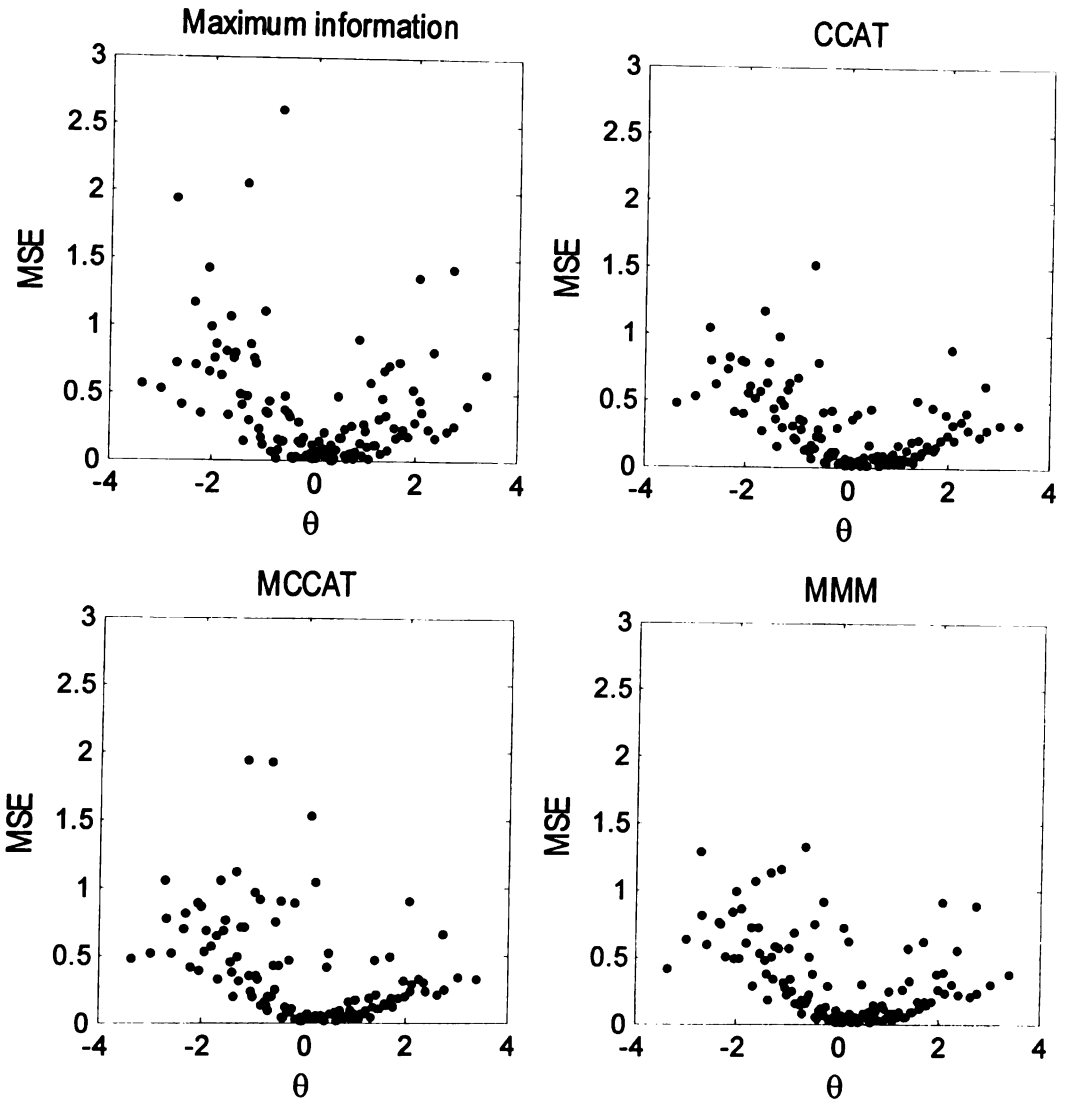


Figure 4.10 Conditional MSEs for the four methods, Difficulty=(-0.6, 0.6, 0), three-dimensional case

4.4 Content balancing

Table 4.8 presents the percentage of tests violating the content-balancing requirement for each method. As expected, the three content-balancing methods perform very well and there is no content violation. In contrast, the maximum information method yields a large number of unbalanced tests. The violation rate ranges from 98.75% to 100% across the conditions. In other words, almost all tests fail to satisfy the content balancing requirement when the maximum information method is used.

Table 4.8 Violation rate of the content-balancing requirement

Difficulty	Correlation	Violation Rate (%)			
		Maximum Information	CCAT	MMM	MCCAT
Panel A: Two-dimensional case					
(0, 0, 0)	0	100	0	0	0
	0.4	98.75	0	0	0
	0.8	98.95	0	0	0
(-0.6, 0.6, 0)	0	100	0	0	0
	0.4	100	0	0	0
	0.8	100	0	0	0
Panel B: Three-dimensional case					
(0, 0, 0)	0	100	0	0	0
	0.4	99.1	0	0	0
	0.8	99.55	0	0	0
(-0.6, 0.6, 0)	0	100	0	0	0
	0.4	100	0	0	0
	0.8	100	0	0	0

Table 4.9 reports the average number of items selected from each content area for the maximum information method. Because two dimensional structures provide very similar results, only the results for the two-dimensional case are discussed.

First, in the two-dimensional case, Content 3 clearly dominates. For example, with even difficulty and uncorrelated abilities, there are on average 2.87 items selected from Content 1, 3.92 items from Content 2, and 23.21 items from Content 3. Intuitively, Content 3 measures the composite of the two dimensions, and hence it is closer to the orientation of the calibrated unidimensional θ-scale. Therefore, the unidimensional discrimination estimates are higher for Content 3 items, which make those items more likely to be administrated when the maximum information method is used.

Table 4.9 The mean number of items selected from each content area for maximum information method

Difficulty	Correlation	Content 1	Content 2	Content 3
Panel A: two-dimensional case				
(0, 0, 0)	0	2.87	3.92	23.21
	0.4	4.39	5.56	20.04
	0.8	6.31	6.36	17.33
(-0.6, 0.6, 0)	0	4.32	6.71	18.97
	0.4	4.81	6.23	18.96
	0.8	5.67	7.51	16.82
Panel B: Three-dimensional case				
(0, 0, 0)	0	2.19	3.47	24.34
	0.4	3.62	4.36	22.02
	0.8	5.30	6.41	18.29
(-0.6, 0.6, 0)	0	3.79	5.45	20.76
	0.4	4.53	6.58	18.90
	0.8	5.48	6.82	17.70

Second, when the three content areas have uneven difficulty levels, relatively more items are chosen from Content 1 and 2, although Content 3 still dominates. For example, with an ability correlation of 0, there are on average 4.32 items selected from Content 1, 6.71items from Content 2, and 18.97 items from Content 3. To understand this change, recall that the mean difficulty is decreased by 0.6 for Content 1 and increased by 0.6 for

Content 2. Therefore, items from Content 1 or 2 become more informative now for those who have very low or high ability and thus have higher chance of being selected.

Finally, as abilities become more correlated, items are selected more evenly across the content areas.

4.5 Item pool usage

Table 4.10 compares the item pool usages for the four methods.

In the two-dimensional case, when the difficulty levels for the three content areas are the same, the maximum information method results in the highest scaled χ^2 statistics and hence has the most unbalanced item pool utilization. For instance, it yields a scaled χ^2 statistics of 103.13 when the correlation between true abilities is zero. In comparison, the three content balancing methods produce the scaled χ^2 statistics no larger than 94.11. This is not surprising, because imposing content constraints forces items to be selected more evenly from all three content areas. As a result, more items are likely to be used. Among the three content balancing methods, MMM performs the best with the lowest scaled χ^2 value, while CCAT does the worst.

The results are different when the difficulty levels for the three content areas are uneven. The maximum information method now becomes more efficient in item pool usage. It yields similar scaled χ^2 value as MMM except for the condition with high ability correlation. This is consistent with the previous founding that uneven difficulty

leads to more items selected from Content 1 and 2. Among the three content balancing methods, MMM still performs the best.

In addition, when the true abilities become more correlated, the scaled χ^2 statistics drop for all methods, which implies a more even item pool utilization when multidimensionality is reduced.

Panel B presents the results for the three-dimensional case, which shows similar patterns as the two-dimensional case. In particular, the maximum information method leads to a more balanced item pool usage when the three content areas have uneven difficulty levels. It yields an even lower scaled χ^2 statistics than the three content balancing methods, although the differences are relatively small.

Table 4.10 item pool usage for four methods

Difficulty	Correlation	Chi-square			
		Max Information	CCAT	MMM	MCCAT
Panel A: Two-dimensional case					
(0, 0, 0)	0	103.13	94.11	89.03	90.3
	0.4	89.45	83.15	77.82	79.62
	0.8	76.68	71.45	67.81	70.74
(-0.6, 0.6, 0)	0	88.85	94.56	88.53	90.15
	0.4	84.02	89.6	85.3	87.42
	0.8	75.71	74.67	70.71	74.44
Panel B: Three-dimensional case					
(0, 0, 0)	0	105.3	101.17	94.08	95.25
	0.4	86.93	84.4	78.78	80.44
	0.8	70.07	69.42	65.62	67.7
(-0.6, 0.6, 0)	0	98.79	106.32	99.81	101.43
	0.4	79.93	88.28	83.94	87.5
	0.8	68.01	74.36	70.37	73.02

4.6 Percentages of underexposed and overexposed items

In this study, we also calculate the exposure rate of each item for the four methods. An item is classified as underexposed if its exposure rate is less than 0.02, and overexposed if larger than 0.2. The percentages of underexposed and overexposed items for different simulation conditions are summarized in Table 4.11 and 4.12. The observations for the two dimensional structures are very similar. For brevity, only the results for the two-dimensional case are discussed.

In the two-dimensional case, when the three content areas have the same difficulty level, the maximum information method tends to produce more underexposed and overexposed items than the three content balancing methods. For example, with a zero ability correlation, 66.5% of items are underexposed for the maximum information method, while the percentage drops to 64.5% for CCAT, 63% for MMM, and 63.5% for MCCAT. Among the three content balancing methods, MMM seems to be the best, although the differences are rather small.

For other conditions, there are no significant differences among these four methods. Generally, about 60% of items are underexposed and about 15% of items are overexposed in the item pool. The high percentages of underexposed and overexposed items are mostly due to the fact that no exposure control technique is employed in this study.

Table 4.11 Percentages of underexposed and overexposed items for the four methods in the two-dimensional case

Correlation	Method	UnderExposed	OverExposed
Panel A: Difficulty= (0, 0, 0)			
0	Max Information	66.5	15.5
0	CCAT	64.5	14.3
0	MMM	63.0	14.0
0	MCCAT	63.5	14.8
0.4	Max Information	64.8	17.0
0.4	CCAT	61.5	15.5
0.4	MMM	60.3	15.3
0.4	MCCAT	61.3	17.0
0.8	Max Information	61.3	15.8
0.8	CCAT	58.8	15.5
0.8	MMM	57.3	15.8
0.8	MCCAT	58.0	16.5
Panel B: Difficulty= (-0.6, 0.6, 0)			
0	Max Information	62.0	15.5
0	CCAT	64.5	15.0
0	MMM	63.3	15.3
0	MCCAT	65.0	16.3
0.4	Max Information	63.0	16.3
0.4	CCAT	63.3	14.3
0.4	MMM	61.5	15.0
0.4	MCCAT	63.5	15.0
0.8	Max Information	61.3	17.0
0.8	CCAT	60.0	15.0
0.8	MMM	58.8	14.8
0.8	MCCAT	60.0	15.3

Table 4.12 Percentages of underexposed and overexposed items for the four methods in the three-dimensional case

Correlation	Method	UnderExposed	OverExposed
Panel A: Difficulty= (0, 0, 0)			
0	Max Information	68.5	15.5
0	CCAT	65.8	14.5
0	MMM	64.5	14.8
0	MCCAT	65.5	15.3
0.4	Max Information	64.3	15.0
0.4	CCAT	60.5	16.3
0.4	MMM	61.3	15.5
0.4	MCCAT	61.5	16.3
0.8	Max Information	59.0	17.5
0.8	CCAT	57.5	17.3
0.8	MMM	56.5	17.3
0.8	MCCAT	57.8	16.5
Panel B: Difficulty= (-0.6, 0.6, 0)			
0	Max Information	64.8	14.5
0	CCAT	65.8	15.3
0	MMM	64.0	14.5
0	MCCAT	65.3	15.0
0.4	Max Information	62.0	15.5
0.4	CCAT	62.5	15.8
0.4	MMM	61.0	15.3
0.4	MCCAT	62.5	16.3
0.8	Max Information	59.3	16.5
0.8	CCAT	60.0	15.5
0.8	MMM	58.5	14.5
0.8	MCCAT	59.0	15.8

Chapter 5

Conclusions and Discussions

Most of the current CAT programs are based on the assumption that a unidimensional IRT model represents the interactions between persons and test items. However, many researchers have argued that this assumption rarely holds in the real world and multiple abilities are required to account for the performance on a test. Meanwhile, content balancing is also a practical consideration in CAT, since lack of content comparability could pose a threat to the validity of scores, and may not be acceptable to test takers and test score users. The purpose of this study is to investigate the effect of fitting a unidimensional IRT model to multidimensional data in a content-balanced CAT. Specifically, unconstrained CAT with maximum information item selection method is chosen as the baseline, and the performances of the three content balancing procedures, the constrained CAT (CCAT), the modified multinomial model (MMM), and the modified constrained CAT (MCCAT), are evaluated in terms of measurement precision, item pool utilization and item exposure control.

5.1 Conclusions

Prior research has shown that when test data is unidimensional, unconstrained CAT with the maximum information method gives the best measurement precision (Kingsbury and Zara, 1991; Cheng, Chang, & Yi, 2007). In contrast, the use of content balancing increases the acceptance of the tests by practitioners, but may cause some loss in the measurement accuracy.

However, the present study shows that when test data is multidimensional, the content balancing methods actually result in similar or even better accuracy than the maximum information method. The reason might be that controlling the percentages of items from individual content areas insures adequate representation of each dimension of the data, and hence improves the measurement precision.

The results also show that the difficulty level of the content areas is a significant factor that affects the performances of the four methods. When the content areas have the same difficulty level, the content balancing methods yield comparable measurement precisions to the maximum information method. In particular, they produce similar biases, MSEs, and correlations between the true and estimated ability. In addition, the content balancing methods tend to result in more efficient item pool utilization, and slightly lower percentages of underexposed items.

On the other hand, when the content areas have uneven difficulty levels, the content balancing methods outperform the maximum information method in terms of measurement precision. However, the differences shrink as the correlation between true abilities increases. Moreover, the maximum information method becomes relatively more efficient in item pool usage. In terms of the percentages of underexposed and overexposed items, there is no significant difference between the four methods.

The study also shows that the results for the two dimensional structures are generally consistent, which indicates the results might be generalized to a higher dimensional space.

Finally, there is no significant difference between the three content balancing methods. They perform similarly in terms of measurement precision and item exposure rate. However, MMM appears to have the most efficient item pool utilization. It yields the smallest scaled χ^2 statistics among the three methods across all conditions.

Table 5.1 Comparison between the maximum information method and the three content balancing methods.

Difficulty	Method	Measurement Precision	Content balancing	Item pool usage	Under exposed	Over exposed
(0, 0, 0)	Max Information	similar				similar
	Content Balancing		√	√	√	
(-0.6, 0.6, 0)	Max Information			similar	similar	similar
	Content Balancing	√	√			

Table 5.2 Comparison between the three content balancing methods.

Difficulty	Method	Measurement Precision	Item pool usage	Under exposed	Over exposed
(0, 0, 0)	CCAT MMM MCCAT	√ (Smaller MSE)	√	similar	similar
(-0.6, 0.6, 0)	CCAT MMM MCCAT	similar	√	similar	similar

Table 5.1 and 5.2 summarize the conclusions of this study. Overall, the content balancing methods are better than the maximum information method, especially for tests with low correlations in the constructs. They not only produce content-balanced tests for examinees and increase the acceptance of the adaptive test by practitioners, but also

improve the measurement efficiency, particularly when the content areas have uneven difficulties. On the other hand, the three content balancing methods perform similarly, but MMM gives the most efficient item pool usage.

The current study has an important practical implication for CAT. Previous literature has shown that using content balancing may induce a loss in measurement precision for unidimensional data. In contrast, we show that it may improve measurement precision when a unidimensional model is fit to multidimensional data. In real testing programs, unidimensional IRT models are often used because of their simplicity and popularity. Meanwhile, many studies found evidences of multidimensionality in real data. For example, Reckase et al. (1988) examined a test from the ACT Assessment Battery and showed that the test was clearly multidimensional. Therefore, if the context in this study resembles the reality more closely, then content balancing is recommended for its improved acceptance by practitioners and better measurement precision.

5.2 Future research

There are several potential extensions to this simulation study. First, we only examine a limited set of item pool structures and the item parameters are simulated from commonly used distributions. Alternatively, we can examine more general and realistic item pool structures generated from real test data. Those "real data" evidences can serve as a good complement to our "pure simulation" results.

Second, the results from this study only apply to fixed-length CATs. Fixed-length CATs are easy to implement in practice, but they might lead to aberrant response patterns

(Chen & Ankenmann, 2004). Therefore, examining variable-length CATs can provide an important robustness check.

Third, in the current simulation study, the three content balancing methods, CCAT, MMM, and MCCAT, are fixed content balancing methods. That is, the number of items from each content area is fixed. However, flexible content balancing is used in several large-scale CAT programs. It allows the number of items from each content area to be between a lower bound and an upper bound (Stocking & Swanson, 1993). Many methods have been developed to handle flexible content-balancing control, such as the weighted deviation model (WDM) (Stocking and Swanson, 1993), the shadow test approach (van der Linden, 2005b; van der Linden and Chang, 2003), and the weighted penalty model (Shin, Chien, Way, & Swanson, 2009). These methods can handle many practical constraints, including item content and item type. It would be interesting to investigate how these flexible content balancing methods perform in the current context.

Finally, this study yields a large percentage of underexposed and overexposure items, because no exposure control is applied. However, in addition to content balancing, item exposure control is another important practical consideration in CAT. Using item selection purely based on maximum information, some items may be administrated too frequently and become known to test takers. As a result, test security and reliability can be threatened. At the same time, when a small proportion of items are over-selected, there are also a large number of items in the item pool rarely used. Therefore, to increase test efficiency and security, mechanics needs to be imposed on the item selection procedure to control the exposure rate of items. This issue has been addressed in great detail in the literature (Chang & Ansley, 2003; Georgiadou, Triantafillou & Economides, 2007; Hetter

& Sympson, 1997; Pastor, Dodd & Chang, 2002; Stocking & Lewis, 1998, 2000; Way, 1998), and a number of strategies for controlling item exposure have been developed (e.g., alpha-stratified design, Chang, Qian & Ying, 2001; Sympson & Hetter's method, 1985). So a natural extension would be incorporating those item exposure control methods into the study. This can make our CAT procedure more realistic, and it would be also interesting to examine the interactions between content balancing and item exposure control.

REFERENCES

Ackerman, T.A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15(1),* 13-24.

Ansley, T.N. (1984). Using a unidimensional latent trait model with multidimensional data: An empirical investigation of robustness. Unpublished doctoral dissertation, University of Iowa, Iowa city, IA.

Bejar, I.I., & Weiss, D.J. (1979). *Computer programs for scoring test data with item characteristic curve models* (Research Rep. No. 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Binet, A., & Simon, T.A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191-244.

Bock, R.D., & Mislevy, R.J. (1981). *Data quality analysis of the Armed Services Vocational Aptitude Battery.* Chicago: National Opinion Research Center.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Measurement in Education, 6,* 431-444.

Bolt, D.M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement,27(6),* 395-414.

Chang, H., & Ying, Z. (1999). A-stratified computerized adaptive testing. *Applied Psychological Measurement, 23,* 211-222.

Chang, H., Qian, J., & Ying, Z. (2001). A-stratified computerized adaptive testing with b-blocking. *Applied Psychological Measurement, 24,* 333-341.

Chang, S.W., & Ansley, T.N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40(1),* 71-103.

Chen, S., & Ankenmann, R.D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41(2),* 149-174.

Cheng, Y., Chang, H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Measurement in Education, 31(6),* 467-482.

Eignor, D.R., Stocking, M.L., Way, W.D., & Steffen, M. (1993). *Case studies in computerized adaptive test design through simulation* (Research Rep. No. RR-93-56). Princeton, NJ: Educational Testing Service.

Educational Testing Service (2007). *Test and score data summary for TOEFL computer-based and paper-based test.* Princeton, NJ: Educational Testing Service.

Fang, Y. (2008). *Using a projection method to estimate subscores from tests with multidimensional structures.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. . *Journal of Technology, Learning, and Assessment, 5(8).* Retrived [data] from from http://www.jtla.org

Green, B. G., Bock, R.D., Humphries, L. G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

Hetter, R. D., & Sympson, J.B. (1997). Item exposure control in CAT-ASVAB. In W.A. Sands, B.K.Waters & J.R. MacBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Ip, E.H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63,* 395-416.

Kinsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

Kinsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4(3),* 241-261.

Leung, C.K., Chang, H., & Hau, K. (2000). *Content balancingin stratified computerized adaptive testing designs.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Leung, C.K., Chang, H., & Hau, K. (2003a). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63,* 257-270.

Leung, C.K., Chang, H., & Hau, K. (2003b). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment, 2(5).* Available from http://www.jtla.org

Liu, J. (2007). Comparing multi-dimensional and uni-dimensional computer adaptive strategies in psychological and health assessment. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associations, Inc.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameters estimation in item response theory. *Journal of Educational Measurement, 23,* 157-162.

Luecht, R.M. (1998). *A framework for exploring and controlling risks associated with test item exposure over time.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Mckinley, R.L., & Reckae, M.D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Rep. No. 83-2). Iowa city, IA: ACT, Inc.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of American Statistical Association, 70,* 351-356.

Pastor, D.A., Dodd, B., & Chang, H.H. (2002). A comparison of item exposure selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Measurement in Education,26(2),* 147-163.

Patsula, L.N., & Steffan, M, (1997). *Maintaining item and test security in a CAT environment: A simulation study.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9(4),* 401-412.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8(3),* 11-15.

Reckase, M.D. (2009). *Multidimensional item response theory.* New York, Springer.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25(3),* 193-203.

Reckase, M.D., & Mckinley, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15(4),* 361-373.

Roussos, L.A., Stout, W.F., & Marden, J.L. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35(1),* 1-30.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

Segall, D.O., Moreno, K.E., & Hetter, R.D. (1997). Item pool development and evaluation. In W.A. Sands, B.K.Waters & J.R. MacBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117-130). Washington, DC: American Psychological Association.

Shin, C.W., Chien, Y.M., Way, W.D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CAT.* Iowa city, IA: Pearson.

Stocking, M.L. (1998). *A framework for comparing adaptive test designs.* Unpublished manuscript.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23(1),* 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston: Kluwer.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.

Swaminathan, H., & Rogers, J.L. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 361-370.

Sympson, J. (1978). A model for testing with multidimensional items. In D.J.Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference.* Minneapolis MN: University of Minnesota.

Sympson, J.B., & Hetter, R.D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* In Proceedings of the 27th annual meeting of the Military Testing Association, (pp.973-977). San Diego CA: Navy Personnel Research and Development Center.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, British Columbia: Educational Research Institute of British Columbia.

U.S. Department of Defense. (1982). *Armed Services vocational aptitude battery.* North Chicago, IL: U.S. Military Entrance Processing Command.

van der Linden, W.J. (2005a). *Linear models for optimal test design.* New York: Springer-Verlag.

van der Linden, W.J. (2005b). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42(3),* 283-302.

van der Linden, W.J., & Chang, H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27(2),* 107-120.

Wainer, H. (1993). Some practical considerations when converting a linearly administrated test to an adaptive format. *Educational Measurement: Issues and Practices, 12,* 15-20.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 6,* 473-492.

Walker, C.M., & Beretvas, S.N. (2003). Comparing multidimensional and unidimensional proficiency classifications: multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40(3),* 255-275.

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effect of latent space misspecification on the application of IRT* (Research Rep. No. NW: 6-24-85). Iowa city, IA: University of Iowa.

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the ONR contractors conference, Gatlinburg, TN.

Wang, T., & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35(2),* 109-135.

Warm, T.A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54,* 427-450.

Way, D.W. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17,* 17-27.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weiss, D.J., & Suhadolnik, D. (1982). Robustness of adaptive testing to multidimensionality. In D.J. Weiss (Ed.), *Proceedings of the 1982 item response theory and computerized adaptive testing conference.* Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Zimwoski, M.M., Muraki, E., Mislevy, R.J., & Bock, D.J. (2003). *BILOG-MG for Windows*. Scientific Software International, Inc., Lincolnwood, IL.