



142  
210  
THS

LIBRARY  
Michigan State  
University

This is to certify that the  
dissertation entitled

COMPARISON OF ABILITY ESTIMATION AND ITEM  
SELECTION METHODS IN MULTIDIMENSIONAL  
COMPUTERIZED ADAPTIVE TESTING

presented by

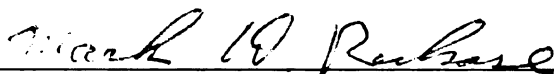
Qi Diao

has been accepted towards fulfillment  
of the requirements for the

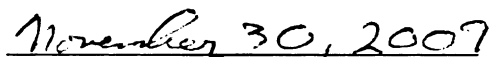
Doctoral

degree in

Measurement and Quantitative  
Methods



Major Professor's Signature



Date

**PLACE IN RETURN BOX** to remove this checkout from your record.  
**TO AVOID FINES** return on or before date due.  
**MAY BE RECALLED** with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
10 20 12 20016 200		

**COMPARISON OF ABILITY ESTIMATION AND ITEM  
SELECTION METHODS IN MULTIDIMENSIONAL  
COMPUTERIZED ADAPTIVE TESTING**

**By**

**Qi Diao**

**A DISSERTATION**

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

**DOCTOR OF PHILOSOPHY**

**Measurement and Quantitative Methods**

**2009**



# **ABSTRACT**

## **COMPARISON OF ABILITY ESTIMATION AND ITEM SELECTION METHODS IN MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING**

By

Qi Diao

The impetus of the study is the lack of guidance in the literature of multidimensional computerized adaptive testing (MCAT) in terms of which item selection and ability estimation methods to use and under what condition. This study did a comprehensive comparison of ability estimation and item selection methods in MCAT. Two ability estimation methods included maximum likelihood estimation and Bayesian estimation method. The item selection methods can be divided into three categories, item selection methods associated with maximum likelihood, item selection with Bayesian with Fisher's information, and item selection method with Kullback-Leibler information. The comparison was made conditioning on such factors as test length, use of priors, etc. Simulations were based on real data from 2005 Michigan Educational Assessment Program. As the result of the study, recommendations were made which method should be used under certain condition. It is believed that the results of the study can help future researchers in selecting ability estimation and item select methods when conducting their own research in MCAT and help the construction of operational MCAT procedures.

To my dear husband Hao Ren, and my parents

## **ACKNOWLEDGEMENTS**

I would like to express my sincere appreciation to my major dissertation advisor, Dr. Mark Reckase; for his guidance, support, and encouragement throughout my dissertation study. Dr. Reckase inspired me through his passion for research, and excellence in teaching and professionalism.

I would also like to express my sincere gratitude and thanks to Dr. Kimberly Maier, for her continuous support and guidance in my course selections, study and my research. I would also like to give my thanks to Dr. Sharif Shakrani for his encouragement and guidance throughout my Ph.D. study and my assistant work. His passion in applying research to aid the improvement of education inspired my choice of career. I would like to thank Dr. Lijian Yang for his consistent guidance of my statistical studies.

I would like to thank CTB/McGraw-Hill companies for their support of this study. This study is partially funded by CTB and without all the support from the research department of CTB, I would not have been able to complete this study while working as a research scientist at CTB. I would like to give my special thanks to Dr. Wim van der Linden for his support and guidance throughout this study.

I would like to thank my husband Hao Ren for his overwhelming support, encouragement and love. Without him, none of this would have been possible. Finally, I would like to thank my beloved parents and friends. Thank you!

# TABLE OF CONTENTS

<b>LIST OF TABLES .....</b>	<b>vii</b>
-----------------------------	------------

<b>LIST OF FIGURES .....</b>	<b>vii</b>
------------------------------	------------

<b>CHAPTER 1. INTRODUCTION.....</b>	<b>1</b>
-------------------------------------	----------

1.1 Multidimensional Item Response Model.....	1
---	---

1.2 Components of a CAT procedure.....	3
--	---

<b>CHAPTER 2. ABILITY ESTIMATION AND ITEM SELECTION METHODS IN MCAT.....</b>	<b>6</b>
--	----------

2.1 Ability Estimation Methods .....	6
--------------------------------------	---

2.1.1 Maximum Likelihood Method.....	6
--------------------------------------	---

2.1.2 Bayesian Estimation Method.....	7
---------------------------------------	---

2.1.3 Other Ability Estimation Methods .....	8
--	---

2.2 Item Selection Methods .....	9
----------------------------------	---

2.2.1 Maximizing the Determinant of the Fisher Information Matrix (D-optimality) .....	9
---	---

2.2.2 Minimizing the Trace of Inverse of Fisher Information Matrix (A-optimality) .....	10
--	----

2.2.3 Largest Decrement in the Volume of Bayesian Credibility Ellipsoid .....	11
---	----

2.2.4 Maximizing the Kullback-Leibler Information .....	12
---	----

2.2.5 Other Item Selection Methods.....	13
---	----

<b>CHAPTER 3. RESEARCH QUESTIONS AND METHODS.....</b>	<b>14</b>
---	-----------

3.1 Research Questions .....	14
------------------------------	----

3.2 Research Methods.....	17
---------------------------	----

3.2.1 Real Data Used .....	17
----------------------------	----

3.2.2 Simulation.....	20
-----------------------	----

<b>CHAPTER 4. RESULTS AND DISCUSSION.....</b>	<b>26</b>
---	-----------

<b>CHAPTER 5. CONCLUSIONS AND FUTURE RESEARCH DIRECTION.....</b>	<b>103</b>
--	------------

## LIST OF TABLES

Table 3.1 All conditions in comparing different ability estimation method.....	16
Table 3.2 All conditions in comparing item selection methods.....	17
Table 3.3 MIRT item parameters for Grade 7 Michigan Educational Assessment Program (MEAP) Mathematics test from Li (2006). ....	18
Table 3.4 Correlation coefficients among 3 dimensions on Grade 7 Michigan Education Assessment Program (MEAP) Mathematics Test.....	20
Table 3.5 Distributions for multidimensional item parameter generation mimicking Grade 7 Michigan Education Assessment Program (MEAP) Mathematics test. 100 items were generated for each dimension. ....	21
Table 3.6 All 11 simulated conditions. All simulations are for 27 true ability points, 50 replicates at each point. ....	23
Table 4.1 Items administered, responses and updated ability estimates after each item for one examinee. The combination was maximum likelihood and D-optimality. Initial estimate was (0, 0, 0) and the true location was (1, 1, 1).....	34
Table 4.2 Computation time for each examinee (Unit: second).....	102

## LIST OF FIGURES

Figure 4.1 Mean biases and RMSEs for maximum likelihood as the ability estimation method and D-optimality as the item selection method, at test length =20 and at test length =50.....	27
Figure 4.2 Mean and standard deviation of Euclidean distance for maximum likelihood as the ability estimation method and D-optimality as the item selection method, at test length =20 and at test length =50.....	31
Figure 4.3 Successive progress plot of updated ability estimates and true location point after administering each item for maximum likelihood method and D-optimality. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	32
Figure 4.4 Euclidean distance of between updated ability estimates and true location point after administering each item for maximum likelihood method and D-optimatlity. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50. ....	33
Figure 4.5 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and identity matrix as variance covariance matrix, at test length =20 and at test length =50.....	36
Figure 4.6 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and identity matrix as variance covariance matrix, for test length =20 and for test length =50.....	40
Figure 4.7 Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with identity matrix. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	42
Figure 4.8 Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with identity matrix. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50. ....	43
Figure 4.9 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and diag (9) matrix as variance covariance matrix, at test length =20 and at test length =50.....	44

Figure 4.10 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and diag (9) matrix as variance covariance matrix, at test length =20 and at test length =50.....	48
Figure 4.11 Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with diag(9) variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	49
Figure 4.12 Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with diag(9) variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	50
Figure 4.13 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and true variance covariance matrix as variance covariance matrix, at test length =20 and at test length =50.....	51
Figure 4.14 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and true variance covariance matrix as variance covariance matrix, at test length =20 and at test length =50.....	55
Figure 4.15 Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with true variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	56
Figure 4.16 Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with true variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	57
Figure 4.17 Mean biases and RMSEs for Bayesian as the ability estimation method and Kullback-Leibler information as the item selection method, at test length =20 and at test length =50.....	58
Figure 4.18 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and Kullback-Leibler information as the item selection method, at test length =20 and at test length =50 .....	62

Figure 4.19 Successive progress plot of updated ability estimates and true location point after administering each item for Kullback-Leibler. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	63
Figure 4.20 Euclidean distance of between updated ability estimates and true location point after administering each item for Kullback-Leibler. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.....	64
Figure 4.21 Mean biases and RMSEs for maximum likelihood as the ability estimation method, with D-optimality and A-optimality at the item selection methods, test length =50.....	65
Figure 4.22 Mean and standard deviation of Euclidean distance for the combination with maximum likelihood as ability estimation method, comparison of D-optimality and A-optimality as item selection methods.....	69
Figure 4.23 Mean biases and RMSEs for Bayesian as the ability estimation method, comparison of prior variance covariance matrix as: 1) identity matrix; 2) diag (9) and 3) true variance covariance matrix. Test length=20 .....	71
Figure 4.24 Mean biases and RMSEs for Bayesian as the ability estimation method, comparison of prior variance covariance matrix as: 1) identity matrix; 2) diag (9) and 3) true variance covariance matrix. Test length=50 .....	75
Figure 4.25 Mean and standard deviation of Euclidean distance of Bayesian as the ability estimation method, comparison among prior variance covariance matrix: 1) identity matrix, 2) diag(9), and 3) true variance covariance matrix.....	79
Figure 4.26 Mean biases and RMSEs for comparison of maximum likelihood method and Bayesian method. Test length=20.....	82
Figure 4.27 Mean biases and RMSEs for comparison of maximum likelihood method and Bayesian method. Test length=50.....	85
Figure 4.28 Means and standard deviations of Euclidean distance, comparison of maximum likelihood method and Bayesian method. Test length=20 and Test length=50.....	89
Figure 4.29 Mean biases and RMSEs of the comparison of Kullback-Leibler and Volume decrement in Bayesian. Variance covariance of priors is identity matrix. Test length=20.....	92
Figure 4.30 Mean biases and RMSEs of the comparison of Kullback-Leibler and Volume decrement in Bayesian. Variance covariance of priors is identity matrix. Test length=50.....	95



Figure 4.31 Means and standard deviations of Euclidean distance, comparison of  
Kullback-Leibler information and volume decrement in Bayesian with Fisher's  
information .....99

# **CHAPTER 1. INTRODUCTION**

Computerized adaptive testing (CAT) has been widely used in many testing programs (e.g. the Graduate Management Admission Test and the Armed Services Vocational Aptitude Test Battery). It is based on the principle of selecting items to match the current proficiency estimate of an examinee. Adaptive tests have many potential advantages, such as improved measurement precision, reduced test time, and flexible individual testing time. Ample research has been done on unidimensional CAT (e.g., van der Linden & Glas 2000; Wainer 2000; Bock & Mislevy 1988). However, only a few studies have been done on multidimensional computerized adaptive testing (MCAT) (e.g., Segall, 1996; Veldkamp & van der Linden, 2002, Reckase 2009). This study will compare methods used in two important parts in adaptive testing: ability estimation and item selection methods under different conditions. It is believed that the results of the study can help future researchers in selecting ability estimation and item select methods when conducting their own research in MCAT and help the construction of operational MCAT procedures.

## **1.1 Multidimensional Item Response Model**

A basic unidimensional model for dichotomously scored response is the three parameter logistic (3PL) model (Birnbbaum, 1968). In this model, the probability of person  $j$  with ability  $\theta_j$  answers item  $i$  correctly is:

$$P(U_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \quad (1.1)$$

where  $a$  is the discrimination parameter;  $b$  is item difficulty parameter; and  $c$  is pseudo guessing parameter. More detailed description of the parameters can be found in McDonald (1999).

One basic assumption of any item response model is local independence. Local independence means that given one examinee, his/her answer to one test item does not influence the probability of his/her answer to another item except through parameter  $\theta$ . Also, one examinee's answer to one test item does not influence another examinee's answer. In unidimensional cases, the assumption is the same as:

$$P(U_1 = u_1, U_2 = u_2, \dots, U_n = u_n | \theta) = \prod_{i=1}^n P_i(\theta)^{u_i} (1 - P_i(\theta))^{1-u_i}, \quad (1.2)$$

So the probability of an examinee getting a set of observed responses  $u_1, u_2, \dots, u_n$  is only a function of item parameters and examinee's ability parameter  $\theta$ .

However, if more than one ability dimensions are measured in the test, the unidimensional models may not fit and multidimensional response models are needed in order to satisfy the local independence assumption. The MIRT model used in this study in a generalization of model (1.1) into multidimensional space:

$$P(U_{ij} = 1 | \theta_j, \mathbf{a}_i, c_i, d_i) = c_i + (1 - c_i) \frac{e^{1.7(\mathbf{a}_i \theta'_j + d_i)}}{1 + e^{1.7(\mathbf{a}_i \theta'_j + d_i)}}, \quad (1.3)$$

where  $\theta$  is a  $1 \times m$  vector of examinee  $j$ 's ability coordinates with  $m$  is the number of dimensions in the coordinate space.  $\mathbf{a}$  is  $1 \times m$  discrimination parameter.  $c$  and the intercept term  $d$  are scalars.

In general, there are at least three motivations for developing MCAT. The first one is the same as said above: for many operational tests, the unidimensional models may not fit. Multidimensional response models are needed in order to satisfy the assumption of local independence. The second motivation is that for testing for diagnostic purposes we want to extract as much information as possible and for correlated ability dimensions information from one dimension can help measure ability in another dimension. The second motivation also leads to the third: efficiency. Because we can use information from correlated abilities, multidimensional adaptive testing can further make the ability estimation process more efficient.

## **1.2 Components of a CAT Procedure**

For any adaptive test, five key questions need to be answered: 1) which model to use; 2) how to select the first item; 3) how to update ability estimate after an examinee gives the response; 4) how to select the next item; 5) how to end the test. So in order to develop any adaptive test, ability estimation and item selection methods are very fundamental. This research is targeted at investigating them in multidimensional cases.

There has been some research done in unidimensional CAT to investigate the properties of ability estimation and item selection methods (e.g. Weiss & McBride,

1984; van der Linden & Pashley, 2000). However, in the current literature on multidimensional adaptive testing, most studies are done using a single ability estimation and item selection method because they focus on other aspects of adaptive testing (e.g. Li Ip & Fuh, 2008). The only study that concentrated on a comparison of different ability estimation and item selection methods for multidimensional adaptive testing was Tam (1992). But that was before most currently used methods (e.g. Segall, 1996; Veldkamp & van der Linden, 2002) were developed. Also, most of the research done in multidimensional adaptive testing used two-dimensional cases, but we believe for the purpose of multidimensional tests, at least three dimensions are needed to give a rigorous evaluation of the procedures. Therefore, in order to have a better understanding of MCAT, this study conducted a comparative study of ability estimation and item selection methods in MCAT under different conditions.

The first attempt to extend unidimensional adaptive testing methods to multidimensional cases was Bloxom and Vale (1987). As mentioned above, Tam (1992) worked on comparing adaptive estimation for multidimensional tests and he also developed an iterative maximum likelihood ability estimation procedure himself. But all studies in those times were limited by computer power, which is not a problem for the computers now. Several current studies have investigated ability estimation methods and item selection methods. Segall (1996, 2000) applied maximum likelihood estimation and item selection methods and Bayesian estimation and item selection methods. Luecht (1996) examined the benefits of applying multidimensional adaptive testing methods in a licensing/certification context. Another item selection method, Kullback-Leibler Information, was first introduced to adaptive testing by

Chang & Ying (1996). Veldkamp & van der Linden (2002) further developed it for the multidimensional case.

In this study, ability estimation methods: maximum likelihood (Segall 1996, 2000, Reckase 2009) and Bayesian methods (Segall 1996, 2000) were investigated. Item selection methods: maximizing Fisher's information (Segall 1996, 2000, Mulder & van der Linden, 2008), including D-optimality, A-optimality, and maximizing Kullback-Leibler information (Veldkamp & van der Linden, 2002) were compared. The objective of the study is to compare the above methods for various conditions, such as test lengths and priors used.

## **CHAPTER 2. ABILITY ESTIMATION AND ITEM SELECTION METHODS IN MCAT**

### **2.1 Ability Estimation Methods**

In this study, we assume the number of dimensions and the multidimensional coordinate space has been determined. The item bank exists and all item parameters  $\mathbf{a}$ ,  $c$ ,  $d$  have been calculated. So the focus here is to administer the test and estimate examinees' ability parameters  $\theta$ .

In any CAT procedure, an initial estimate of a person's location in the coordinate system  $\theta_0$  is specified, and then an item is selected and administered to the examinee. Based on the examinee's answer to the item, an updated location estimate is calculated. Then another item is selected based on the updated location estimate, this procedure is repeated until the end of the test. The final location estimate for this examinee is given. In this section, two methods of how to estimate persons' locations in the coordinate system are shown. The two methods are maximum likelihood and Bayesian methods. The algorithm of each method is briefly introduced.

#### **2.1.1 Maximum Likelihood Method**

Maximum likelihood method was first applied in MCAT by Segall (1996, 2000). It begins with the likelihood function. Assumed  $n$  items have been administered, from

the local independence assumption, the likelihood of an examinee with ability  $\theta$  observes a vector of responses  $\mathbf{u}$  is:

$$L(\mathbf{u} | \theta) = L(u_{v_1}, u_{v_2}, \dots, u_{v_n} | \theta) = \prod_{i \in \mathbf{v}} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}, \quad (2.1)$$

where  $P_i(\theta)$  is defined by (1.3),  $Q_i(\theta) = 1 - P_i(\theta)$ , and  $\mathbf{v}$  is a vector containing the identifiers of the administered items.

The maximum likelihood estimates are the solution to the set of  $m$  simultaneous equations given by:

$$\frac{\partial}{\partial \theta} \ln L(\mathbf{u} | \theta) = \mathbf{0}, \quad (2.2)$$

This set of equations does not have a closed form solution, so Segall (1996, 2000) suggested using an iterative numerical procedure, e.g. Newton-Raphson procedure, to obtain the estimates. A more detailed description of the method can be found in Segall (1996, 2000).

### 2.1.2 Bayesian Estimation Method

This Bayesian estimation method is introduced by Segall (1996). From Bayes Theorem, the posterior density function of  $\theta$  is:

$$f(\theta | \mathbf{u}) = L(\mathbf{u} | \theta) \frac{f(\theta)}{f(\mathbf{u})}, \quad (2.3)$$

where  $L(\mathbf{u} | \theta)$  is defined as in (2.1),  $f(\theta)$  is the prior distribution of  $\theta$ , and  $f(\mathbf{u})$  is the marginal probability of  $\mathbf{u}$ .

In most of the studies, we assume the prior distribution of  $\theta$  is multivariate normal with mean  $\mu$  and variance covariance matrix  $\Phi$ :



$$f(\boldsymbol{\theta}) = (2\pi)^{-m/2} |\boldsymbol{\Phi}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]. \quad (2.4)$$

There are two ways of obtaining the point estimates of ability: the mode of the posterior distribution (MAP) or the mean of the posterior distribution (EAP). MAP is used more often simply because it requires far less computation. But with the increase of the computer power, EAP is also applicable.

MAP can be obtained from the solution to the system of equations:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta} | \mathbf{u}) = \mathbf{0}. \quad (2.5)$$

The same as in the case of solving the equation (2.2), no explicit solution can be found. So an iterative numerical procedure such as Newton-Raphson procedure must be applied to find the solution.

EAP is calculation by:

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} | \mathbf{u}), \quad (2.6)$$

where the expectation is taken according to the posterior distribution of  $\boldsymbol{\theta}$ . More detailed description of this method can be found in Segall (1996, 2000).

### 2.1.3 Other Ability Estimation Methods

Bloxom and Vale (1987) developed Owen's (1975) unidimensional sequential updating procedure into a multivariate extension through a series of normal approximations. Tam (1992) developed an iterative maximum likelihood ability estimation method for the two dimensional normal ogive model. Some combinations of maximum likelihood and Bayesian methods are proposed in Reckase (2009). One

example would be to use Bayesian ability estimation method at the beginning of the test and when the ability location estimates become finite, maximum likelihood method in 2.1.1 can be used.

## **2.2 Item Selection Methods**

After each time the ability location estimates are updated, the next item needs to be selected for the examinee. There are several methods for choosing the next item. All of them are based on either maximizing or minimizing some criteria at the most recently updated location estimates. The difference among all item selection methods are the kind of criterion chosen. This section will describe several item selection methods that can be found in the research literature.

### **2.2.1 Maximizing the Determinant of the Fisher Information Matrix (D-optimality)**

This method was proposed in MCAT setting in Segall (1996). For unidimensional cases, the largest reduction in the sampling variance of  $\hat{\theta}$  is achieved by selecting the item with the largest information value. However, in MCAT, information is no longer a scalar but a  $m \times m$  matrix. It is defined that information based on previous administered items and updated ability estimate  $\hat{\theta}$ ,  $\{r\text{-th}, s\text{-th}\}$  elements of the information matrix is:

$$\mathbf{I}_{rs}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = -E\left[\frac{\partial^2 \ln L}{\partial \theta_r \partial \theta_s}\right], \quad (2.7)$$

and the  $\{r\text{-th}, s\text{-th}\}$  elements of an item information matrix is defined:

$$\mathbf{I}_{rs}(\boldsymbol{\theta}, u_i) = \frac{\frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_r} \times \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_s}}{P_i(\boldsymbol{\theta})Q_i(\boldsymbol{\theta})}, \quad (2.8)$$

This method selects the next item which can achieve the largest decrement in the volume of the confidence ellipsoid. In order to realize that, a criterion of maximize:

$$\arg \max \det(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, u_k)), \quad (2.9)$$

is set where  $\hat{\boldsymbol{\theta}}_{k-1}$  is the ability estimate update after  $k-1$  items have been administered and  $k$ th item needs to be selected. More details about the method of maximizing the determinant of Fisher information matrix is shown in Segall (1996, 2000).

### **2.2.2 Minimizing the Trace of Inverse of Fisher Information Matrix (A-optimality)**

Mulder & van der Linden (2008) introduce the method of minimizing the Fisher information matrix as the standard for selecting the next item. Mulder & van der Linden (2008) observed that in the optimal design literature, usage of determinant or trace of an information matrix or a covariance matrix is the standard practice. While using the determinant can select items that lead to the smallest generalized variance of the ability estimators, using the trace may select a different set of items because it only focuses on the variances of the ability estimators. But the results in Mulder & van der Linden (2008) showed that the precision of using trace of the inverse of the Fisher information matrix was comparable to using the determinant of the Fisher

information matrix in most cases. So this method is also included in this study. The criterion for selection is:

$$\arg \min \text{trace}(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k))^{-1} \quad (2.10)$$

In Mulder & van der Linden (2008), a more detailed description of this criterion can be found. A 3-dimensional case example was given in Mulder & van der Linden (2008). Let eigenvalues of  $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k)$  be  $x_1, x_2, x_3$  ( $x_1, x_2, x_3 \neq 0$ ). It has:

$$\text{trace}((\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k))^{-1}) = \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}$$

$$\det(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k)) = x_1 x_2 x_3$$

So the criterion of A-optimality:

$$\begin{aligned} \arg \min \text{trace}(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k))^{-1} &= \arg \min \frac{x_1 x_2 + x_1 x_3 + x_2 x_3}{x_1 x_2 x_3} \\ &= \arg \max \frac{x_1 x_2 x_3}{x_1 x_2 + x_1 x_3 + x_2 x_3} = \arg \max \frac{\det(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k))}{\sum_{l=1}^3 \det(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, \mathbf{u}_k))_{[l,l]}} \end{aligned}$$

A-optimality contains the criterion of D-optimality as an import part. So the behavior of D-optimality and A-optimality should be similar.

### 2.2.3 Largest Decrement in the Volume of Bayesian Credibility Ellipsoid

Based on the criterion in section 2.2.1, Segall (1996) developed another criterion for item selection. This Bayes Theorem based criterion selects the next item that leads to the largest decrement in the volume of the Bayesian credibility ellipsoid. When Bayesian methods are used, prior information about the population ability distribution is available. Then the criterion given in section 2.2.1 (Segall 1996) changes to:

$$\arg \max \det(\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{k-1}) + \mathbf{I}(\boldsymbol{\theta}, u_k) + \boldsymbol{\Phi}^{-1}), \quad (2.11)$$

and  $\boldsymbol{\Phi}$  is the same as defined in (2.4), which is the variance-covariance matrix of the prior multivariate normal ability distribution. More details about the method of maximizing the decrement in the volume of the Bayesian credibility ellipsoid is shown in Segall (1996, 2000).

#### 2.2.4 Maximizing the Kullback-Leibler Information

The information most used in CAT research is Fisher's information. All the above methods are based on Fisher's information. Kullback-Leibler information was first introduced for the unidimensional CAT by Chang and Ying (1996). Veldkamp and van der Linden (2002) further generalized it to the multidimensional cases. Kullback-Leibler information is suggested to perform better than Fisher information, especially during the beginning stage of the test (Chang & Ying 1996). It measures the distance between two likelihoods over the same parameter space (Lehmann & Casella, 1998). As suggested by Veldkamp and van der Linden (2002), it is desirable to select the next item that yields a likelihood at the true ability value maximally from those at any other  $\boldsymbol{\theta}$ .

For one single item  $i$ , Kullback-Leibler information is defined as:

$$K_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = E\left[\ln \frac{L(u_i | \boldsymbol{\theta}_0)}{L(u_i | \boldsymbol{\theta})}\right], \quad (2.12)$$

where  $\boldsymbol{\theta}_0$  is the true ability value of the examinee.

After administering  $n$  items, for a set of response vector  $\mathbf{u}$ , the measure is defined:

$$K_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = E\left[\ln \frac{L(\mathbf{u} | \boldsymbol{\theta}_0)}{L(\mathbf{u} | \boldsymbol{\theta})}\right]. \quad (2.13)$$

Because  $\boldsymbol{\theta}_0$  is unknown and  $\boldsymbol{\theta}$  is unspecified, Veldkamp and van der Linden (2002) based their item selection on the posterior expected Kullback-Leibler information that the most recent updated ability estimate  $\hat{\boldsymbol{\theta}}^{k-1}$  after  $k-1$  items. So the criterion of selecting the next item is to maximize:

$$K_i^B(\hat{\boldsymbol{\theta}}^{k-1}) = \int_{\boldsymbol{\theta}} K_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{k-1}) f(\boldsymbol{\theta} | \mathbf{u}) d\boldsymbol{\theta}. \quad (2.14)$$

Chang and Ying (1996) and Veldkamp and van der Linden (2002) have more details on the item selection method of maximizing Kullback-Leibler information.

### 2.2.5 Other Item Selection Methods

The above item selection methods are the most often used ones in the recent MCAT studies. However, there are other item selection methods in MCAT. For example, Mulder & van der Linden (2009) introduced three criteria for item selection based on Kullback-Leibler Information: 1) posterior expected Kullback-Leibler Information; 2) Kullback-Leibler distance between subsequent posteriors; 3) mutual information. Details about those criteria can be found in Mulder & van der Linden (2009). Reckase (2009) also proposed to select the next item that maximized the information in the direction that had the least information. More details about this item selection method can be found in Reckase (2009).

## CHAPTER 3. RESEARCH QUESTIONS AND METHODS

### 3.1 Research Questions

The goal of this research is to compare several ability estimation and item selection methods. To compare the performance of the methods, the criterion is which of the estimates yields values that are closer to the true value of the location for an examinee after a fixed number of items have been administered. Mean bias and Root Mean Squared Error (RMSE) were used as the standards to measure the precision of the estimates. Even though computer power is not a problem nowadays, it would still be interesting to compare the computation time of each method to find the balance between computation time and the precision of the estimates.

For the research questions on ability estimation methods, first of all, as mentioned above, one problem with maximum likelihood estimation (Segall 1996) is that it may not converge at the beginning of the test. No research has shown how many items need to be administered before the estimates converge and when the estimates are near the true ability value. So the first research question is what test length is needed to have a converging results for maximum likelihood ability estimation method. The trend for the mean biases and RMSEs will help to decide whether the estimation is converging or not. Plots of successive estimates of the location as the test processes will be drawn to determine whether the test is converging at certain number of items.

Both maximum likelihood (Segall 1996) and Bayesian methods (Segall 1996) are used in MCAT research literature. So the second research question is which one of them performs better and under what conditions. The performance of those two methods was compared with different test lengths. The hypothesis is that when the test length is short, Bayesian methods would outform the maximum likelihood method. However, when the test length is long, those two methods should not differ much.

Different priors for Bayesian method may be used, whether they are informative or not so informative. The research question is whether priors have any impact on the estimation results. Three priors were selected by this study: strong prior, relatively relaxed prior, and true priors. We define true prior here as the prior with variance covariance matrix from the whole examinee population. The hypothesis is that when true priors are used, more accurate estimates are expected. Stronger priors' estimates are better than relatively relaxed priors for the students whose ability distribution nearer to the prior but worse for the students whose ability distribution further away from the priors'. Argument for the usage of relatively relaxed priors is to be objective in administration of the tests.



Table 3.1 All conditions in comparing different ability estimation methods

Ability Estimation Methods	Prior	Test Length
MLE	N/A	short long
Bayesian	Strong prior	short long
	Relatively relaxed prior	short long
	True prior	short long

For research questions on item selection methods, the first one is to compare the performance of D-optimality (maximizing the determinant (Segall 1996)) and A-optimality (minimizing the trace (Mulder & van der Linden 2008)) when maximum likelihood method is used. From the literature of optimal design, those two methods should be comparable. In this study, the two item selection methods were compared at the long test length and research hypothesis is their performance is comparable.

The second comparison of item selection methods is to compare the performance of Bayesian method based on Fisher's information (maximizing decrement in Bayesian) with the one based on Kullback-Leibler information. The comparison is conditioning on test length and we would also use plots of successive estimates to see how fast each method converges.

All methods will also be compared with themselves conditioning on test length. The results can provide evidence for each combination of ability update and item selection method, when it converges to the true ability point in the dimension and has stable and accurate estimates.

Table 3.2 All conditions in comparing item selection methods

<b>Item Selection Methods</b>	<b>Ability Estimation method</b>	<b>Test Length</b>
D-optimality	MLE	long
A-optimality	MLE	long
Bayesian Volume Decrease	Bayesian	short long
Kullback-Leibler	Bayesian	short long

Note: Bayesian Volume Decrease refers to the item selection method of maximizing the decrement in volume in Bayesian credibility ellipsoid described in section 2.2.3.

## **3.2 Research Methods**

### **3.2.1 Real Data Used**

In this study, item pool was simulated based on real data from Michigan Educational Assessment Program (MEAP). Li (2006) used the data from 2005 MEAP mathematics test for the 7<sup>th</sup> graders. This real data set included 8562 examinees and 50 multiple choice items. From the dimensionality analysis results of Li (2006), this data set measured three ability dimensions: the first dimension measured ability to abstract math concepts; the second dimension measured vocabulary and operations ability; the third dimension measured problem solving ability. The estimated item parameters from Li (2006) for all items are listed in table 3.

Table 3.3 MIRT item parameters for Grade 7 Michigan Educational Assessment Program (MEAP) Mathematics test from Li (2006).

Dimensi on	Items*	c*	a1*	a2*	a3*	mdiff*
1	01 (N-FL)	0.13	0.48	0.09	-0.09	-2.15
1	03 (N-FL)	0.11	0.28	0.17	-0.07	-1.14
1	07 (A-FO)	0.04	2.54	-0.58	-0.23	-1.14
1	08 (A-FO)	0.05	3.03	-0.59	-0.53	-1.11
1	09 (N-FL)	0.07	0.59	0.06	0.04	-1.76
1	11 (D-PR)	0.08	0.57	-0.12	0.23	-1.78
1	12 (A-RP)	0.11	0.48	-0.10	0.15	-1.22
1	14 (A-FO)	0.05	0.62	0.16	0.18	-1.53
1	16 (A-RP)	0.11	0.51	0.18	0.05	-1.29
1	20 (G-GS)	0.08	0.38	0.10	-0.13	0.58
1	22 (A-RP)	0.13	0.54	0.12	0.05	-1.55
1	26 (A-FO)	0.14	0.22	0.14	-0.06	-1.23
1	38 (G-GS)	0.11	0.39	0.00	-0.01	-0.39
1	40 (A-FO)	0.28	0.37	0.07	-0.10	3.67
2	06 (N-FL)	0.21	-0.16	1.22	0.14	0.42
2	17 (A-FO)	0.20	0.56	0.48	0.06	-1.01
2	21 (A-FO)	0.30	0.07	0.77	0.14	0.04
2	24 (N-FL)	0.21	0.34	0.76	-0.41	0.01
2	25 (A-PA)	0.18	0.32	0.40	0.48	-0.58
2	33 (N-ME)	0.11	-0.13	0.98	0.56	-0.07
2	35 (N-FL)	0.22	-0.03	0.57	0.60	0.24
2	36 (G-TR)	0.22	-0.01	1.08	0.20	0.64
2	37 (A-FO)	0.28	0.08	0.74	0.26	0.38
2	39 (N-FL)	0.15	0.26	0.78	-0.42	0.13
2	43 (N-ME)	0.12	-0.37	1.18	0.46	0.52
2	44 (G-TR)	0.25	0.04	0.48	0.17	-0.02
2	47 (G-GS)	0.20	-0.17	1.05	0.06	0.74
2	50 (A-FO)	0.24	0.04	0.37	0.17	1.70
2	52 (N-MR)	0.25	-0.45	1.43	0.03	0.72
2	53 (N-FL)	0.26	-0.29	0.76	0.30	1.09
2	55 (N-FL)	0.21	-0.02	0.82	-0.07	0.81
2	59 (N-FL)	0.22	0.06	0.54	0.41	-0.26
2	60 (G-TR)	0.14	-0.07	0.48	0.14	1.53
3	05 (N-FL)	0.23	-0.50	0.56	1.33	0.47
3	10 (N-ME)	0.18	0.58	-0.01	0.45	-1.43
3	13 (N-ME)	0.14	0.52	-0.47	1.31	-0.89
3	15 (N-FL)	0.12	0.51	0.19	0.40	-1.42
3	18 (N-FL)	0.16	0.23	-0.10	0.84	-0.73

Table 3.3 (Continue)

3	19 (N-ME)	0.09	0.46	0.12	0.37	-1.24
3	23 (N-ME)	0.10	0.27	0.12	0.29	-1.02
3	27 (D-PR)	0.14	0.49	-0.04	0.79	-1.25
3	28 (A-PA)	0.19	0.04	0.20	0.48	-0.44
3	31 (M-UN)	0.23	-0.12	0.29	0.93	-0.03
3	32 (N-FL)	0.19	-0.10	0.43	0.85	0.05
3	34 (N-MR)	0.15	-0.46	0.53	3.12	-0.11
3	41 (A-PA)	0.10	-0.23	0.19	1.39	0.11
3	45 (N-MR)	0.10	0.00	-0.05	1.11	-0.22
3	46 (N-FL)	0.22	-0.20	0.15	1.06	0.36
3	49 (D-PR)	0.11	0.08	0.10	0.55	0.45
3	51 (A-FO)	0.21	0.01	0.30	0.48	0.75

Note:

\*.  $c$  is the pseudo guessing parameter specified in equation 1.3;  $a_1$ ,  $a_2$ ,  $a_3$  are discrimination parameters for each of the three ability dimensions, where  $\mathbf{a}=(a_1, a_2, a_3)$ ,  $\mathbf{a}$  as in equation 1.3;  $mdiff$  is the difficulty parameter in MIRT with negative value representing easy items. It is different from  $d$  as in equation 1.3. More details will be given in the section 3.2.2.

\*\*. The 'items' column contains a two-digit number for each item, representing the position of the item in actual test administration. Abbreviations for content classifications are listed at the number.

Also, Li's study showed that the test had simple structure, which means each item mainly loaded on one dimension. As shown in Table 3, the first 14 rows are the 14 items that mainly measure dimension 1, abstracting math concepts. From row 15 to row 33 are items mainly measuring dimension 2, vocabulary and operations ability. The last 17 rows represent 17 items that measure mainly dimension 3, problem solving ability. More details about the dimensional structures of test items can be found in Reckase (2009).

In Li (2006), all correlations among the three  $\theta$ -scales were about 0.5. To be more specific, for all 8562 examinees, the variance-covariance matrix among dimensions is as in Table 4. This was used as the true prior in the simulation for Bayesian ability estimation method.

Table 3.4 Correlation coefficients among 3 dimensions on Grade 7 Michigan Education Assessment Program (MEAP) Mathematics Test

	Dimension 1	Dimension 2	Dimension 3
Dimension 1	1	0.5104	0.5117
Dimension 2	0.5104	1	0.5675
Dimension 3	0.5117	0.5675	1

### 3.2.2 Simulation

The study was simulated based on compensatory MIRT model as in equation 3.1 with all  $c$  parameters set as 0. In Li (2006), instead of generating  $\mathbf{a}$  and  $d$ , other derived MIRT statistics  $mdisc_i$ ,  $mdiff_i$ , and  $dcos_{jk}$  were generated first. Parameters  $\mathbf{a}$  and  $d$  were derived as the functions of those statistics. The relationship between them is represented by equation 3.2 and 3.2.

$$P(U_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i, c_i, d_i) = \frac{e^{1.7(\mathbf{a}_i \boldsymbol{\theta}'_j + d_i)}}{1 + e^{1.7(\mathbf{a}_i \boldsymbol{\theta}'_j + d_i)}}. \quad (3.1)$$

$$d_i = -mdiff_i \times mdisc_i \quad (3.2)$$

where  $mdiff_i$  is the difficulty parameter, and  $mdisc_i$  is the discriminating power of

the item for the most discriminating combinations of dimensions,  $mdisc_i = \sqrt{\sum_{k=1}^m a_{ik}^2}$ .

$$a_{ik} = dcos_{jk} \times mdisc_i, \quad (3.3)$$

where  $dcos_{jk}$  is the directional cosine that reflects how well an item measures each dimension.

Based on Li (2006)'s item parameters, 300 items were generated with 100 items mainly measuring each dimension. The item parameters were generated from distributions derived from Li (2006) for mimicking Grade 7 Michigan Education Assessment Program (MEAP) Mathematics test.

Table 3.5 Distributions for multidimensional item parameter generation mimicking Grade 7 Michigan Education Assessment Program (MEAP) Mathematics test. 100 items were generated for each dimension.

**A. mdiff: difficulty parameter (negative value represents easy item)**

	<b>Distribution</b>
Dimension 1	normal (mean=-0.8, var=0.6)
Dimension 2	normal (mean=0.37, var=0.4)
Dimension 3	normal (mean=-0.39, var=0.4)

**B. mdics: discriminating power of the items at direction of best measurement**

<b>Distribution</b>
lognormal (mean=1, var=0.03)

**C. dcos: directional cosine determining the direction an item are measuring**

		<b>Distribution</b>
Dimension 1	dcos1	$\sqrt{1 - (d \cos 2^2 + d \cos 3^2)}$
	dcos2	beta (mean=0.0246, var=0.002)
	dcos3	beta (mean=0.1694, var=0.003)
Dimension 2	dcos1	beta (mean=0.1366, var=0.005)
	dcos2	$\sqrt{1 - (d \cos 1^2 + d \cos 3^2)}$
	dcos3	beta (mean=0.0846, var=0.004)
Dimension 3	dcos1	beta (mean=0.1161, var=0.006)
	dcos2	beta (mean=0.0507, var=0.002)
	dcos3	$\sqrt{1 - (d \cos 1^2 + d \cos 2^2)}$

Because the data set was three dimensional, 50 replications were simulated for each combination of  $\theta_1 = -1, 0, 1$ ,  $\theta_2 = -1, 0, 1$  and  $\theta_3 = -1, 0, 1$ . If Bayesian methods were used, all interim ability estimates were MAP estimates and the final ability estimates were EAP estimates. Mean bias and root mean squared errors (RMSE) were used as measures of estimation precision. Mean biases and root mean squared errors (RMSE) were calculated for each dimension. Euclidean distance was also calculated as another index of the precision of the estimates. Euclidean distance in three-dimensional space between the estimate and true location point was calculated in as in equation 3.4.

$$D = \sqrt{(\hat{\theta}_1 - \theta_1)^2 + (\hat{\theta}_2 - \theta_2)^2 + (\hat{\theta}_3 - \theta_3)^2}, \quad (3.4)$$

where  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$  is the current updated ability location point in the space, and  $\theta = (\theta_1, \theta_2, \theta_3)$  is true ability location point.

For general maximum likelihood method, a limit of  $\pm 3$  was set in order to provide the estimate updates when they were not converging. So whenever the estimates were larger than 3, the value 3 was set to the estimates. If the estimates were smaller than -3, the value -3 was set for the estimates.

Total of 13 conditions were simulated to make a comprehensive comparison of each ability update and item selection method. Table 4 shows all the conditions.

Table 3.6 All 11 simulated conditions. All simulations are for 27 true ability points, 50 replicates at each point.

<b>Ability Estimation Methods</b>	<b>Item Selection Methods</b>	<b>Prior</b>	<b>Test Length</b>
MLE	D-optimality	N/A	20
			50
	A-optimality	N/A	50
Bayesian	Bayesian Volume Decrease	Mean= <b>0</b> , var-cov=identity matrix	20
			50
	Bayesian Volume Decrease	Mean= <b>0</b> , var-cov=diag(9)	20
			50
	Bayesian Volume Decrease	Mean= <b>0</b> , var-cov=true ability distribution	20
			50
	Kullback-Leibler	Mean= <b>0</b> , var-cov=identity matrix	20
			50

Note: Bayesian Volume Decrease here refers to the item selection method of maximizing the decrement in volume in Bayesian credibility ellipsoid described in section 2.2.3.

The test length of 20 was chosen to represent short tests (e.g. Electronics Information Test in ASVAB). The test length of 50 was chosen to represent long test (e.g. 2007 MEAP Mathematics Test). The test lengths of 20 and 50 were generated for each condition (combination of an ability update and item selection method). For each condition, 50 replicated were simulated for all 27 true ability points.

In order to answer the research question about the convergence problems of the general maximum likelihood method, test lengths of 20 and test length of 50 for the combination of MLE and D-optimality were simulated and compared. Estimates for



each dimension were calculated and compared to the true values. Euclidean distance was calculated and successive plot was draw to see the converging speed.

To compare the performance of maximum likelihood and Bayesian as the ability estimation methods, the combination of D-optimality and maximum likelihood and the combination of Fisher's information and Bayesian were simulated at the test lengths of 20 and 50. Their final estimates were compared to the true values. And Euclidean distances were calculated and successive plots were draw to compare the convergence rate.

When ability estimation method of maximum likelihood was used, item selection methods of A-optimality and D-optimality were simulated and compared at the test length of 50. They were compared in terms of convergence rate and accuracy of the final estimates.

One of the research questions is to evaluate the impact of priors used when Bayesian methods were used. Three priors were selected for the simulation. All of them were multinormal distributions with mean 0. The first one was with identity matrix as the variance covariance matrix. This represented a strong prior. The second one was with  $\text{diag}(9)$ , that is, all the diagonal elements were 9 and all the off-diagonal elements were 0. This represented a relatively weak prior. The last one was with true ability variance covariance matrix as specified as in Table 4, which was calculated from all 7<sup>th</sup> graders of 2005 MEAP test. Test lengths of 20 and 50 were simulated for each prior. Their final estimation results were compared to measure the impact of the priors.

In order to compare the performance of item selection methods of Bayesian Volume Decrease (short term used here for maximizing decrement volume in Bayesian) and Kullback-Leibler information, test lengths of 20 and 50 were simulated and the final estimates were compared. The prior used was multinormal with mean  $\mathbf{0}$  and identity matrix as the variance covariance matrix for both methods. Euclidean distance was calculated and successive plots were draw to compare the convergence rate.

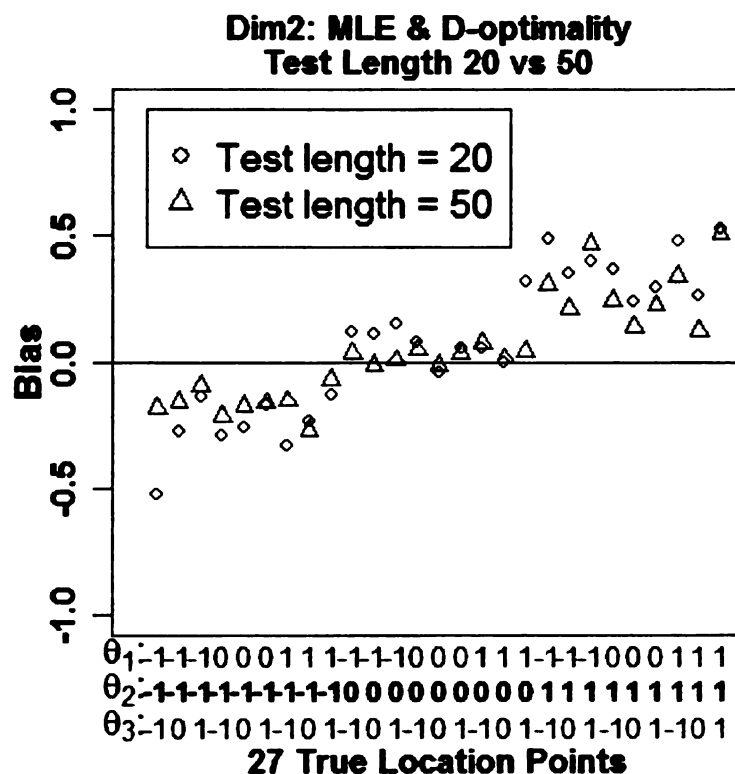
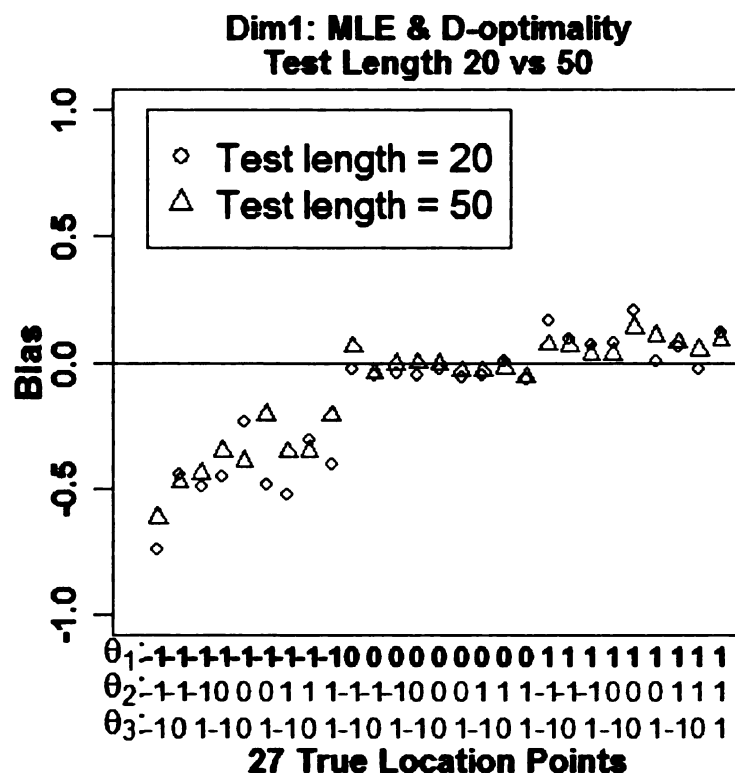
## **CHAPTER 4. RESULTS AND DISCUSSIONS**

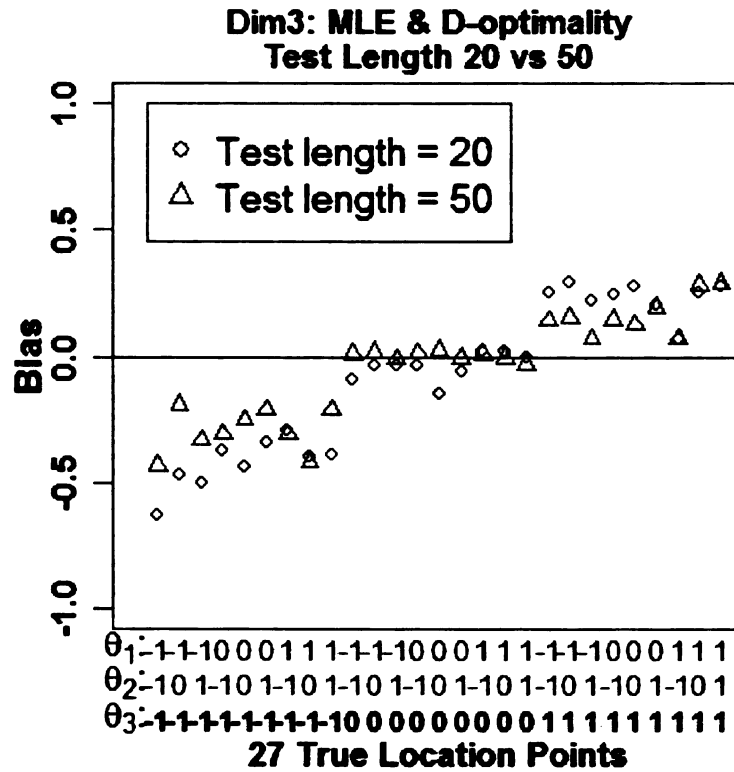
In order to measure the impact of non-convergence problems at the beginning of the test when maximum likelihood was used as the ability estimation method, the results of test lengths of 20 and 50 were compared for D-optimality.

First, the comparison of biases and RMSEs are shown in Figure 4.1. When the test was short (test length=20), the estimation was not stable: size for both the biases and RMSEs was large. But when the test was longer (test length=50), the size for both the biases and RMSEs became smaller. The estimates were more stable and accurate for all three dimensions.

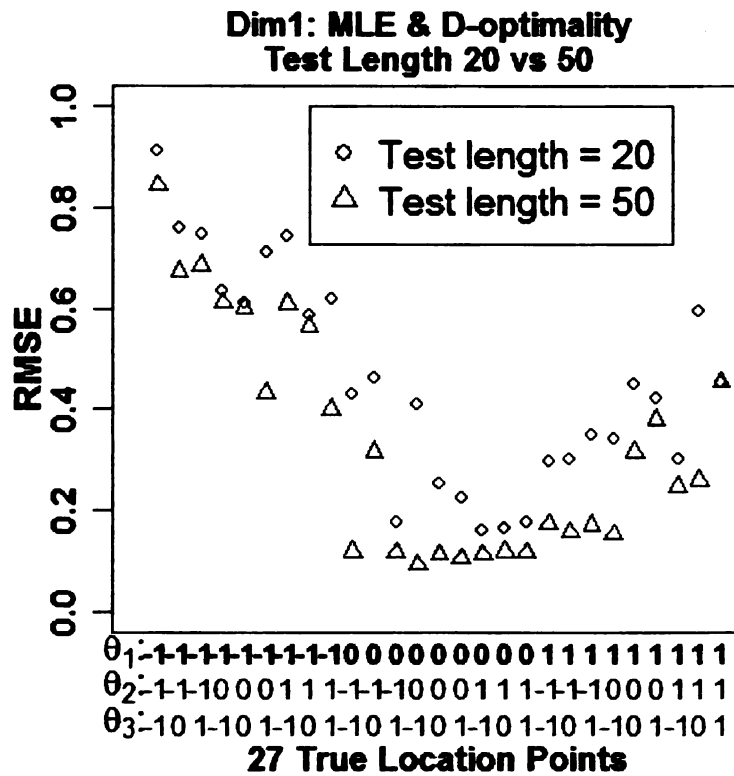
Figure 4.1 Mean biases and RMSEs for maximum likelihood as the ability estimation method and D-optimality as the item selection method, at test length =20 and at test length =50.

A: Biases





**B: RMSEs**



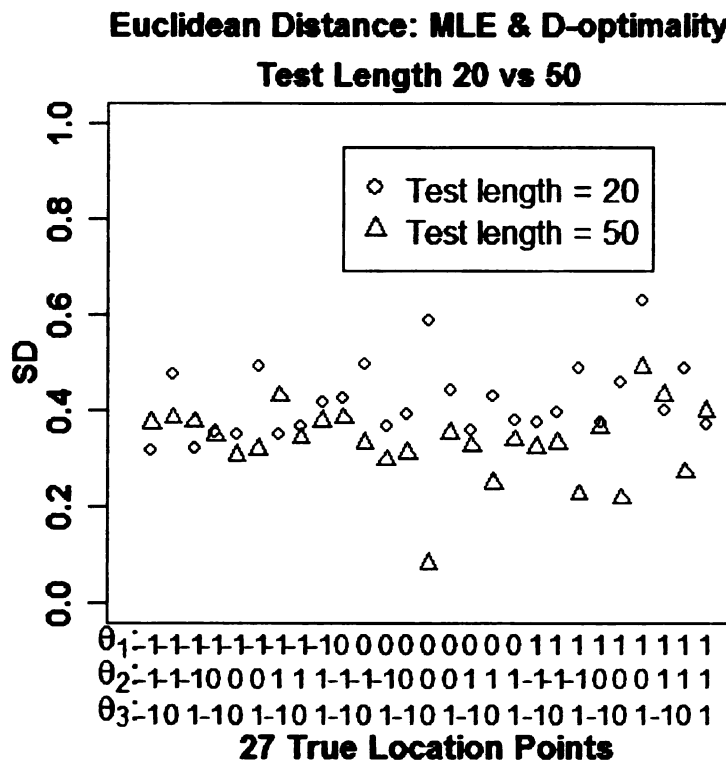
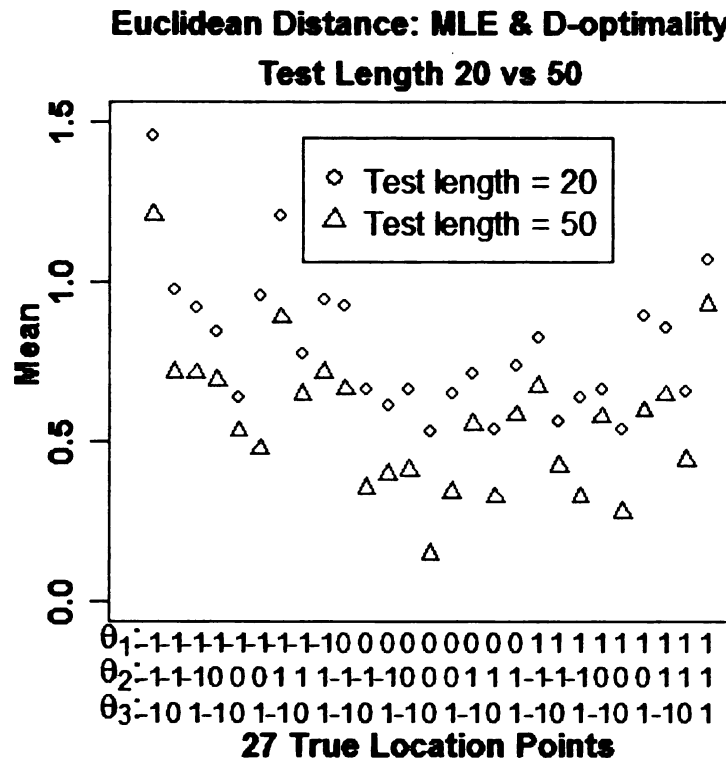
A scatter plot showing the Root Mean Square Error (RMSE) on the y-axis (ranging from 0.0 to 1.0) against 27 True Location Points on the x-axis. The x-axis labels are binary strings:  $\theta_1$ -1-1-1-1-0-0-0-1-1-1-1-1-1-1-0-0-0-1-1-1-1-1-1-1-1-1-1,  $\theta_2$ -1-1-1-1-1-1-1-1-1-1-0-0-0-0-0-0-0-0-0-1-1-1-1-1-1-1-1-1-1, and  $\theta_3$ -1. The legend indicates that open diamonds represent a Test length of 20, and open triangles represent a Test length of 50. The plot shows that for most points, the RMSE is higher for a test length of 20 compared to 50, with some points showing a significant decrease in RMSE as the test length increases.

A scatter plot titled "RMSE vs. 27 True Location Points". The y-axis is labeled "RMSE" and ranges from 0.0 to 1.0. The x-axis is labeled "27 True Location Points" and has three rows of labels:  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ . Each row contains 27 binary digits (0 or 1). A legend indicates that diamonds represent "Test length = 20" and triangles represent "Test length = 50". The plot shows that RMSE generally decreases as the number of true location points increases, particularly for test length 20.

True Location Point	$\theta_1$	$\theta_2$	$\theta_3$	Test length = 20 (RMSE)	Test length = 50 (RMSE)
1	-1	-1	-1	0.86	0.75
2	-1	-1	-1	0.63	0.47
3	-1	-1	-1	0.74	0.58
4	0	-1	-1	0.65	0.58
5	0	-1	-1	0.60	0.50
6	0	-1	-1	0.56	0.45
7	0	-1	-1	0.60	0.48
8	1	-1	-1	0.58	0.48
9	1	-1	-1	0.65	0.57
10	1	-1	-1	0.58	0.48
11	1	-1	-1	0.50	0.47
12	1	-1	-1	0.52	0.10
13	1	-1	-1	0.16	0.10
14	1	-1	-1	0.18	0.10
15	1	-1	-1	0.40	0.10
16	0	-1	-1	0.58	0.10
17	0	-1	-1	0.40	0.10
18	0	-1	-1	0.34	0.10
19	0	-1	-1	0.45	0.12
20	1	-1	-1	0.56	0.16
21	1	-1	-1	0.48	0.11
22	1	-1	-1	0.56	0.33
23	1	-1	-1	0.48	0.35
24	1	-1	-1	0.56	0.35
25	1	-1	-1	0.52	0.35
26	1	-1	-1	0.56	0.42
27	1	-1	-1	0.68	0.23
28	1	-1	-1	0.58	0.53
29	1	-1	-1	0.52	0.53
30	1	-1	-1	0.68	0.53

The same results can be found in calculating the Euclidean distance between the final estimates and true ability location points for test lengths of 20 and 50. Euclidean distance between the estimates and true ability location points were as specified as in equation 3.4. For all 27 true ability points, both the means and the standard deviations of the Euclidean distance of test length of 20 were larger than those of test length of 50. The estimates of test length of 20 were not stable, while at the test length of 50, the estimates were more stable and accurate. The U-shape of means of the Euclidean distance were observed, which showed that the estimation precision were more accurate for examinees with location 0. However, for examinees whose positions in the three dimensional space were away from the origin, the precision was not as good. Means and standard deviations of the Euclidean distances were show in Figure 4.2 for the combination of maximum likelihood and D-optimality.

Figure 4.2 Mean and standard deviation of Euclidean distance for maximum likelihood as the ability estimation method and D-optimality as the item selection method, at test length =20 and at test length =50.





From the results above, the estimates for maximum likelihood for short test (test length=20) were not very reliable and accurate. The plot of successive estimates for one examinee with true location point (1, 1, 1) was drawn in Figure 4.3. The initial estimate was (0, 0, 0) and the test length was 50. It showed that at the beginning of the test, the estimates were not converging. They hit the ceiling we set  $\pm 3$  when the estimate was not converging. After several items, estimate converged and became nearer and nearer to the true location point. The Euclidean distance at each estimate updated point for this particular examinee was also drawn and showed in Figure 4.4. The pattern was the same: at the beginning of the test, the estimation was not converging and it took several items till it converged.

Figure 4.3 Successive progress plot of updated ability estimates and true location point after administering each item for maximum likelihood method and D-optimality. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

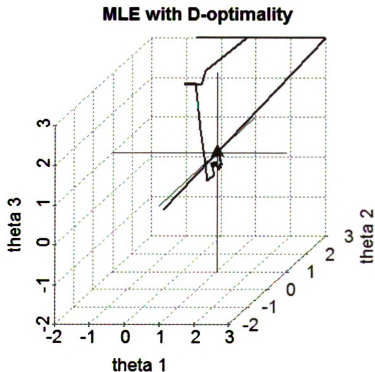


Figure 4.4 Euclidean distance of between updated ability estimates and true location point after administering each item for maximum likelihood method and D-optimatlity. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

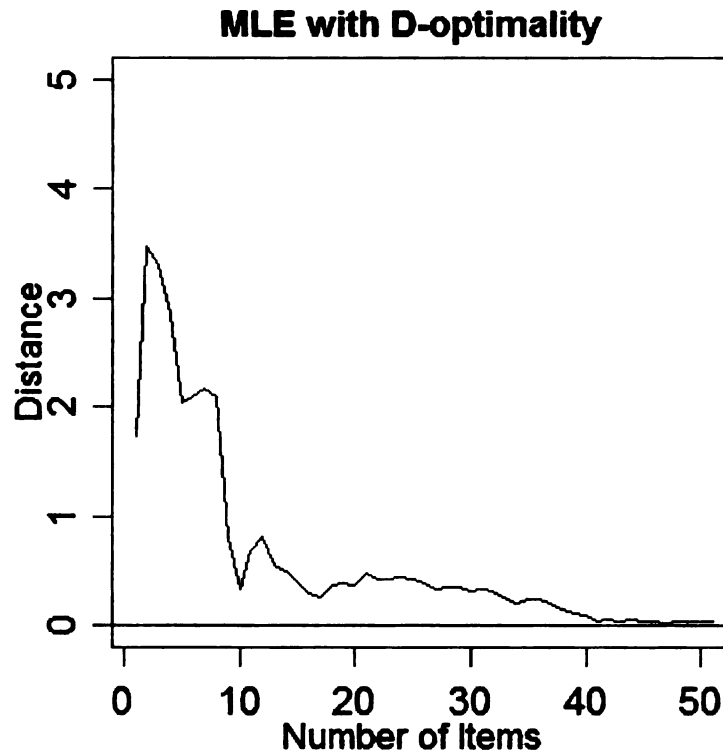


Table 4.1 showed all the items this particular examinee took and his updated ability estimates after administering each item. It can be seen that after 7 items, location estimates on all three dimensions converged and instead of assigning ceiling value 3 or flooring value -3, maximum likelihood estimates were set as the estimates. After about 38 items being administered, this examinee's location estimates were near the true location point (1, 1, 1). And it stayed around the true points afterwards.

Table 4.1 Items administered, responses and updated ability estimates after each item for one examinee. The combination was maximum likelihood and D-optimality. Initial estimate was (0, 0, 0) and the true location was (1, 1, 1).

Items Administered	Item ID	Response	$\theta_1$	$\theta_2$	$\theta_3$
1	30	1	3.00	3.00	3.00
2	128	0	-0.72	3.00	3.00
3	178	1	0.74	3.00	3.00
4	152	0	0.60	1.13	3.00
5	28	0	0.89	0.37	3.00
6	143	0	0.39	0.42	3.00
7	114	1	0.67	0.47	3.00
8	162	0	0.65	0.74	1.69
9	296	0	0.73	0.82	0.97
10	220	0	0.81	0.83	0.37
11	68	1	0.74	1.32	0.30
12	201	1	0.70	1.26	0.61
13	141	1	0.83	1.24	0.60
14	64	0	0.86	1.03	0.62
15	236	1	0.85	1.00	0.75
16	183	1	0.97	1.00	0.75
17	242	0	0.98	1.02	0.64
18	115	1	1.09	1.02	0.63
19	132	0	1.02	1.03	0.64
20	214	0	1.02	1.05	0.53
21	235	1	1.02	1.03	0.58
22	56	1	1.02	1.07	0.58
23	123	1	1.09	1.08	0.58
24	144	0	1.05	1.07	0.58
25	213	1	1.05	1.06	0.62
26	251	1	1.04	1.04	0.67
27	197	1	1.11	1.05	0.66
28	137	0	1.08	1.04	0.67
29	256	1	1.08	1.03	0.69
30	164	1	1.11	1.02	0.69
31	217	1	1.11	1.03	0.71
32	276	1	1.10	1.02	0.76
33	250	1	1.10	1.00	0.82
34	218	0	1.10	1.01	0.78
35	124	0	1.09	1.01	0.78

Table 4.1 (Continue)

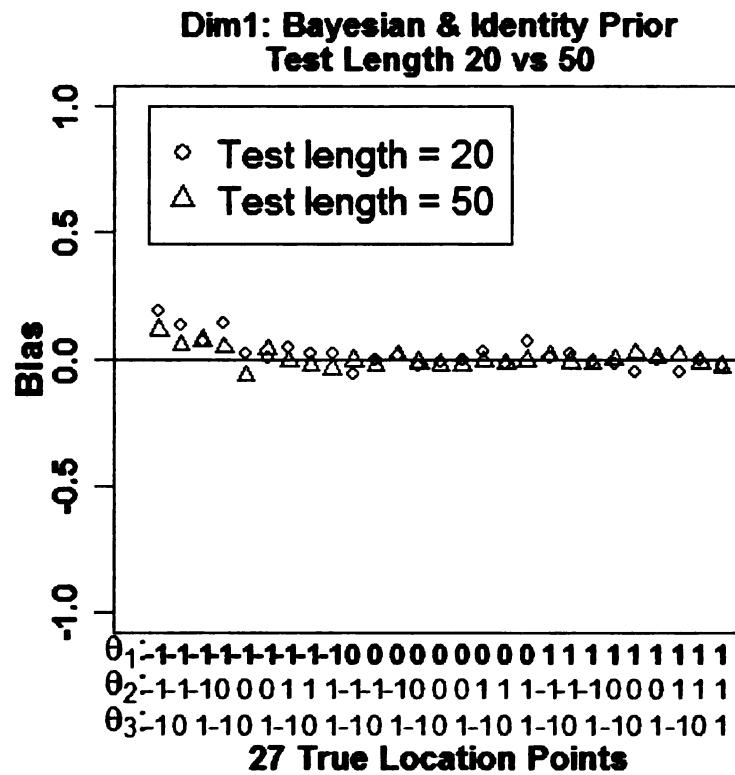
36	267	1	1.08	1.02	0.81
37	285	1	1.08	1.01	0.87
38	239	1	1.07	0.99	0.92
39	270	1	1.07	0.97	0.97
40	157	0	1.02	0.98	0.97
41	126	1	1.03	0.98	0.97
42	215	1	1.03	0.97	1.01
43	196	0	1.01	0.95	1.01
44	57	1	1.01	0.96	1.01
45	272	0	1.01	0.97	0.99
46	36	1	1.01	0.98	0.99
47	109	1	1.04	0.98	0.99
48	174	0	1.03	0.97	0.99
49	289	0	1.03	0.97	0.98
50	184	0	1.00	0.97	0.98

When the combination of Bayesian ability estimation method and Bayesian volume decrement item selection method was used, the comparison of test lengths of 20 and 50 was made for each prior to determine what test length was needed to have accurate estimates. The first comparison was made when the prior is a multinormal distribution with mean  $\mathbf{0}$ , and identity matrix as the variance-covariance matrix. This was used in the study as an example of strong prior. Mean biases and RMSEs for each dimension were compared for the test lengths of 20 and 50 and the results were shown in Figure 4.5. For all dimensions, the estimation precision at the test length of 50 was slightly better than that of the test length of 20. However, the differences were small and at the test length of 20, the estimation was already stable and near the true values. Both the mean biases and RMSEs were small. When the test length increased to 50, the biases and RMSEs became slightly smaller. But the difference was not as big as those

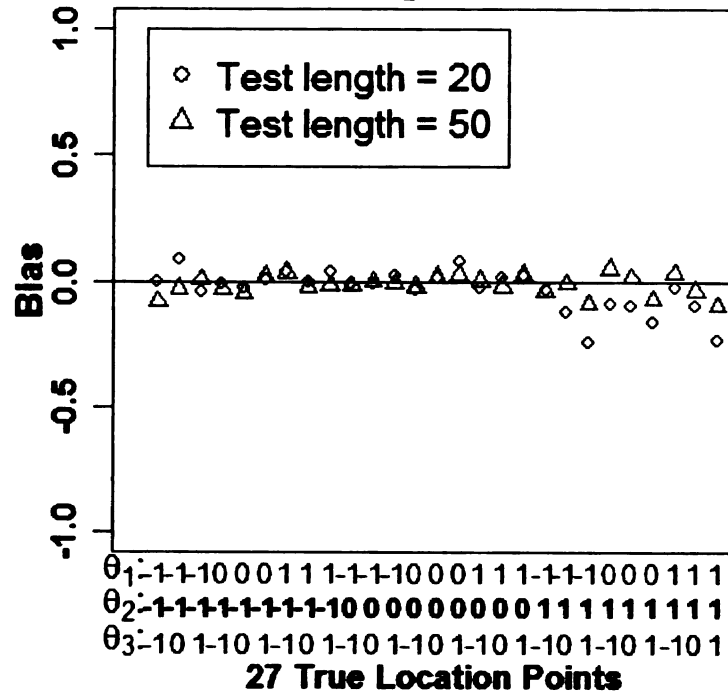
in the combination methods when maximum likelihood method was used. Mean biases and RMSEs are shown in Figure 4.5 for all three dimensions.

Figure 4.5 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and identity matrix as variance covariance matrix, at test length =20 and at test length =50.

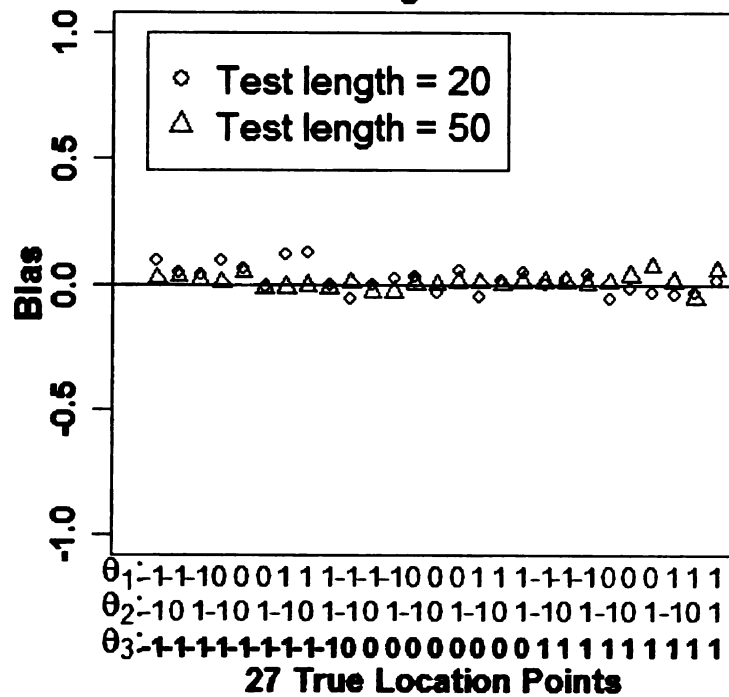
#### A: Biases



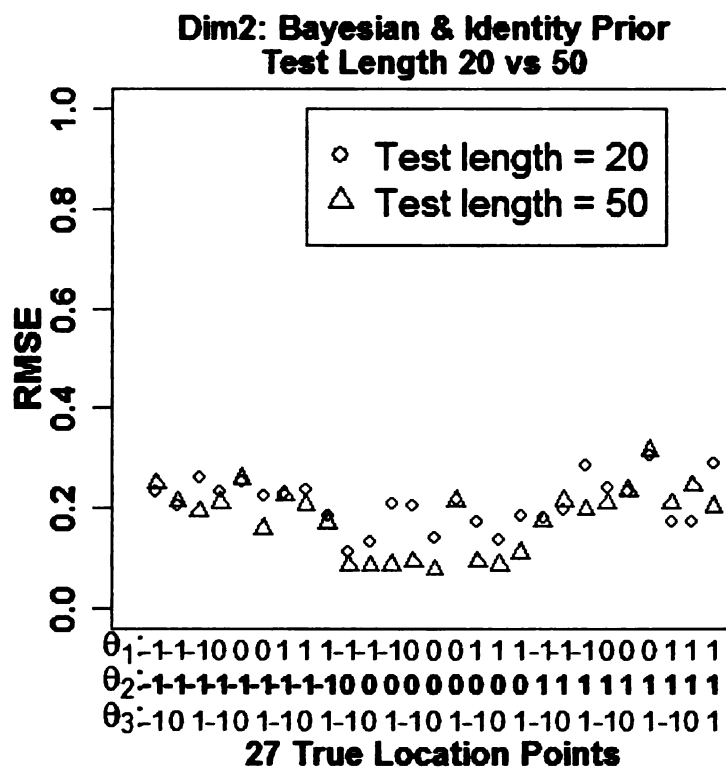
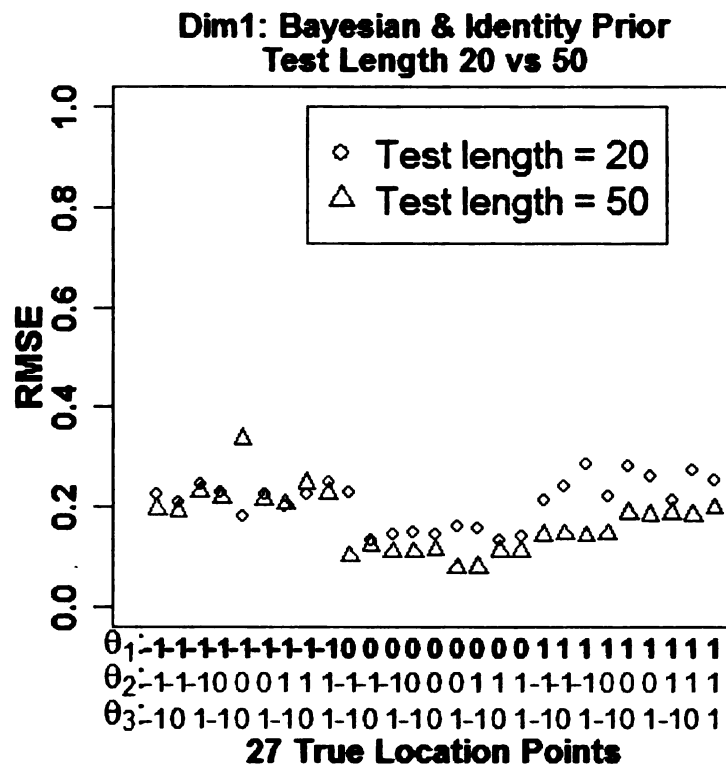
**Dim2: Bayesian & Identity Prior  
Test Length 20 vs 50**



**Dim3: Bayesian & Identity Prior  
Test Length 20 vs 50**



## B: RMSEs



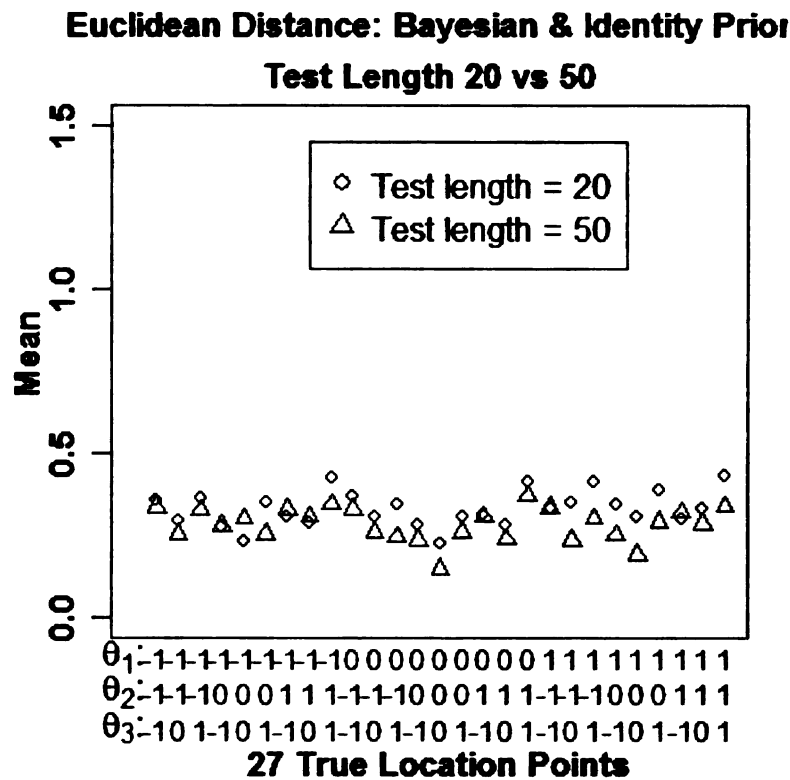
A scatter plot showing the Root Mean Square Error (RMSE) on the y-axis (ranging from 0.0 to 1.0) against 27 True Location Points on the x-axis. The x-axis labels are binary strings:  $\theta_1$ : -1 -1 -1 0 0 0 1 1 1 -1 -1 -1 0 0 0 1 1 1 -1 -1 -1 0 0 0 1 1 1 -1 -1 -1 0 0 0 1 1 1,  $\theta_2$ : -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1 -1 0 1, and  $\theta_3$ : -1 -1 -1 -1 -1 -1 -1 -1 -1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1. The legend indicates that open diamonds represent a Test length of 20 and open triangles represent a Test length of 50. The RMSE values for both test lengths are generally low, mostly between 0.1 and 0.3, with some variation across the 27 points.

39

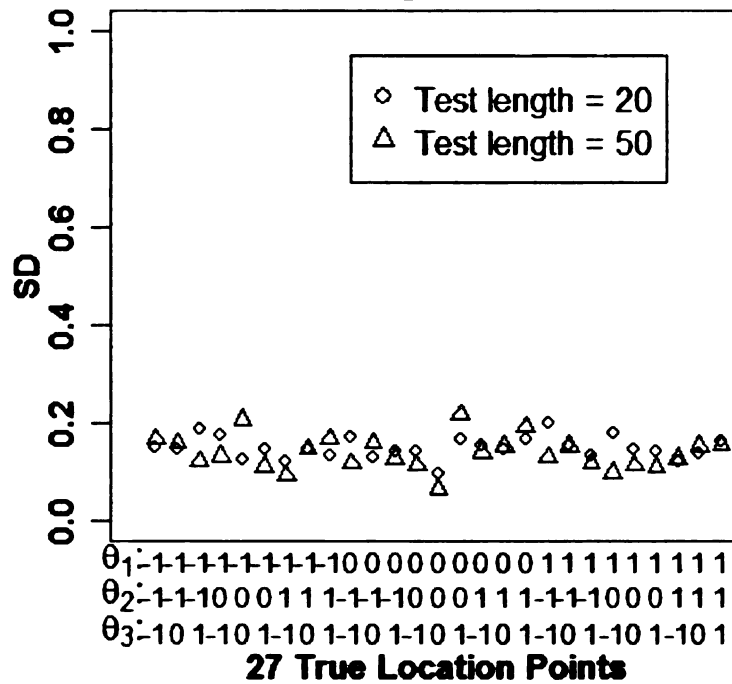


method and maximizing decrement volume in Bayesian as the item selection method was already good with the test length of 20. Means and standard deviations of the Euclidean distance were shown in Figure 4.6.

Figure 4.6 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and identity matrix as variance covariance matrix, for test length =20 and for test length =50.



**Euclidean Distance: Bayesian & Identity Prior**  
**Test Length 20 vs 50**



As in the results for maximum likelihood method, plot of successive estimates of the combination of Bayesian and Bayesian volume decrement with identity matrix as prior was shown in Figure 4.7 for one example of examinee with true location point of (1, 1, 1) and initial estimate of (0, 0, 0). This figure showed there was no non-converging issue with Bayesian method and the estimates quickly converged to the true location. This was evidence why at the test length of 20, the estimate was already accurate for this combination of method. Euclidean distance between the estimates and true location (1, 1, 1) was calculated and shown in Figure 4.8. The results corresponded to the results in Figure 4.7.

Figure 4.7 Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with identity matrix. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

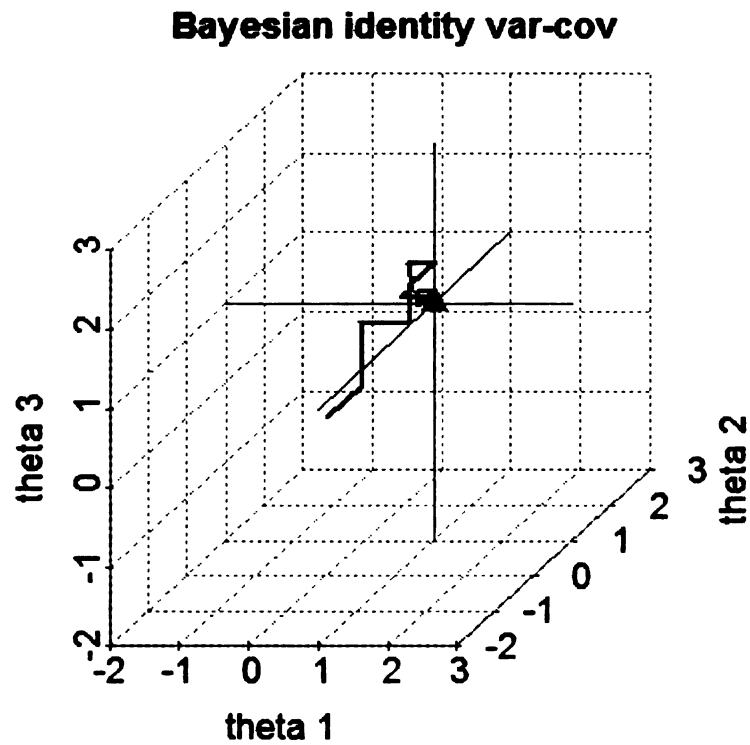
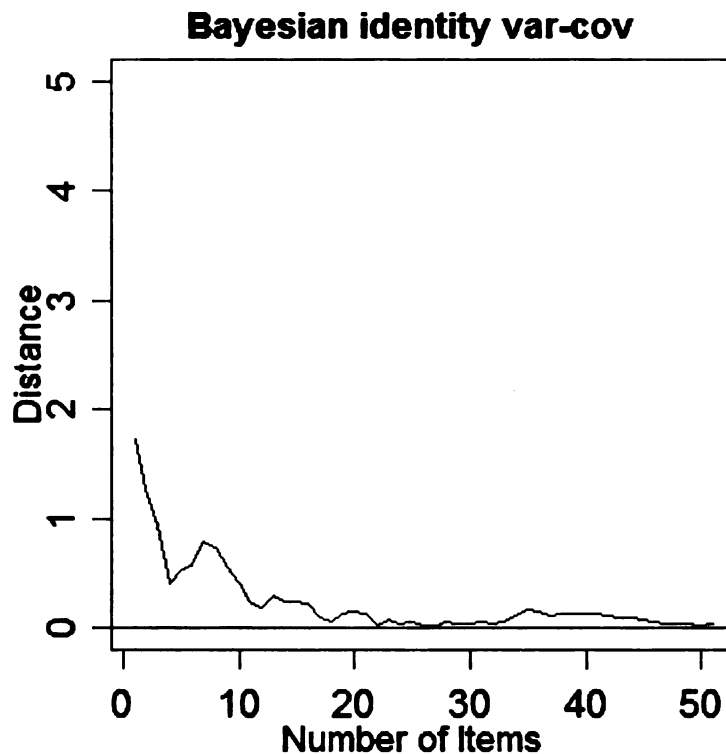


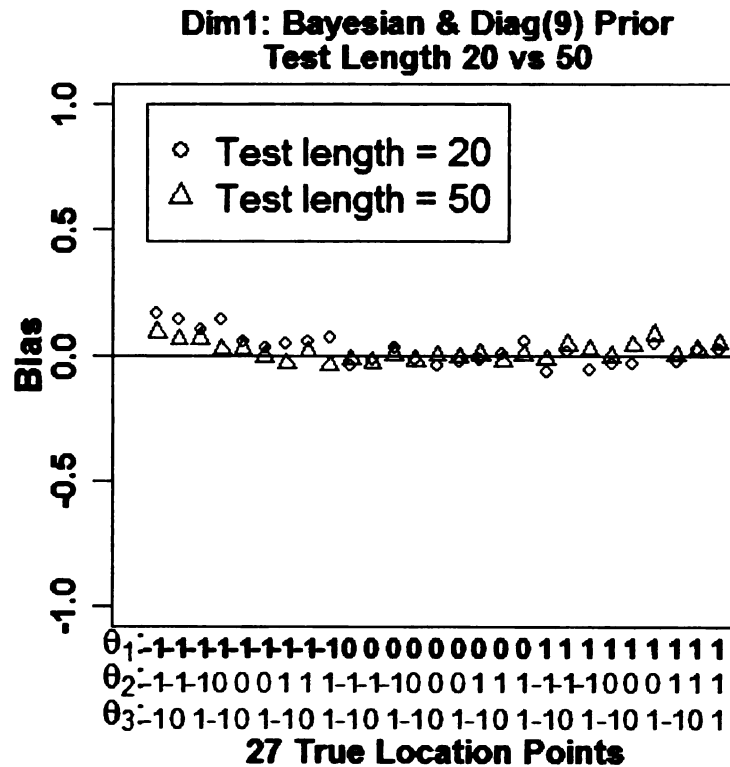
Figure 4.8 Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with identity matrix. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.



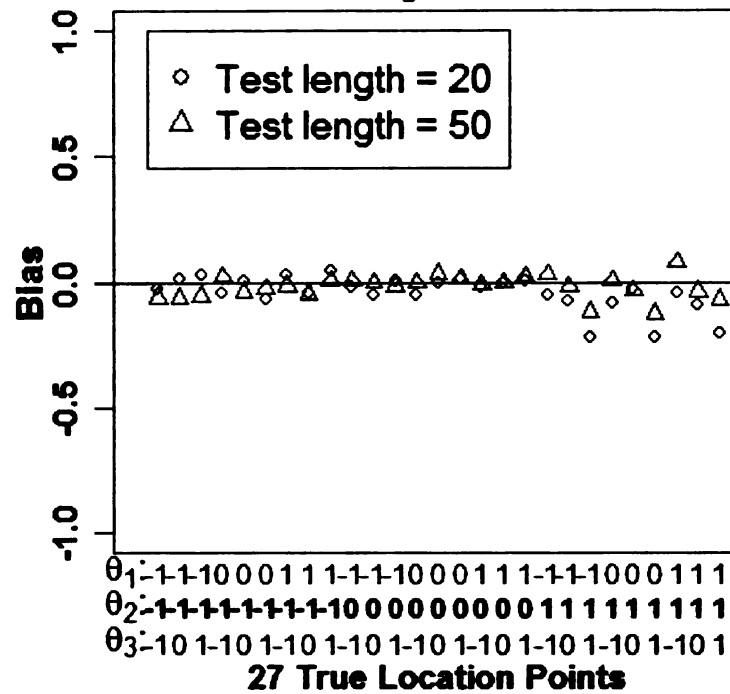
The study also compared the performance of different test lengths for the combination of Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with mean 0, and prior set as diag (9). The results were similar to the conditions that had identity matrix as the variance covariance matrix prior. Both the mean biases and RMSEs showed evidence that the estimation precision at the test length of 50 was slightly better than that of the test length of 20. However, the differences were small and at the test length of 20, the estimation was already stable and precise. Both the mean biases and RMSEs were small. When the test length increased to 50, the biases and RMSEs became slightly smaller. But the difference was not much. Mean biases and RMSEs for each dimension were shown in Figure 4.9.

Figure 4.9 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and diag (9) matrix as variance covariance matrix, at test length =20 and at test length =50.

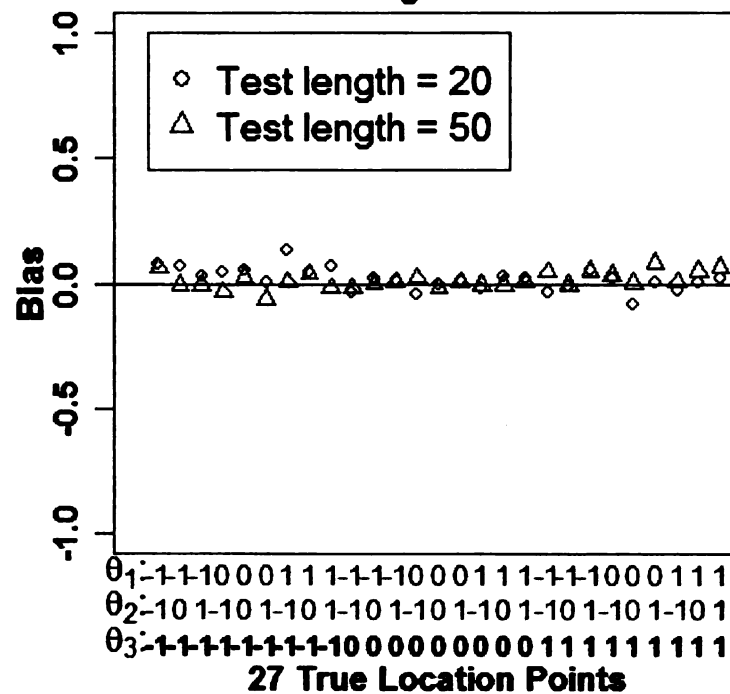
**A: Biases**



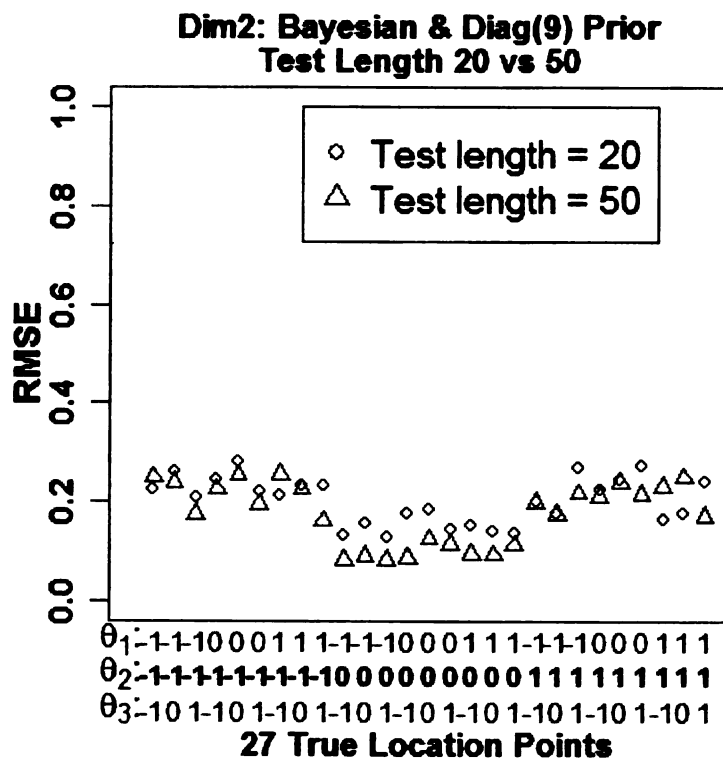
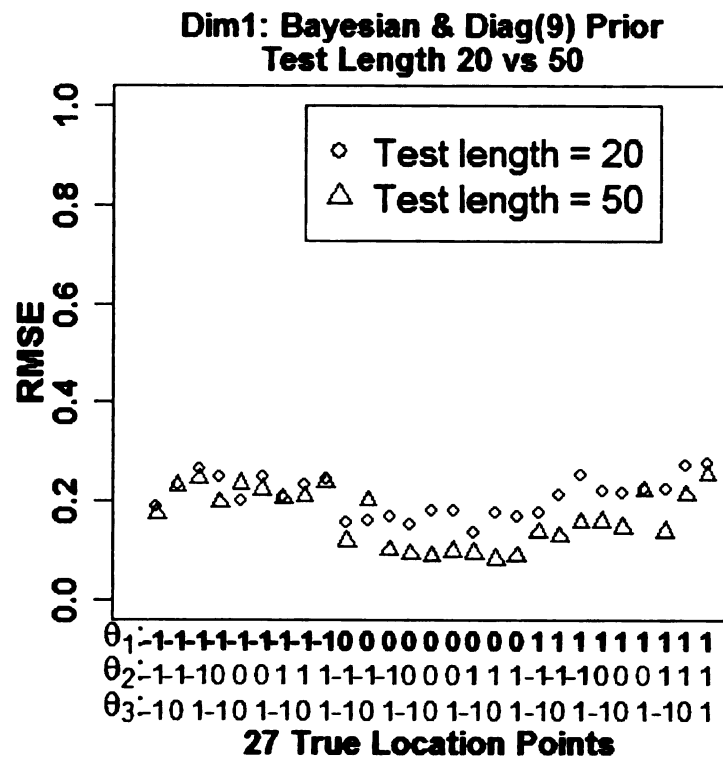
**Dim2: Bayesian & Diag(9) Prior  
Test Length 20 vs 50**

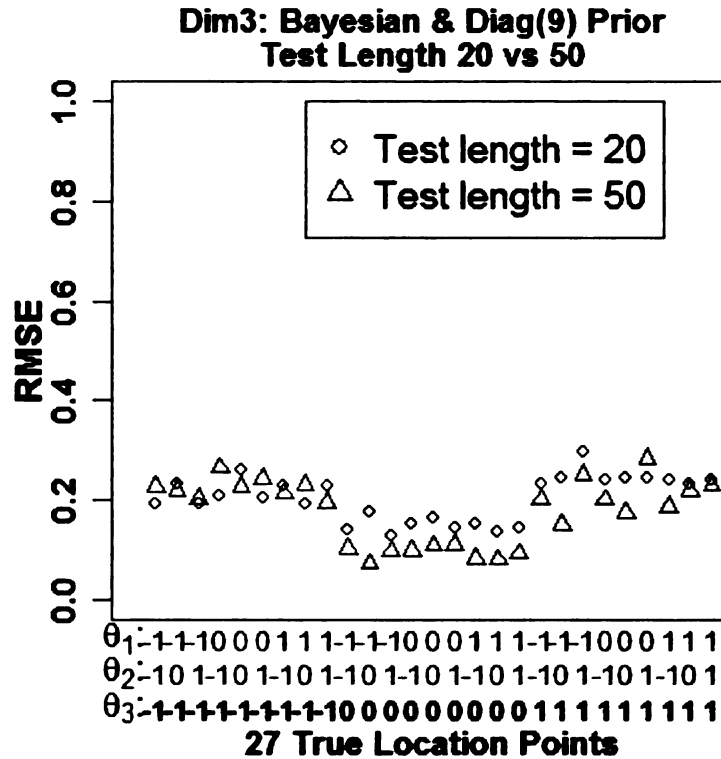


**Dim3: Bayesian & Diag(9) Prior  
Test Length 20 vs 50**



## B: RMSEs

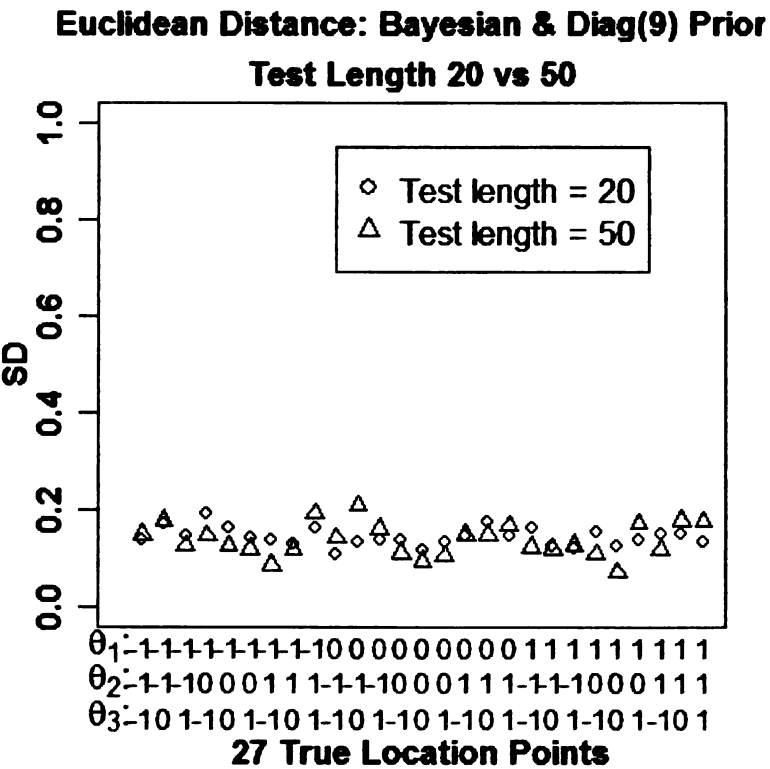
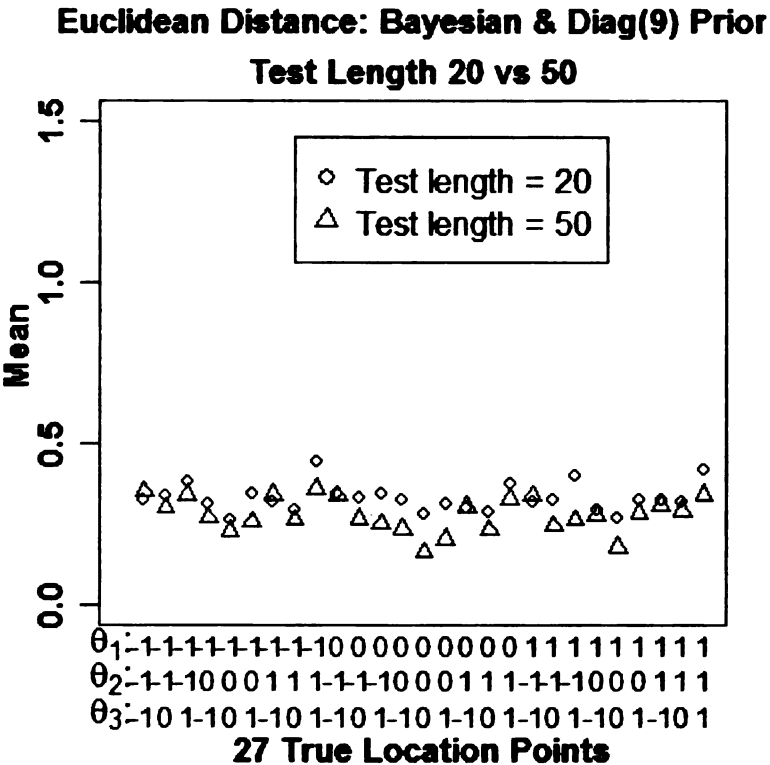




For this combination with diag(9) as prior variance covariance matrix, the Euclidean distances for the test length of 20 were slightly larger than those of the test length of 50. However, the differences were small. At the test length of 20, the final estimates were very near the true ability location points. When the test length increased to 50, the mean final estimates were closer to the true ability location points. The change was not much though. This corresponded to the result for the mean biases and RMSEs: the performance of the combination method of Bayesian as the ability update method and maximizing decrement volume in Bayesian, diag (9) prior as the item selection method was already good with the test length of 20. Mean and standard deviation of the Euclidean distance were shown in Figure 4.10.



Figure 4.10 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and diag (9) matrix as variance covariance matrix, at test length =20 and at test length =50.



Plot of successive estimates of one examinee with true location point (1, 1, 1) and initial estimate (0, 0, 0) with this combination of diag(9) as the prior variance covariance matrix is shown in Figure 4.11. Successive Euclidean distance between the updated estimates and the true ability location (1, 1, 1) was shown in Figure 4.12. The same as in conditions with identity matrix as the prior variance covariance matrix, the results showed that there was no non-convergence problems at the beginning of the test and the estimates quickly moved from initial estimate to the true location point.

Figure 4.11 Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with diag(9) variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

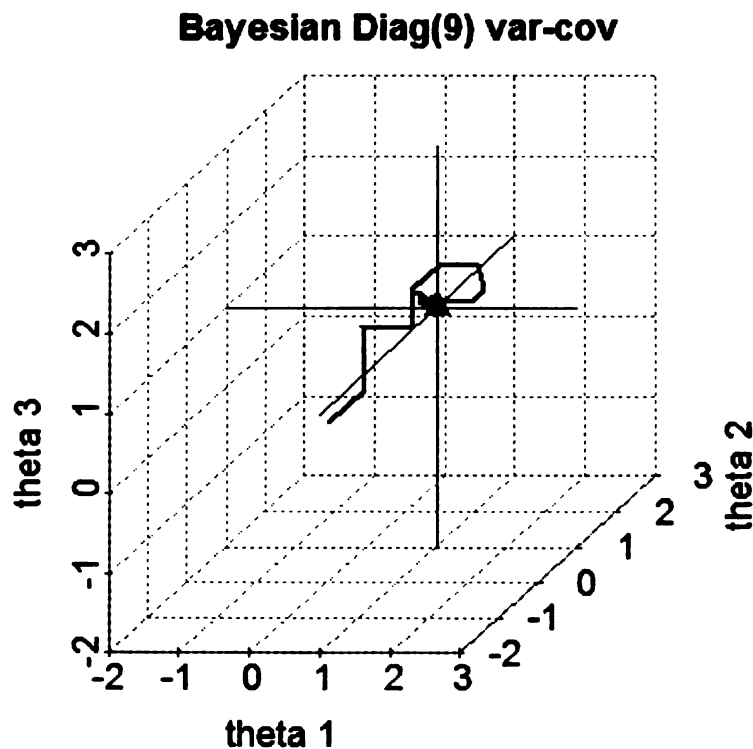
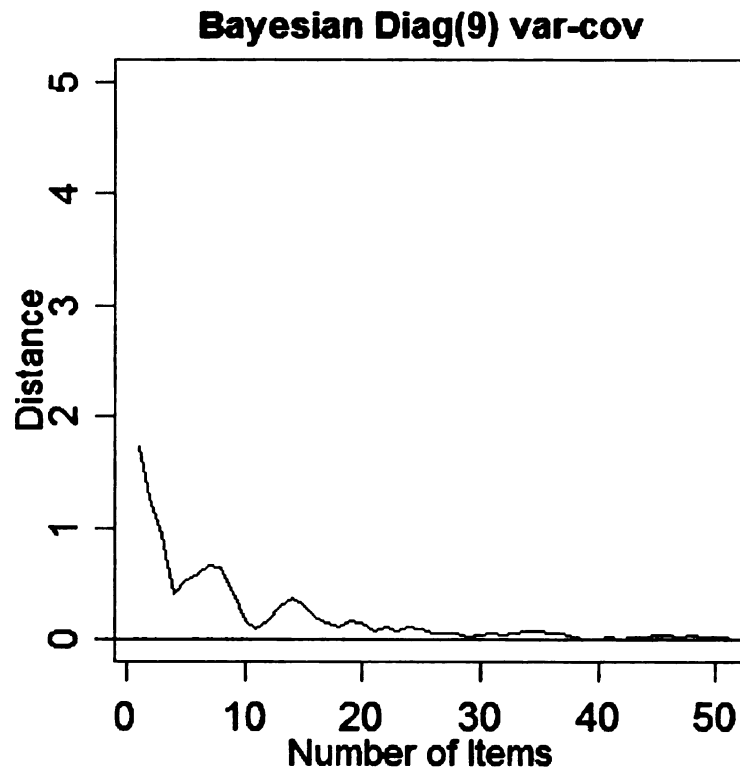


Figure 4.12 Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with diag(9) variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

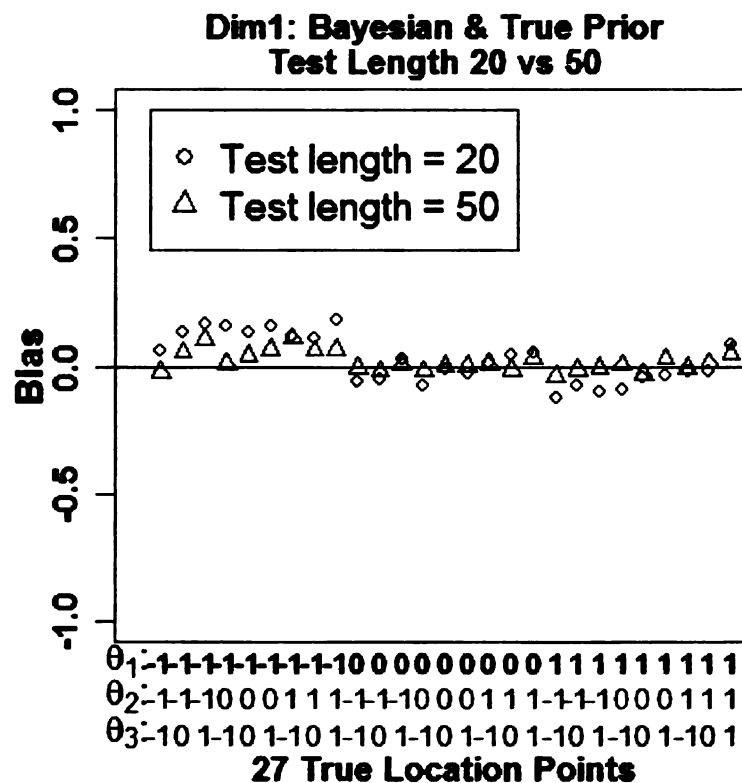


The third and final prior set in the study was true variance covariance matrix as the prior's variance covariance. Comparison between the test length of 20 and test length of 50 was made to access the converging speed of the combination and whether the estimates were accurate for a short test. For this combination, Bayesian was the ability estimation method and maximizing decrement volume in Bayesian was the item selection method with mean 0, and prior set as true variance covariance matrix. The true variance covariance matrix was the correlation matrix calculated from all 8562 7<sup>th</sup> grade examinees of 2005 Michigan Education Assessment Program (MEAP) Mathematics test. The true variance covariance matrix was given in Table 3.4. When using true variance covariance matrix as the prior's variance covariance matrix, both the mean biases and RMSEs showed evidence that the estimation precision at the test

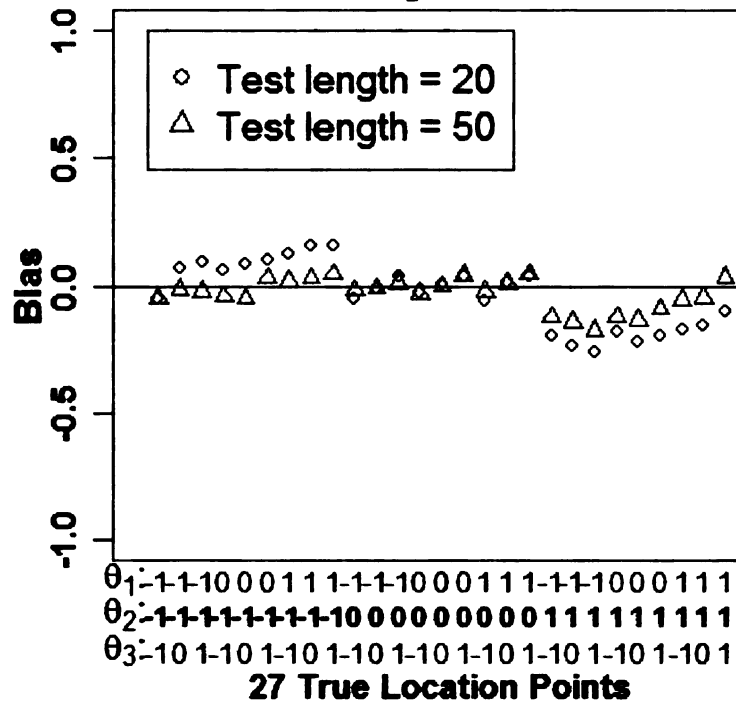
length of 50 was slightly better than that of the test length of 20. It was the same as for the conditions that had the identity matrix and diag (9) matrix as the variance covariance matrix prior. However, the differences were small and at the test length of 20, the estimation was already stable and accurate. When the test length increased to 50, the biases and RMSEs became slightly smaller. But the difference was not big as for practical purposes. Mean biases and RMSEs for each dimension were shown in

Figure 4.13 Mean biases and RMSEs for Bayesian as the ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and true variance covariance matrix as variance covariance matrix, at test length =20 and at test length =50.

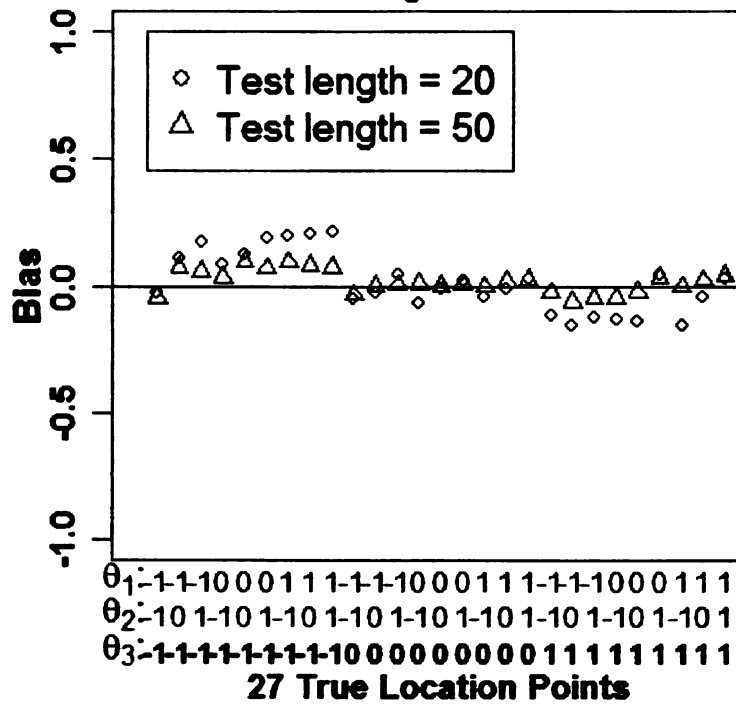
#### A: Biases



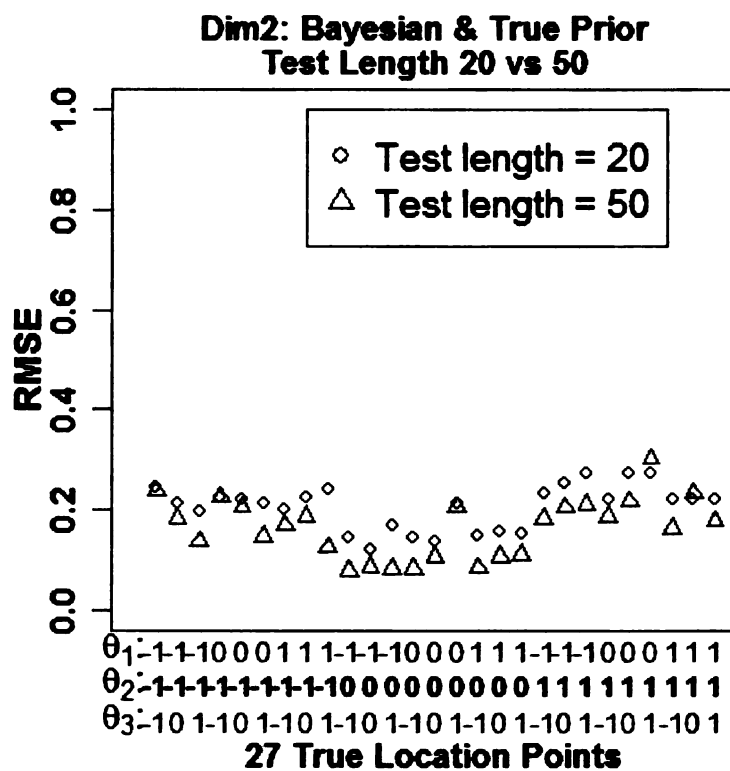
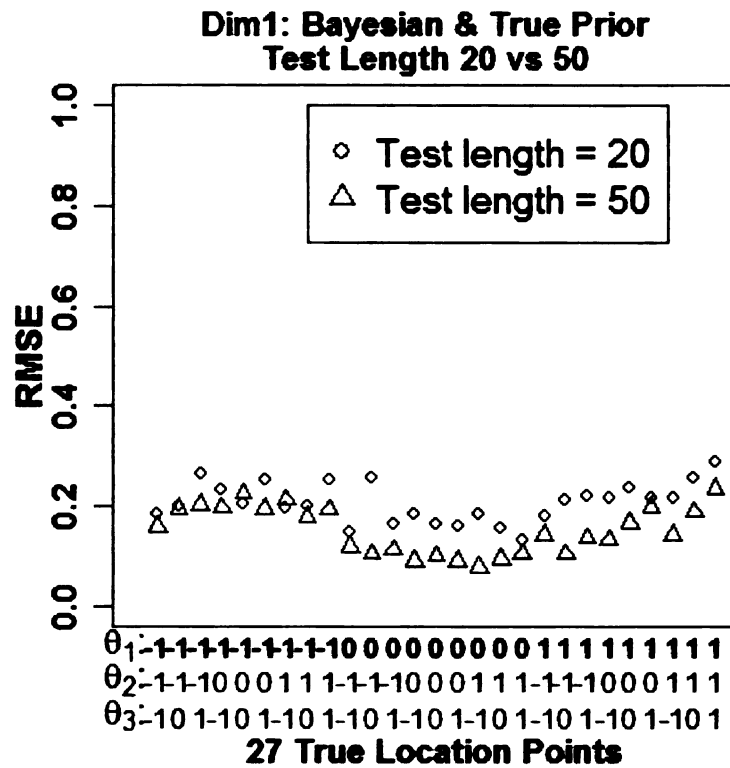
**Dim2: Bayesian & True Prior  
Test Length 20 vs 50**

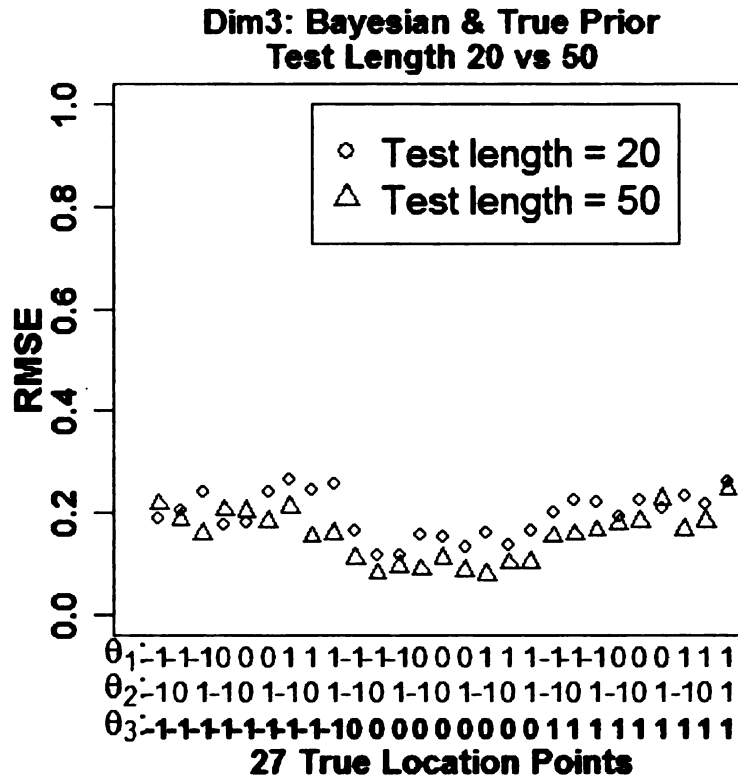


**Dim3: Bayesian & True Prior  
Test Length 20 vs 50**



### B: RMSEs

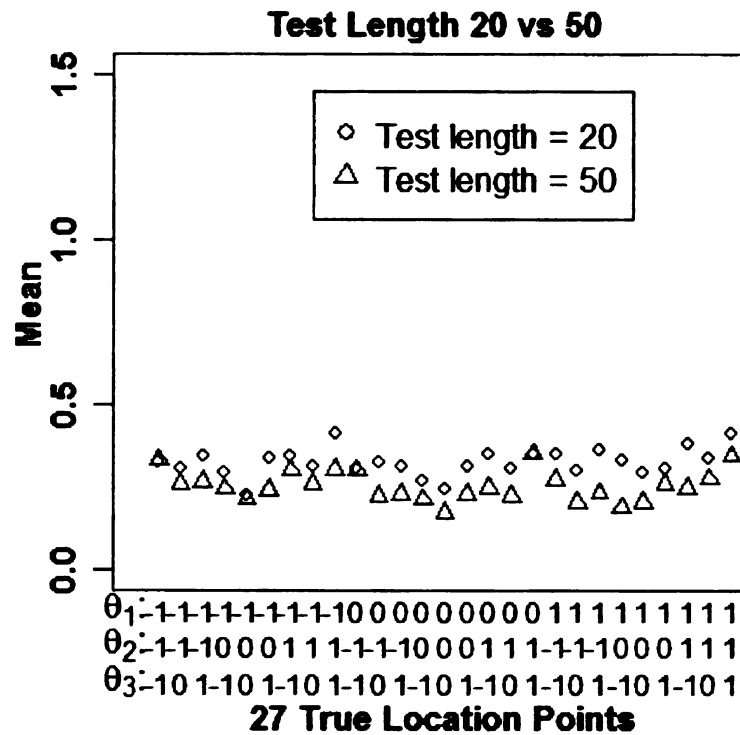




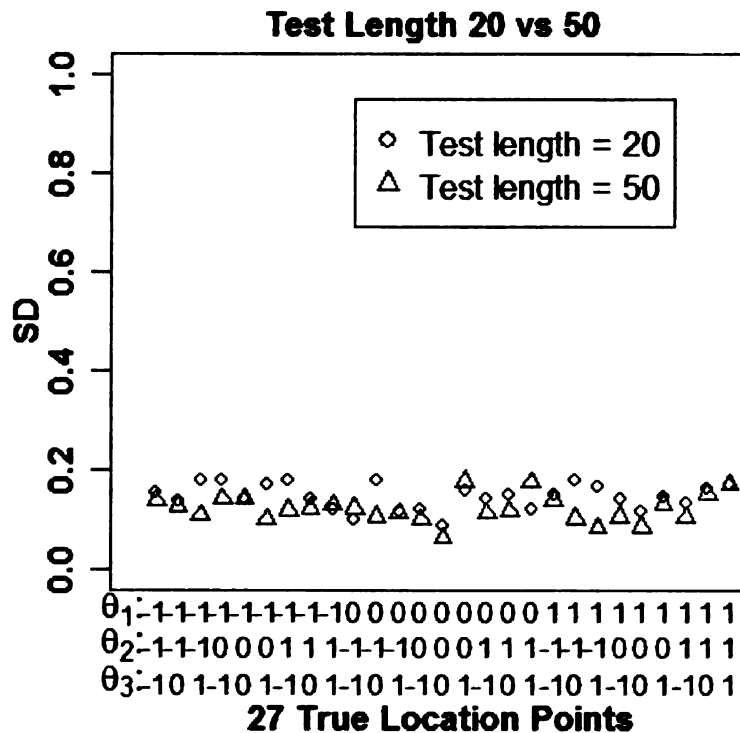
For the combination with true variance covariance matrix as the prior variance covariance matrix, the means and standard deviations of the Euclidean distance for the test length of 20 were slightly bigger than those of the test length of 50. However, the difference was small. At the test length of 20, the final estimates were already very near the true ability location points. When the test length increased to 50, the mean final estimates became closer to the true ability location points. The change was not so big as for practical purposes. The performance of the combination method of Bayesian as the ability update method and maximizing decrement volume in Bayesian, with true variance covariance matrix as the item selection method was already good with the test length of 20. Means and standard deviations of the Euclidean distance were shown in Figure 4.14.

Figure 4.14 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and maximizing decrement volume in Bayesian as the item selection method with prior set as mean 0 and true variance covariance matrix as variance covariance matrix, at test length =20 and at test length =50.

### Euclidean Distance: Bayesian & True Prior



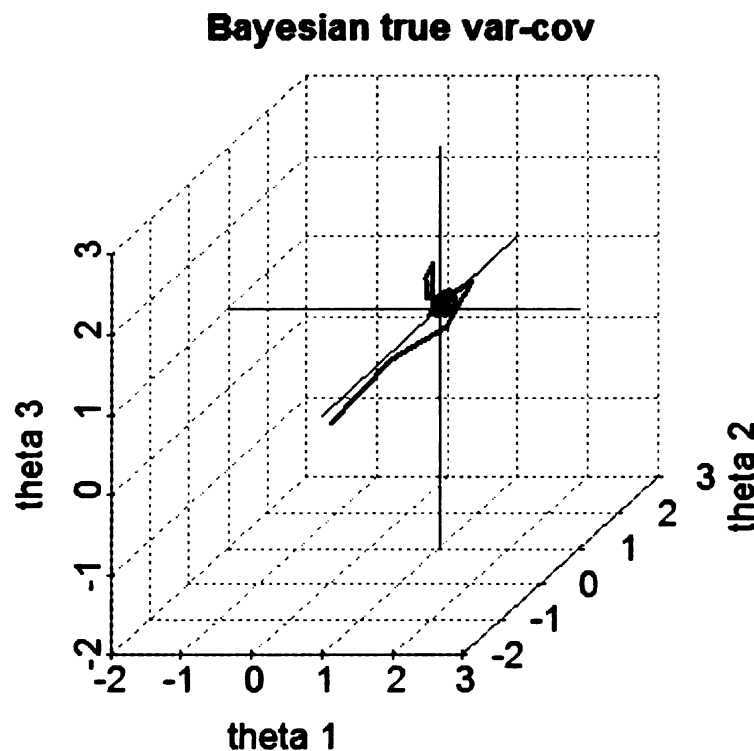
### Euclidean Distance: Bayesian & True Prior



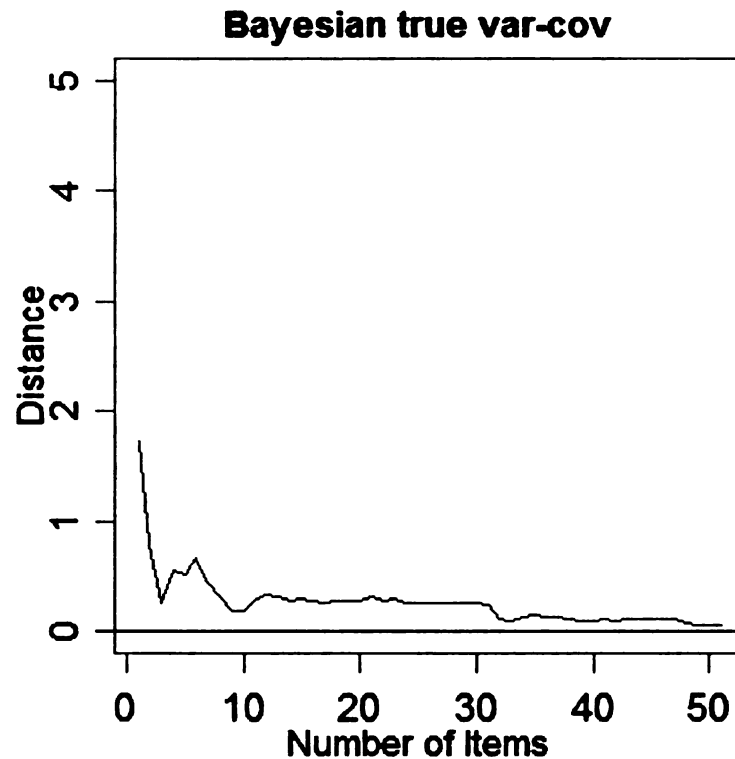


Plot of successive estimates of one examinee with true location point (1, 1, 1) and initial estimate (0, 0, 0) with this combination of true variance covariance as the prior variance covariance matrix is shown in Figure 4.15. Successive Euclidean distance between the updated estimates and the true ability location (1, 1, 1) are shown in Figure 4.12. The same as in conditions with identity matrix as the prior variance covariance matrix, the results show that there was no non-convergence problems at the beginning of the test and the estimates moved quickly from initial estimate to the true location point.

**Figure 4.15** Successive progress plot of updated ability estimates and true location point after administering each item for Bayesian method with true variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.



**Figure 4.16** Euclidean distance of between updated ability estimates and true location point after administering each item for Bayesian method with true variance covariance matrix as prior. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

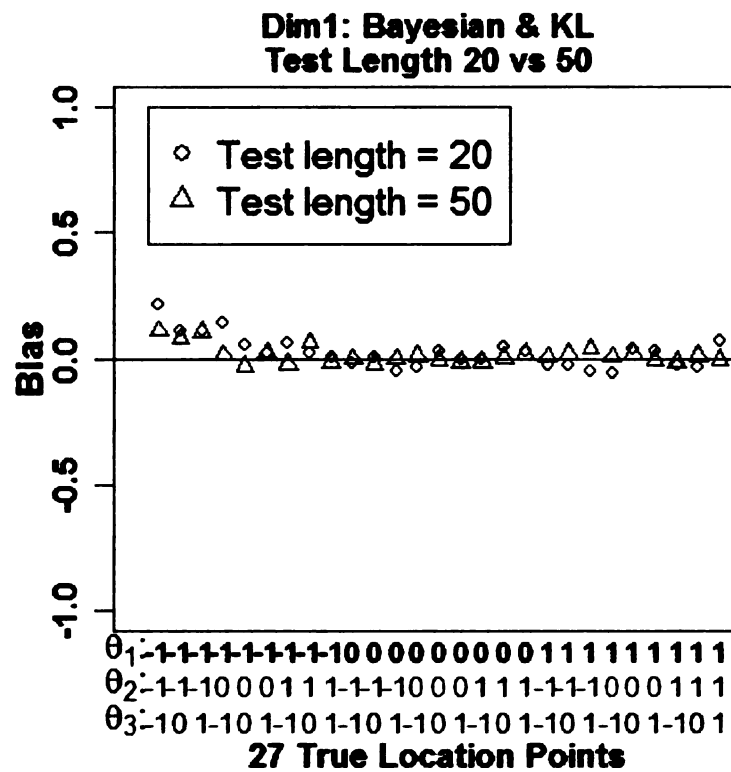


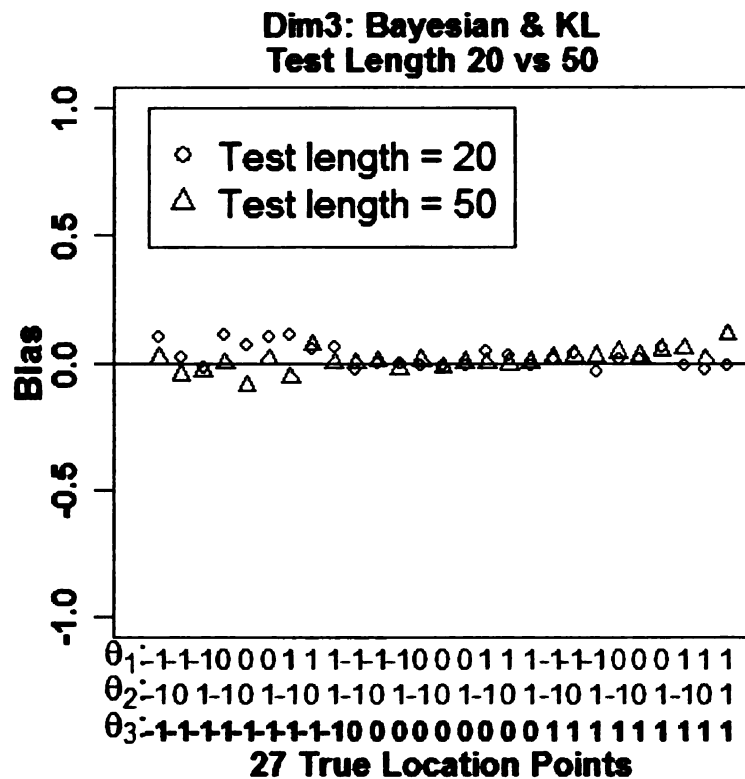
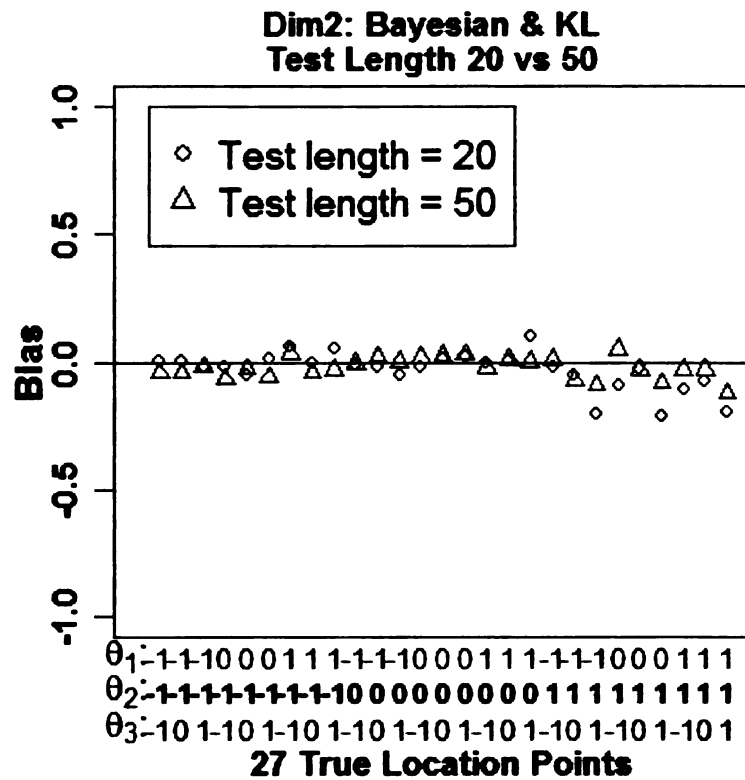
When Kullback-Leibler information was used instead of Fisher's information, the comparison between test length of 20 and test length of 50 was made to test the performance conditioning on different test lengths for this combination. For this combination, Bayesian was the ability estimation method and maximizing Kullback-Leibler information was the item selection method. The mean biases and RMSEs showed evidence that the estimation precision at the test length of 50 was better than that of the test length of 20. However, the differences were small and at the test length of 20, the estimation was already stable and precise. When the test length increased to 50, the biases and RMSEs became slightly smaller. But the difference was not so big as for practical purposes. This result was similar to the

combinations in which Bayesian was used as ability update method and Bayesian volume decrement used as item selection method.

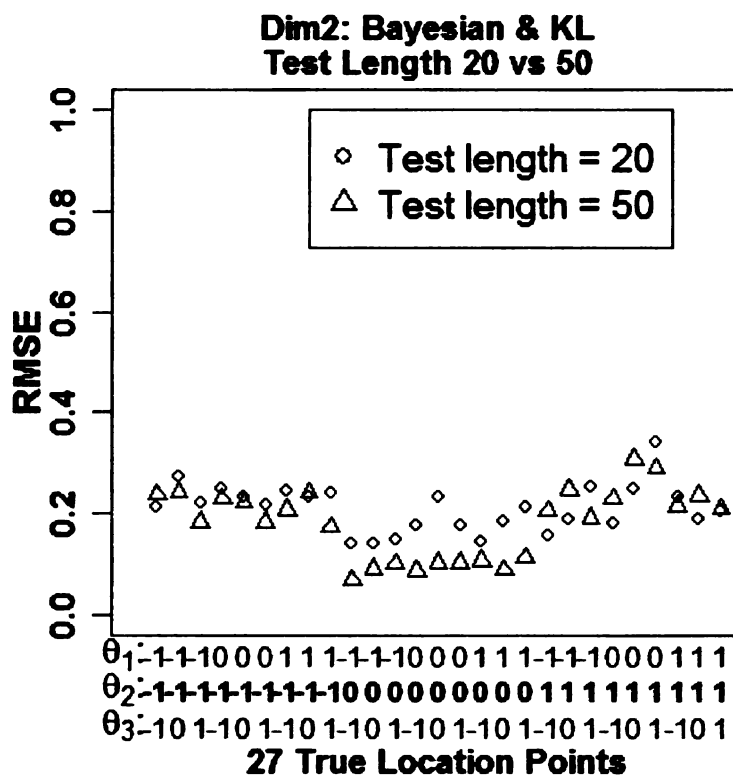
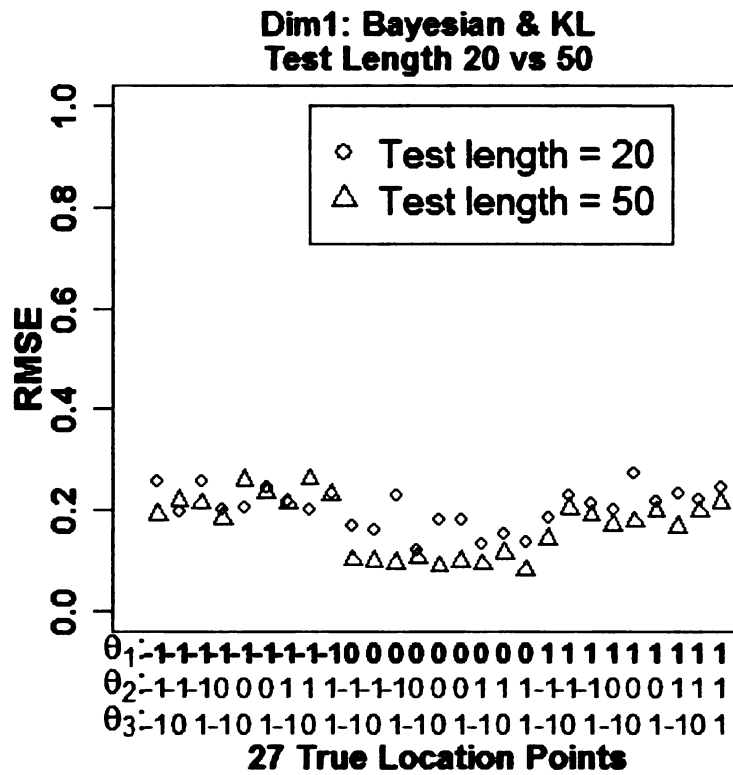
**Figure 4.17** Mean biases and RMSEs for Bayesian as the ability estimation method and Kullback-Leibler information as the item selection method, at test length =20 and at test length =50.

**A: Biases**





## B: RMSEs

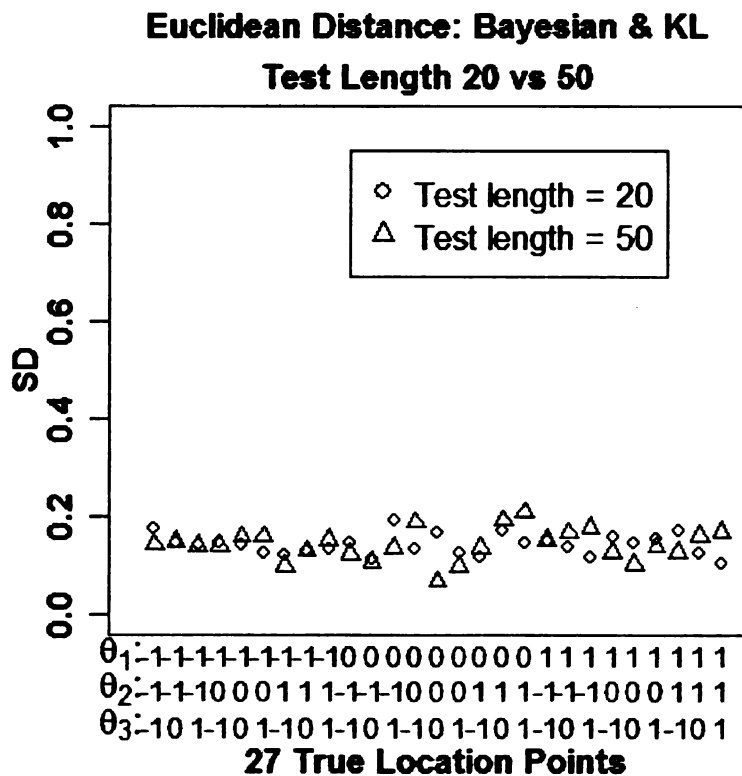
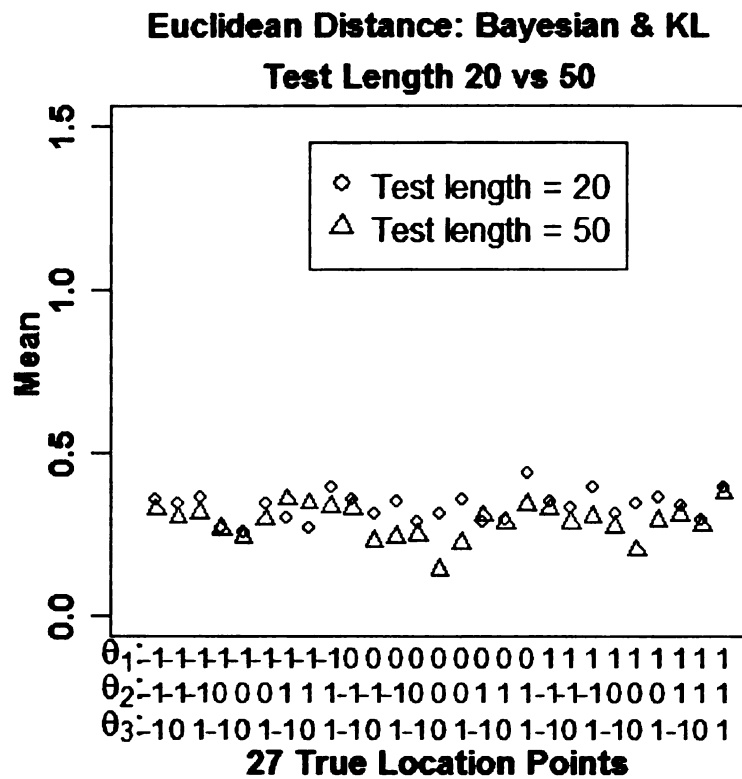


A scatter plot showing the Root Mean Square Error (RMSE) on the y-axis (ranging from 0.0 to 1.0) against 27 True Location Points on the x-axis. The x-axis labels are binary strings for three parameters:  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ . The legend indicates two data series: Test length = 20 (represented by diamonds) and Test length = 50 (represented by triangles). The RMSE values generally fluctuate between 0.1 and 0.3, with some points showing higher error (up to 0.3) and others showing lower error (down to 0.1). The error tends to be higher for the first 10 points and the last 10 points, and lower for the middle 7 points.

True Location Point	$\theta_1$	$\theta_2$	$\theta_3$	Test length = 20 (RMSE)	Test length = 50 (RMSE)
1	1	1	1	0.20	0.18
2	1	1	1	0.22	0.20
3	1	1	1	0.18	0.18
4	1	1	1	0.22	0.22
5	1	1	1	0.20	0.22
6	1	1	1	0.22	0.25
7	1	1	1	0.20	0.22
8	1	1	1	0.22	0.25
9	1	1	1	0.20	0.22
10	1	1	1	0.22	0.20
11	1	1	1	0.18	0.18
12	1	1	1	0.15	0.15
13	1	1	1	0.18	0.15
14	1	1	1	0.15	0.15
15	1	1	1	0.18	0.15
16	1	1	1	0.15	0.15
17	1	1	1	0.18	0.15
18	1	1	1	0.15	0.15
19	1	1	1	0.18	0.15
20	1	1	1	0.15	0.15
21	1	1	1	0.18	0.15
22	1	1	1	0.15	0.15
23	1	1	1	0.18	0.15
24	1	1	1	0.15	0.15
25	1	1	1	0.18	0.15
26	1	1	1	0.15	0.15
27	1	1	1	0.18	0.15

61

Figure 4.18 Mean and standard deviation of Euclidean distance of Bayesian as ability estimation method and Kullback-Leibler information as the item selection method, at test length =20 and at test length =50.



Plot of successive estimates of one examinee with true location point (1, 1, 1) and initial estimate (0, 0, 0) with maximizing Kullback-Leibler information is shown in Figure 4.19. Successive Euclidean distance between the updated estimates and the true ability location (1, 1, 1) are shown in Figure 4.20. The same as in conditions with all other Bayesian methods, the results for Kullback-Leibler showed that there was no non-convergence problems at the beginning of the test and the estimates quickly moved from initial estimate to the true location point. The results also corresponded to results from the analysis of mean biases and RMSEs that at the short test (test length=20) the estimates were already accurate.

Figure 4.19 Successive progress plot of updated ability estimates and true location point after administering each item for Kullback-Leibler. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.

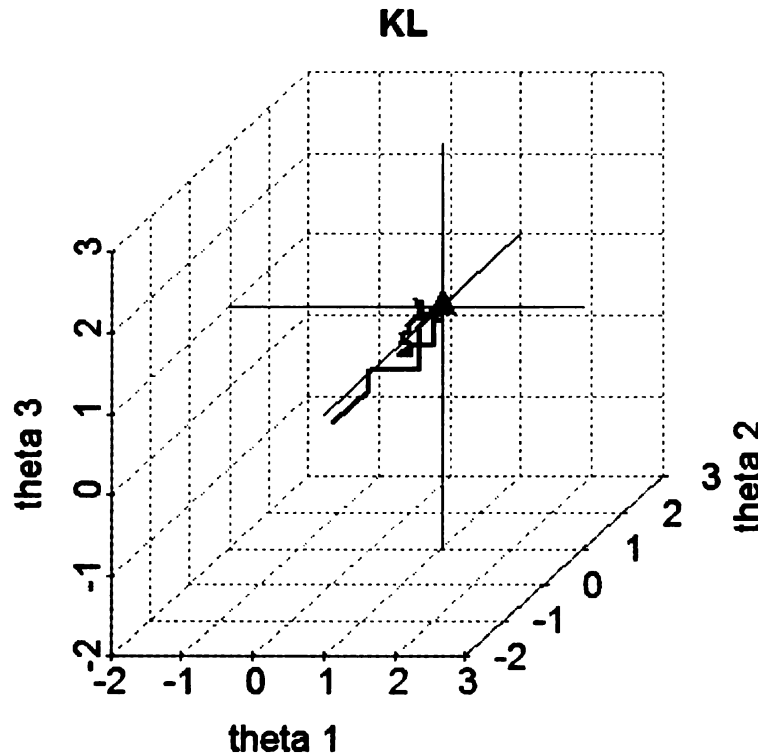
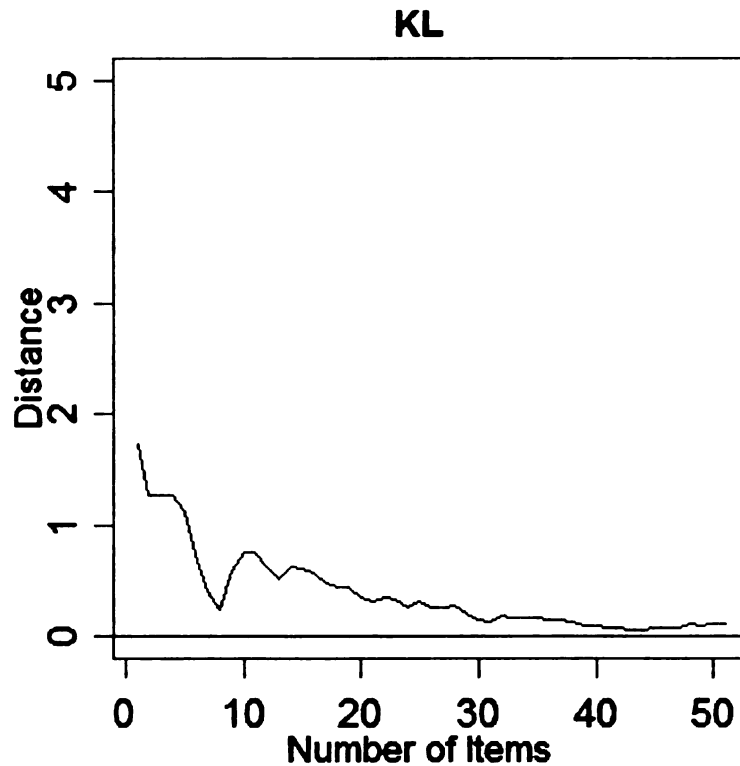




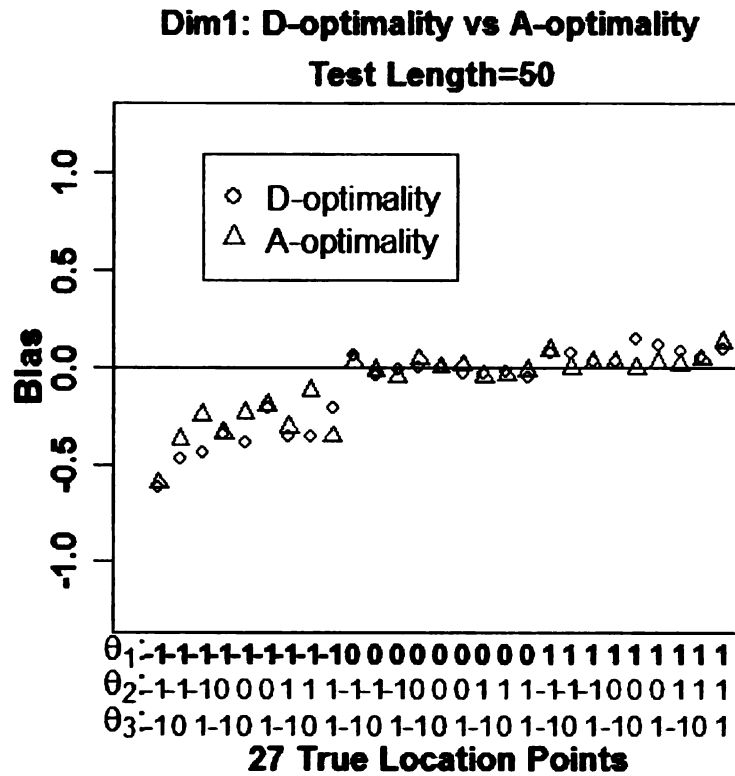
Figure 4.20 Euclidean distance of between updated ability estimates and true location point after administering each item for Kullback-Leibler. Initial estimate (0, 0, 0). True location point (1, 1, 1). Test length=50.



One of the research questions was to compare the performance of A-optimality with D-optimality as the item selection method when maximum likelihood was used as the ability estimation method. The hypothesis was that their performance was comparable. Mathematical aspects of this comparability are given in the Chapter 5. Mean biases and RMSEs of the final estimates of both methods were compared at the test length of 50 for each dimension. Means and standard deviations of Euclidean distance between the final estimates and true ability location points were also compared.

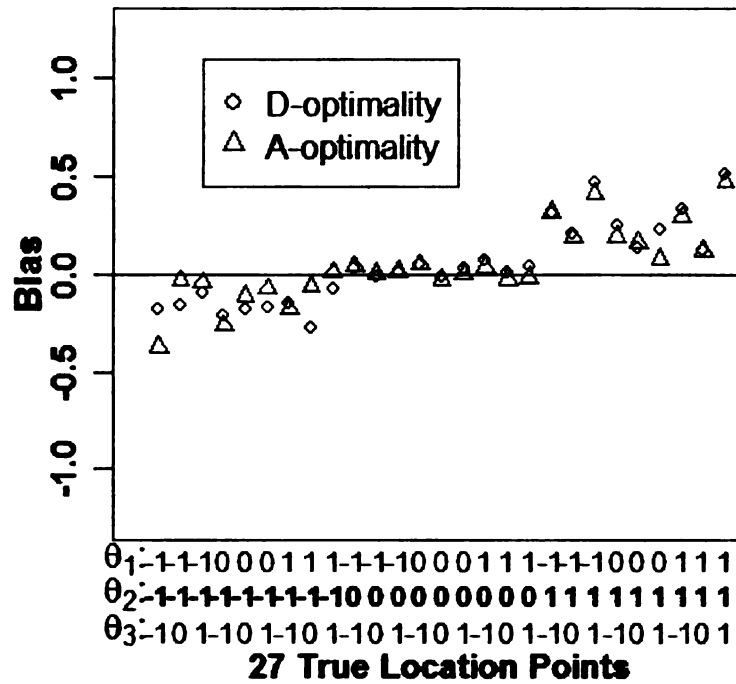
Figure 4.21 Mean biases and RMSEs for maximum likelihood as the ability estimation method, with D-optimality and A-optimality at the item selection methods, test length =50.

**A: Biases**



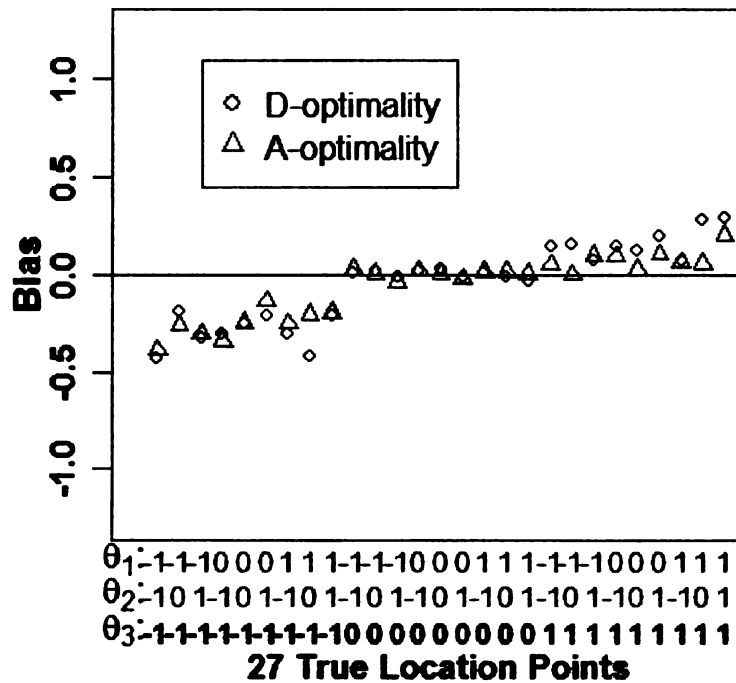
### Dim2: D-optimality vs A-optimality

Test Length=50



### Dim3: D-optimality vs A-optimality

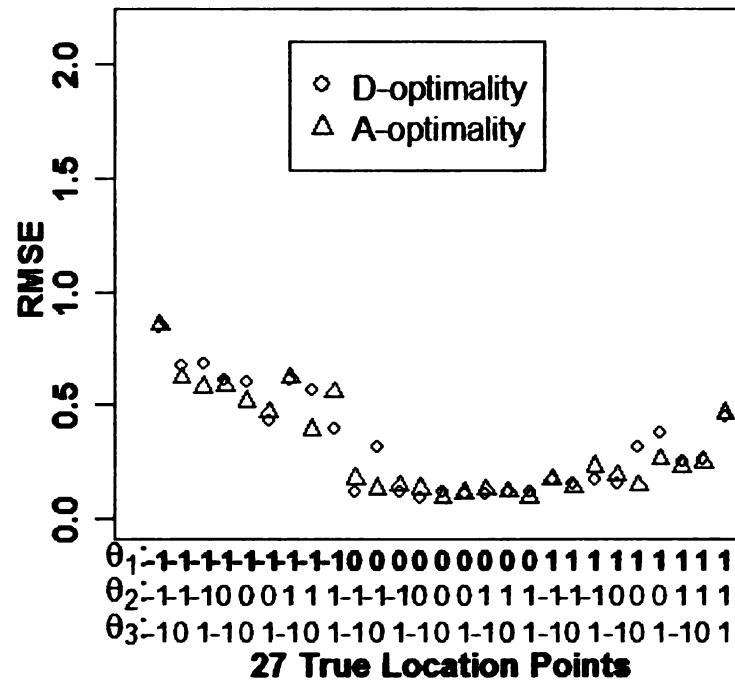
Test Length=50



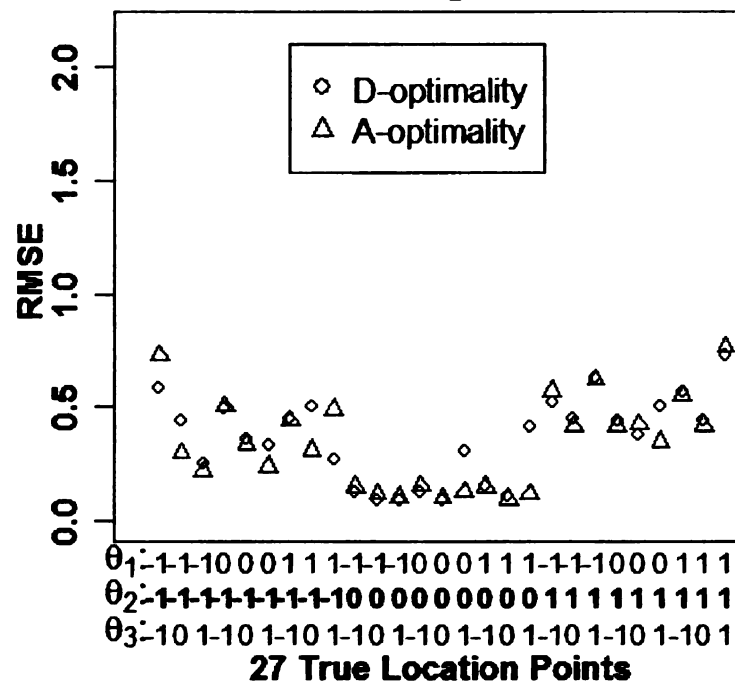
### B: RMSEs

### Dim1: D-optimality vs A-optimality

Test Length=50

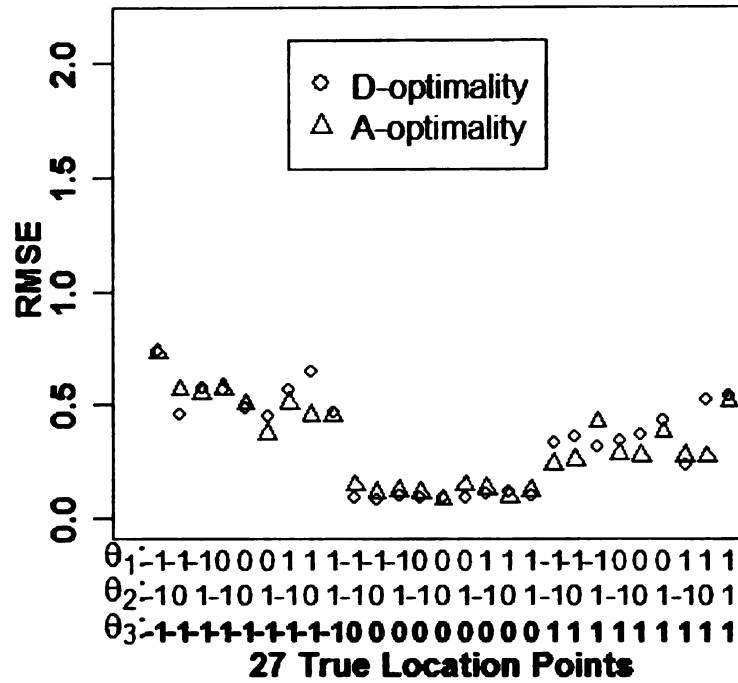


**Dim2: D-optimality vs A-optimality**  
**Test Length=50**



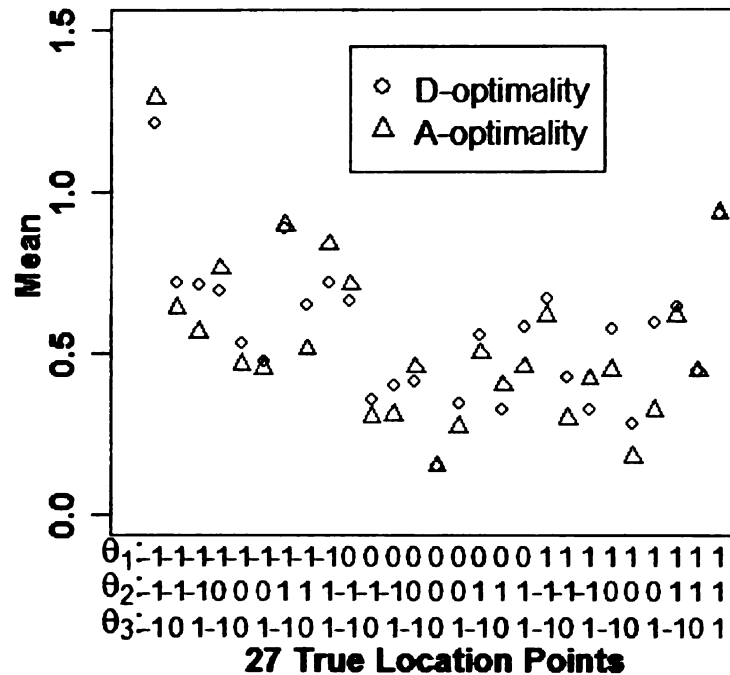
### Dim3: D-optimality vs A-optimality

Test Length=50

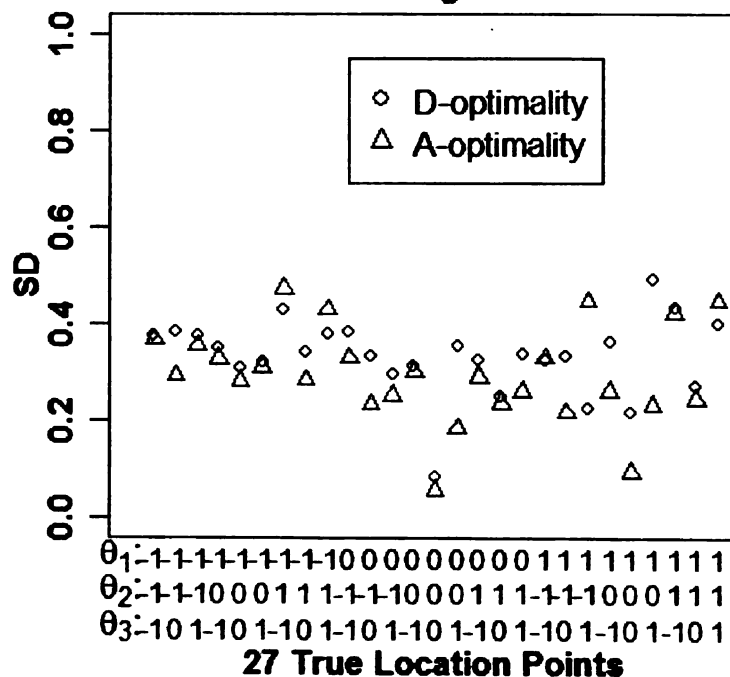


The mean biases and RMSEs were very similar for D-optimality and A-optimality. It showed that at the test length of 50, the two item selection methods were comparable, which was the same as in the research hypothesis.

**Test Length=50**



**Test Length=50**

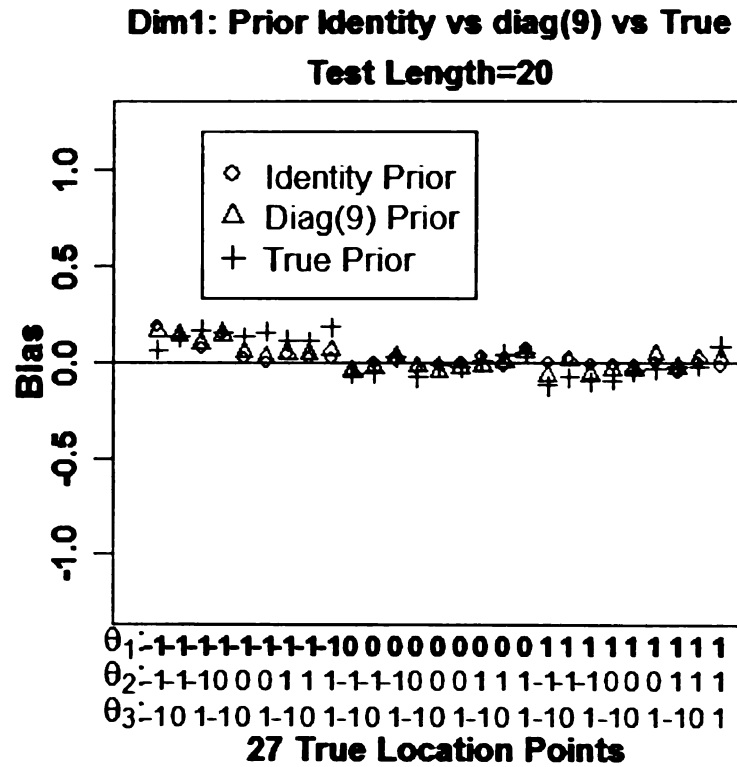


The results of the means and standard deviations of Euclidean distance between the final estimates and true location points were measures of estimation precision over dimensions. Figure 4.22 shows that over all three dimensions, the estimation precision of D-optimality and A-optimality was similar.

For the research question on the evaluation of the impact of priors on the performance of using Bayesian as the item selection method and maximizing decrement volume in Bayesian as the item selection method, the comparisons were made for the test length of 20 and test length of 50 and variance covariance were identity matrix,  $\text{diag}(9)$ , and true variance covariance matrix calculated from the population. Mean biases and RMSEs are shown in Figure 4.23 for test length of 20 and in Figure 4.24 for the test length of 50.

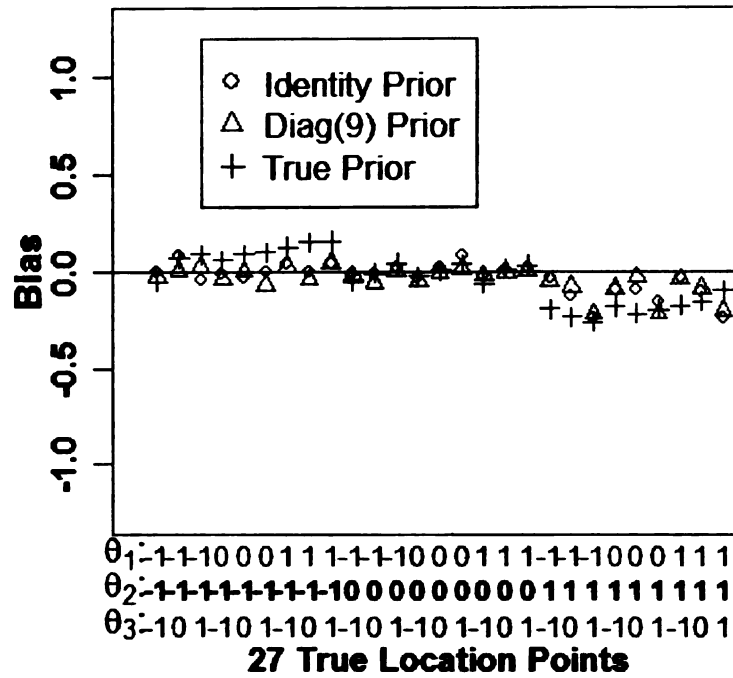
Figure 4.23 Mean biases and RMSEs for Bayesian as the ability estimation method, comparison of prior variance covariance matrix as: 1) identity matrix; 2) diag (9) and 3) true variance covariance matrix. Test length=20.

**A: Biases**

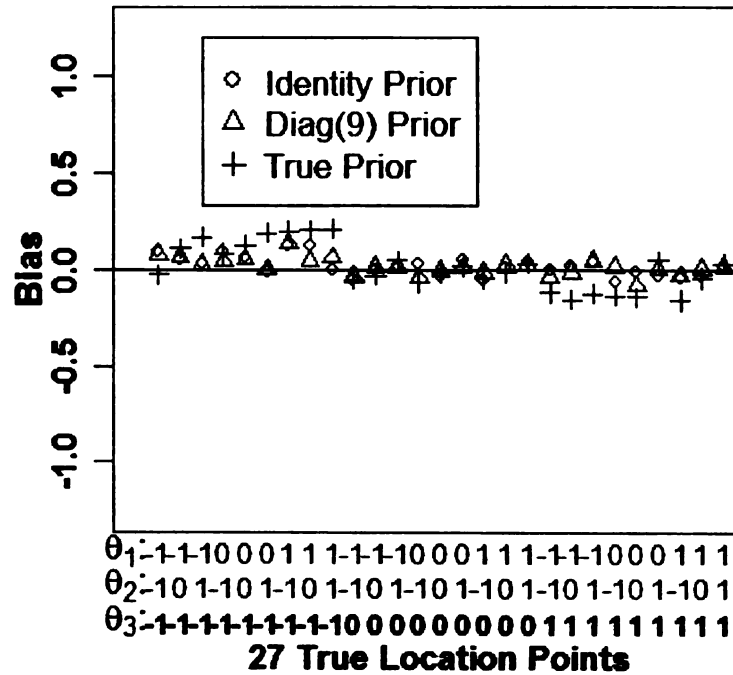




**Dim2: Prior Identity vs diag(9) vs True**  
**Test Length=20**

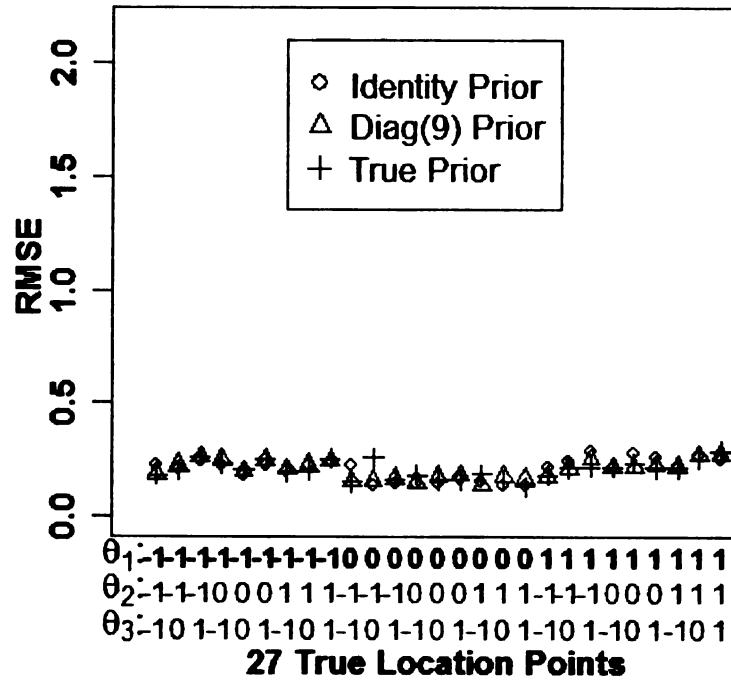


**Dim3: Prior Identity vs diag(9) vs True**  
**Test Length=20**

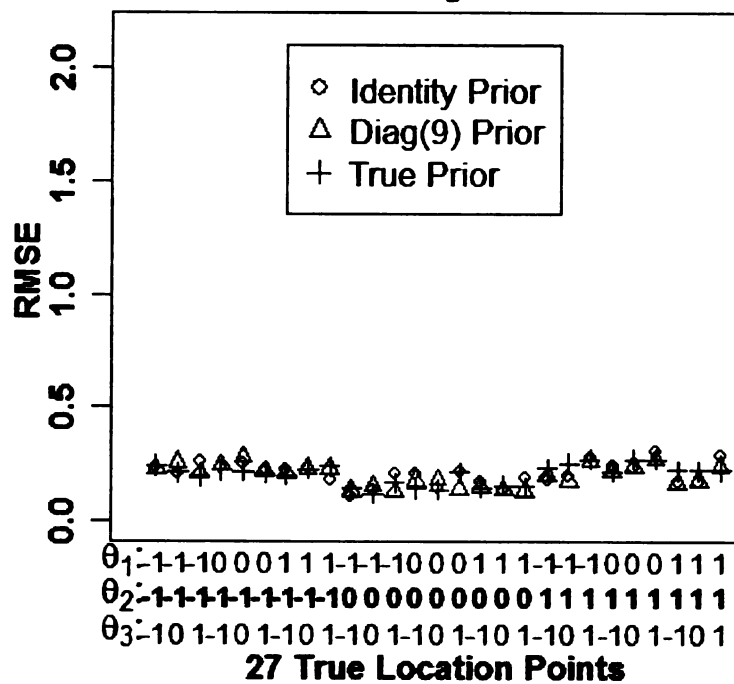


## B: RMSEs

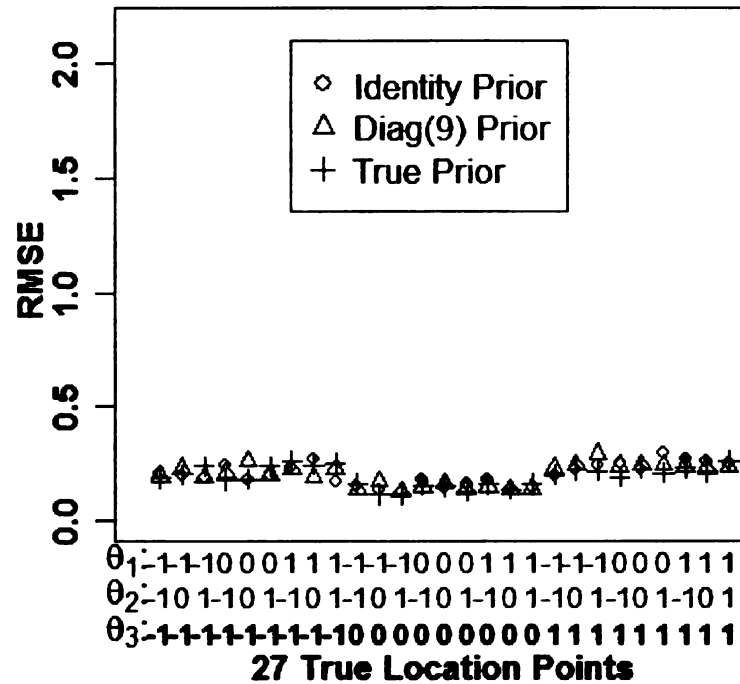
**Dim1: Prior Identity vs diag(9) vs True**  
**Test Length=20**



**Dim2: Prior Identity vs diag(9) vs True**  
**Test Length=20**



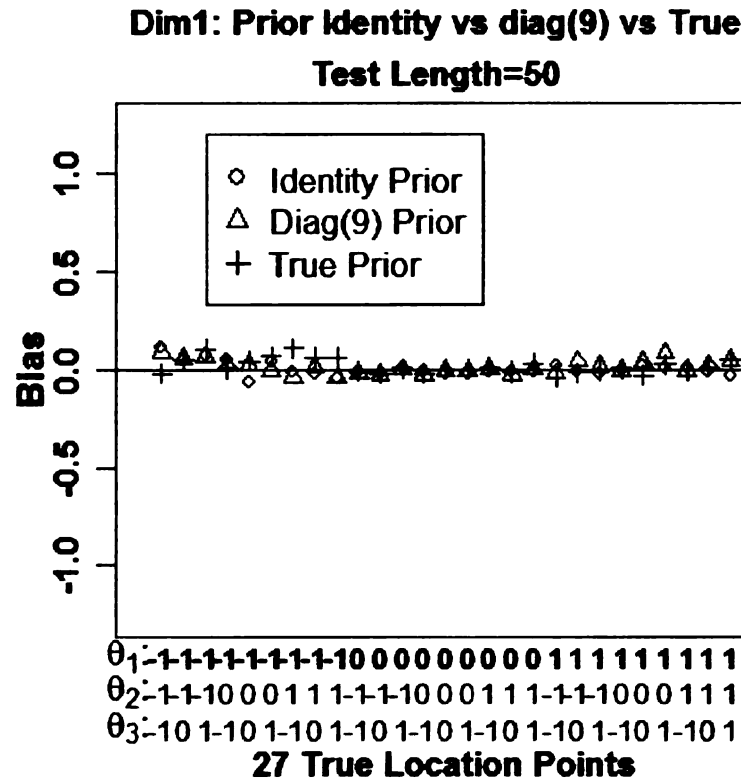
**Dim3: Prior Identity vs diag(9) vs True**  
**Test Length=20**



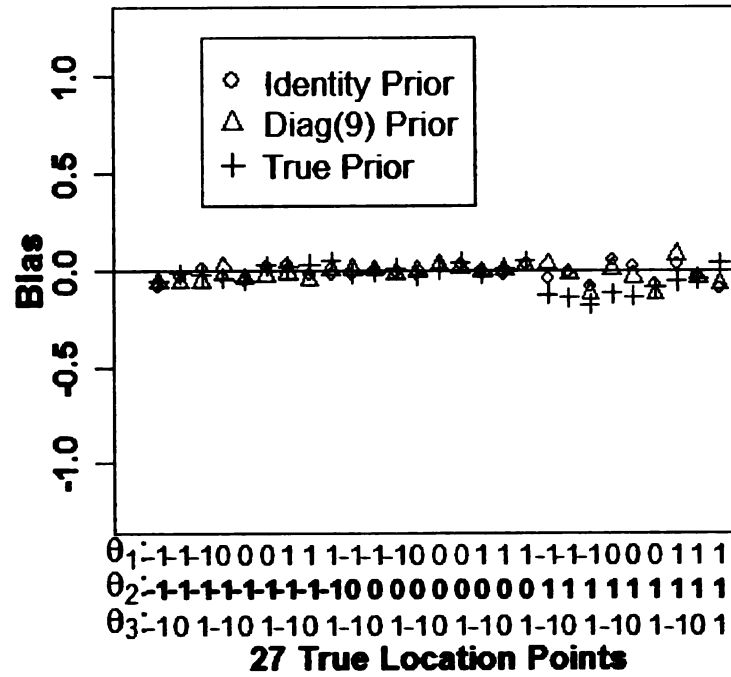
At the test length of 20, if the value of true location points on the dimension on which the biases were calculated was 0, the biases of all three priors were very close. When the true value was either 1 or -1, among the three priors, the biases for the true variance covariance matrix were the biggest. Prior variance covariance matrix as identity matrix and diag(9) were comparable. However, overall, the biases for all three priors on all three dimensions were very small and comparable, even though the true variance covariance had the largest biases for true values away from 0. The comparison from RMSEs showed that all three priors were comparable and there was no big difference at the test length of 20.

Figure 4.24 Mean biases and RMSEs for Bayesian as the ability estimation method, comparison of prior variance covariance matrix as: 1) identity matrix; 2) diag (9) and 3) true variance covariance matrix. Test length=50.

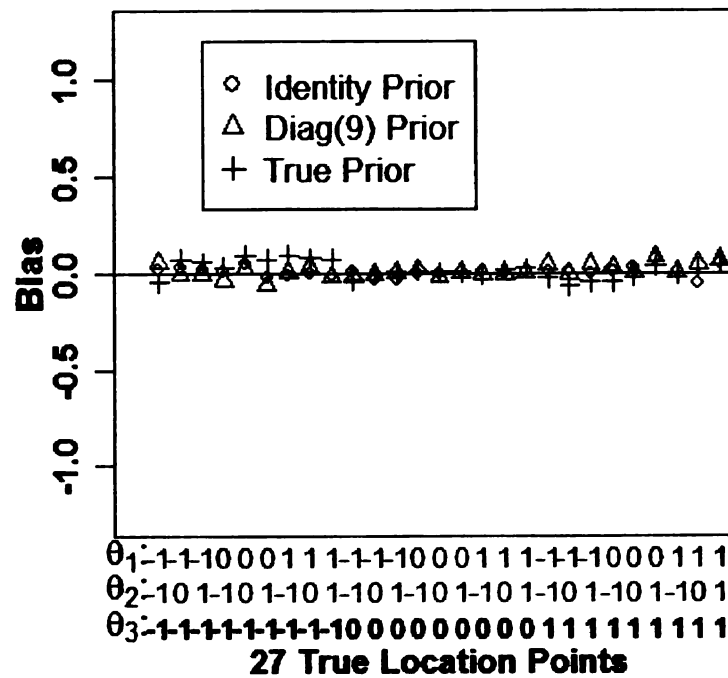
#### A: Biases



**Dim2: Prior Identity vs diag(9) vs True**  
**Test Length=50**



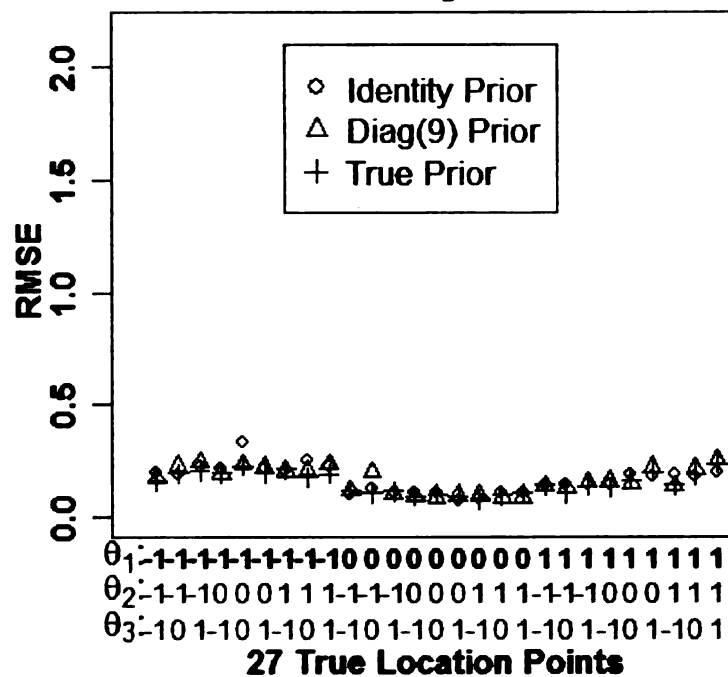
**Dim3: Prior Identity vs diag(9) vs True**  
**Test Length=50**



## B: RMSEs

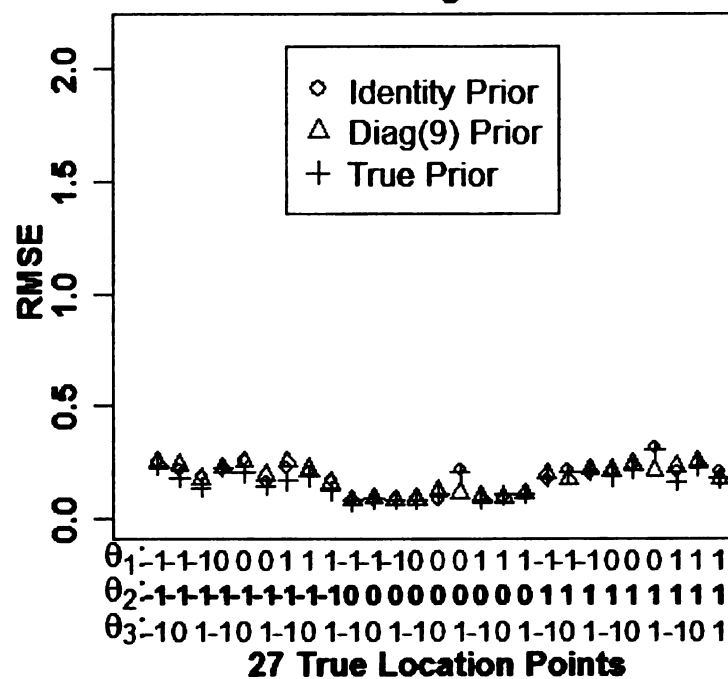
**Dim1: Prior Identity vs diag(9) vs True**

**Test Length=50**



**Dim2: Prior Identity vs diag(9) vs True**

**Test Length=50**



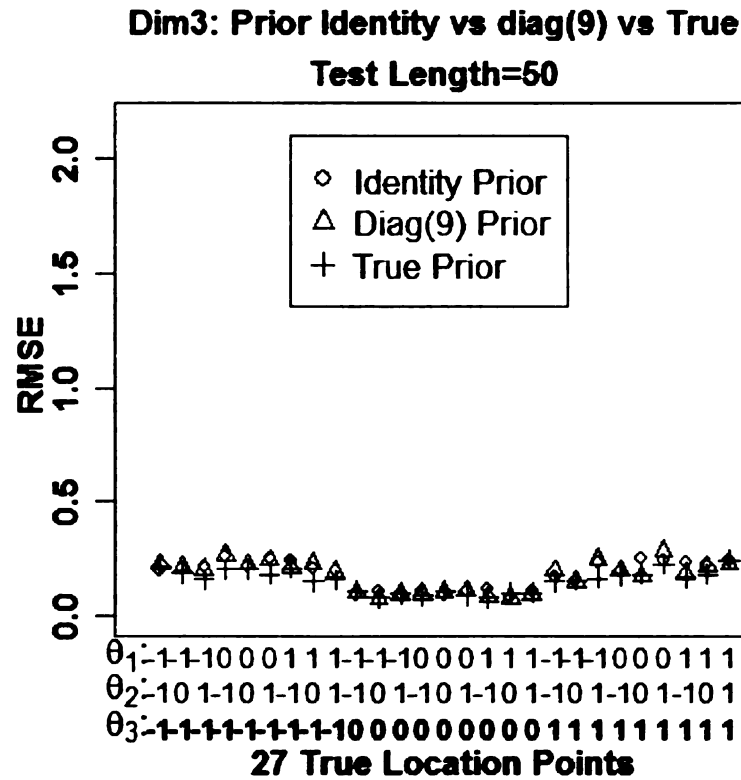


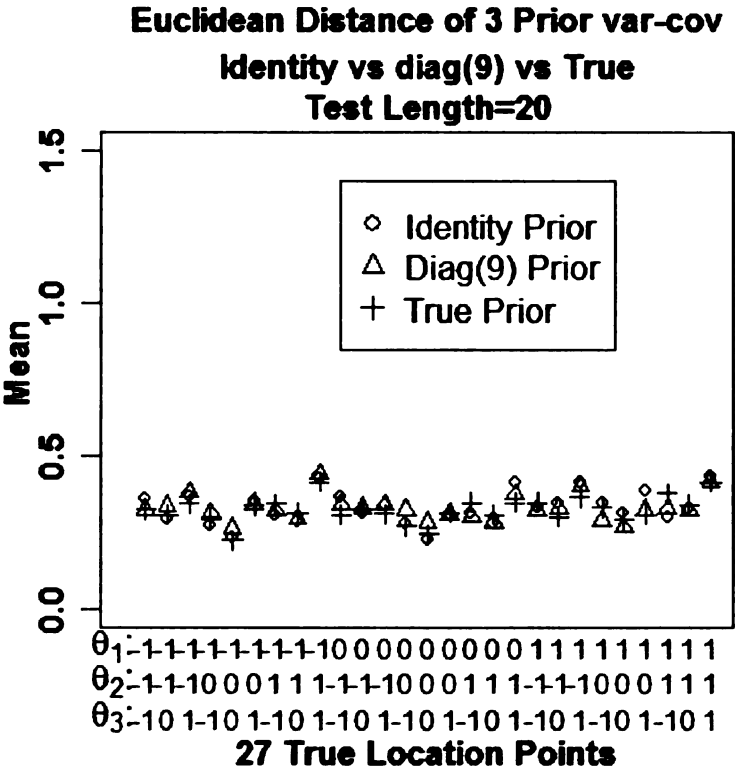
Figure 4.24 was the comparison of mean biases and RMSEs for all three priors when the test length was 50. It showed that at the test length of 50, both biases and RMSEs were very small and estimates were very accurate for all three priors. Therefore, when the test was long (test length=50), the impact of prior was small for the combination of Bayesian as the ability estimation method and maximizing volume decrement in Bayesian as the item selection method. This combination for all three prior, that is, strong prior, relatively weak prior, and true prior, produced accurate estimates at the end of the tests.

The above results were drawn for each dimension. An overall measure, Euclidean distance was also calculated and shown in Figure 4.25. Mean and standard deviation of Euclidean distance between the final estimates and true location points for both the

test length of 20 and test length of 50 showed that the impact of priors was small and all three priors were comparable.

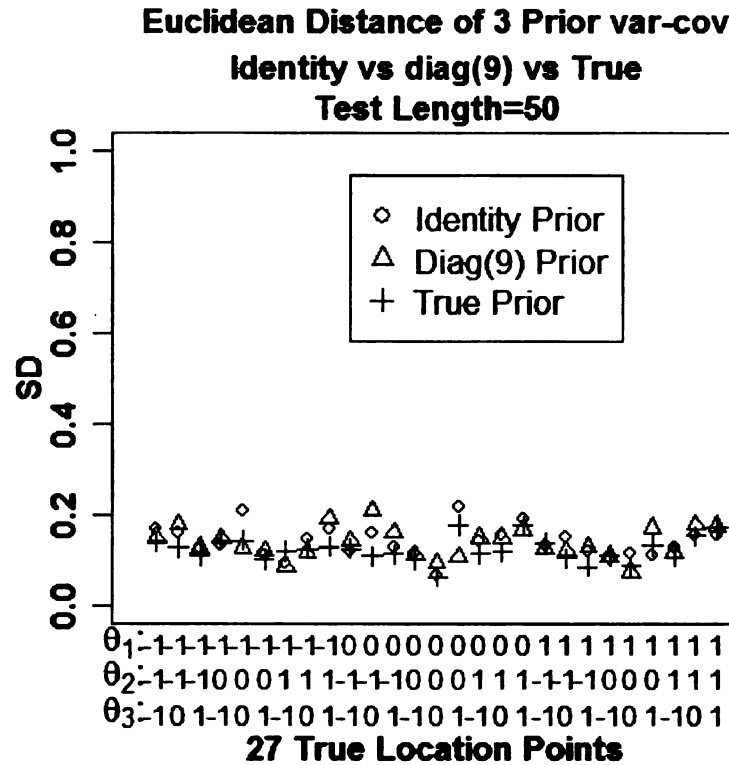
Figure 4.25 Mean and standard deviation of Euclidean distance of Bayesian as the ability estimation method, comparison among prior variance covariance matrix: 1) identity matrix, 2) diag(9), and 3) true variance covariance matrix.

Test length =20





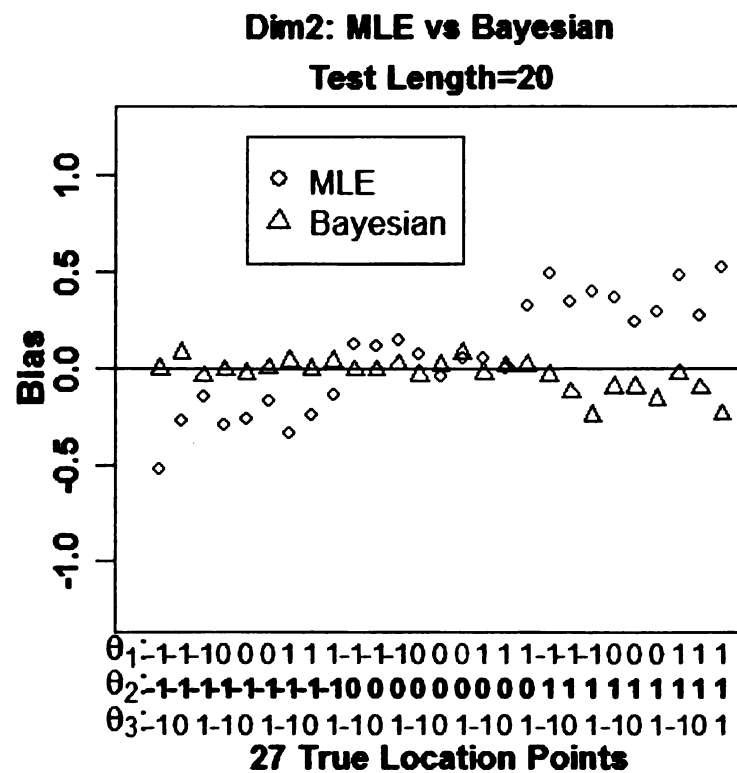
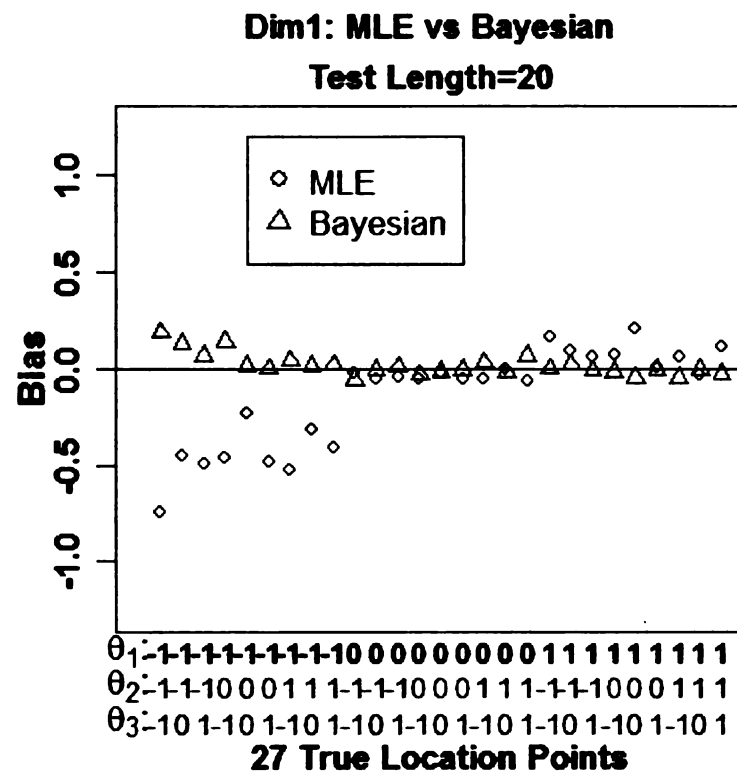




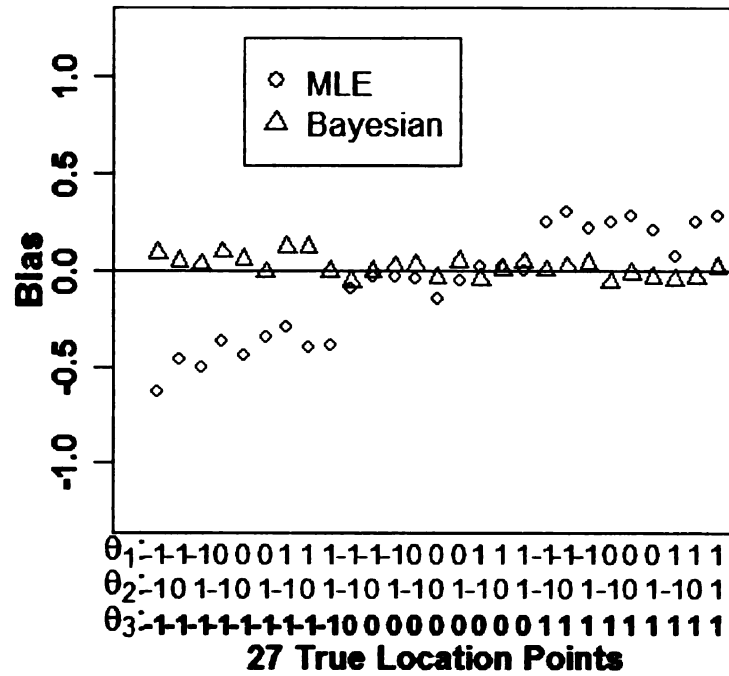
Another research question was which ability estimation methods performed better, maximum likelihood or Bayesian. In order to make the comparison, the combination of maximum likelihood and D-optimality, and the combination Bayesian with maximizing volume decrement in Bayesian with identity matrix as the prior were compared at the test lengths of 20 and 50. The mean biases and RMSEs were compared and the results at the test length of 20 are shown in Figure 4.26 and the results at the test length of 50 are shown in Figure 4.27.

Figure 4.26 Mean biases and RMSEs for comparison of maximum likelihood method and Bayesian method. Test length=20.

**A: Biases**

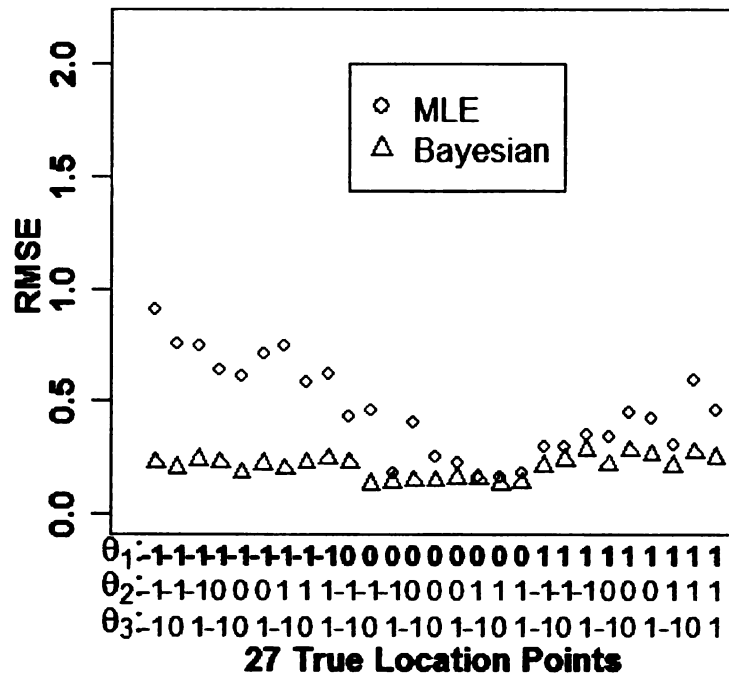


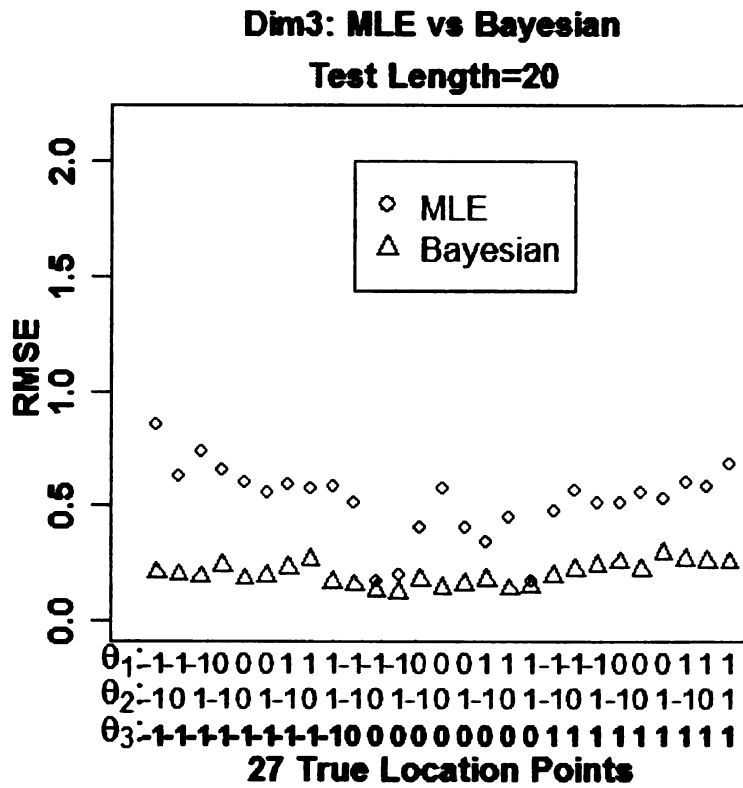
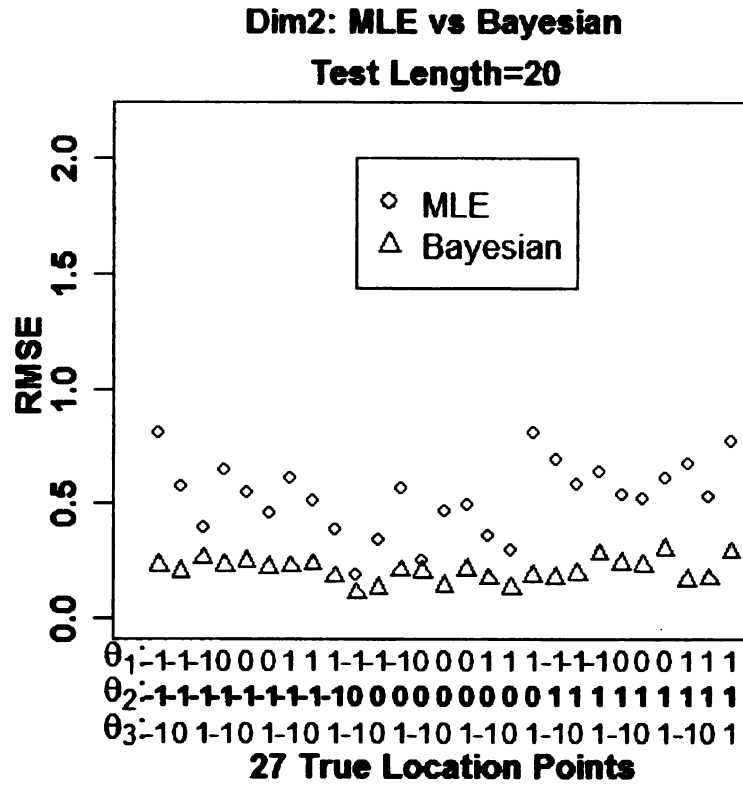
**Dim3: MLE vs Bayesian**  
**Test Length=20**



**B: RMSEs**

**Dim1: MLE vs Bayesian**  
**Test Length=20**



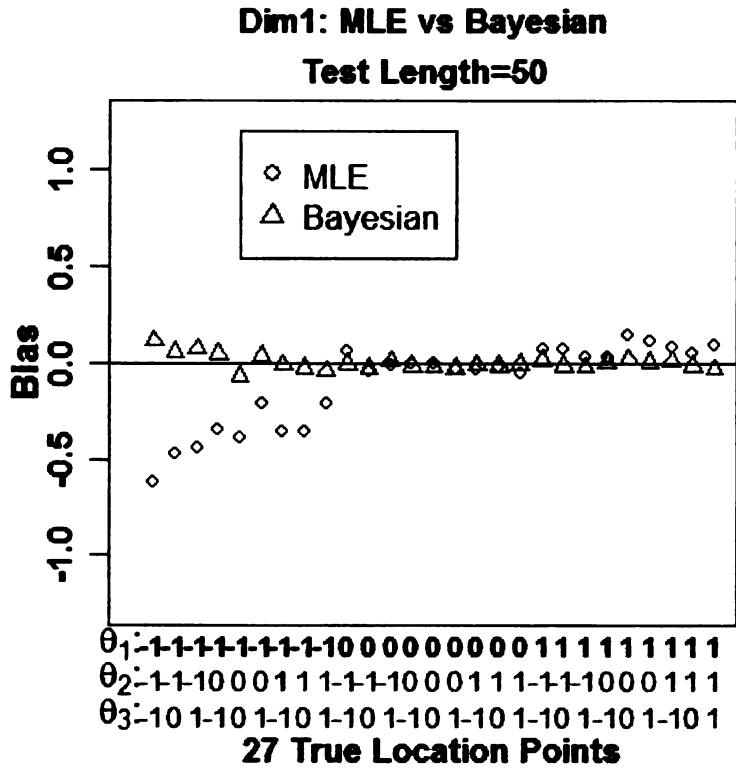


At the test length of 20, it can be seen from Figure 4.26 that the mean biases of maximum likelihood were much larger than those of Bayesian method. The

comparison of RMSEs also confirmed that Bayesian ability estimation method outperformed maximum likelihood at a short test (test length=20). Another interesting thing that could be found in the above graph was that for the true ability values that were negative, the mean biases for the maximum likelihood method were negatively biases while for Bayesian method, they were positive. When the true ability values were positive, the mean biases for the maximum likelihood method were positive and for the Bayesian method, they were negative.

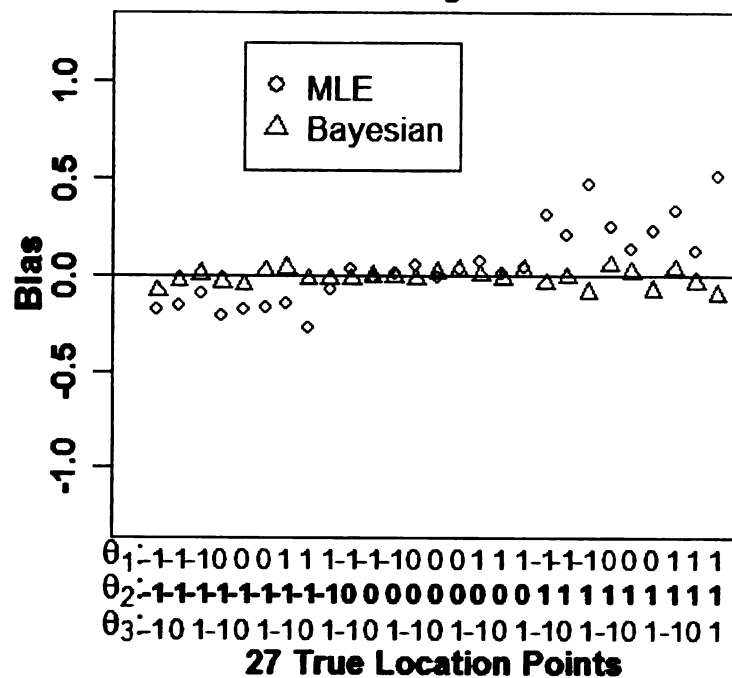
Figure 4.27 Mean biases and RMSEs for comparison of maximum likelihood method and Bayesian method. Test length=50.

**A: Biases**



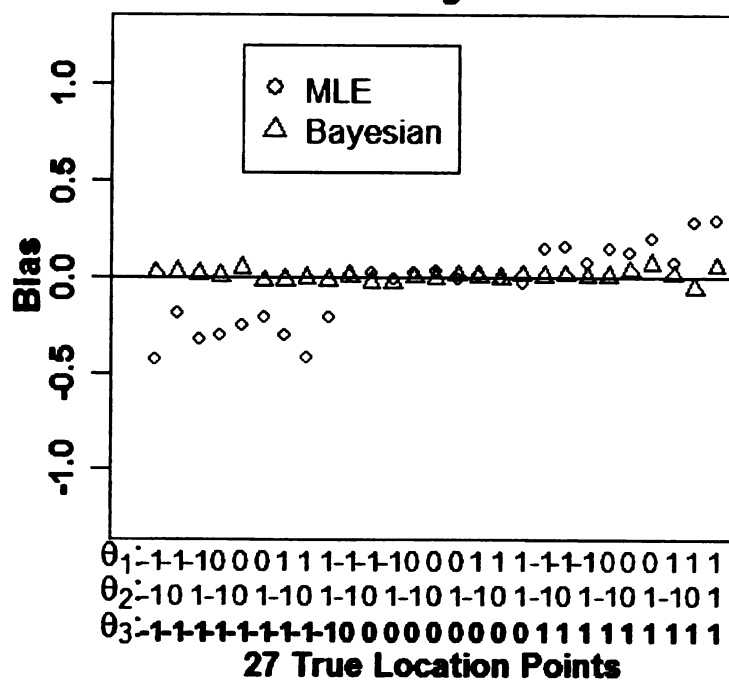
### Dim2: MLE vs Bayesian

Test Length=50

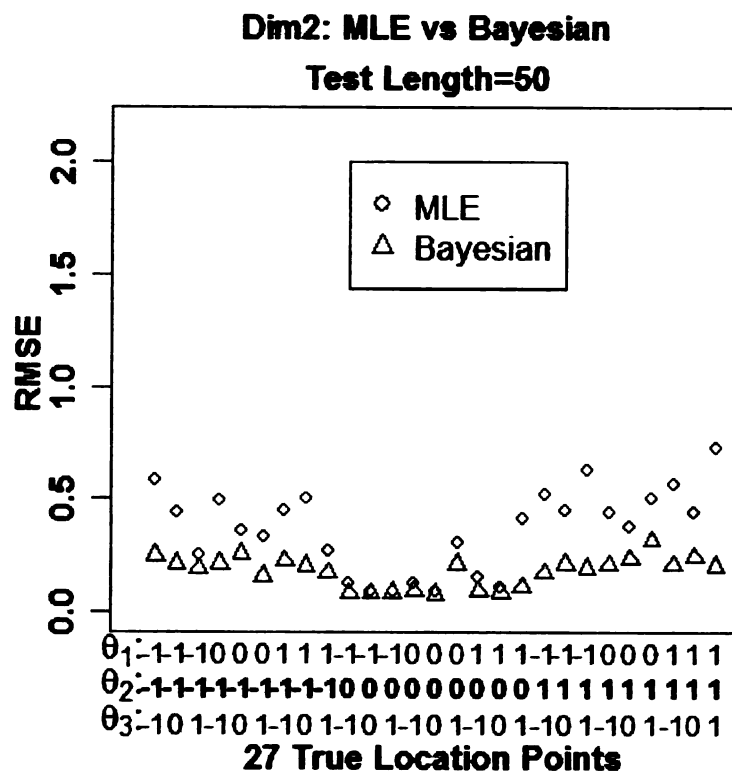
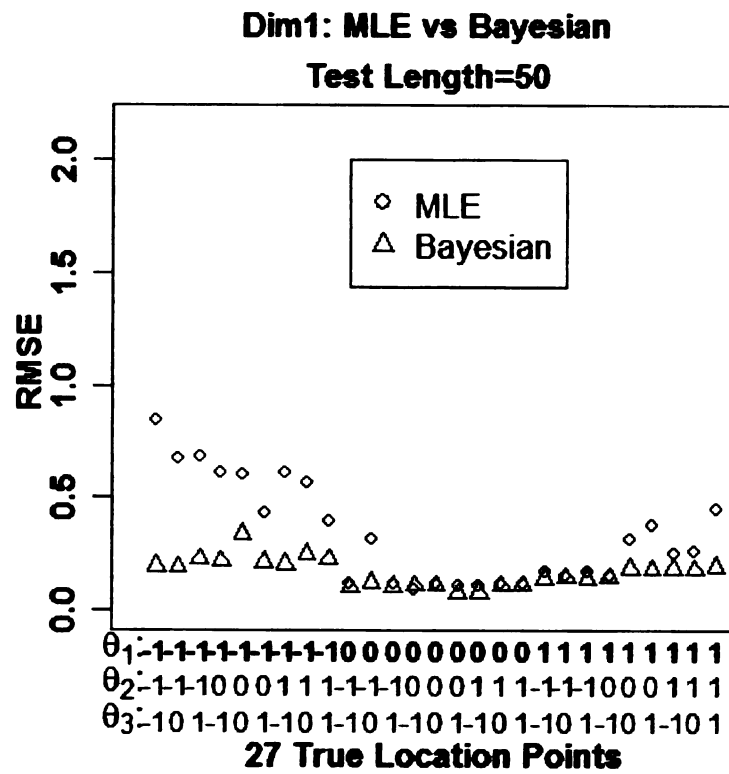


### Dim3: MLE vs Bayesian

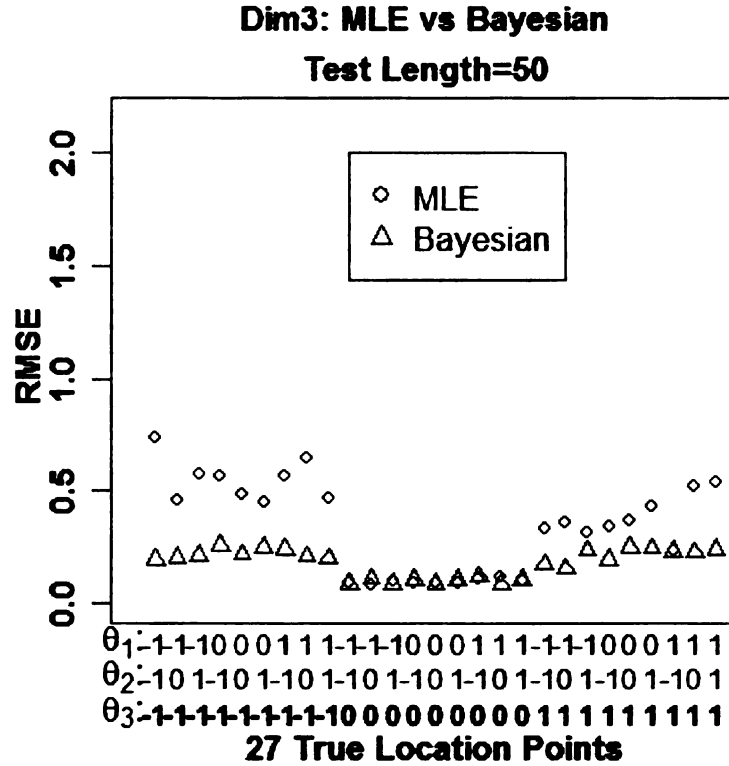
Test Length=50



## B: RMSEs



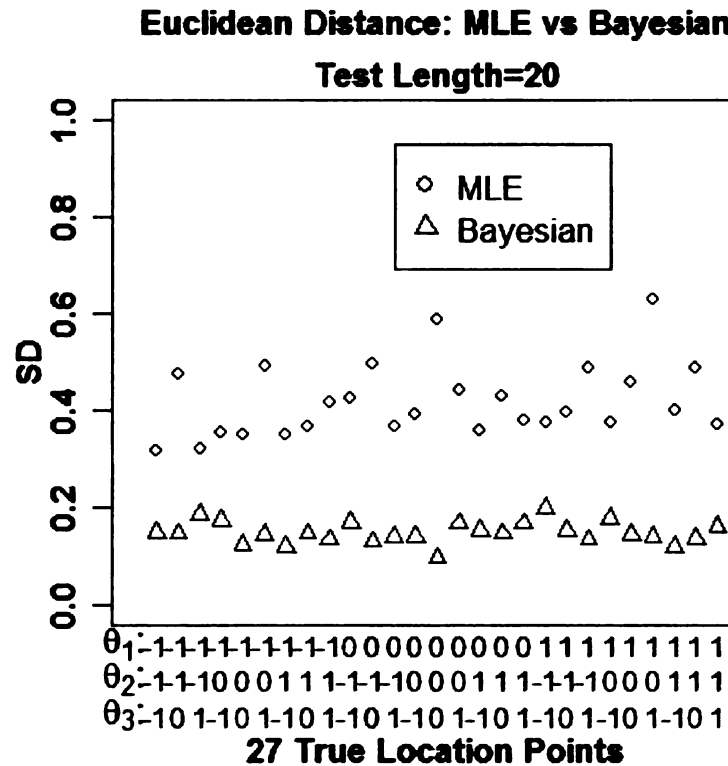
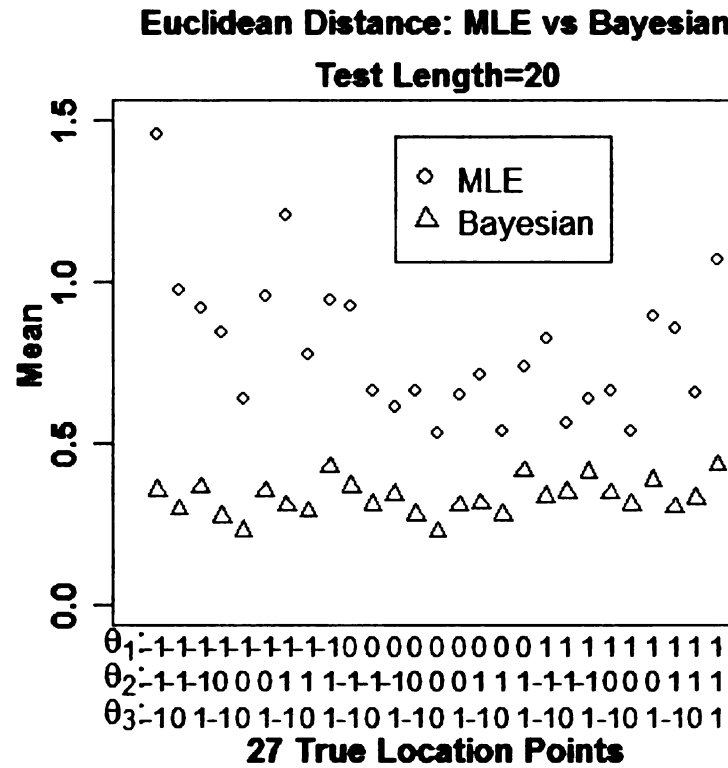




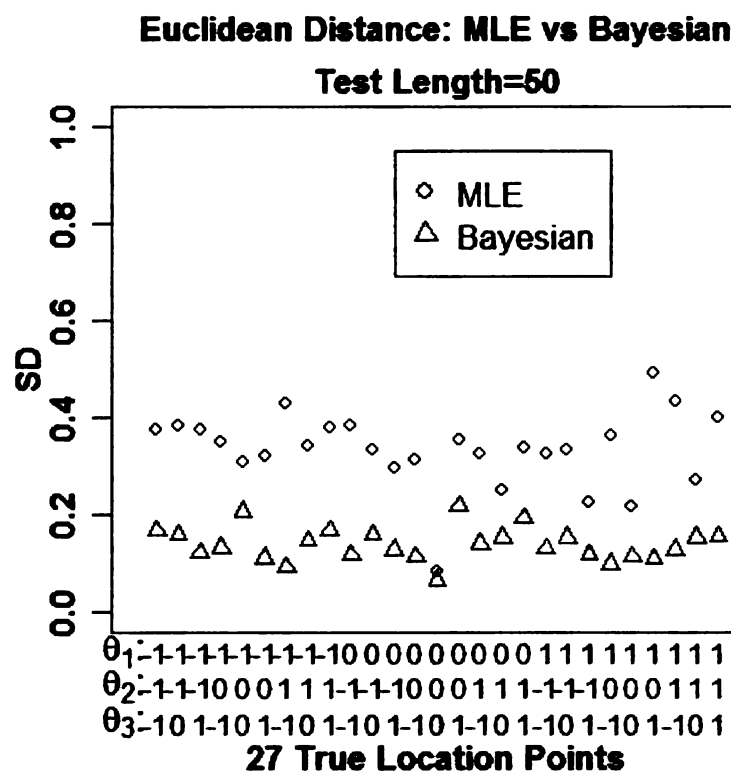
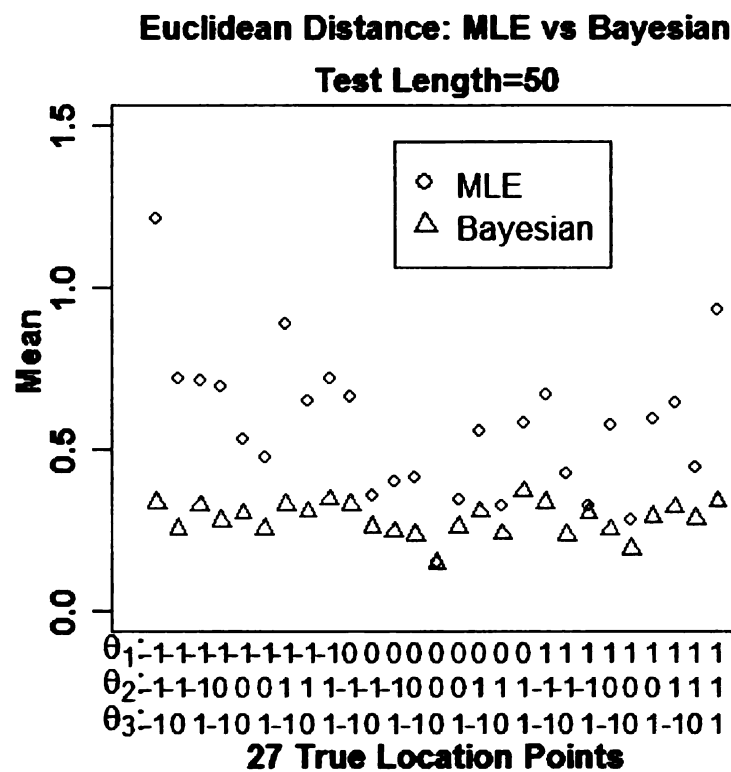
The results from Figure 4.27 showed that at the test length of 50, the mean biases of the maximum likelihood were still larger than those of Bayesian method. RMSEs were also larger for the maximum likelihood method than the Bayesian method. Therefore, even for long test (test length=50), Bayesian ability estimation method still outperformed the maximum likelihood ability estimation method.

The mean biases and RMSEs in Figure 4.26 and Figure 4.27 were measures for the precision of each dimension. The study also used means and standard deviations of the Euclidean distance between the final estimates and true location points as the measure of overall precision. Figure 4.28 shows the results of the means and standard deviations of the Euclidean distance for the comparison of maximum likelihood and Bayesian ability estimation methods at both the test length of 20 and the test length of 50.

Figure 4.28 Means and standard deviations of Euclidean distance, comparison of maximum likelihood method and Bayesian method. Test length=20 and Test length=50.  
**Test length=20**



Test length=50

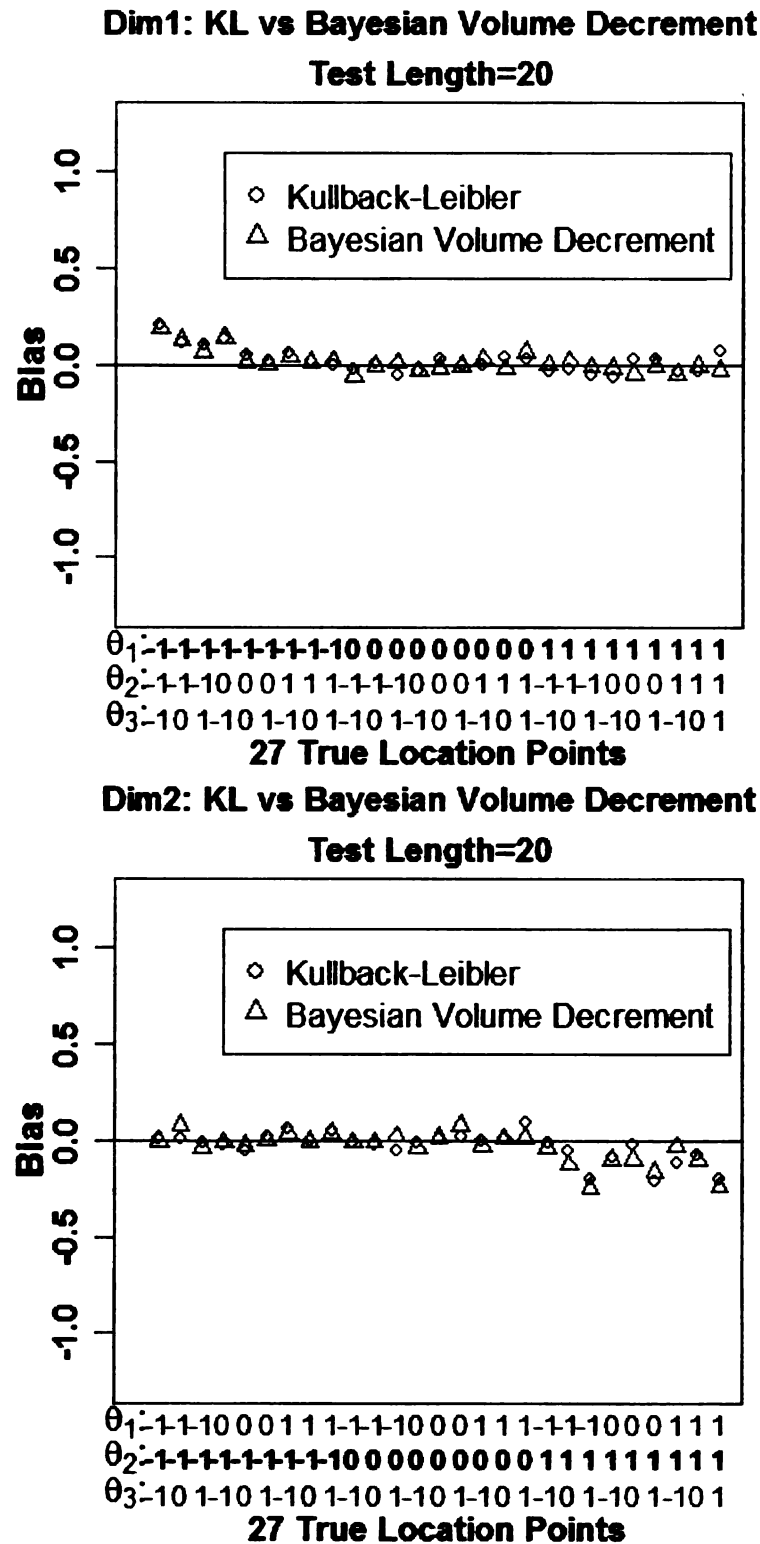


At the test length of 20, the mean Euclidean distances between the final estimates and true ability points were much larger for maximum likelihood method than for Bayesian ability estimation method. The standard deviations of the Euclidean distance were also much larger for the maximum likelihood estimation method. So over all three dimensions, the estimation precision for the maximum likelihood method was not good. Bayesian estimation method outperformed it in a large degree at the short test length (test length=20). When the test length increase to 50, the accuracy of both methods became better and the gap of the precision between the two methods became smaller. However, from the results of means and standard deviations of Euclidean distance, the overall estimation accuracy was still better for Bayesian than maximum likelihood estimation method. The results also showed that the final estimates for Bayesian method were also more stable than maximum likelihood method.

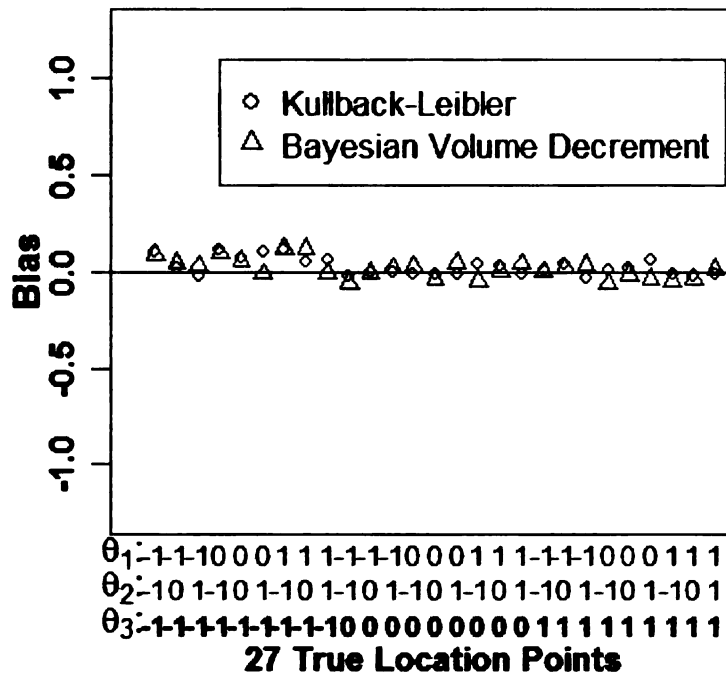
The last research question was to compare the performance of volume decrement in Bayesian with Fisher's information and the performance of maximizing Kullback-Leibler information. In order to make such comparison, both methods used the prior with mean  $\mathbf{0}$ , and identity matrix as the variance covariance matrix. The comparison was conditioning on test lengths. The mean biases and RMSEs for the final estimates of each dimension were calculated and Figure 4.29 shows the comparison at the test length of 20 and Figure 4.30 shows the comparison of the two methods at the test length of 50.

Figure 4.29 Mean biases and RMSEs of the comparison of Kullback-Leibler and Volume decrement in Bayesian. Variance covariance of priors is identity matrix. Test length=20.

A: Biases

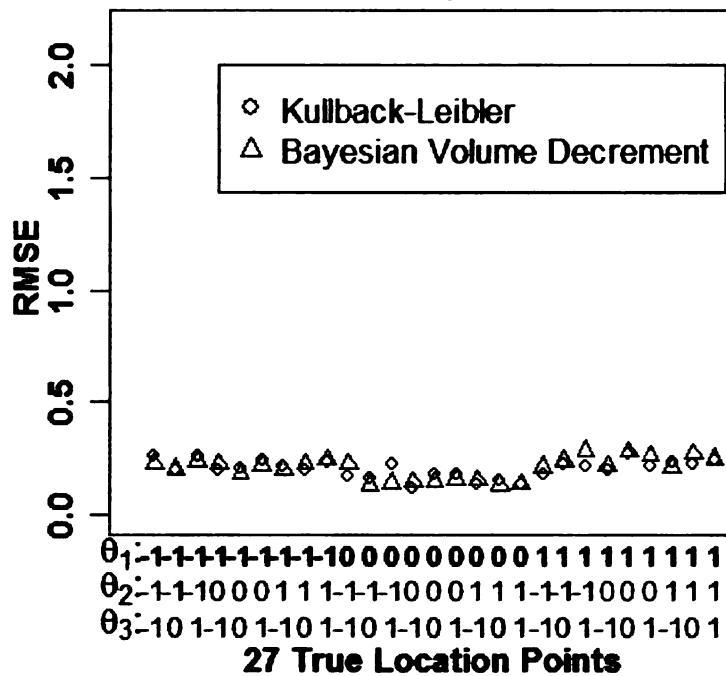


**Dim3: KL vs Bayesian Volume Decrement**  
**Test Length=20**



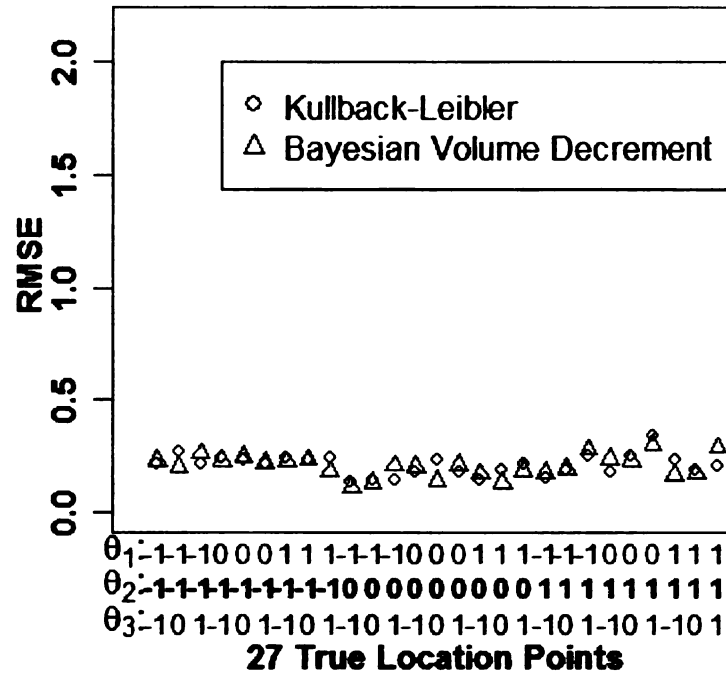
**B: RMSEs**

**Dim1: KL vs Bayesian Volume Decrement**  
**Test Length=20**



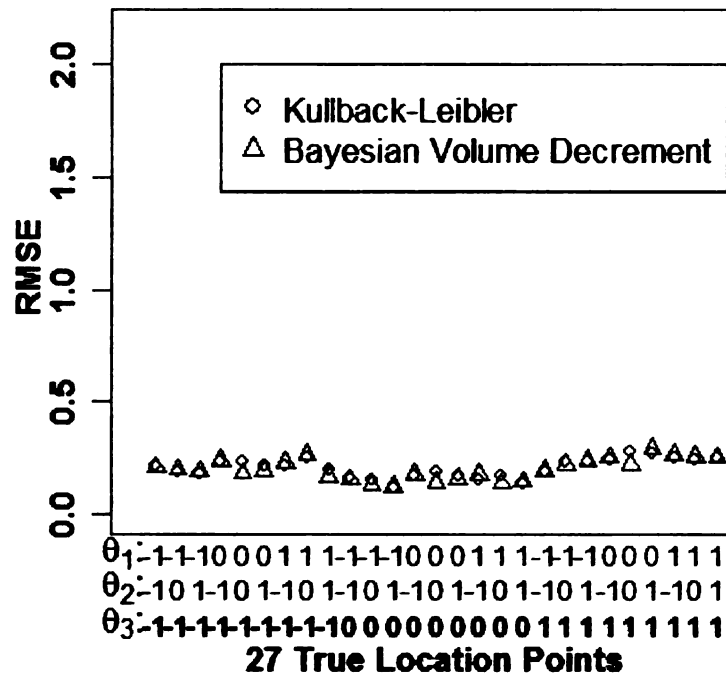
### Dim2: KL vs Bayesian Volume Decrement

Test Length=20



### Dim3: KL vs Bayesian Volume Decrement

Test Length=20

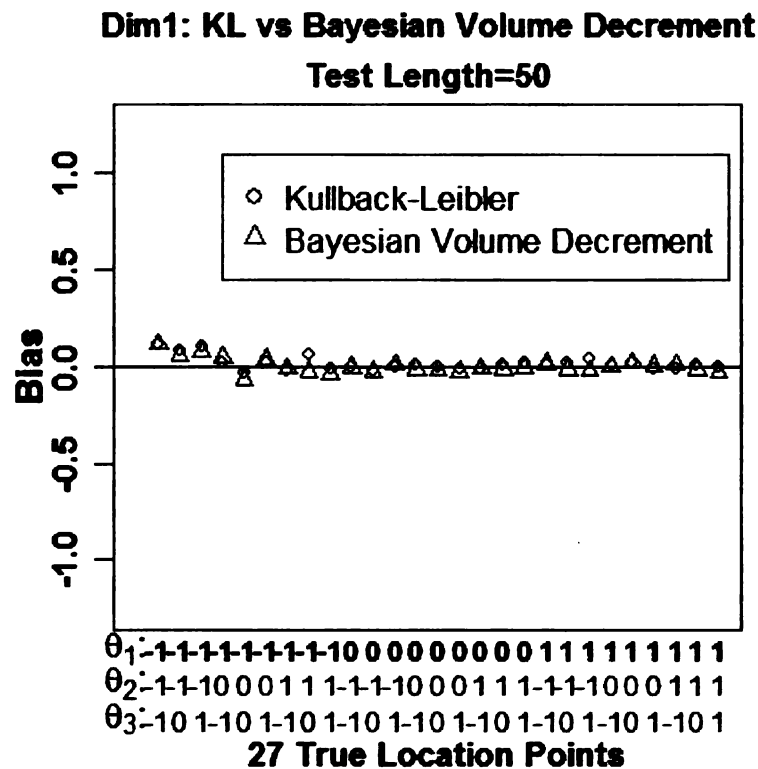


At the test length of 20, the mean biases were small for both the volume decrement in Bayesian with Fisher's information method and maximizing Kullback-Leibler information item selection method. Those two methods produced accurate final

estimates at the test length of 20. RMSEs were also small for both methods. So the estimates of both methods were already stable at the test length of 20. From both the mean biases and RMSEs, it can be seen that Kullback-Leibler information and Bayesian method using Fisher's information were comparable at the test length of 20.

Figure 4.30 Mean biases and RMSEs of the comparison of Kullback-Leibler and Volume decrement in Bayesian. Variance covariance of priors is identity matrix. Test length=50.

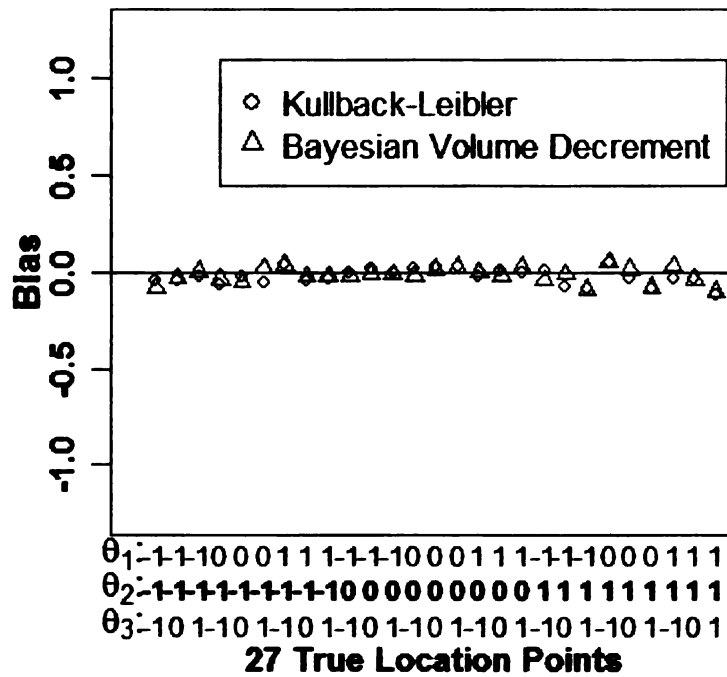
#### A: Biases





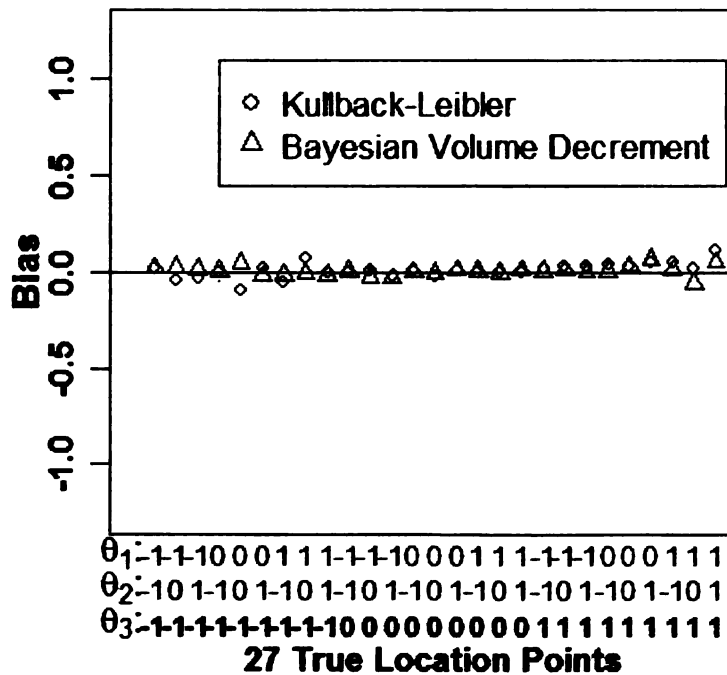
## Dim2: KL vs Bayesian Volume Decrement

Test Length=50



## Dim3: KL vs Bayesian Volume Decrement

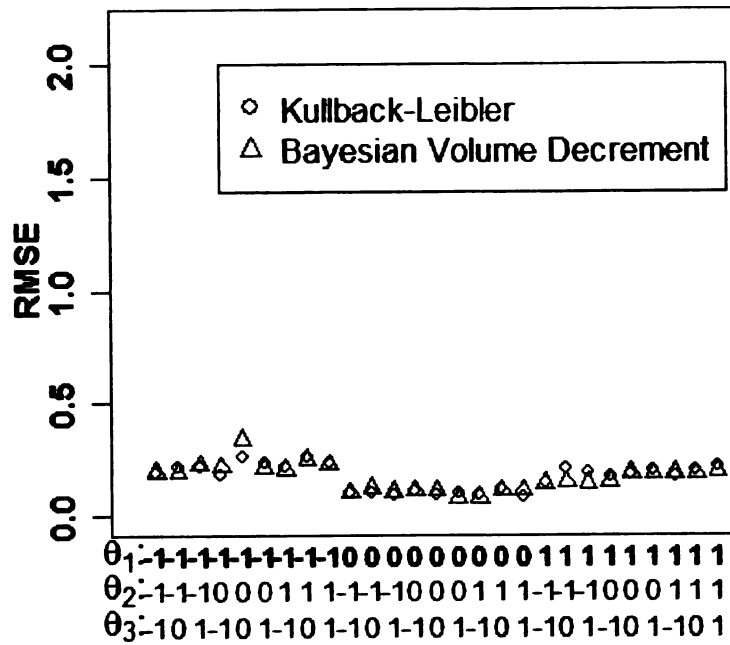
Test Length=50



B: RMSEs

### Dim1: KL vs Bayesian Volume Decrement

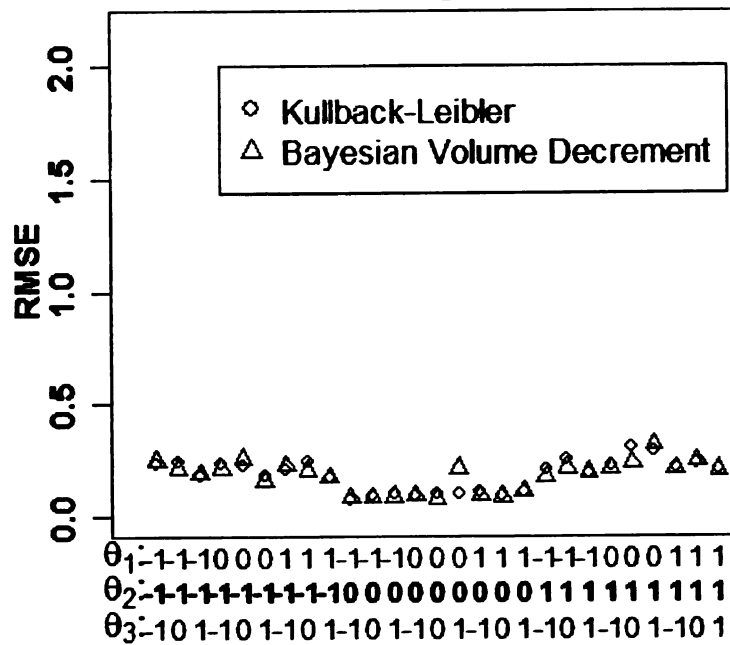
Test Length=50



27 True Location Points

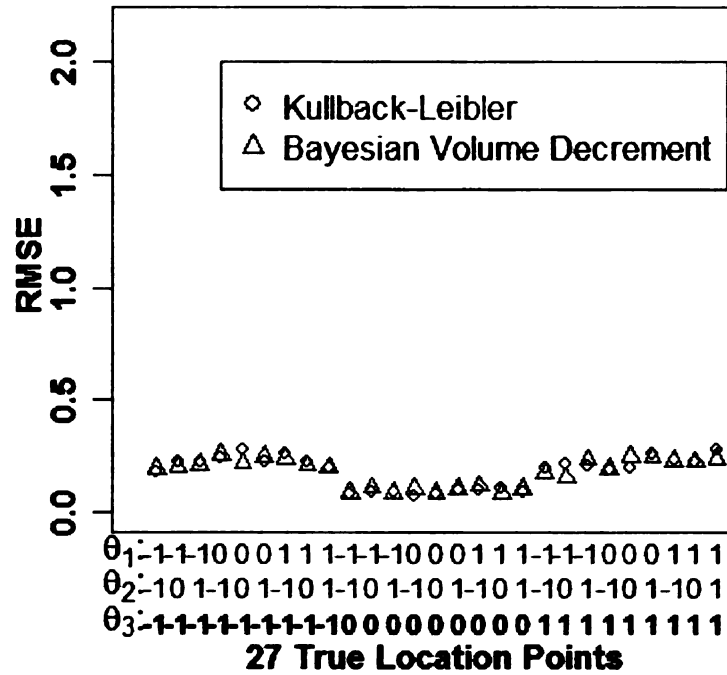
### Dim2: KL vs Bayesian Volume Decrement

Test Length=50



27 True Location Points

**Dim3: KL vs Bayesian Volume Decrement**  
**Test Length=50**

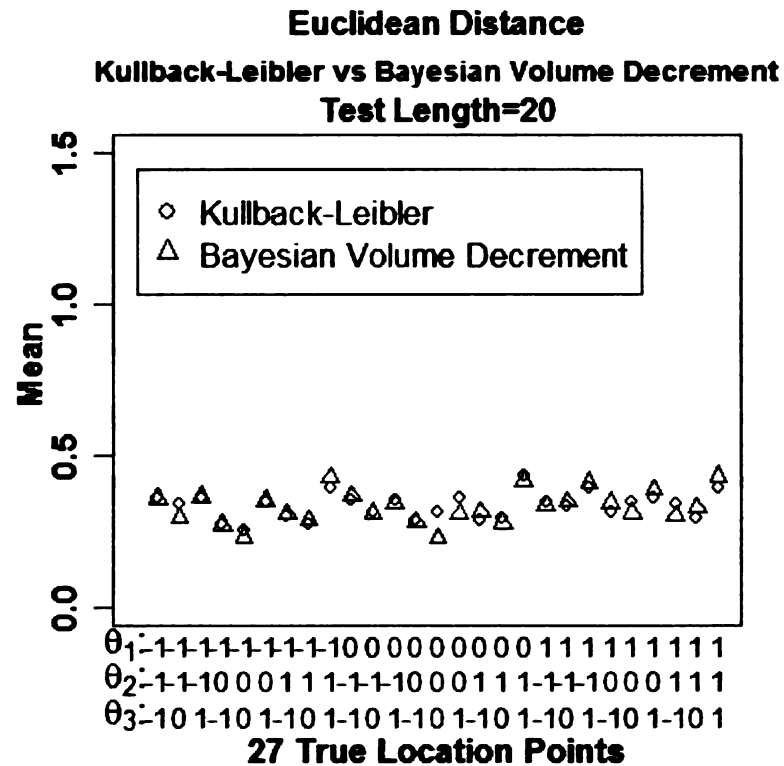


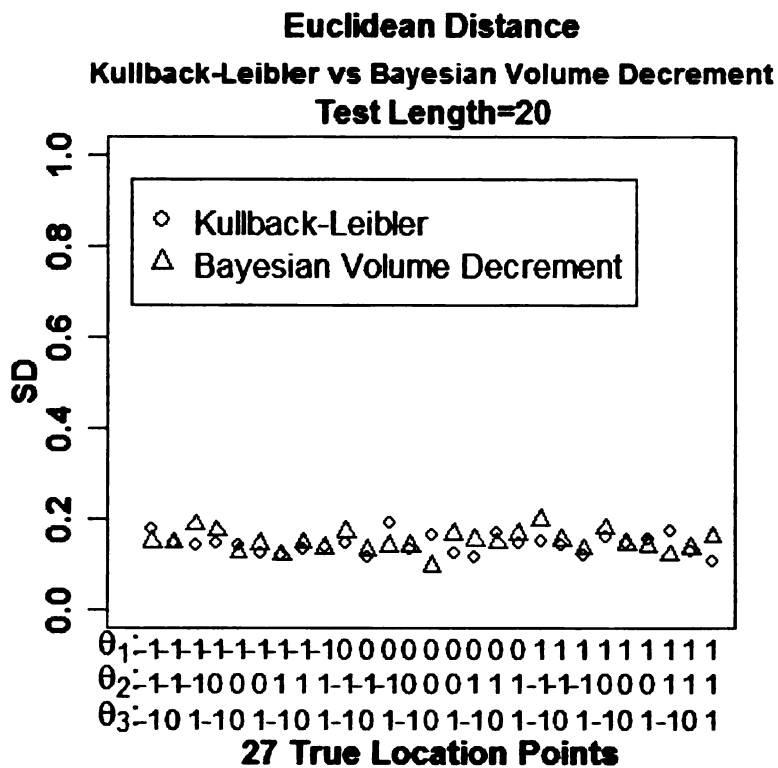
When the test length increased to 50, from the results of mean biases and RMSEs, the precision of the two methods: volume decrement in Bayesian using Fisher's information and maximizing Kullback-Leibler information, were good and those two methods were comparable in terms of estimation accuracy and stability.

The overall measure, Euclidean distance was also calculated and shown in Figure 4.31. From the comparison of means and standard deviations of the Euclidean distance, it can be seen that the performance of Kullback-Leibler information of the volume decrement in Bayesian with Fisher's information was comparable both at the test lengths of 20 and 50.

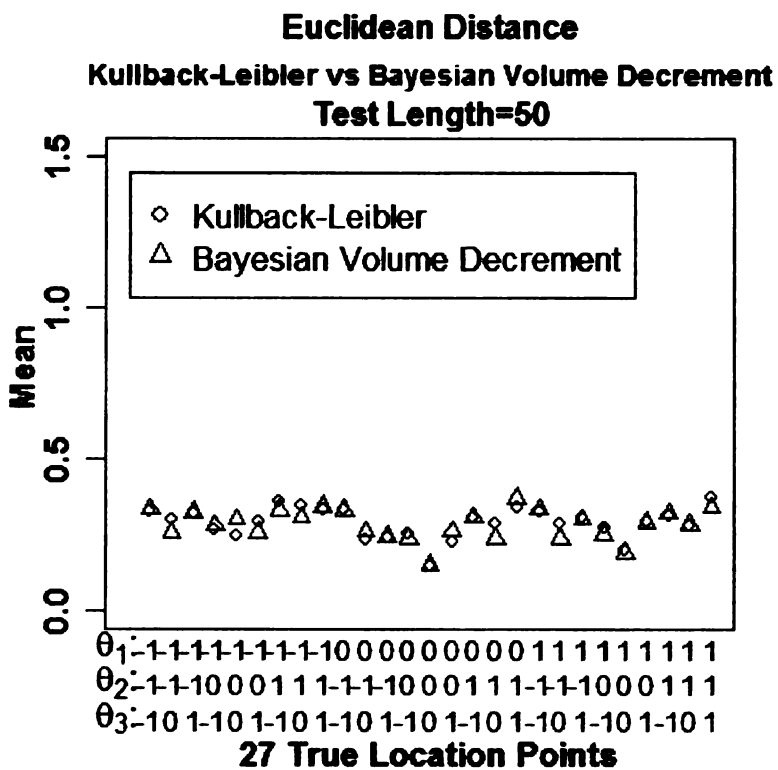
Figure 4.31 Means and standard deviations of Euclidean distance, comparison of Kullback-Leibler information and volume decrement in Bayesian with Fisher's information.

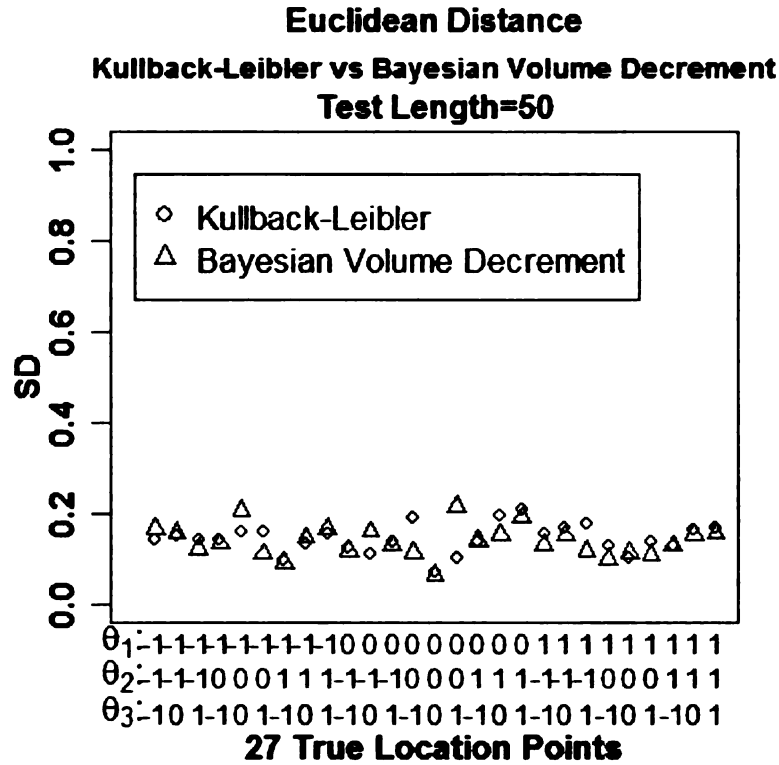
Test Length=20





Test Length=50





As stated in the research questions, even though computation time was not as important as before, it was so interesting to calculate the computation time for each method to have a balance between estimation precision and computation time. For each combination of ability estimation and item selection method, the computation time was calculated at the examinee level using second as the unit for time. The times shown in Table 4.2 are how many seconds were needed to administer test to one examinee. In the simulation study, the examinees' response time was set as 0. Except for the item selection method using Kullback-Leibler information, the computation time was around 2 seconds for the test length of 20 and around 9 seconds for the test length of 50. When Kullback-Leibler information was used, the computation increased about 10 times for both test lengths.

Table 4.2 Computation time for each examinee (Unit: second)

<b>Ability Estimation Method</b>	<b>Item Selection Method</b>	<b>Prior</b>	<b>Test Length=20</b>	<b>Test Length=50</b>
MLE	D-optimality	N/A	2.765	8.696
	A-optimality	N/A	1.889	6.580
Bayesian	Bayesian Volume Decrement	Identity	2.442	9.429
		diag(9)	2.460	9.490
		True	2.441	9.418
Bayesian	Kullback-Leibler	Identity	20.694	99.624

Note: All the integrating calculations were programmed in FORTRAN and all other programming was done in R. All computation time was calculated on a PC with a 3.0 GHz AMD Athlon 64 Dual Core processor and 2.00 GB RAM.

## **CHAPTER 5. CONCLUSION, RECOMMENDATION, AND FUTURE RESEARCH DIRECTION**

This study did a comprehensive comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. Two ability estimation methods included maximum likelihood estimation and Bayesian estimation method. The item selection methods can be divided into three categories, item selection methods associated with maximum likelihood, item selection with Bayesian with Fisher's information, and item selection method with Kullback-Leibler information. D-optimality (maximizing the determinant of Fisher's information) and A-optimality (minimizing the trace of the inverse of Fisher's information) were included for item selection methods that were associated with maximum likelihood method. Three priors of Bayesian method with maximizing the volume decrement with Fisher's information were selected to measure the impact of the priors in Bayesian. Different test lengths were selected (test length=20 and test length=50). In total, 11 combinations of ability estimation and item selection methods were simulated and compared in the study.

The initial estimate for all examinees was 0 and the mean of all priors was 0. This led to one trend for all biases. For Bayesian estimation, all biases were "inward bias". Estimators of positive values of  $\theta_i$  ( $i=1, 2, 3$ ) were negatively biased and the estimators of negative values were positively biased. In the opposite, when maximum likelihood estimation was used, the biases were "outward bias". Estimators of positive



values of  $\theta$  were positively biases and the estimators of negative values were negatively biases.

From the results of mean biases and RMSEs of final estimates for each dimension, and means and standard deviations of Euclidean distance, it can be seen that maximum likelihood ability estimation method did have non-convergence problems at the beginning of the test and it affected the estimation precision of the method. Plots of successive progress of updated estimates also supported this conclusion. Therefore, it was recommended that a longer test should be used when maximum likelihood ability estimation method was applied.

When Bayesian ability estimation method was applied, for all the combinations with the item selection methods, the comparison of test lengths of 20 and 50 showed that the precision difference was small. The final estimates were already stable and accurate. Therefore, if Bayesian ability estimation method was used, a short test (test length=20 or more) could be used.

The comparison of maximum likelihood and Bayesian ability estimation methods showed that Bayesian ability estimation method outperformed maximum likelihood method, especially for short test length. In general, Bayesian ability estimation was recommended as the ability estimation method. But with Bayesian, the test designers need to select the priors, which might not be as objective as the maximum likelihood method. So all factors need to be taken into considerations when choosing the ability estimation method. In theory, if the test length is very long (estimates for both methods converged and were stable), the estimation of the two methods should be comparable.

The study also evaluated the impact of priors when Bayesian method was used. Three priors: a strong prior, a relative weak prior, and a true prior calculated from the population were compared. When the true ability value on the dimension was 0, all three priors were comparable and the mean biases were small. When the true ability value was negative or positive, and opposite to the research hypothesis, the true prior did not perform as well as the other two priors. It was because mean of multinormal distributions for all priors was 0, the priors pulled the estimates towards the mean 0. With the true prior, the force of pulling was the strongest. So the biases were the biggest. But for all three priors and conditioning on both short and long test lengths, the performance of Bayesian estimation was good and the final estimates were stable and accurate. More studies need to be done on how to utilize the collateral information for priors to assist a better estimation with Bayesian method. Instead of the population prior, as was used in the study, an individual prior may be used or hierarchical models could be tested to see if that can lead to better final estimation.

All the priors used in study had the same values on the diagonal respectively. There was more regular compared to cases like variances were quite different and correlations more varied. More studies need to be done to investigate such priors to assess the impact of items selection and ability estimation methods under such conditions.

The Kullback-Leiber information was relatively new compared to the Fisher's information in multidimensional adaptive testing. The comparison of the two in the study showed that the performance of the two was comparable for both the short and long test lengths. However, the Kullback-Leiber information did cost much longer

computation time than other methods. And if computation time is one of the concerns for one test application, then volume decrement of Bayesian with Fisher's information was recommended rather than the Kullback-Leibler information. Also, the cases studied here were three dimensional. With the increase of the dimensions, it was expected that the computation time would also increase. Therefore, extra care should be taken if higher dimensions were studied.

Multidimensional computerized adaptive testing is a relatively new area of research. This study was a comparison of ability estimation and item selection methods to make recommendation and guidance in terms of what ability estimation and item selection methods to use when designing a multidimensional computerized adaptive testing. The conclusions of this study were limited to the conditions of item pool, test lengths and priors used. Also, during the work of this study, more and more ability estimation and item selection methods are being developed. So in future, more research needs to be done to compare the new methods with all the methods in this study. There are also other issues in multidimensional CAT, such as how to select the first item and how to end the test, which needs more research on.

## REFERENCES

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Lord, F. M., & Novick, M. R. (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Addison-Wesley, Reading, MA.
- ✓ Bloxom, B. M., & Vale, C. D. (1987). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta*. Paper presented at the meeting of the Psychometric Society, Montreal.
- Bock, R. D. & Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement* 6, 431-444.
- ✓ Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Chang, H.-H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Hetter, R. R., & Sympon, J. B. (1997). Item-exposure in CAT-ASVAB. In W. A. Sands, J. R. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.
- Kim, J. P. (2001). *Proximity measures and cluster analyses in multidimensional item response theory*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.
- Lee, Y.-H, Ip, E. H., & Fuh, C.-D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, 68, 215-232.
- ✓ Li, T. (2006). *The effect of dimensionality on vertical scaling*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404.
- Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York, NY: Springer-Verlag.
- McDonald, R. P. (1997) Normal –ogive multidimensional model. In W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258-270). New York: Springer.

- McDonald, R. P. (1999). *Test Theory: a unified treatment*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- o Mulder, J. & van der Linden, W. J. (2008). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*. Tentatively accepted for publication.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous items response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.
- Reckase, M. D. (2009?) Multidimensional Item Response Theory??
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.
- o Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-73). Boston: Kluwer.
- Sympon, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27<sup>th</sup> annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
- Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait*. Unpublished doctoral dissertation, Columbia University, New York City, NY.
- o van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston: Kluwer.
- van der Linden, W. J., & Glas, C. A. W (Eds) (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer.
- o van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29, 273-291.

- ^ van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, 32, 398-418.
- ✓ Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67, 575-588.
- Wainer, Howard (2000). *Computerized adaptive testing: a primer*, 2<sup>nd</sup> Ed. Mahwah, NJ: Lawrence Erlbaum Association.



MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03063 2677