141
319
THS

THESIS
2
2009

LIBRARY
Michigan State
University

This is to certify that the
dissertation entitled

USING A PROJECTION METHOD TO ESTIMATE
SUBSCORES FROM TESTS WITH MULTIDIMENSIONAL
STRUCTURES

presented by

YU FANG

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Measurement and Quantitative Methods

Mark D. Reckase
Major Professor's Signature

12/3/08
Date

MSU is an Affirmative Action/Equal Opportunity Employer

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
030513 [illegible]		

5/08 K:/Proj/Acc&Pres/CIRC/DateDue.indd

USING A PROJECTION METHOD TO ESTIMATE SUBSCORES FROM TESTS WITH MULTIDIMENSIONAL STRUCTURES

By

Yu Fang

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2008

ABSTRACT

USING A PROJECTION METHOD TO ESTIMATE SUBSCORES FROM TESTS WITH MULTIDIMENSIONAL STRUCTURES

By

Yu Fang

A long time problem in factor analysis is the rotational indeterminacy of solutions. This same problem also exists for multidimensional item response theory (MIRT) model analyses. Commonly, the widely used mathematical criteria from factor analysis, such as the Varimax and Promax methods, are adopted for the MIRT model to match the item discrimination matrix to the simple structure for better interpretation of item characteristics. However, no substantial steps have been taken to provide a better explanation and solution to the estimation of correlated proficiency estimates from MIRT calibration procedures. These steps are often ignored in some popular software which provides only the uncorrelated proficiency estimates. This study uses the MIRT item and person parameter estimates when the proficiencies are assumed uncorrelated, and projects the uncorrelated proficiency estimates onto the most discriminating directions of item clusters to get the subscore estimates. This solution provides the correlated construct scores related to different item clusters, explains the relationship and difference between the model dimensionality and the number of item clusters, and is useful for subscore reporting.

Dedicated to my beloved wife: Yang Lu

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my dissertation advisor, Dr. Mark Reckase, for his continued wisdom and advice through my dissertation writing. The constructive conversation with him greatly inspires my interest in the measurement field and in MIRT. He is always readily available for questions, and without his timely and insightful feedback, I would not have understood the essence of MIRT or been so efficient with my analysis.

I am especially grateful to Dr. Kimberly Maier, my academic advisor and chair of my guidance committee, who always trusts me, encourages me, and guides me through my graduate study.

Also special thanks go out to Dr. Ken Frank, who teaches me how one can enjoy the research, and guides me to think of questions from perspectives of social networks and sensitivity analysis.

In addition, I appreciate the support from Dr. Connie Page, who helps me with funding opportunities for most of my graduate study. The statistical consulting work really enhances my skill in explaining complicated things in a simple way. I am very fortunate to work with her.

Thanks again to these four kind and knowledgable committee members for their suggestions and encouragements towards my dissertation and program study, as well as my career life. They are always teachers and friends to me.

I also gratefully acknowledge Dr. Joseph Martineau and Steve Viger, both of whom are working at the Michigan Department of Education, for their generous help with the MEAP data for my analysis.

Finally, I owe a huge debt of gratitude to my parents, my wife and my friends, for their unfailing and patient support in all aspects. Without them, my life would not be so interesting and rewarding.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

1 Introduction to MIRT 1
- 1.1 Dimensionality 1
- 1.2 Importance of Subscores 4
- 1.3 Compensatory MIRT model 6
- 1.4 Indeterminacies in MIRT Models 10
- 1.5 Solution to Rotational Indeterminacy in Factor Analysis 12
- 1.6 Interdependency between Proficiency Correlation and Item Discrimination 13

2 Construct Estimation Using Projection 18
- 2.1 Transformation in the Orthogonal Coordinate System 18
- 2.2 Construct Estimation 20
  - 2.2.1 Step 1: Calibration 20
  - 2.2.2 Step 2: Cluster Analysis 24
  - 2.2.3 Step 3: Reference Composite 26
  - 2.2.4 Step 4: Projection 27
- 2.3 Dimensionality and Number of Clusters 28
- 2.4 Characteristics of Construct Estimates 29

3 Simulation Study 32
- 3.1 Methods 32
  - 3.1.1 Parameter Simulation 32
  - 3.1.2 Simulation Design 33
  - 3.1.3 Calibration and Projection 34
  - 3.1.4 Evaluation Criteria 34
- 3.2 Results 36
  - 3.2.1 Parameter Estimation 36
  - 3.2.2 Item and Person Parameter Recovery 41
  - 3.2.3 Coordinate Axes Recovery 48
  - 3.2.4 Item Grouping 49
  - 3.2.5 Stability of Construct Estimates 52
  - 3.2.6 Relationship with NC Subscores and Unidimensional Estimates 54
  - 3.2.7 Accuracy of Construct Estimates 55
  - 3.2.8 Correlation Recovery 62

4 Real Data Applications **64**

4.1 Test Information and Analysis Procedure 64

4.2 Dimensionality and Cluster Detection 69

4.3 MIRT Calibration and Subscore Reporting 79

5 Conclusions, Implications, Limitations and Future Research **86**

5.1 Conclusions and Implications 86

5.2 Limitations and Future Research 91

REFERENCES **96**

LIST OF TABLES

3.1 MIRT Item Parameters for Design 1 37

3.2 MIRT Item Parameters for Design 2 38

3.3 Descriptive Statistics of MIRT Person Parameters 40

3.4 Recovery of Item Parameters 42

3.5 Recovery of Person Parameters 45

3.6 True and Estimated Proficiency Correlation Matrix 47

3.7 Reference Composite Vectors from Parameters and Replication 1 Estimates 48

3.8 Summary of Standard Deviations of Raw Proficiency and Construct Estimates 54

3.9 Average Correlation Among Construct Estimates, Unidimensional Estimates and NC Subscores 55

3.10 Construct Recovery from Unidimensional Estimates and Construct Estimates 58

3.11 Hypothesis Testing on the Difference of Correlations for Construct Recovery by Construct Estimates and Unidimensional Estimates 59

3.12 Correlation Recovery by Unidimensional Estimates, the Promax Method and Construct Estimates 63

4.1 Strands and Domains for Michigan Mathematics Content Expectation 65

4.2 Item Content of the Fall 2007 Grade 7 MEAP Mathematics Test . . . 66

4.3 Strand and Domain Percentage in the Test 66

4.4 Summary Statistics for the Population Data 67

4.5 Results from the DIMTEST Software 70

4.6 Eigenvalues for the Population Data and Ten Random Datasets . . . 71

4.7 Model Fit Indices from the NOHARM Software 73

4.8 Item Cluster Content of the Fall 2007 Grade 7 MEAP Mathematics Test 78

4.9 Varimax Transformed NOHARM Item Estimates for the Population Data 81

4.10 Summary Statistics for Item Parameter Estimates 82

4.11 Reference Composite Vectors from the Population Data Estimation . 82

4.12 Eigenvalues of $A_l'A_l$ Matrix 82

4.13 Correlations among NC Subscores and Construct Estimates 82

LIST OF FIGURES

1.1 Representation of the Characteristics of 40 Items in a Two-Dimensional Space 9

1.2 Representation of the Characteristics of 40 Items in a Two-Dimensional Space when the Proficiency Correlation is 0.0, 0.2, 0.4 and 0.6 15

3.1 Plots of Item Vectors and Reference Composite Vectors 39

3.2 Plots of Item Parameters versus Bias 43

3.3 Plots of Proficiency Parameters versus Bias 46

3.4 Plots of Reference Composite Vectors from Parameters and Replication 1 Estimates 50

3.5 Dendrogram for Replication 1 in Design 1 51

3.6 Plots of Construct Parameters versus Estimated Standard Deviations 53

3.7 Plots of Average Construct Estimates versus Average Unidimensional Estimates across all Replications 56

3.8 Plots of Average Construct Estimates versus Average NC Subscores across all Replications 57

3.9 Plots of Recovery Bias by Construct Estimates and Unidimensional Estimates in Design 1 60

3.10 Plots of Recovery Bias by Construct Estimates and Unidimensional Estimates in Design 2 61

4.1 Plot of Eigenvalues for the Population Data and Ten Random Datasets 71

4.2 Dendrogram for the Population Data Calibrated with 8 Dimensions . 72

4.3 Dendrogram for the Population Data Calibrated with 4 Dimensions . 75

4.4 Dendrogram for the 5000 Sample Data Calibrated with 4 Dimensions 76

4.5 Dendrogram for the 10000 Sample Data Calibrated with 4 Dimensions 77

4.6 Plots of NC Subscores versus Construct Estimates for the Population Data Calibrated with 4 Dimensions 84

4.7 Plots of NC Subscores versus Raw Proficiency Estimates for the Population Data Calibratcd with 4 Dimensions 85

CHAPTER 1

Introduction to MIRT

1.1 Dimensionality

The unidimensional IRT model has become increasingly important, generally accepted and widely used in the educational measurement field, especially in test construction, equating and person proficiency estimation. The underlying assumption which most model-users seldom cast any doubt on is that all the test items measure the unidimensional or vaguely general proficiency of a person, which is intended to be measured by test developers. For this reason, the person proficiency and item difficulty can be put on the same continuum, and the probability of correctly answering each item is actually determined by the difference between the values of these two indices.

However, in one test, there can be items measuring the knowledge of different subareas. For example, there can be algebra and geometry items in one mathematics test. Furthermore, Reckase (1985) pointed out that, for most test items, more than one hypothetical construct is needed for people to get the answer correct. His well-known argument to support this is that verbal, mathematical computation and problem solving skills are indispensable to get the mathematical story-problem item correct. Hence, if the person proficiency is supposed to be defined for these items, it should be related to the proficiencies on these three skills within that person, maybe a weighted composite of them. The tests including any of the above items are called multidimensional tests, and the multidimensionality in these two cases are later distinguished as between-item multidimensionality and within-item multidimensionality (W.-C. Wang et al., 1997).

Generally speaking, for any testing situation, people are assumed to have one set of proficiencies, denoted as $\boldsymbol{\theta}_{person}$ vector, where correlation can be allowed between any two vector elements. The whole test requires another set of proficiencies from people, denoted as $\boldsymbol{\theta}_{test}$ vector, and each item is most discriminating along one proficiency direction, or the direction of a composite of several proficiencies. Since the responses are interactions between person proficiencies and item characteristics, the proficiency set required for the responses, $\boldsymbol{\theta}_{person,test}$ vector, should be the intersection of $\boldsymbol{\theta}_{person}$ and $\boldsymbol{\theta}_{test}$. That is to say, the responses are determined by the proficiencies that are not only owned by persons but also measured by items.

Although $\boldsymbol{\theta}_{person,test}$ is required for the test, it does not mean that all these proficiencies are intended to be measured by test developers. For example, test developers want to evaluate high school students' performance on mathematics, sometimes they cannot avoid the descriptive sentences that require students' reading skills. They will try to make the words and sentences as simple as possible to make sure all the students can understand these questions. In other words, these tests are made "insensitive" to students' reading skills, which are needed but not intended for the test. Therefore, it is actually $\boldsymbol{\theta}_{person,test,sensitive}$ vector that determines the dimensionality of the response matrix generated by the person-item interaction. Another way to explain the difference between $\boldsymbol{\theta}_{person,test}$ and $\boldsymbol{\theta}_{person,test,sensitive}$ is that the first one is defined from the psychological view to include all the required proficiencies, while the second one is determined from the statistical view to model the proficiencies that not only vary among persons but also can be discriminated by items.

The formal definition of dimensionality is the total number of proficiencies that are required to meet the local independence assumption for the IRT model (Lord & Novick, 1968). The reasoning for $\boldsymbol{\theta}_{person,test,sensitive}$ makes sure that it is the critical factor explaining the variation in the data matrix.

The estimation and interpretation of dimensions for $\boldsymbol{\theta}_{person,test,sensitive}$ are based on the empirical data analysis and expert judgement from the perspectives of psychometrics and psychology. However, after the idea of multidimensional tests first emerged, lots of efforts have been made to find the reasoning of approximating the widely used unidimensional model to the test that requires more than one hypothetical construct (Reckase, 1979; Drasgow & Parsons, 1983; M. Wang, 1985, 1986; Reckase et al., 1988; Ackerman, 1989; Luecht & Miller, 1992), while at the same time, some research also shows cautions and warnings to this approximation (Reckase et al., 1986; Ackerman, 1991).

As the MIRT model became more and more popular and acceptable, determining the dimensionality for it turns out to be an interesting topic. Lots of literature is trying to justify the number of dimensions by using different model fit indices and criteria (Hambleton & Rovinelli, 1986; Stout, 1987; Roznowski et al., 1991; Gessaroli & Champlain, 1996; Stone & Yeh, 2006; Kao, 2007), infer it by analyzing the number of item clusters from the MIRT calibration result (Miller & Hirsch, 1992; Roussos et al., 1998), or use parallel analysis to compare the observed eigenvalues with those from randomly simulated data (Ledesma & Valero-Mora, 2007).

Furthermore, the determination of dimensionality is complicated by the fact that the data can be fitted better by a more complex model. Hirsch and Miller (1991) showed that overfactoring does not lead to serious negative consequences. Reckase and Hirsch (1991) pointed out that overfactoring is very useful to avoid projection problems which occur when fewer numbers of dimensions are used to analyze the response matrix which is supposed to have more dimensions involved. Actually, the suggestion for overfactoring leads to one important difference between factor analysis and the MIRT analysis: as a data reduction procedure, factor analysis tries to use fewest factors possible to explain the relationship among items or tests; however, the MIRT analysis needs to identify all the dimensions, which differentiate persons or even a subgroup (e.g. high proficiency group) of persons, from the test data (Reckase et al., 1986; Reckase, 2009). The disadvantage for overfactoring is that extracting too many factors may cause serious estimation errors since more parameters need to be estimated based on the same dataset. Thus, choosing a suitable number of dimensions for the MIRT model can also be regarded as finding a good balance between the number of dimensions and parameter estimation.

All in all, it is still by no means certain how to best determine the number of elements in $\boldsymbol{\theta}_{person,test,sensitive}$ or choose the number of dimensions for the MIRT model. The difficulty greatly lies in the model-data fit definition, sampling and estimation errors, and complex situations in the real world.

1.2 Importance of Subscores

It is common for testing programs to purposely design items measuring different constructs in one test; accordingly, there is an increasing demand to extract more information from the test, namely to report the subscores in addition to or in replacement of the overall score. Through this way, test takers or policy makers can know the strength and weakness of proficiencies in specific areas besides the ambiguous general proficiency, and get more detailed diagnostic information for remediation.

As is well known, for the multidimensional tests where more than one hypothetical construct is required, it is only when all the test items measure the same weighted composite of proficiencies or all the proficiencies are highly correlated that the unidimensional model can be used to calibrate the whole response matrix (Reckase et al., 1988; Yao & Boughton, 2007). Only in these situations, the use of one overall general proficiency score is justified and there is no need for the subscore estimation. Otherwise, there is no reasoning to assume the proficiencies measured by different test items are exactly the same.

The biggest challenge to report subscores is that they are less reliable than the total score; therefore, some research shows caution in reporting these unreliable subscores and doubts on any added value of these subscores over the total score (Sinharay et al., 2007). However, they also pointed out that, on the positive side, the subscore seems very useful when there are a reasonably large number of items for each subcategory to ensure reliability, and also there can be moderate but not high correlation between the subscore and the total score to ensure the added value.

For the purpose of subscore reporting, either the commonly used Number-Correct (NC) subscore or unidimensional estimate can be calculated separately for each item cluster, where all items in the same cluster are assumed to measure the same construct proficiency that can be put on the same continuum scale. These item clusters are either predefined by the careful item selection and content scrutiny by item developers and content experts or postdefined by some empirical data analysis, such as the cluster analysis based on item estimates (Luecht & Miller, 1992).

If there is no or low correlation between person proficiencies for different item clusters, separate analysis for items in each cluster seems to be all that needs to be done. As long as there is moderate correlation between different proficiencies, in order to increase the reliability of the subscore, its estimate can be post-hoc adjusted by borrowing information from the total score or the estimates of other subscores, and some research has already shown this with classical test theory (Yen, 1987; Wainer et al., 2000; Haberman, 2008).

However, by allowing simultaneous estimation of parameters for all dimensions, the MIRT model is a growing methodology for the calibration of multidimensional test data. Under the framework of the MIRT, the proficiency can be generalized to be the weighted composite, which is a linear combination of proficiencies on several hypothetical constructs. Items are flexible to have different discrimination power to each proficiency dimension, and even items in the same cluster are not supposed to most discriminate along exactly the same direction. After the dimensionality of the MIRT model is empirically and theoretically chosen, the model calibration can provide a proficiency vector for each person, which is useful for subscore reporting.

1.3 Compensatory MIRT model

The formula for the commonly used compensatory MIRT model is

$$P(u_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i, c_i) = c_i + (1 - c_i)\frac{\exp\left(1.7(\boldsymbol{a}_i'\boldsymbol{\theta}_j + d_i)\right)}{1 + \exp\left(1.7(\boldsymbol{a}_i'\boldsymbol{\theta}_j + d_i)\right)} \tag{1.1}$$

where u_{ij} is the response of j^{th} person to i^{th} item, $\boldsymbol{\theta}_j$ is a column vector of j^{th} person proficiency coordinates in a m-dimensional space, $\boldsymbol{a}_i$ is a column vector that specifies the discrimination power of the i^{th} item for each of the m dimensions, d_i is a scalar parameter that is related to the item difficulty, c_i is a scalar parameter for guessing, and the constant 1.7 is used to approximate the logistic function to the normal ogive one with the input as $\boldsymbol{a}_i'\boldsymbol{\theta}_j + d_i$.

This compensatory model follows the logic of factor analysis and assumes the probability of the person-to-item response is related to a linear combination of several proficiencies. Accordingly, this similarity allows the compensatory MIRT model to borrow some analysis and estimation methods from factor analysis.

There is also a noncompensatory version of the MIRT model (Sympson, 1978),

$$P(u_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i, c_i) = c_i + (1 - c_i)\prod_{k=1}^{m}\frac{\exp\left(1.7a_{ik}(\theta_{jk} - d_{ik})\right)}{1 + \exp\left(1.7a_{ik}(\theta_{jk} - d_{ik})\right)} \tag{1.2}$$

where a_{ik}, θ_{jk} and d_{ik} indicate the item discrimination, person proficiency, and item difficulty for the k^{th} dimension.

With this model, Sympson (1978) argued that the probability of getting an item correct cannot exceed the probability of getting each dimension correct. When a person with low reading skill cannot understand the story problem, he has no chance to get this item correct, how can the deficit be compensated by his high mathematical skill? From this argument, the noncompensatory model seems more realistic, since the probability for this version of MIRT models is related to the product of probabilities for each dimension. However, several problems enormously hinder the development and use of this model, such as more parameters to be estimated, inefficient algorithm for parameter estimation and the proficiency rescaling issue when the number of dimensions increases (Bolt & Lall, 2003; Reckase, 2009).

For these two model versions, the study by Spray et al. (1990) concluded that the difference between the compensatory model and the noncompensatory model could be considered unimportant in practice, especially when the proficiencies are correlated. Besides the above reasons, the compensatory model is preferred due to its comparably easy estimation procedure and interpretation.

In order to find the analogous counterparts in MIRT as in IRT, Reckase (1985) and Reckase and Mckinley (1991) defined the multidimensional generalized discrimination and direction cosines for each item. Through this way, the Cartesian coordinates for each discrimination vector are converted to the polar coordinates in the Cartesian coordinate system.

$$MDISC_i = \sqrt{\boldsymbol{a}_i'\boldsymbol{a}_i} = (\sum_{k=1}^{m} a_{ik}^2)^{\frac{1}{2}} \tag{1.3}$$

$$\begin{aligned} \cos\boldsymbol{\alpha}_i &= (\cos\alpha_{i1}, \cdots, \cos\alpha_{im})' \\ &= (\frac{a_{i1}}{(\sum_{k=1}^{m} a_{ik}^2)^{\frac{1}{2}}}, \cdots, \frac{a_{im}}{(\sum_{k=1}^{m} a_{ik}^2)^{\frac{1}{2}}})' \end{aligned} \tag{1.4}$$

The MIRT generalized discrimination index $MDISC_i$ is actually the length of the discrimination vector, which is an overall measure of the capacity of an item that distinguishes persons in the multidimensional space. The direction cosines vector satisfies the constraint that $(\cos\boldsymbol{\alpha}_i)'\cos\boldsymbol{\alpha}_i = 1$. Therefore, this vector can also be regarded as a normalized version for the discrimination vector. It is well known that this vector plays a very important role in MIRT models, where it determines not only the most discriminating but also the most non-discriminating direction for all items with this direction cosines vector. On the one hand, these items can most differentiate the people whose positions are parallel to the direction cosines line. On the other hand, these items give no discrimination power to the people positioned in the plane, which is orthogonal to the direction cosines line. The reason is that all points in the plane provide the same value for the linear combination, $\boldsymbol{a}_i'\boldsymbol{\theta}$, which is sufficient to determine the probability of person's response to these items. The line or plane with the form of $\boldsymbol{a}_i'\boldsymbol{\theta} = Constant$ in the $\boldsymbol{\theta}$ space is defined as the contour line corresponding to some specific probability.

The multidimensional difficulty is a generalized version of the unidimensional IRT difficulty, and its definition is given by the following formula:

$$B_i = -\frac{d_i}{(\sum_{k=1}^{m} a_{ik}^2)^{\frac{1}{2}}} = -\frac{d_i}{MDISC_i} \tag{1.5}$$

The generalized discrimination, generalized difficulty and direction cosines can be represented by arrowed lines in the multidimensional $\boldsymbol{\theta}$ space. The length of the arrowed line stands for $MDISC_i$, the distance from the origin to the base of the arrowed line is B_i, and the direction of the arrowed line is the same as represented by the direction cosines. One example of the item vector plot is shown in Figure 1.1.

Another detailed index for determining the discrimination power of an item along a certain direction in the $\boldsymbol{\theta}$ space is the information function. For each item, the value of information varies with different values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the latter of which is defined as a certain direction in the multidimensional $\boldsymbol{\theta}$ space (Reckase & Mckinley, 1991).

$$I_i(\boldsymbol{\theta})_{\boldsymbol{\beta}} = 1.7^2 P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta}) (\sum_{k=1}^{m} a_{ik} \cos\beta_k)^2 \tag{1.6}$$

For the fixed item and person, the largest information is given when $\boldsymbol{\beta} = \boldsymbol{\alpha}_i$, since

$$\begin{aligned}
(\sum_{k=1}^{m} a_{ik} \cos\beta_k)^2 &= (\sum_{k=1}^{m} a_{ik}^2)(\sum_{k=1}^{m} \cos\alpha_{ik} \cos\beta_k)^2 \\
&\leq (\sum_{k=1}^{m} a_{ik}^2) \sum_{k=1}^{m} (\cos\alpha_{ik})^2 \sum_{k=1}^{m} (\cos\beta_k)^2 \\
&= \sum_{k=1}^{m} a_{ik}^2
\end{aligned}$$

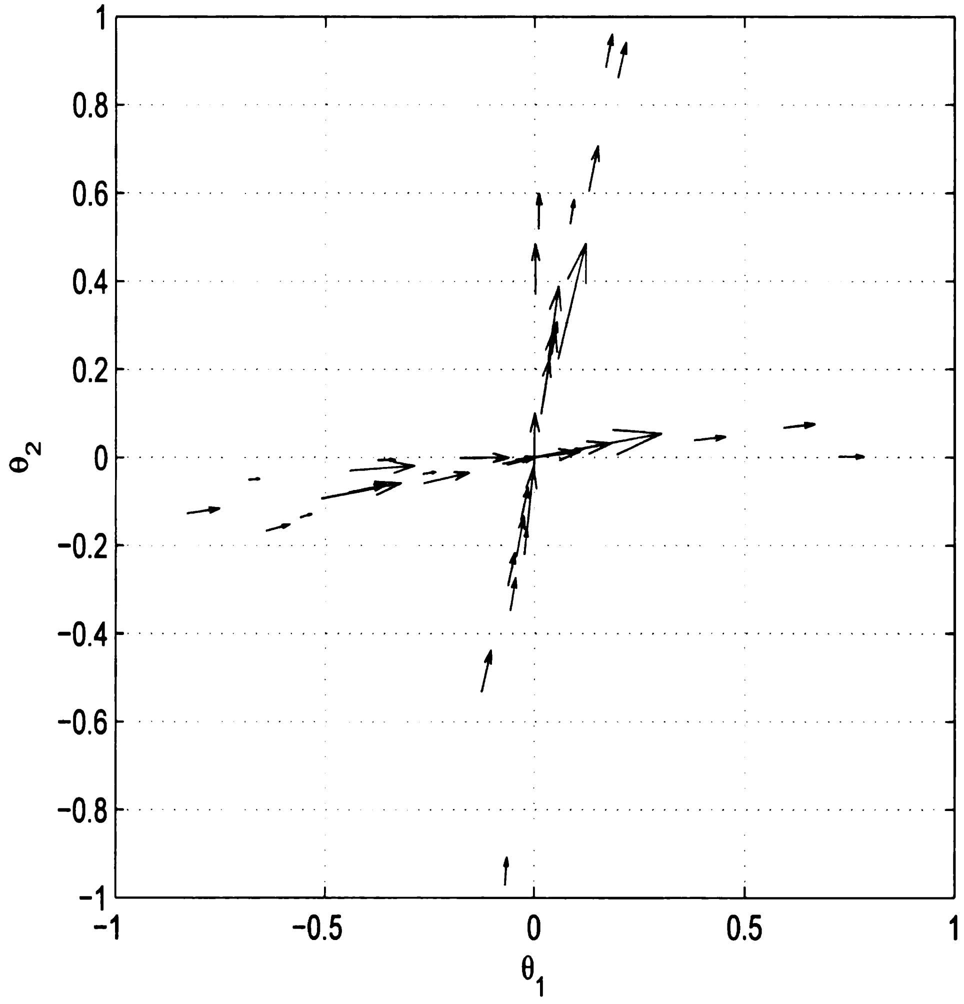

Figure 1.1. Representation of the Characteristics of 40 Items in a Two-Dimensional Space

The information for this direction is simplified as

$$I_i(\boldsymbol{\theta})_{\beta=\boldsymbol{\alpha}_i} = 1.7^2 P_i(\boldsymbol{\theta}) Q_i(\boldsymbol{\theta}) \sum_{k=1}^{m} a_{ik}^2 \tag{1.7}$$

Thus, oftentimes $\boldsymbol{\alpha}_i$ is also called the most discriminating direction of the item.

The above formula can be further maximized when $P_i(\boldsymbol{\theta}) = Q_i(\boldsymbol{\theta}) = 0.5$. The maximum value is $0.85^2 \sum_{k=1}^{m} a_{ik}^2$, which comes out just as a generalized version for the maximum information provided by the unidimensional IRT model. In words, the information for each item is largest for the people who are positioned in the 0.5 probability contour line, with the differentiating direction provided by $\boldsymbol{\alpha}_i$.

The test information along a certain direction is the sum of all the item information along that direction.

$$I(\boldsymbol{\theta})_\beta = \sum_{i=1}^{I} I_i(\boldsymbol{\theta})_\beta \tag{1.8}$$

1.4 Indeterminacies in MIRT Models

This compensatory MIRT model is a combination of the IRT model and factor analysis model; as a result, it suffers not only from the origin and unit indeterminacies in both models, but also from the rotational indeterminacy that especially plagues factor analysis. With the formula, the MIRT indeterminacies come from the fact that there are infinite new sets with $\boldsymbol{\theta}^* = KT\boldsymbol{\theta} + O$, $(\boldsymbol{a}^*)' = \boldsymbol{a}'T^{-1}K^{-1}$ and $d^* = d - \boldsymbol{a}'T^{-1}K^{-1}O = d - (\boldsymbol{a}^*)'O$, which satisfy

$$\begin{aligned}(\boldsymbol{a}^*)'\boldsymbol{\theta}^* + d^* &= (\boldsymbol{a}'T^{-1}K^{-1})(KT\boldsymbol{\theta} + O) + d - \boldsymbol{a}'T^{-1}K^{-1}O \\ &= \boldsymbol{a}'\boldsymbol{\theta} + \boldsymbol{a}'T^{-1}K^{-1}O + d - \boldsymbol{a}'T^{-1}K^{-1}O \\ &= \boldsymbol{a}'\boldsymbol{\theta} + d\end{aligned}$$

$K_{m\times m}$ is a diagonal matrix used for adjusting the unit length for each dimension, $T_{m\times m}$ is a rotational matrix with the uniqueness defined by the row normalization, $O_{m\times 1}$ is a column vector for the origin change.

For these new sets, the order for describing the transformation is arranged as rotation, unit change, and finally origin change. As this order is changed, the formula for the new sets will be changed too; however, it can still do the transformation between the same old and new sets. For this transformation, only one matrix, which is the product of K and T, can do both the rotation and unit change. However, the K matrix is oftentimes separated as the rescaling matrix after the rotation. This matrix will not be emphasized here, since this study mostly focuses on the rotational indeterminacy, namely the $T_{m \times m}$ matrix.

To find a suitable rotation matrix is very important in the MIRT as in factor analysis. The common factor analysis terms the rotation as searching for the best simple loading structure to explain intercorrelations among items or tests; however, in the MIRT modeling, this process is also important for finding an explainable coordinate system for person proficiency estimates.

The common way for partly solving the indeterminacies is to put some assumption constraints on the $\boldsymbol{\theta}$ vector: $E(\boldsymbol{\theta}) = \mathbf{0}_{m \times 1}$ and $\mathrm{cov}(\boldsymbol{\theta}) = \boldsymbol{I}_{m \times m}$. These constraints greatly simplify the parameter estimation procedures of the commonly used NOHARM and TESTFACT software for the MIRT calibration (Fraser, 1988; Bock et al., 2003). However, the zero correlation assumption can be easily violated, since most proficiencies are correlated in reality. For example, most people believe a person's algebraic skill is highly correlated with his geometric skill. Some people may even doubt the zero correlation assumption between reading skill and mathematical skill. Hence, the proficiency estimates with these constraints enforced cannot be directly used for the interpretation of person proficiencies.

1.5 Solution to Rotational Indeterminacy in Factor Analysis

Varimax and Promax are two widely used methods for solving the rotational indeterminacy in factor analysis, and the first one is for orthogonal rotation, while the second one is designed for oblique rotation, which is based on the result from the Varimax rotation. Both rotations try to rotate the item discrimination matrix to a simple structure matrix (Thurstone, 1947).

Suppose the item discrimination matrix is $A_{I \times m}$, with I items and m dimensions. The Varimax method aims at searching for the orthogonal rotation to maximize the sum of squared loading variance across all factors, with adjustments from the item communalities. Mathematically, it resulted in the matrix $A_{Varimax} = AT$, with the constraint as $T'T = I_{m \times m}$ and the following criterion maximized for the $A_{Varimax}$ (Kaiser, 1958).

$$\begin{aligned} v &= \sum_{k=1}^{m} \Big((I \sum_{i=1}^{I} (a_{ik}^2/h_i^2)^2 - (\sum_{i=1}^{I} (a_{ik}^2/h_i^2))^2)/I^2 \Big) \\ &= \frac{1}{I} \sum_{k=1}^{m} (\sum_{i=1}^{I} (a_{ik}/h_i)^4 - I(\sum_{i=1}^{I} (a_{ik}/h_i)^2)^2) \qquad (1.9) \end{aligned}$$

a_{ik} is the loading for the i^{th} item and k^{th} dimension, h_i^2 is the communality of the i^{th} item.

The Promax method is based on the result from the Varimax method and the new axes are free to take any position in the multidimensional space (Hendrickson & White, 1964). First, a target matrix P is defined using the t^{th} power of each element in $A_{Varimax}$, where t is commonly from two to four.

$$p_{ik} = |a_{ik}^{t+1}|/a_{ik} \qquad (1.10)$$

Then the least-squares fit of $A_{Varimax}$ to the target matrix is found by the following formula, which is similar to the regression coefficient estimation.

$$T = (A'_{Varimax}A_{Varimax})^{-1}A'_{Varimax}P \tag{1.11}$$

Finally, after the columns of T are normalized, $A_{Promax} = A_{Varimax}T$.

Finch (2006) pointed out that both methods are effective in identifying which item is associated with which factor; however, the Promax rotation performs better in matching the simple structure.

Both methods are popular in practice and they are the pure mathematical criteria for matching the rotated loading matrix to the simple structure. If these methods are applied to the MIRT model, the disadvantage is that the results from these fixed procedures can seldom be adjusted by the evidence from item contents. Furthermore, the rotated loadings may serve for identifying the grouping of different items; however, it cannot be guaranteed that the loading matrix after the rotation recovers the true item discrimination power for the MIRT Model.

1.6 Interdependency between Proficiency Correlation and Item Discrimination

M. Wang (1986) pointed out that for the general case, the new interpretable $\boldsymbol{\theta}^* = \boldsymbol{L\theta} + \boldsymbol{\mu}$ can be used such that $E(\boldsymbol{\theta}^*) = \boldsymbol{\mu}_{m\times 1}$ and $\mathrm{cov}(\boldsymbol{\theta}^*) = \boldsymbol{\Sigma}_{m\times m}$, if $\boldsymbol{LL}' = \boldsymbol{\Sigma}$ and one possible choice of $\boldsymbol{L}$ can be obtained by the Cholesky decomposition. Accordingly, $(\boldsymbol{a}^*)' = \boldsymbol{a}'\boldsymbol{L}^{-1}$ and $d^* = d - (\boldsymbol{a}^*)'\boldsymbol{\mu}$.

For the test calibration where there is no reference group, it is reasonable to set the zero mean and unit variance for the proficiency of each dimension, just like the origin and unit solution for the unidimensional model. In this situation, the constraints for $\boldsymbol{\theta}^*$ actually should be $E(\boldsymbol{\theta}^*) = \boldsymbol{0}_{m\times 1}$ and $\mathrm{cov}(\boldsymbol{\theta}^*) = \boldsymbol{R}_{m\times m}$, where $\boldsymbol{R}_{m\times m}$ is the correlation matrix among proficiencies.

Now the biggest problem is that there is no presumed value for the $\boldsymbol{R}_{m\times m}$ matrix, which is unknown for most of the time. Obviously, if the item and person parameter estimation is based on the convenient constraints as $E(\boldsymbol{\theta}) = 0_{m\times 1}$ and $\text{cov}(\boldsymbol{\theta}) = \boldsymbol{I}_{m\times m}$, their estimates are essentially already adjusted for the correlation matrix, since $\boldsymbol{a}' = (\boldsymbol{a}^*)'\boldsymbol{L}$ and $\boldsymbol{\theta} = \boldsymbol{L}^{-1}\boldsymbol{\theta}^*$ can satisfy the above constraints. As pointed out by Reckase (1997), in this case, "the observed correlations among the item scores will be accounted for solely by the a-parameters" (p.275).

In situations where $\boldsymbol{R}_{m\times m}$ is unknown, it is even harder to obtain the $\boldsymbol{a}^*$ and $\boldsymbol{\theta}^*$ estimates. Researchers should be cautious when interpreting the $\boldsymbol{a}$ vector as the item discrimination power for the proficiencies, since $\boldsymbol{a}$ is paired with $\boldsymbol{\theta}$, not with the realistically correlated construct estimate $\boldsymbol{\theta}^*$, which is of interest to test developers. Especially, it is very common to use $\boldsymbol{a}$ vectors to define item clusters for the inference of dimensionality and the computation of composite scores (Miller & Hirsch, 1992; Luecht & Miller, 1992). Since the correlation-adjusted $\boldsymbol{a}$ depends on both $\boldsymbol{a}^*$ and $\boldsymbol{L}$, we are never certain that the item clustering inferred from $\boldsymbol{a}$ is due to similar item direction cosines or highly correlated proficiencies for the person population.

Figure 1.2 gives the item vector plots in a two-dimensional space when the proficiency correlation is on different levels. In order to ensure the same product as $\boldsymbol{a}'\boldsymbol{\theta}$, namely the invariance property of the MIRT model, when the proficiency coordinates are transformed, so are the item vectors. From the figure, it is also easy to see that the angle between item clusters increases as the proficiency correlation increases.

Commonly, two assumptions can be adopted to interpret and use the information from correlation-adjusted $\boldsymbol{a}$ vectors. The first one is to assume there is no correlation between multidimensional proficiencies. With this assumption, the $\boldsymbol{a}$ vector may be the same as $\boldsymbol{a}^*$, which can be interpreted as the weight of the orthogonal proficiencies and used for the composite score calculation. The other one is to assume the simple structure for $\boldsymbol{a}^*$, and try to find the correlation matrix among the elements of

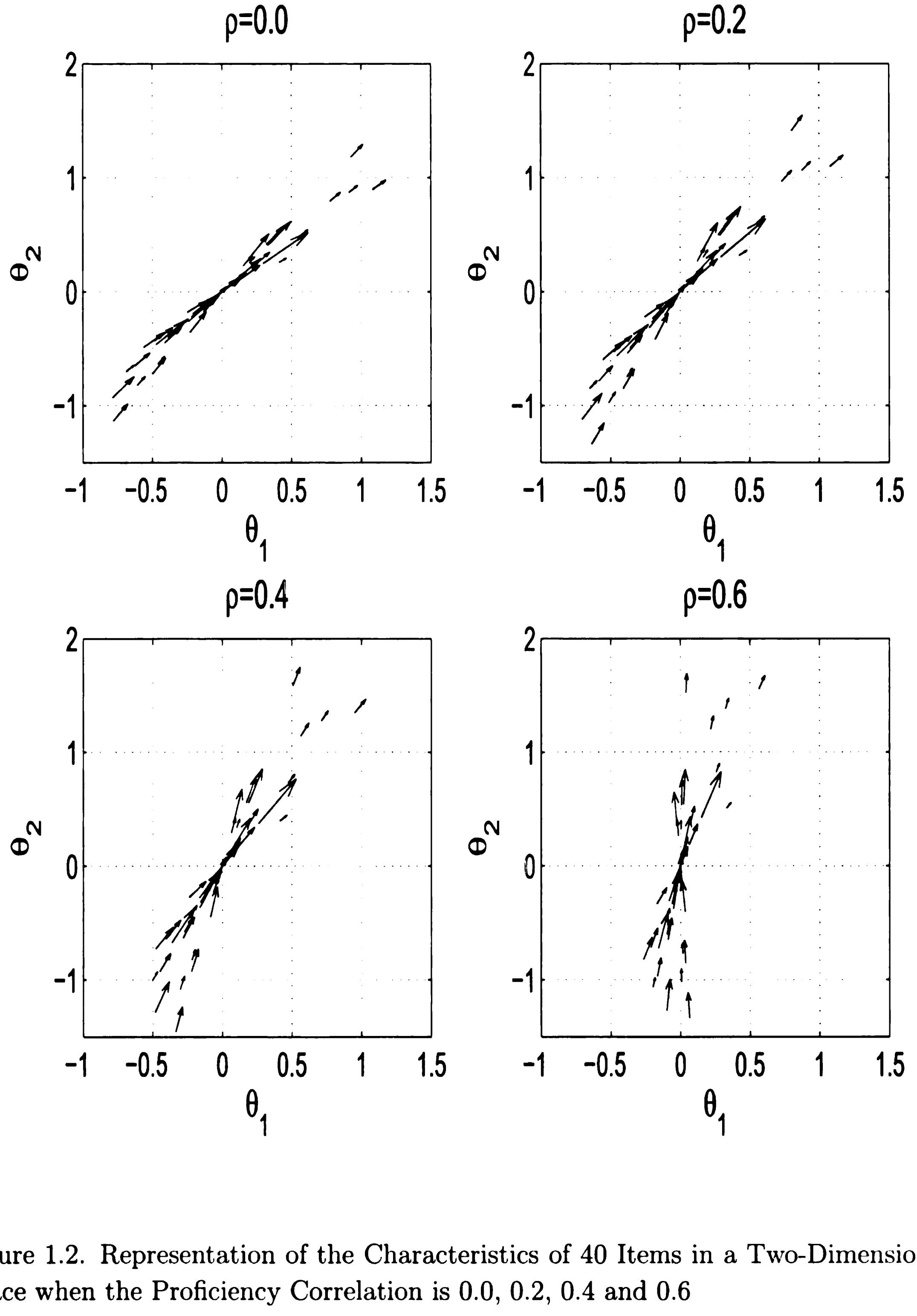

Figure 1.2. Representation of the Characteristics of 40 Items in a Two-Dimensional Space when the Proficiency Correlation is 0.0, 0.2, 0.4 and 0.6

$\boldsymbol{\theta}^*$. The use of these two assumptions is very similar to those for the Varimax and Promax methods in factor analysis, except that $\boldsymbol{\theta}^*$ may only be interpreted as the primary factor score in factor analysis while a more general view in the MIRT is to interpret $\boldsymbol{\theta}^*$ as the construct score, which may be a weighted composite of several raw proficiencies. Without these two useful assumptions, it is hard to separate the proficiency correlation and item discrimination matrix, because they are unknown but dependent during the calibration for the person-by-item response matrix.

As is well known, the first assumption above is very useful in data generation and parameter estimation, while the second one can better serve for the purpose of interpretation and score reporting. Now there comes an interesting research question: how can the construct estimate $\boldsymbol{\theta}^*$ be obtained for better interpretation when the proficiency correlation matrix is unknown? Can it be converted from the uncorrelated $\boldsymbol{\theta}$ estimate?

The subscore estimation method given by Luecht and Miller (1992) avoids this transformation problem by obtaining proficiency estimates only from several separate unidimensional IRT calibrations. They used a two-stage approach; they first did the MIRT calibration on the data and identified the item clusters through the angular analysis for item pairs, and then used the unidimensional model to calibrate the items within each cluster. In their study, the MIRT calibration was only used to group items from the empirical evidence instead of the subjective judgement by experts. Therefore, except the item grouping, this method is no different from the commonly used unidimensional estimation.

The study by Yao and Boughton (2007) proposed the MCMC algorithm to simultaneously estimate parameters for the confirmatory version of the MIRT model, where the correlation matrix among proficiencies is known and all the items are assumed to have discrimination power only on one dimension. Since the coordinate system is already set up with these constraints, the problem of rotational indeterminacy doesn't exist in their study.

In real test settings, the proficiency correlation is seldom known and it is common that within-dimensionality items exist in the test. This situation leads to the common use of the general exploratory version of the MIRT model, whose item and person parameters are all free to be estimated. In order to solve the rotational indeterminacy in the exploratory version, this study uses the item and person estimates from the MIRT model calibration, and projects the uncorrelated $\boldsymbol{\theta}$ estimates onto the most discriminating direction for each item cluster to get $\boldsymbol{\theta}^*$ solution. The simulation study was conducted to detect the effect of balance/unbalanced item design and sampling errors on this $\boldsymbol{\theta}^*$ estimation. Finally, the empirical analysis using the Michigan Educational Assessment Program (MEAP) test data are shown.

CHAPTER 2

Construct Estimation Using Projection

2.1 Transformation in the Orthogonal Coordinate System

Commonly, the orthogonal Cartesian coordinate system is used for positioning each person's proficiency vector in the m-dimensional space. Although the axes in the coordinate system are orthogonal to each other, correlation can be allowed among the coordinates; therefore, besides the uncorrelated proficiencies, the correlated ones can also be fully represented in this orthogonal system. That is to say, there is no need to turn to the oblique system, where axes are not restricted to be orthogonal to each other.

Linear transformation can be performed for points represented by this coordinate system. In the proficiency context, it is defined as the transformation, which can be denoted as the T matrix, from $\boldsymbol{\theta}$ coordinates to $\boldsymbol{\theta}^*$ coordinates, and the transformation is only involved with the elements of $\boldsymbol{\theta}$ that are of the first degree. If the number of elements in $\boldsymbol{\theta}^*$ is the same as in $\boldsymbol{\theta}$, this transformation can also be regarded as the rotation of the coordinate system. Linear transformation and the rotation of the coordinate system function for the same purpose, and the difference is whether the rotation is made on points relative to fixed axes or on axes relative to fixed points. For this reason, the new proficiency coordinates, defined as $\boldsymbol{\theta}^* = T_{m\times m}\boldsymbol{\theta}$, can be interpreted as the linear transformations of points in the old system, or as the fixed points represented by the new system.

If $\mathrm{cov}(\boldsymbol{\theta}) = I$ is assumed,

$$\begin{aligned} \mathrm{cov}(\boldsymbol{\theta}^*) &= T\mathrm{cov}(\boldsymbol{\theta})T' \\ &= TT' \end{aligned} \tag{2.1}$$

The diagonal entry in TT' matrix determines the variance for each element of $\boldsymbol{\theta}^*$, while the off-diagonal entry provides information for the covariance between elements of $\boldsymbol{\theta}^*$. If $TT' = K$ and K is a diagonal matrix, this transformation is called an orthogonal transformation, and it retains the zero correlation between elements of $\boldsymbol{\theta}^*$ in the above situation. Moreover, when $TT' = I$, it is called an orthonormal transformation. The advantage for this transformation is that it is an isomorphic transformation, which preserves the structure between vectors or points. More specifically, the distance of points to the origin does not change, and the angle or scalar product between vectors remains the same, because the configuration is not altered by applying this transformation matrix (Thurstone, 1947). The existence of this transformation leads to the fact that, besides the zero mean and identity variance-covariance matrix constraints on person proficiency estimates, additional constraint is necessary in order for the parameters to be uniquely identified. Some examples are the QR decomposition for the NOHARM software (Stoer & Bulirsch, 2002) and the Varimax criterion to match the simple structure. Both of them are constraints forced on the item discrimination matrix.

The orthogonal projection to vectors is another concept that is very useful for defining the transformation matrix in this study. Suppose there are two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in an m-dimensional space, α is the angle between them, the projection from $\boldsymbol{v}$ to $\boldsymbol{u}$ is calculated by

$$\begin{aligned}
P_{\boldsymbol{v}\rightarrow\boldsymbol{u}} &= (\sqrt{\boldsymbol{v}'\boldsymbol{v}}\text{cos}\alpha)\frac{\boldsymbol{u}}{\sqrt{\boldsymbol{u}'\boldsymbol{u}}} \\
&= (\sqrt{\boldsymbol{v}'\boldsymbol{v}}\frac{\boldsymbol{u}'\boldsymbol{v}}{\sqrt{\boldsymbol{u}'\boldsymbol{u}}\sqrt{\boldsymbol{v}'\boldsymbol{v}}})\frac{\boldsymbol{u}}{\sqrt{\boldsymbol{u}'\boldsymbol{u}}} \\
&= \frac{\boldsymbol{u}'\boldsymbol{v}}{\boldsymbol{u}'\boldsymbol{u}}\boldsymbol{u} \qquad (2.2)
\end{aligned}$$

If $\boldsymbol{u}$ is of unit length, the above equation can be further simplified as

$$P_{\boldsymbol{v}\rightarrow\boldsymbol{u}} = (\boldsymbol{u}'\boldsymbol{v})\boldsymbol{u} \qquad (2.3)$$

where $\boldsymbol{u}'\boldsymbol{v}$ is the length of the projection and $\boldsymbol{u}$ is a unit-length vector defining the direction of the projection.

This projection separates vector $\boldsymbol{v}$ into two unique vectors based on $\boldsymbol{u}$: one is in the direction of $\boldsymbol{u}$, and the other is orthogonal to it. This is not only very useful in constructing the orthonormal basis, but also serves as the best prediction of vector $\boldsymbol{v}$ from vector $\boldsymbol{u}$.

The following section is focused on how to use the projection to define the T matrix, which can be used to transform the orthogonal proficiency estimates to the correlated and interpretable construct scores.

2.2 Construct Estimation

2.2.1 Step 1: Calibration

With the local independence assumption in the MIRT, the joint probability for the person-by-item data matrix can be obtained by multiplying the probability of each person-by-item interaction across all items and persons. Therefore, the likelihood function for the response data based on the model is

$$L(\boldsymbol{U}|\boldsymbol{A},\boldsymbol{d},\boldsymbol{c},\Theta) = \prod_{j=1}^{J}\prod_{i=1}^{I} P(u_{ij}=1|\boldsymbol{\theta}_j)^{u_{ij}}(1-P(u_{ij}=1|\boldsymbol{\theta}_j))^{1-u_{ij}} \qquad (2.4)$$

where J stands for the number of persons, I for the number of items, and $P(u_{ij} = 1|\boldsymbol{\theta}_j)$ is defined by Equation 1.1.

According to maximum likelihood theory, the estimate of $\boldsymbol{A}$,$\boldsymbol{d}$,$\boldsymbol{c}$,$\boldsymbol{\Theta}$ is the set which can maximize the likelihood function in Equation 2.4. However, in order to avoid heavy computation, some variations of estimation procedures are implemented in the two commonly used MIRT software: NOHARM and TESTFACT. Both software uses the constraint of $E(\boldsymbol{\theta}) = \mathbf{0}_{m\times 1}$ and the normal ogive model configuration, and need to input $\boldsymbol{c}$ vector into the software if guessing is assumed. The guessing vector can be estimated from the BILOG software (Zimowski et al., 2003) by the unidimensional IRT calibration to the complete multidimensional data (Bock et al., 2003). Both software gives the unrotated item discrimination matrix as the default and also can provide the Varimax solution for the orthogonal proficiency structure and Promax solution for the oblique proficiency structure. All these three versions of the item discrimination matrix are actually linear transformations of each other.

There are also some differences, in the estimation constraint, estimation procedure and initial coordinate system setup between these two software (Fraser & McDonald, 1988; Bock et al., 1988; Reckase, 2009).

In the NOHARM software, the default setting for cov($\boldsymbol{\theta}$) is $\boldsymbol{I}_{m\times m}$; however, it is also flexible to be changed to other configurations. This software uses polynomials to approximate the normal ogive values, and applies the unweighted least-squares criterion and quasi-Newton algorithm to find the best match between the observed and model-predicted values for the population estimate of the joint probability of correctly answering item pairs. It constrains the estimate of the item discrimination matrix with the first m items to a lower triangle matrix structure, through which the coordinate system for proficiencies is constructed. The resulting discrimination matrix is similar to the R matrix when the QR decomposition method is applied to the original item discrimination matrix (Stoer & Bulirsch, 2002). The disadvantage is that the item discrimination matrix estimate is not very stable or accurate when some of the first m items seem to measure the same construct, and this software does not provide any $\boldsymbol{\theta}$ estimate.

The TESTFACT software directly uses the constraint of $\text{cov}(\boldsymbol{\theta}) = \boldsymbol{I}_{m\times m}$ and simplifies the estimation procedures by applying the EM algorithm to maximize the marginal probability (Bock et al., 1988).

$$L(\boldsymbol{U}|\boldsymbol{A},\boldsymbol{d}) = \prod_{j=1}^{J} \int L(\boldsymbol{u}_j|\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{2.5}$$

where $L(\cdot)$ is the likelihood function, $\boldsymbol{u}_j$ is the response vector by j^{th} person, and $g(\boldsymbol{\theta})$ is the distribution of $\boldsymbol{\theta}$, which is usually assumed to be the multivariate standard normal distribution.

It is common that many people may have the same response strings, so the likelihood can also be written as the multinomial form:

$$L(\boldsymbol{U}|\boldsymbol{A},\boldsymbol{d}) = \frac{N!}{r_1!r_2!\cdots r_s!}\Big[\int L(\boldsymbol{u}_1|\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\Big]^{r_1}\cdots\Big[\int L(\boldsymbol{u}_s|\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\Big]^{r_s} \tag{2.6}$$

where N is the total number of persons, s is the number of unique response strings, and $r_1, \cdots, r_s$ stand for the observed frequency for each unique response string.

Clearly, the person proficiency $\boldsymbol{\theta}$ is integrated out in Equation 2.6, and only item discrimination and difficulty are the unknown parameters that influence the value of the likelihood function. The EM algorithm is applied to maximize the *log* version of the likelihood and the starting values are obtained from the principal component analysis for the guessing-adjusted tetrachoric correlation matrix of responses. The integration in Equation 2.6 can only be approximated numerically, and different numbers of quadrature points lead to different degrees of accuracy for the integral value, which will result in different parameter estimation errors. The disadvantage of this software is that the result is much influenced by the quality of starting values, and the EM algorithm takes some time to converge. For example, the computer for the later simulation study was equipped with the Intel Pentium D processor of CPU 3.49 GHZ speed and 1.99 GB RAM, and it took more than half an hour for each calibration run.

After the item parameters are estimated, they are regarded as fixed and the person estimates are calculated under the Bayesian framework. Two score options are available in the TESTFACT software: The MAP (Maximum A Posteriori) score is calculated as the mode of $L(\boldsymbol{u}_j; \boldsymbol{\theta})g(\boldsymbol{\theta})$, and the EAP (Expected A Posteriori) score, which will be used in this study, is calculated as

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}_j &= \int \boldsymbol{\theta} \widetilde{f}(\boldsymbol{\theta}; \boldsymbol{u}_j) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{\int \boldsymbol{\theta} L(\boldsymbol{u}_j; \boldsymbol{\theta}) g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int L(\boldsymbol{u}_j; \boldsymbol{\theta}) g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \end{aligned} \tag{2.7}$$

The EAP score is much preferred because it only needs easy computation, incorporates the prior information, and avoids infinite scores when response strings are all 0's, all 1's, or inconsistent with the model. The EAP score is also much more reliable than the maximum likelihood estimate (Muraki & Engelhard, 1985).

The characteristic of the EAP score is that the prior normal density information gives large weights to the $\boldsymbol{\theta}$ values close to the center, so the EAP score is biased toward the mean when the number of items are finite (Muraki & Engelhard, 1985; Li & Lissitz, 2000). This leads to the fact that the mean of the estimated scores is approximately the same as the mean of the proficiency distribution; however, the standard deviation of the estimated scores underestimates the standard deviation of the proficiency distribution (DeMars, 2006). Although the EAP score is not comparable to the person parameter as the widely used unbiased maximum likelihood estimation (MLE) score, these two scores give roughly the same rank ordering for people. These EAP scores are treated as $\boldsymbol{\theta}$ estimates for the later projection step.

It should be noted that, in this data calibration step, it is assumed that the model fits the data well, and the number of dimensions has also been confirmed based on some data analysis and expert judgement on item contents.

2.2.2 Step 2: Cluster Analysis

This step is to identify item clusters based on the item discrimination estimates from the MIRT calibration (Miller & Hirsch, 1992). The purpose of cluster analysis is to allocate all the items into different clusters, with the within-cluster variation minimized while the between-cluster variation maximized.

The reason that some items are clustered together is that they are assumed to measure the same construct. Levine and Drasgow (1982) suggested that items in one test can be analyzed into interrelated blocks for the appropriateness measurement. The article by Miller and Hirsch (1992) also pointed out that "each item set can be treated as a different unidimensional composite of the abilities represented in the space, and the amount of spread among the vectors in the same set reflects the degree to which unidimensionality holds for that set".

The study by Reckase et al. (1988) already showed that items which have the same angles with the coordinate axes meet the unidimensionality assumption. In this item cluster context, the orientations of item vectors in the same cluster should be very similar in the multidimensional space so that these items can be assumed to measure the same construct. Accordingly, items in different clusters are supposed to measure different constructs.

The cluster analysis in this step is to group items according to the similarity and dissimilarity among these item vectors. For the cluster analysis, both parametric (Miller & Hirsch, 1992) and nonparametric proximity measures (Roussos et al., 1998) can be used to construct the dissimilarity matrix as the input of cluster analysis. The difference is that the first one uses the angles between item vectors while the second one is based on the contingency table between item pairs after people have been partitioned into groups of equal proficiency. After the dissimilarity matrix is obtained, the distance between hypothetical clusters also needs to be defined as a criterion to group similar elements. Kim (2001) found that the Ward's method, which uses the minimum variance as the distance, by employing the parametric proximity measure yields stable classifications under various conditions as opposed to other methods or nonparametric proximity measures. For this reason, the Ward's method and parametric proximity measure are specified for the cluster analysis.

The cosine of the angle between item vector pairs is calculated by Equation 2.8, and then converted to the degree angle. If $\cos\alpha_{ii'} = 1$, these two items measure exactly the same construct; however, if $\cos\alpha_{ii'}$ is close to 0, the discrimination vectors for these two items are orthogonal to each other and they actually measure two completely different constructs. The angular distances for all item pairs form the dissimilarity matrix as the input matrix, and then the cluster analysis results in a dendrogram showing the hierarchical tree relationship among items, from which the number of clusters and the item grouping are determined by "eyeballing" the hierarchy and linkage of these items.

$$\begin{aligned} \cos\alpha_{ii'} &= \frac{\boldsymbol{a}_i'\boldsymbol{a}_{i'}}{|\boldsymbol{a}_i||\boldsymbol{a}_{i'}|} \\ &= (\cos\boldsymbol{\alpha}_i)'(\cos\boldsymbol{\alpha}_{i'}) \end{aligned} \tag{2.8}$$

Because of this subjectivity in cluster analysis and the inevitable sampling errors, the true item grouping is sometimes hard to be recovered, especially when it is only obtained empirically. Therefore, after the cluster analysis, expert judgement on the substantive meaning of clusters can be considered to adjust the items in each cluster.

After this step, all the items are allocated mutually exclusively and exhaustively into different clusters, and the items in each group are regarded as measuring the same construct.

It should be noted that, theoretically, the cluster analysis should be conducted on the actual item discrimination vector $\boldsymbol{a}^*$, which is hard to obtain from the MIRT calibration. However, oftentimes, the item clustering pattern is also obvious even when the correlation-adjusted item discrimination vector $\boldsymbol{a}$ is applied in the analysis, in which case, all the items seem to be more clustered together than that is done with the actual $\boldsymbol{a}^*$.

2.2.3 Step 3: Reference Composite

This step calculates the reference composite vector, also called the "centroid" vector, for the items within each cluster (M. Wang, 1985, 1986). This vector has the minimum average distance to all the item vectors in the cluster and is regarded as representing the most discriminating direction for these items. It has been applied to explain the unidimensional model approximation to the multidimensional model for the whole test data; however, little research has emphasized its use for the items in the same cluster.

In previous studies, the reference composite vector is the eigenvector associated with the largest eigenvalue for the $A'A$ matrix, while A is the discrimination matrix for all items in the test. Similarly, the reference composite vector for the l^{th} cluster, denoted as $\boldsymbol{w}_l$ here, is defined as the eigenvector related to the largest eigenvalue for the $A_l'A_l$ matrix, where A_l is the discrimination matrix for all items in the l^{th} cluster. This reference composite vector is supposed to point in the most discriminating direction for the items in the cluster.

Since the eigenvector is already in the normalized version, it can also be treated as the direction cosines for the reference composite vector

$$\cos\boldsymbol{\varpi}_l = \boldsymbol{w}_l = (w_{l1}, \cdots, w_{lm})' \tag{2.9}$$

2.2.4 Step 4: Projection

This step is to project the uncorrelated $\boldsymbol{\theta}$ estimate onto the reference composite vector for each cluster and get the correlated $\boldsymbol{\theta}^*$ solution. According to Equation 2.3, the construct estimate based on the reference composite for the l^{th} cluster is calculated by the following formula

$$\theta_l^* = \boldsymbol{w}_l'\boldsymbol{\theta} \tag{2.10}$$

Suppose the cluster analysis at Step 2 results in m^* clusters, all the m^* elements in the $\boldsymbol{\theta}^*$ solution can be obtained by linear transformations of the $\boldsymbol{\theta}$ estimate. The transformation matrix is the eigenvector matrix or the direction cosines matrix, with dimensions of $m^* \times m$.

$$\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_{m^*}' \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} (\cos\boldsymbol{\varpi}_1)' \\ \vdots \\ (\cos\boldsymbol{\varpi}_{m^*})' \end{pmatrix} \boldsymbol{\theta} \tag{2.11}$$

Therefore, the variance-covariance matrix for $\boldsymbol{\theta}^*$ is given by

$$\text{cov}\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_{m^*}' \end{pmatrix} \text{cov}\boldsymbol{\theta} \left(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{m^*}\right) \tag{2.12}$$

If $\text{cov}\boldsymbol{\theta} = I_{m\times m}$ as the assumption,

$$\text{cov}\boldsymbol{\theta}^* = (\boldsymbol{w}_i'\boldsymbol{w}_j)_{ij} = ((\cos\boldsymbol{\varpi}_i)'(\cos\boldsymbol{\varpi}_j))_{ij} \tag{2.13}$$

Ideally, the diagonal elements in the $\text{cov}\boldsymbol{\theta}^*$ matrix are all 1's and the covariances are only determined by the closeness between reference composite vectors. However, due to the sampling errors and biased EAP score, $\text{cov}\boldsymbol{\theta}$ may not be the identity matrix as the assumption. Although the rescaling matrix K can possibly be applied to adjust the unit length to one for each $\boldsymbol{\theta}$ dimension, there may still be a small amount of correlation between the elements in the $\boldsymbol{\theta}$ estimate.

2.3 Dimensionality and Number of Clusters

The number of item clusters can be different from the number of proficiency dimensions for the MIRT model (Miller & Hirsch, 1992; Reckase, 2009), which is also the reason that they are denoted by m^* and m separately.

However, the empirical detections of item clusters and proficiency dimensions are dependent on each other. On the one hand, the angular distance matrix for the cluster analysis is based on the item discrimination estimates from the MIRT calibration with a certain number of dimensions assumed for person proficiencies, which should at least result in a good fit between model and data. M. Wang (1985, 1986) showed that if the unidimensional model is used to analyze the multidimensional test, it actually estimates one composite score for the cluster consisting of all items in the test. The implicit assumption for this approximation is that only one cluster results from the cluster analysis based on the multidimensional calibration. On the other hand, Miller and Hirsch (1992) and Roussos et al. (1998) gave examples of using the item cluster analysis to infer the dimensionality of person proficiencies required by the test. Reckase (2009) points out that substantive meaning should be carefully scrutinized when these two numbers are interpreted, which means that expert judgment on item contents is indispensable in the dimension and cluster determination process.

When $m^* < m$, this can be the case when the test requires reading and mathematical computation skills; however, all the items measure a similar weighted composite of these two skills. In this situation, the high dimensional proficiencies are projected to a low dimensional space, and there is definitely some loss of information, which Reckase and Hirsch (1991) already warned against. However, if the direction cosines of item vectors are very similar or the proficiencies are highly correlated, the use of low dimensional solution to the high dimensional test data is justified.

When $m^* = m$, this is the case when some items in the test measure the reading skill and the other items measure the computation skill. More generally, it can also be the case when some items measure one weighted composite of these two skills, while the other items require the same skills but with different weights. The projection solution in this situation actually chooses a different but interpretable coordinate system for construct estimates, and there is no information loss for this projection.

When $m^* > m$, the first step is to check whether the dimensionality of the MIRT model needs to be increased or not. Since this study assumes there is a good model-data fit and experts also confirm only these proficiency dimensions are required for the test after scrutinizing the item contents, the possibility of increasing proficiency dimensions will be skipped here. Now this $m^* > m$ situation can be the case when more items are added to the test in the previous $m^* = m$ situation, and they measure a composite score with weights different than other items. In this case, the elements in $\boldsymbol{\theta}^*$ are linearly dependent, and any element can be inferred from any other m elements in the $\boldsymbol{\theta}^*$ solution. Therefore, there is no information increase or loss for this projection. Moreover, each of these m^* construct scores corresponds specifically to each item cluster.

In short, all these three situations are possible and common in real test situations. Attention should be paid when the $\boldsymbol{\theta}^*$ estimate is calculated and interpreted.

2.4 Characteristics of Construct Estimates

Each element in $\boldsymbol{\theta}^*$ can be regarded as the subscore for items in the same cluster and can be used for score reporting after the scaling process. There are several advantages of using this $\boldsymbol{\theta}^*$ solution.

First, $\boldsymbol{\theta}^*$ is invariant to any orthonormal rotation of the coordinate system representing $\boldsymbol{\theta}$ estimates. As mentioned previously, the relative distance between points or vectors is not altered by this transformation. In theory, any set of item and person estimates with the constraints of $E(\boldsymbol{\theta}) = \mathbf{0}_{m\times 1}$ and $\text{cov}(\boldsymbol{\theta}) = \boldsymbol{I}_{m\times m}$ leads to the same $\boldsymbol{\theta}^*$ solution, no matter what orthonormal rotations the coordinate system takes. This can be easily shown in Equation 2.14, where $\widetilde{\boldsymbol{\theta}^*}$ indicates the values in the new coordinate system. Therefore, there is no need to decide which rotation of the item discrimination matrix needs to be used for the MIRT calibration: the one with the special lower diagonal, the one after the Varimax rotation, or any other choice.

$$\begin{aligned}\widetilde{\boldsymbol{\theta}^*} &= \begin{pmatrix} \widetilde{\boldsymbol{w}}_1' \\ \vdots \\ \widetilde{\boldsymbol{w}}_{m^*}' \end{pmatrix} \widetilde{\boldsymbol{\theta}} = \begin{pmatrix} (T\boldsymbol{w}_1)' \\ \vdots \\ (T\boldsymbol{w}_{m^*})' \end{pmatrix} (T\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_{m^*}' \end{pmatrix} T'T\boldsymbol{\theta} \\ &= \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_{m^*}' \end{pmatrix} \boldsymbol{\theta} = \boldsymbol{\theta}^* \end{aligned} \tag{2.14}$$

Second, $\boldsymbol{\theta}^*$ is a vector containing several scores with regard to subsets of items, and these subscores give meaningful orderings for people with the interpretation specific to the constructs measured by different item clusters. These subscores are very important for people to know their strength and weakness in each subarea, which cannot be achieved by the simple unidimensional calibration to the whole test data if there is more than one item cluster.

Third, $\boldsymbol{\theta}^*$ can be interpreted as a composite score, and it allows correlations among proficiencies. Due to the interdependency between proficiency correlation and item discrimination, it is hard to separate them to recover the true parameters. The solution here is to use the eigenvector matrix as a possible oblique transformation matrix and obtain the projected construct estimates in the most discriminating directions of different item clusters. This solves the problem that the orthogonal $\boldsymbol{\theta}$ estimates cannot be directly used for interpretation or score reporting, since it is difficult to give substantial meanings to these uncorrelated proficiencies.

Fourth, the elements in $\boldsymbol{\theta}^*$ solution borrow information from each other, especially when these proficiencies are correlated. This actually should give credit to the advantage of MIRT over IRT, since the MIRT estimates these proficiencies simultaneously rather than estimating them separately with the unidimensional IRT calibration. Hence, the $\boldsymbol{\theta}^*$ solution is more reliable than the estimates given by the unidimensional model.

Fifth, the $\boldsymbol{\theta}^*$ solution depends on the grouping of items, which can also include expert judgement to reduce the effect of sampling errors on the item clustering. Therefore, this method is preferred to some pure mathematical criteria, such as the Varimax and Promax, which have fixed procedures to obtain the transformation matrix without any consideration on item contents.

Sixth, the transformation from the $\boldsymbol{\theta}$ estimate to the $\boldsymbol{\theta}^*$ solution clearly explains the relationship and difference between the proficiency dimension and item cluster for multidimensional tests, especially when their numbers are different. It also gives a rationale for when tests that are sensitive to differences on multiple dimensions can be fit by a unidimensional IRT model.

Finally, this projection method is not too much work after the MIRT calibration. It applies the cluster analysis to the angular distance matrix based on the item estimates and then obtains the transformation matrix by calculating the eigenvector that corresponds to the largest eigenvalue of the $A_l^{\prime}A_l$ matrix for each item cluster.

Based on all the above advantages, the $\boldsymbol{\theta}^*$ solution is easy to calculate and ready for interpretation and score reporting.

CHAPTER 3

Simulation Study

This simulation study is aimed at detecting the accuracy and stability of $\boldsymbol{\theta}^*$ estimates and comparing them with NC subscores and unidimensional $\boldsymbol{\theta}_u$ estimates described in Luecht and Miller (1992). Due to the interdependency between proficiency correlation and item discrimination, this simulation study simply assumed the uncorrelated proficiency coordinates and set the direction cosines of the three reference composite vectors as close as possible to the direction of $(1,0,0)$, $(0,1,0)$ and $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$. Therefore, the correlations among $\boldsymbol{\theta}^*$ parameters are built in by the angles between these reference composite vectors instead of into $\boldsymbol{\theta}$ parameters.

3.1 Methods

3.1.1 Parameter Simulation

The item parameters were simulated from the commonly used distribution in simulation studies. The generalized discrimination was assumed to follow a lognormal distribution where $log(MDISC_i)$ had a mean of 0 and a standard deviation of 0.5. The within-cluster angular variation was assumed to be 15°, which was suggested by Roussos et al. (1998) for the approximate simple structure. Therefore, the first two degree angles for each item were assumed to uniformly variate between $0^\circ \sim 15^\circ$ or $-7.5^\circ \sim 7.5^\circ$ around the corresponding reference composite, with the constraint that the sum of cosine squares of the first two angles should be less than 1, and the last angle was determined by $(\cos\boldsymbol{\alpha}_i)'\cos\boldsymbol{\alpha}_i = 1$. Since all the discrimination parameters were assumed to take positive values, the directions for item vectors close to the axis were more restricted than those somewhat distant from all the axes. The generalized difficulty B_i was sampled from a normal distribution with mean of 0 and standard deviation of 0.75, and then b_i was calculated by $-B_i * MDISC_i$. All the c_i's were set to 0, in order to reduce the guessing effect.

The person parameter matrix was simulated from the multivariate normal distribution with zero mean vector and identity variance-covariance matrix, and the same matrix was applied to all replications in both designs.

The probability matrix for the person-item interactions was calculated by Equation 1.1 with c=0, and then it was compared with an equal-size matrix whose elements were randomly generated from the standard uniform distribution on [0, 1]. If the simulated number was less than or equal to the probability, the corresponding response was assigned to 1 and 0 otherwise. This step resulted in the dichotomous item score matrix.

3.1.2 Simulation Design

Each dataset contained responses from 5000 persons and 45 items, and one example of the test with 45 items is the Collegiate Assessment of Academic Proficiency from the ACT. In order to test the effect of balanced/unbalanced item numbers for different clusters, the number of items for the three clusters was set to have two designs: balanced (15-15-15) and unbalanced (10-15-20), and these two designs were labeled as Design 1 and Design 2 respectively. This unbalanced design was considering 10 as the fewest number of items used for any cluster, which followed most simulation analyses in Reckase (2009). In order to reduce the effect of sampling errors, 50 replication datasets were created for each design.

3.1.3 Calibration and Projection

Under the MIRT framework, the TESTFACT software with the Promax rotation option specified in the command was used to calibrate each dataset. As mentioned in the previous chapter, the Promax rotation is based on the result of the Varimax rotation. Although the Promax option was requested, the software still provided the item and person estimates with the Varimax rotated loadings as the starting values; additionally, it provided the estimated proficiency correlation matrix obtained from the Promax rotation.

The cluster analysis was then performed on the angular matrix to determine the grouping of items. Although the true clustering of items was known for the simulation study, the cluster analysis was conducted for the purpose of complete integrity of the whole procedure. Based on the item grouping, the EAP estimates were projected onto the reference composite vector for each item cluster.

According to the Luecht and Miller (1992) method, the BILOG software was used to get the unidimensional estimate separately for each item cluster. 45 items were divided into three subtests and analyzed separately. In order to make the unidimensional scores more comparable to the multidimensional projected EAP scores, the EAP estimates instead of the maximum likelihood estimates were requested for the unidimensional calibration.

In both software, the convergence criterion was set to be 0.005, and the number of quadrature points for both the EM algorithm and EAP score estimation was set as the default.

3.1.4 Evaluation Criteria

The evaluation was mostly based on descriptive statistics, such as mean, standard deviation, correlation, Bias, Root Mean Squared Error (RMSE), and the scatter plot.

First, the recovery of item and person parameters was analyzed by the average correlation across replications and average Bias/RMSE across items and persons. The formulas for Bias and RMSE are given by Equations 3.1 and 3.2, where η is defined as any item/person parameter, $\hat{\eta}_r$ as the estimate from the r^{th} replication and R as the total number of replications. It is easy to see that Bias and RMSE are calculated as the average raw or squared difference between the estimated and true parameter across replications. Large deviation was expected to be found in the person proficiency recovery since the EAP scores obtained under the Bayesian framework are known to be biased toward the mean.

$$\text{Bias}(\eta) = \sum_{r=1}^{R} \frac{(\hat{\eta}_r - \eta_r)}{R} \tag{3.1}$$

$$\text{RMSE}(\eta) = \sqrt{\sum_{r=1}^{R} \frac{(\hat{\eta}_r - \eta_r)^2}{R}} \tag{3.2}$$

Second, the variation of $\boldsymbol{\theta}^*$ estimates across replications was calculated to investigate the stability of these estimates.

Third, the correlations among $\boldsymbol{\theta}^*$ estimates, unidimensional $\boldsymbol{\theta}_u$ estimates and NC subscores were provided. All three scores were assumed to be highly correlated and give roughly the same rank order to people.

Fourth, the recovery of true $\boldsymbol{\theta}^*$ by both $\boldsymbol{\theta}^*$ estimates and $\boldsymbol{\theta}_u$ estimates was calculated according to the same three criteria: correlation, Bias and RMSE. Then the recovery efficiencies by these two estimates were compared separately for each cluster.

Finally, the correlation matrix among the elements of $\boldsymbol{\theta}^*$ estimates was compared with those obtained from the Promax method and unidimensional $\boldsymbol{\theta}_u$ estimates, with reference to the correlation among $\boldsymbol{\theta}^*$ parameters.

3.2 Results

3.2.1 Parameter Estimation

The simulated item discrimination, difficulty, generalized discrimination, generalized difficulty and directional degree angles with each axis are shown in Tables 3.1 and 3.2 respectively for the balanced and unbalanced designs. These parameters were only samples from the commonly used distributions, and they were regarded as the parameters for the following simulation study. It is clear that, in both designs, items in cluster 1 mostly measure the proficiency on the first dimension, and items in cluster 2 mostly discriminate the proficiency on the second dimension. However, different than items in the first two clusters, the items in cluster 3 mostly discriminate the roughly equally-weighted composite of proficiencies from all three dimensions. This is also obvious from the directional degree angles: for items in the first two clusters, the angles with one axis are close to zero; however, for items in the third cluster, none of the angles is close to zero.

It should be noted that this projection method is at the preliminary stage; therefore, this simulation study assumed the mixed structure items but uncorrelated proficiency coordinates to only investigate the efficiency of this method based on the item composite effect.

According to the method in Subsection 2.2.3, the direction cosines for the three reference composite vectors are $(0.996, 0.029, 0.081)$, $(0.101, 0.992, 0.081)$, $(0.614, 0.614, 0.496)$ for Design 1 and $(0.996, 0.040, 0.081)$, $(0.083, 0.992, 0.097)$, $(0.616, 0.559, 0.555)$ for Design 2. These reference composite vectors together with item vectors are shown in Figure 3.1. The arrowed solid lines are the item vectors, and the arrowed dashed lines indicate the reference composite vectors, whose lengths are stretched for better view. In the figure, the number of coordinate axes indicates the dimensionality for the MIRT model, while the number of reference composite vec-

Table 3.1. MIRT Item Parameters for Design 1

Cluster	item	a_1	a_2	a_3	d	$MDSIC$	B	α_1	α_2	α_3
1	1	0.95	0.02	0.04	1.22	0.95	-1.28	3	89	87
1	2	0.69	0.00	0.07	0.39	0.69	-0.57	6	90	84
1	3	0.70	0.01	0.01	0.39	0.70	-0.57	1	89	89
1	4	0.79	0.01	0.07	0.71	0.79	-0.89	5	90	85
1	5	0.57	0.01	0.01	0.12	0.57	-0.21	1	89	89
1	6	1.76	0.03	0.07	1.17	1.77	-0.66	3	89	88
1	7	0.75	0.02	0.15	0.63	0.76	-0.82	11	88	79
1	8	0.77	0.01	0.02	-0.11	0.77	0.14	2	89	88
1	9	1.74	0.13	0.24	-0.11	1.76	0.06	9	86	82
1	10	0.73	0.04	0.00	0.51	0.73	-0.70	3	87	90
1	11	0.94	0.01	0.20	1.14	0.96	-1.19	12	89	78
1	12	0.68	0.01	0.10	0.01	0.68	-0.02	8	90	82
1	13	1.24	0.00	0.02	-1.51	1.24	1.22	1	90	89
1	14	0.75	0.05	0.12	-0.83	0.76	1.09	10	86	81
1	15	0.92	0.00	0.00	0.61	0.92	-0.66	0	90	90
2	16	0.02	1.10	0.02	-1.31	1.10	1.19	89	2	89
2	17	0.00	0.69	0.01	0.02	0.69	-0.03	90	1	89
2	18	0.07	0.81	0.04	-1.45	0.81	1.79	85	6	87
2	19	0.08	0.78	0.03	-0.24	0.79	0.30	84	6	88
2	20	0.02	0.81	0.19	0.26	0.83	-0.32	88	13	77
2	21	0.06	1.07	0.10	-0.99	1.08	0.92	87	6	85
2	22	0.02	1.03	0.01	0.99	1.03	-0.96	89	1	89
2	23	0.05	0.75	0.10	-0.01	0.76	0.02	87	9	82
2	24	0.05	0.58	0.11	0.00	0.59	-0.00	85	12	79
2	25	0.36	1.53	0.13	-0.39	1.58	0.25	77	14	85
2	26	0.07	0.78	0.05	0.04	0.78	-0.05	85	7	86
2	27	0.09	1.09	0.08	1.94	1.09	-1.77	85	6	86
2	28	0.20	1.04	0.14	0.05	1.07	-0.05	79	13	82
2	29	0.12	0.58	0.06	0.39	0.60	-0.65	78	13	84
2	30	0.00	0.71	0.04	-1.13	0.71	1.59	90	3	87
3	31	0.87	0.80	0.51	-0.14	1.29	0.11	47	52	67
3	32	0.33	0.34	0.44	0.64	0.64	-1.00	60	59	47
3	33	0.94	0.93	0.68	-0.20	1.49	0.13	51	51	63
3	34	0.24	0.25	0.21	-0.36	0.40	0.91	54	52	59
3	35	0.26	0.27	0.22	-0.05	0.43	0.12	53	52	60
3	36	0.42	0.45	0.50	0.04	0.79	-0.05	58	56	51
3	37	0.49	0.59	0.43	-0.74	0.88	0.84	56	48	61
3	38	0.46	0.63	0.61	0.35	0.99	-0.36	62	50	52
3	39	0.24	0.23	0.28	0.18	0.43	-0.42	57	58	49
3	40	0.34	0.42	0.44	-0.35	0.70	0.51	61	53	51
3	41	1.38	1.28	0.85	1.37	2.06	-0.67	48	52	66
3	42	0.64	0.49	0.60	-0.05	1.00	0.05	51	61	53
3	43	0.73	0.80	0.66	0.01	1.26	-0.01	55	51	59
3	44	0.18	0.21	0.25	0.23	0.37	-0.62	62	55	48
3	45	0.28	0.36	0.28	-0.16	0.54	0.30	58	48	59
Mean		0.51	0.48	0.20	0.07	0.91	-0.07			
Std		0.45	0.42	0.22	0.73	0.37	0.77			

Table 3.2. MIRT Item Parameters for Design 2

Cluster	item	a_1	a_2	a_3	d	$MDSIC$	B	α_1	α_2	α_3
1	1	0.88	0.14	0.11	0.13	0.90	-0.14	11	81	83
1	2	0.78	0.05	0.08	0.34	0.78	-0.44	7	86	84
1	3	1.24	0.10	0.26	1.56	1.27	-1.23	13	86	78
1	4	0.96	0.07	0.13	0.81	0.97	-0.83	9	86	82
1	5	0.81	0.00	0.14	-0.38	0.82	0.46	10	90	80
1	6	0.83	0.01	0.02	0.76	0.83	-0.91	1	89	89
1	7	2.12	0.01	0.03	-1.35	2.12	0.64	1	90	89
1	8	1.06	0.01	0.02	-0.57	1.06	0.53	1	90	89
1	9	1.08	0.01	0.03	0.02	1.08	-0.01	1	90	89
1	10	0.99	0.11	0.19	-0.45	1.01	0.45	12	84	79
2	11	0.03	0.58	0.10	-0.28	0.59	0.47	87	11	80
2	12	0.00	0.82	0.01	0.34	0.82	-0.41	90	1	90
2	13	0.05	0.67	0.07	-0.05	0.68	0.07	86	7	84
2	14	0.09	0.60	0.09	-0.07	0.62	0.12	81	12	82
2	15	0.17	1.60	0.31	-1.11	1.63	0.68	84	13	79
2	16	0.02	0.66	0.04	0.10	0.66	-0.15	88	4	86
2	17	0.01	1.16	0.09	-0.81	1.16	0.69	89	4	86
2	18	0.03	1.43	0.06	1.21	1.44	-0.85	89	3	88
2	19	0.31	1.93	0.21	-0.07	1.96	0.04	81	11	84
2	20	0.03	0.54	0.10	0.96	0.55	-1.74	87	11	79
2	21	0.04	0.89	0.03	0.45	0.89	-0.50	87	3	88
2	22	0.00	0.56	0.00	-0.47	0.56	0.84	90	0	90
2	23	0.03	0.72	0.03	-0.95	0.72	1.32	88	3	88
2	24	0.03	0.20	0.02	0.31	0.20	-1.55	82	11	83
2	25	0.10	1.05	0.08	0.27	1.06	-0.25	85	7	86
3	26	0.42	0.35	0.48	0.88	0.73	-1.20	55	61	49
3	27	0.90	0.73	0.89	-1.90	1.46	1.30	52	60	53
3	28	1.63	1.58	1.25	-2.62	2.59	1.01	51	52	61
3	29	0.27	0.31	0.39	-0.18	0.57	0.32	62	57	47
3	30	0.96	1.10	0.87	0.79	1.70	-0.46	55	50	59
3	31	0.59	0.67	0.71	-0.94	1.14	0.82	59	54	52
3	32	0.39	0.31	0.28	0.14	0.57	-0.25	47	57	61
3	33	0.87	0.67	0.72	0.15	1.31	-0.11	49	59	57
3	34	0.46	0.47	0.54	-0.38	0.85	0.45	57	56	51
3	35	1.04	0.74	0.89	-0.25	1.56	0.16	48	62	55
3	36	0.46	0.48	0.51	0.66	0.84	-0.78	57	55	53
3	37	0.57	0.65	0.58	-0.34	1.05	0.33	57	51	56
3	38	0.85	0.68	0.74	-1.34	1.31	1.02	50	59	56
3	39	0.28	0.29	0.24	-0.16	0.47	0.34	53	52	59
3	40	0.99	0.70	0.84	-0.77	1.47	0.53	48	62	55
3	41	0.86	0.93	0.71	0.85	1.45	-0.58	54	50	61
3	42	1.11	0.80	0.99	-0.39	1.69	0.23	49	62	54
3	43	0.55	0.50	0.52	-1.50	0.91	1.65	53	56	55
3	44	0.53	0.51	0.67	0.63	0.99	-0.63	58	59	48
3	45	0.17	0.15	0.22	-0.22	0.31	0.69	58	61	46
Mean		0.57	0.59	0.34	-0.14	1.05	0.05			
Std		0.49	0.46	0.34	0.84	0.49	0.77			

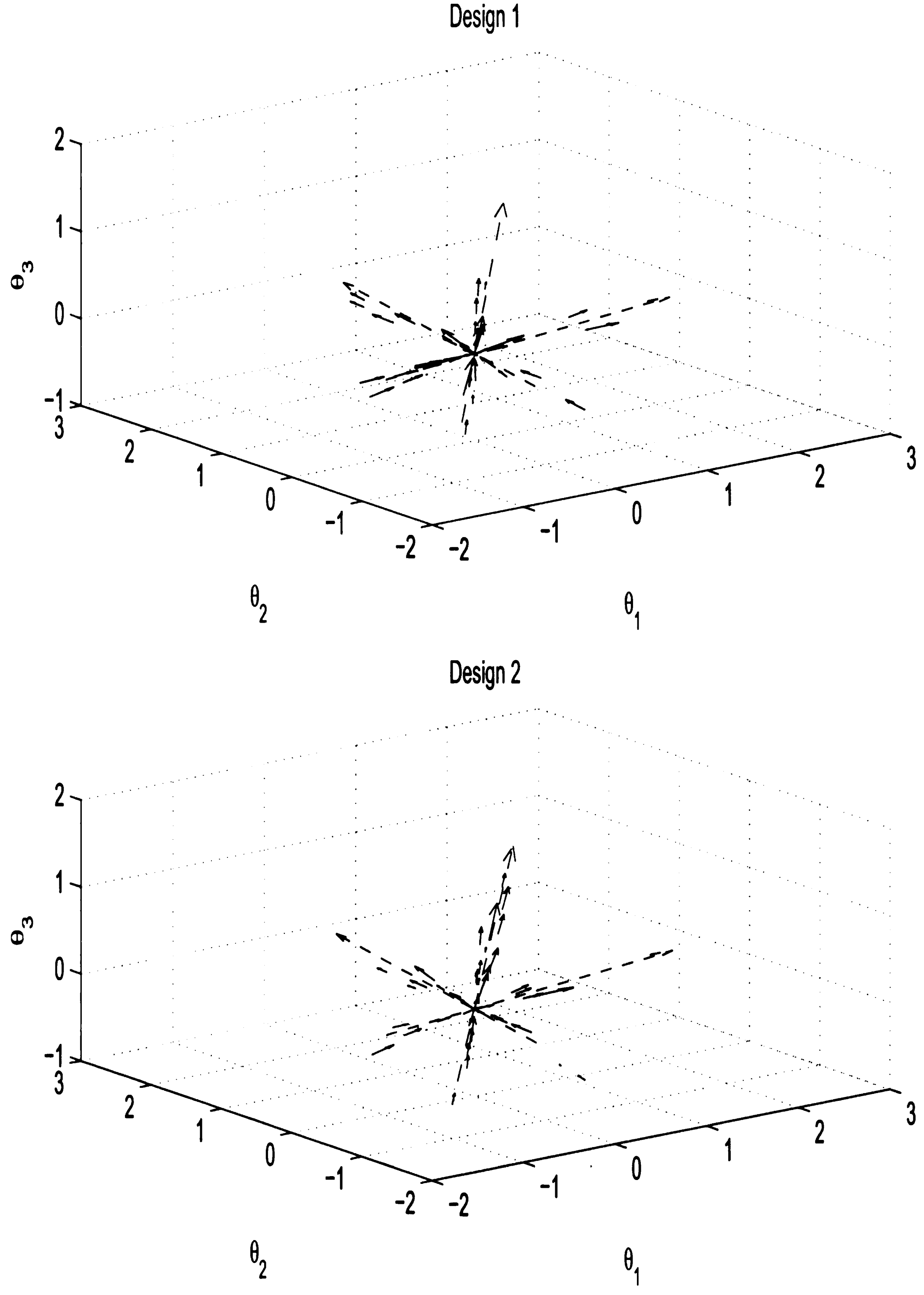

Figure 3.1. Plots of Item Vectors and Reference Composite Vectors

tors shows the number of subscores that are necessary for score reporting. Although it is not shown in this figure, it is worth attention that the dimensionality can be different from the number of item clusters.

The three-dimensional vectors for person proficiencies were simulated from the multivariate standard normal distribution. The means, standard deviations and correlation matrix for this sample are shown in Table 3.3.

Table 3.3. Descriptive Statistics of MIRT Person Parameters

	θ_1	θ_2	θ_3
θ_1	1.0000	0.0128	-0.0020
θ_2	0.0128	1.0000	0.0086
θ_3	-0.0020	0.0086	1.0000
Mean	0.0148	0.0364	0.0243
Std	1.0058	1.0061	1.0147

With these item and person parameters, the dichotomous response matrix was created by comparing the probability matrix with the random matrix generated from the uniform distribution. Then the TESTFACT software was used for the MIRT calibration. The computer for these TESTFACT runs was equipped with the Intel Pentium D processor of CPU 3.49 GHZ speed and 1.99 GB RAM. It took about 33 minutes of CPU time for each run, and most estimations converged at around 25 iterations for the balanced design and a little longer for the unbalanced design, e.g., 75 iterations.

TESTFACT is a data calibration software under the MIRT model, and it aims at estimating item characteristic and person proficiency parameters. This software may not give the estimates of several dimensions in the same order as those in the data generation. For example, in Design 1, the estimates for the first two dimensions were switched in the TESTFACT output. Furthermore, the TESTFACT software may give results as almost all the discrimination estimates for the whole dimension are negative, especially when the Varimax or Promax rotation is specified. These changes are valid under the framework of factor analysis. And this "negative discrimination" phenomenon should not lead to any problem for the later construct estimation if the proficiency coordinates for the corresponding dimension are also estimated as the negated values by fixing these negative a values as item parameters.

However, one principle in the TESTFACT calibration is that in order to make the above-zero scores usually assigned with the above-average achievements, the software sometimes negates the proficiency estimates that are obtained by fixing the current item estimates. This phenomenon is pointed out in the TESTFACT help file: "Factor scores are not unique in the sense that multiplication of any column of factor scores by -1 does not affect the validity of the estimates. It may therefore happen that negative scores are associated with above average percent responses and vice versa for below average responses. TESTFACT attempts to reverse the signs in such a way that scores above zero are usually assigned with above average achievement." Several separate TESTFACT runs confirmed that, no matter whether the item discrimination estimates for the whole dimension were kept the same or negated, the proficiency estimates given by the TESTFACT software never changed.

The solution to avoid this "negative discrimination" effect was to obtain the correlation between the estimated and true item/person parameters. When the correlation was found to be negative, these item/person estimates were taken as the negative values of the estimates from the output. And this is to obtain one set of "correct" item and person estimates from the MIRT calibration and avoid situations where item estimates are negative while person estimates are not negated or vice versa.

3.2.2 Item and Person Parameter Recovery

In the TESTFACT software, item parameters are estimated using the EM algorithm with starting values from the Varimax transformed loadings, which come from the factor analysis on the tetrachoric correlation matrix.

Table 3.4 shows the average correlation across all replications and average Bias/RMSE across all items. From the table, both correlation and RMSE indices are roughly the same for the discrimination estimates of the first two dimensions and for both designs. The recovery for a_3 is a little worse than for a_1 and a_2 in both designs, and especially it is much worse in Design 2, with the average Bias as -0.1549 and RMSE as 0.1741.

Table 3.4. Recovery of Item Parameters

		a_1	a_2	a_3	d
Design 1	Correlation	0.9848	0.9732	0.9186	0.9989
	Bias	-0.0307	0.0064	-0.0384	0.0305
	RMSE	0.0642	0.0731	0.1219	0.0457
Design 2	Correlation	0.9906	0.9854	0.9359	0.9986
	Bias	-0.0394	0.0360	-0.1549	0.0406
	RMSE	0.0672	0.0795	0.1741	0.0534

The Bias in the table is averaged across all items, which may not be a good measure, since the parameters at different value levels may have different degrees of bias in the estimation and Bias values with different signs are cancelled out. Therefore, it is not surprising that these Bias values in the table may seem quite different. To check the discrimination parameter recovery for each specific item, the scatterplots between the parameters and their Bias are shown in Figure 3.2. The pictures in the first three rows illustrate the Bias for the item discrimination parameters. The ideal case is that all the points locate on the $Bias = 0$ line, which means that there is no systematic estimation error for all the item parameters in the long run. From the figure, the a_1 and a_2 parameters with large values are slightly underestimated, while the estimation of almost all the a_3 parameters is negatively biased. This is more serious for items with the high $MDSIC$ index in Design 2. The underestimation of a_3 may indicate that, with the Varimax rotation in the TESTFACT software, the estimation of discrimination parameters for some dimension is negatively biased if there is no item vector close to that dimensional axis.

As is well known, due to the indeterminacy, it is extremely hard to recover the

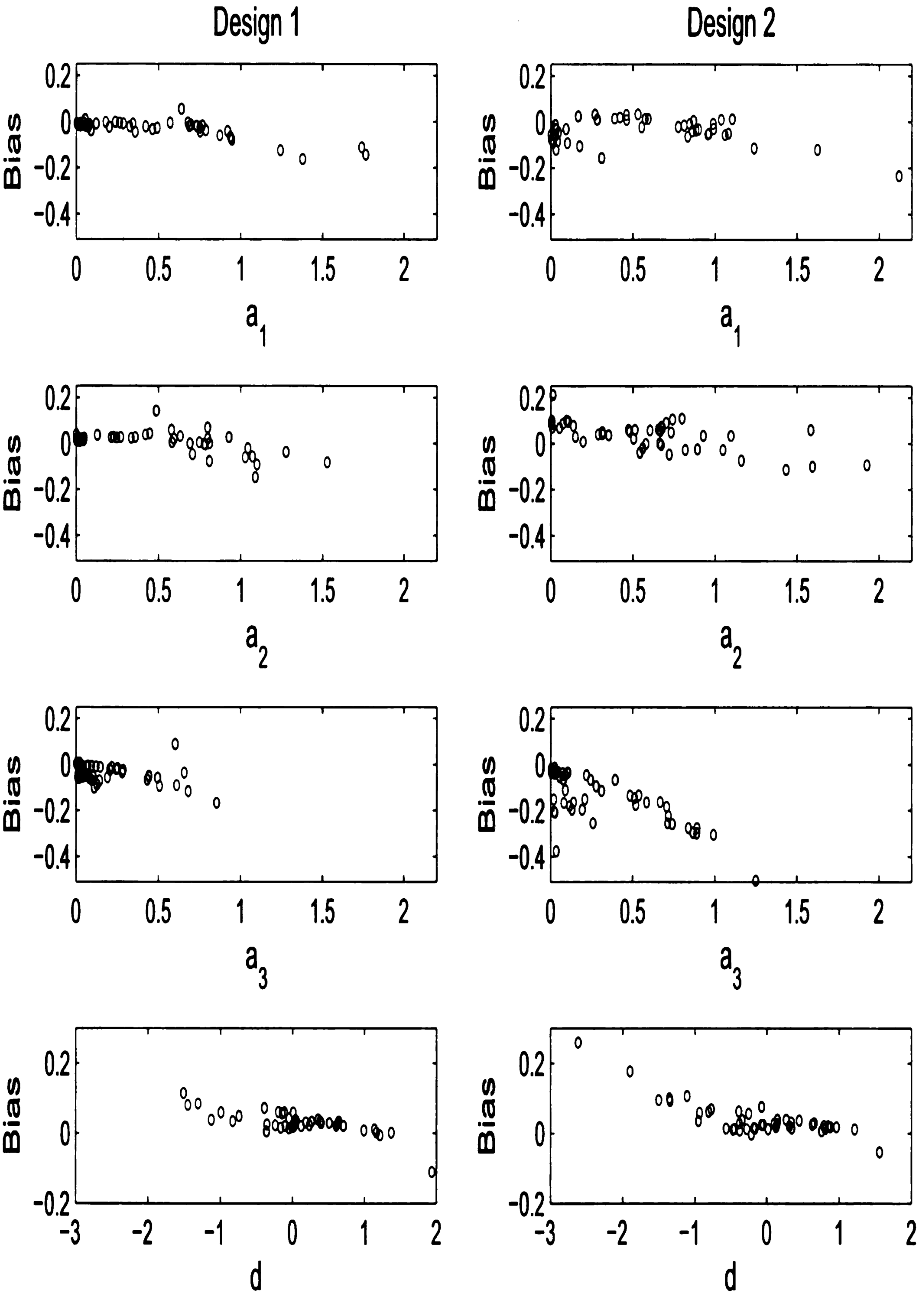

Figure 3.2. Plots of Item Parameters versus Bias

original coordinate system in factor analysis. It is the same here that the orientations of coordinate axes for the discrimination estimates provided by the TESTFACT software may not be the same as those for parameters, and it is especially problematic for the axis where no item most discriminates that proficiency dimension. Furthermore, the alignment of the coordinate axes is complicated by the sampling errors involved with each specific replication.

d is a scalar parameter, whose estimation is not influenced by the rotational indeterminacy, so the recovery for this parameter is much more satisfactory than for the discrimination parameters. From Table 3.4, the correlations between the estimated and true *d* parameters are very close to 1. In addition, with regard to the low values of Bias and RMSE, the recovery for the *d* parameters is acceptable. Observed from the last row in Figure 3.2, the *d* parameters with large values are underestimated, while those with small values are overestimated. The more extreme the true value is, the more bias between the estimated and true parameter there exists. The RMSE for *d* is 0.0511 for Design 1, which is less than 0.0704 for Design 2. The reason may be related to the fact that there are more extreme *d* values in Design 2 than in Design 1.

The same three criterion indices for the recovery of raw proficiency parameters are shown in Table 3.5. Since θ_1 and θ_2 are most influential in determining persons' responses to certain subsets of items, their recovery is much better than that of θ_3. The correlations between the true and estimated proficiencies for the first two dimensions are around 0.9, which indicates that these estimates roughly retain the true ordering of people. The correlation for θ_3 recovery is only around 0.6 for Design 1, and barely reaches 0.7 even when the third dimensional proficiency influentially determines more responses in Design 2. The bad recovery of person parameters for this dimension is not surprising because the coordinate system, especially the orientation of the third axis, is not the same as that for the true parameters, which also results in the bad estimation for a_3.

Table 3.5. Recovery of Person Parameters

		θ_1	θ_2	θ_3
Design 1	Correlation	0.9140	0.9082	0.6180
	Bias	-0.0284	-0.0303	-0.0262
	RMSE	0.3970	0.4113	0.7459
Design 2	Correlation	0.8893	0.9083	0.6982
	Bias	-0.0094	-0.0342	-0.0201
	RMSE	0.4417	0.4082	0.6823

In Table 3.5, every Bias value is averaged across all persons, and they are all of negative values. This is mostly because the mean of proficiency parameters is positive for all three dimensions, or there are more people with positive proficiency parameters for all three dimensions, and the estimation of positive parameters is known to be mostly negatively biased.

The RMSE criterion also shows that there is a big deviation between the estimated EAP score and true proficiency. Its value is much larger for θ_3 recovery than for θ_1 and θ_2 recovery, which confirms again that the parameters related to the third coordinate axis are not well recovered due to its incorrectly recovered orientation in the multidimensional space. The RMSE is 0.4417 for θ_1 in Design 2, which is a little larger than the corresponding value in Design 1. This is reasonable since θ_1 is influential on responses of fewer simple structure items in Design 2 than in Design 1. The similar reason can be used to explain the RMSE difference for θ_3 recovery in both designs.

Figure 3.3 gives the scatterplots between the proficiency parameters and their Bias for each dimension and for each design, in order to check the Bias related to different proficiency levels. From the figure, it is clear that all these proficiencies are underestimated for large values and overestimated for small values. This phenomenon is within expectation, since the EAP score is known to be biased toward the mean of the prior distribution, which is 0 for the default multivariate standard normal distri-

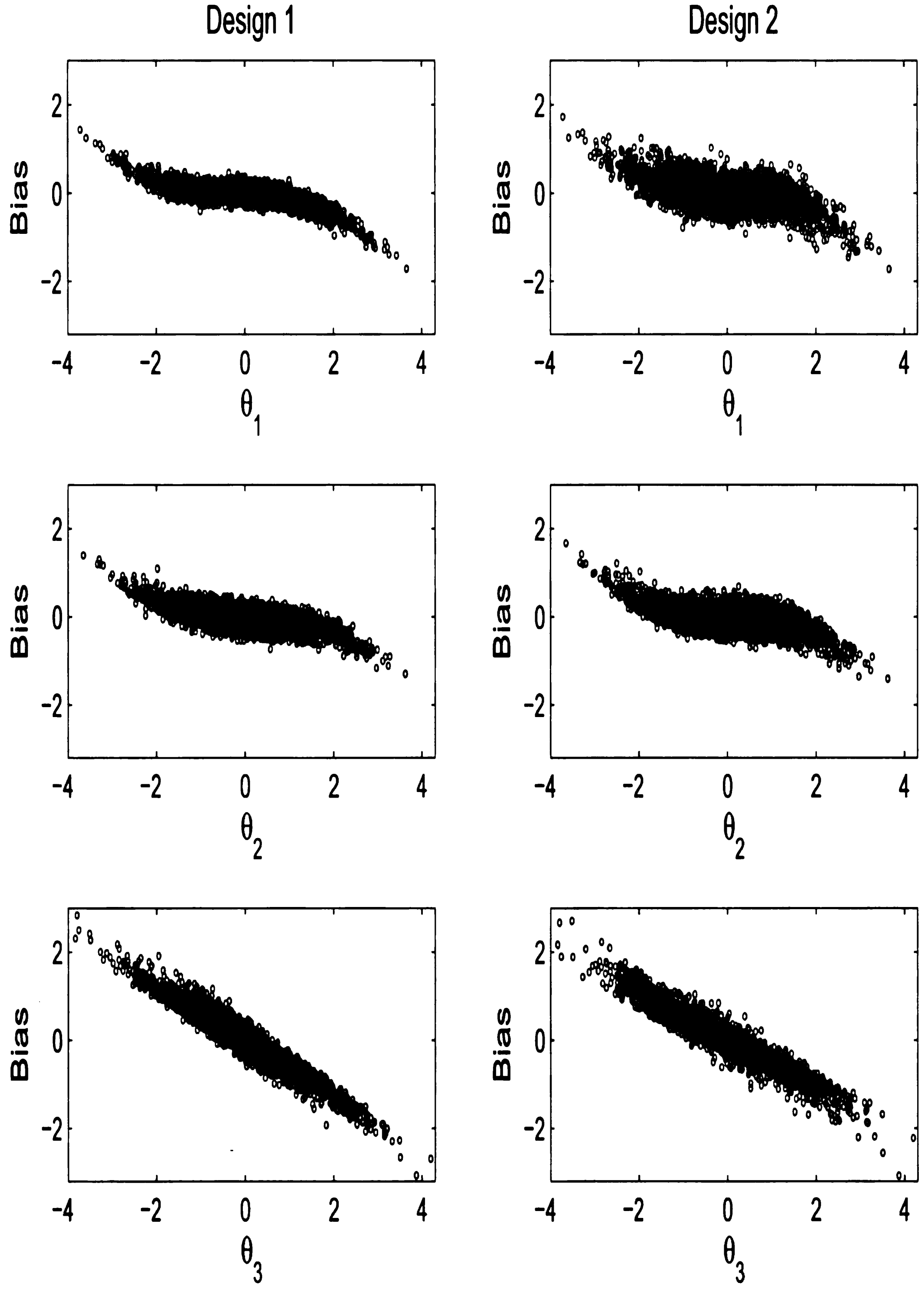

Figure 3.3. Plots of Proficiency Parameters versus Bias

bution in the TESTFACT software. Furthermore, it is easy to see that Bias values for θ_3 recovery dramatically change across different proficiency levels, and the vertical spread of θ_1 Bias in Design 2 is larger than that of the other θ_1 and θ_2 Bias in both designs.

Table 3.6 gives the correlation matrix among raw proficiency parameters and also the average correlation matrix among proficiency estimates across replications. Compared with the roughly zero true correlation, there are small values of correlations among proficiency estimates. This is acceptable for the EAP score since this method shrinks the range of proficiency estimates.

Table 3.6. True and Estimated Proficiency Correlation Matrix

		θ_1	θ_2	θ_3
True	θ_1	1.0000	0.0128	-0.0020
	θ_2	0.0128	1.0000	0.0086
	θ_3	-0.0020	0.0086	1.0000
Recovery from Design 1	θ_1	1.0000	0.0401	0.1713
	θ_2	0.0401	1.0000	0.1504
	θ_3	0.1713	0.1504	1.0000
Recovery from Design 2	θ_1	1.0000	0.0674	0.1057
	θ_2	0.0674	1.0000	0.1656
	θ_3	0.1057	0.1656	1.0000

From all the above descriptive analysis for the parameter recovery, it is obvious that parameters for the third dimension recover much worse than for the first two dimensions, because there is no simple structure item to correctly orient the third coordinate axis. The item characteristic and person proficiency estimates seem misleading especially for this dimension. However, this set of estimates was obtained with the TESTFACT software and it is a valid one among the infinite solutions maximizing the likelihood. The problem is that when there is no simple structure item to define the coordinate axis, it is very difficult to recover the original coordinate system even when the proficiencies are assumed uncorrelated.

3.2.3 Coordinate Axes Recovery

The unsatisfactory recovery of both item and person parameters, especially for the third dimension, reveals the problem that besides the estimation error, the coordinate system may not be well recovered with reference to the original one, which is not surprising due to the indeterminacy of the coordinate system.

Table 3.7 shows the direction cosines of the reference composite vectors from both the item parameters and their estimates from replication 1. One problem found in obtaining the most discriminating direction for the item cluster was how to get the correct direction for the eigenvector. As is well known, the negative of one eigenvector can be regarded as another valid eigenvector for the same matrix. The eigenvector direction is difficult to choose when elements in the eigenvector have mixed signs. In order to get the correctly positioned direction for the projection purpose, the eigenvector was obtained by forcing the vector element with the largest absolute value to be positive.

Table 3.7. Reference Composite Vectors from Parameters and Replication 1 Estimates

		Design 1			Design 2		
		C_1	C_2	C_3	C_1	C_2	C_3
True	θ_1	0.996	0.101	0.614	0.996	0.083	0.616
	θ_2	0.029	0.992	0.614	0.040	0.992	0.559
	θ_3	0.081	0.081	0.496	0.081	0.097	0.555
Repl	θ_1	0.993	0.089	0.585	0.983	0.001	0.650
	θ_2	0.043	0.995	0.643	0.142	0.999	0.641
	θ_3	0.110	0.056	0.495	-0.115	0.046	0.407

Since θ_1 and θ_2 are fairly dominant for the reference composite directions of cluster 1 and cluster 2 respectively, as long as the weights for these proficiencies are still large, both reference composite vectors are well recovered. However, compared with the true reference composite vector for cluster 3 in Design 1, the estimate gives less weight to θ_1 while more weight to θ_2. And for cluster 3 in Design 2, θ_3 is less weighted, while both θ_1 and θ_2 are more weighted. These patterns are consistent across most replications in both designs.

Figure 3.4 shows the graphical presentation of reference composite vectors for both true and estimated parameters in replication 1. The dashed line indicates the vector calculated from the parameters, while the solid line stands for that from the estimates. None of the true and estimated reference composite vectors are exactly overlapped. Despite the large deviation between the true and estimated reference composite vectors for the third cluster, it is not obvious to observe this deviation from this angle.

Therefore, besides the estimation error, one partial explanation for the deviation is that the coordinate system consistently rotates to roughly the same deviated direction during the estimation if the estimated reference composite vectors are regarded as fixed without too much influence from the estimation error or sampling error. The final orientation of the coordinate system chosen by the software may depend on the parameters, the sampling error and the rotation criterion. One extra analysis was conducted to rotate the estimated item discriminations to match their parameters. Although the rotated version and true parameters are not exactly the same due to the nonignorable estimation and sampling errors, this rotated version matches the true parameters much better than the raw estimates given by the software.

Based on the discrepancy between the true and estimated coordinate system in this simulation, one finding is that it is hard to recover the axis orientation of the proficiency dimension without any simple structure item measuring that proficiency.

3.2.4 Item Grouping

Although the true item grouping was known, the cluster analysis on the estimated item discrimination matrix for the first replication in Design 1 was still conducted to show how this analysis works.

Figure 3.5 shows the dendrogram of the linkages among all 45 items for that repli-

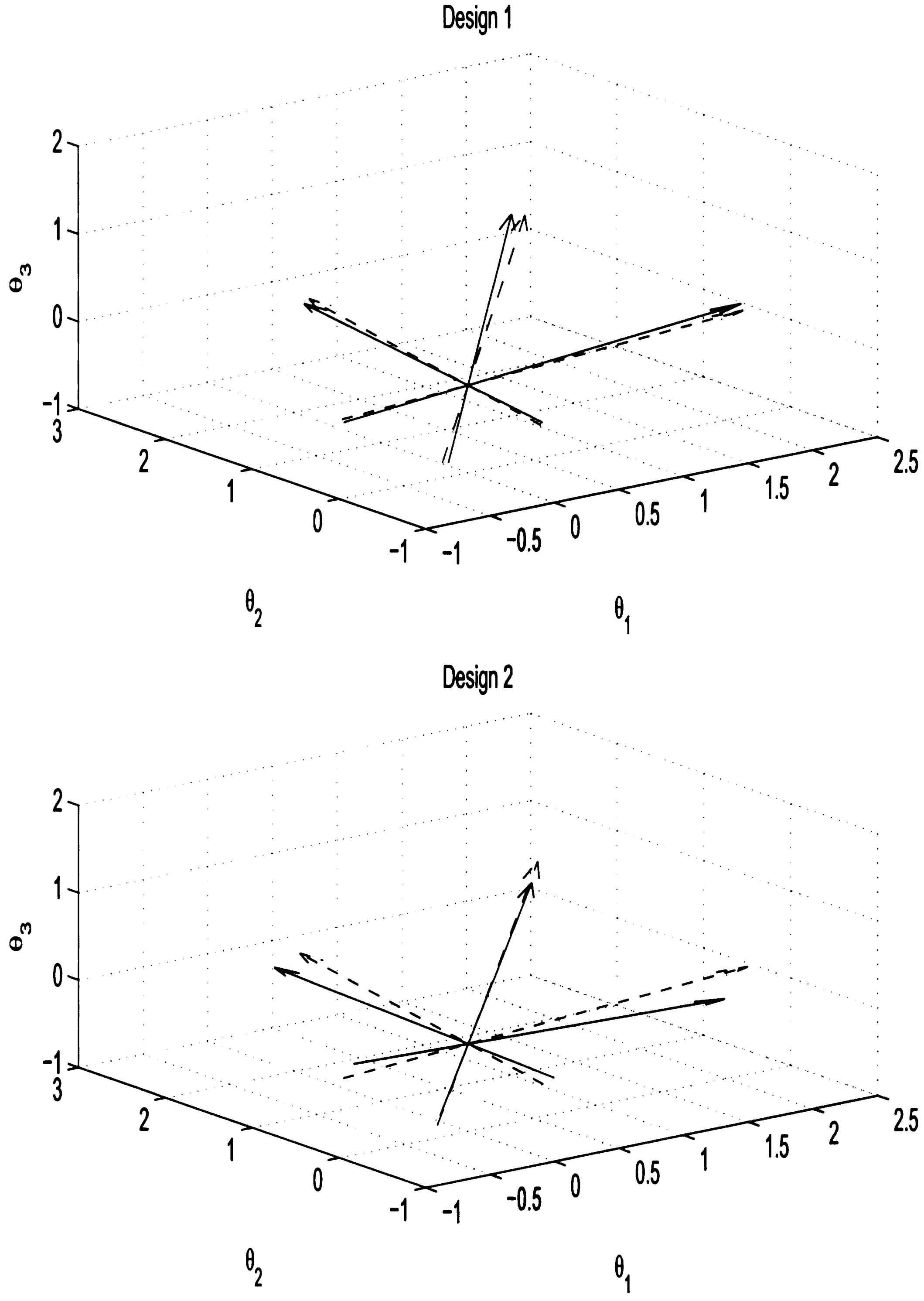

Figure 3.4. Plots of Reference Composite Vectors from Parameters and Replication 1 Estimates

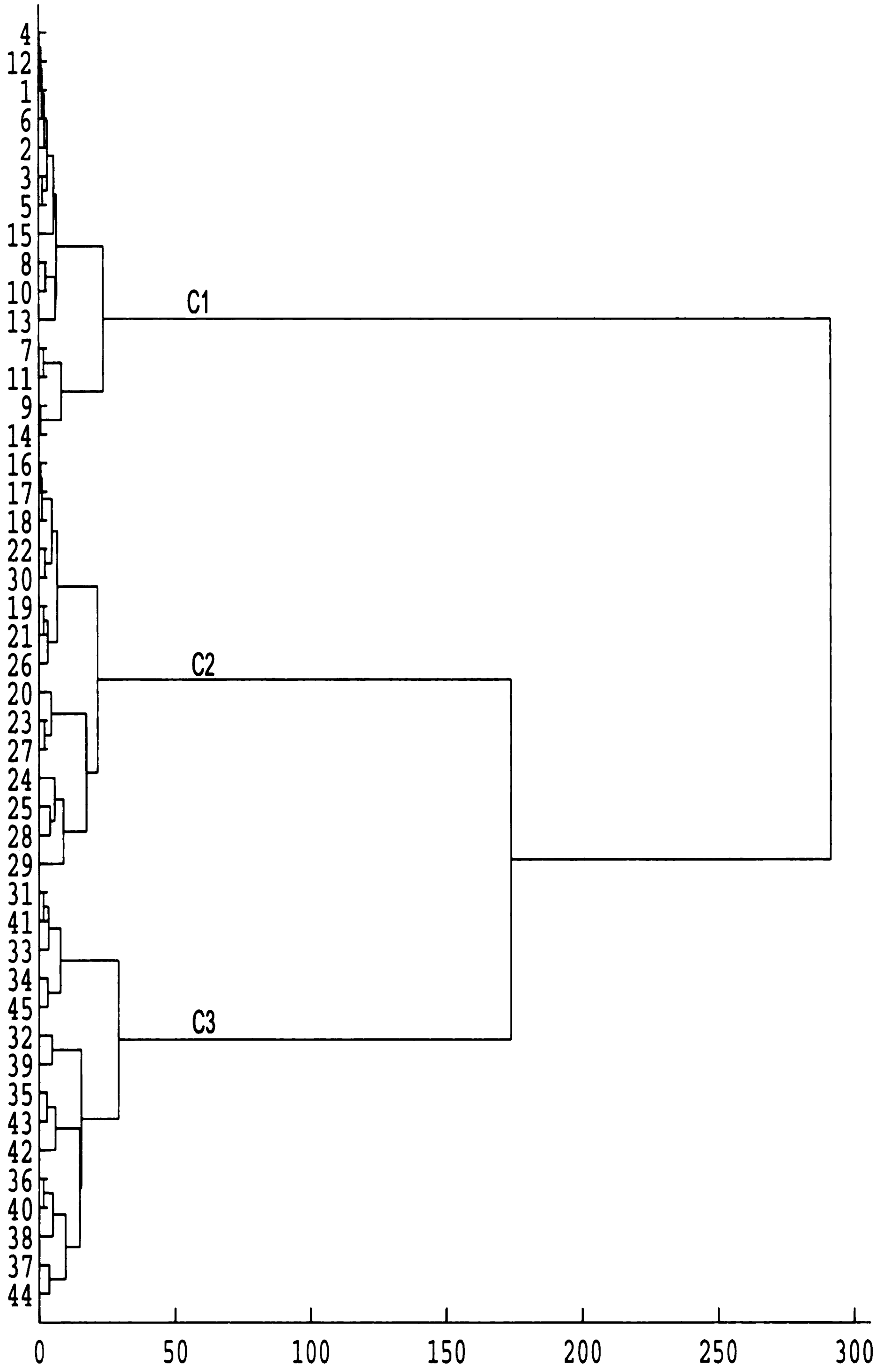

Figure 3.5. Dendrogram for Replication 1 in Design 1

cation. The horizontal axis shows the Ward distance measure, while the vertical axis denotes the item number. The grouping pattern is very clear that items 1-15 belong to the first cluster, items 16-30 to the second cluster, and finally items 31-45 to the third cluster. This result is quite satisfactory; however, it should be kept in mind that this subjective eyeballing may lead to different item grouping, when there are large sampling errors involved with the data or high correlations among the proficiencies. Even combined with the evidence from expert judgement, it may still result in ambiguous item grouping due to different content classification criteria used by these experts.

3.2.5 Stability of Construct Estimates

$\boldsymbol{\theta}^*$ estimates were calculated as the projection of $\boldsymbol{\theta}$ estimates onto the most discriminating direction of each item cluster. Although the orthonormal rotation changes the coordinates of points represented by the coordinate system, their relative position and distance are invariant under this transformation. Therefore, the values of $\boldsymbol{\theta}^*$ estimates do not change whatever orthonormal rotation the coordinate system takes.

Table 3.8 shows the mean, minimum and maximum for the standard deviations of estimated $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ across all persons. From the table, $\boldsymbol{\theta}^*$ estimates seem to be more stable than $\boldsymbol{\theta}$ estimates, especially for the third element. Also the variation of the third construct estimates is smaller than that of the other two construct estimates while the opposite pattern occurs in the raw proficiency estimates. Besides the sampling and estimation error, $\boldsymbol{\theta}$ estimates are largely influenced by the orientation of the coordinate axes. However, the values of $\boldsymbol{\theta}^*$ estimates do not change no matter what orthonormal rotation of the coordinate system representing the $\boldsymbol{\theta}$ estimates. This invariance is especially important in recovering the weighted composite, where more than one element of $\boldsymbol{\theta}$ plays a significant role.

Figure 3.6 illustrates the relationship between the true $\boldsymbol{\theta}^*$ and the variations of

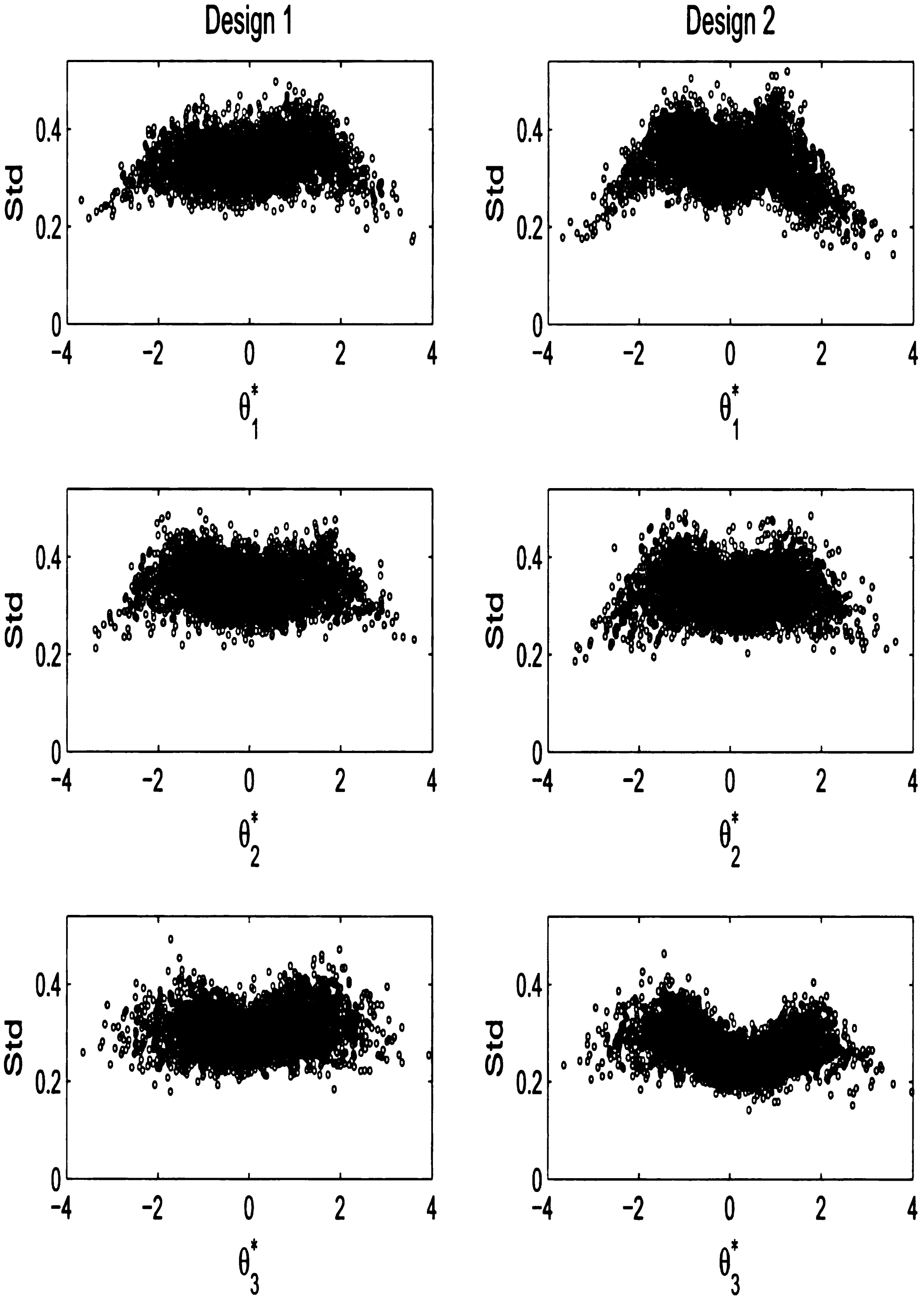

Figure 3.6. Plots of Construct Parameters versus Estimated Standard Deviations

Table 3.8. Summary of Standard Deviations of Raw Proficiency and Construct Estimates

		θ_1	θ_2	θ_3	θ_1^*	θ_2^*	θ_3^*
Design 1	Mean	0.3500	0.3484	0.4975	0.3371	0.3400	0.3028
	Min	0.1764	0.2029	0.2868	0.1706	0.2121	0.1790
	Max	0.5371	0.5144	0.7430	0.4979	0.4937	0.4929
Design 2	Mean	0.3426	0.3312	0.4810	0.3378	0.3308	0.2634
	Min	0.1820	0.1760	0.1758	0.1417	0.1866	0.1424
	Max	0.5330	0.4870	0.6788	0.5200	0.4938	0.4635

their estimates. All the shapes look roughly like the letter "M", which indicates that the variation for the extreme and middle values is slightly smaller than for the values between them. This is reasonable since the EAP scoring method restricts the possible range of proficiency estimates, and the middle-value proficiencies are better estimated. The reason for the latter is that the average test difficulty is also around the middle values and the EAP estimates for these middle-value proficiencies are more stable by incorporating the prior standard normal density. It can also be observed from the figure that points for the third construct are closer to the horizontal axis than those for the other constructs. This phenomenon reveals that the variation of θ_3^* estimates is slightly smaller than for θ_1^* and θ_2^* estimates, and this can also be observed from Table 3.8.

3.2.6 Relationship with NC Subscores and Unidimensional Estimates

As a new method to obtain subscores related to different item clusters, it was assumed that construct estimates should be highly correlated with the commonly used NC subscores and unidimensional estimates, especially when the variation of directional degree angles for the items in the same cluster is very small.

Table 3.9 shows the average correlations among $\boldsymbol{\theta}^*$ estimates, $\boldsymbol{\theta}_u$ estimates and NC subscores across all replications. Obviously, all the correlations are very high, with the lowest value as 0.9450. For $\boldsymbol{\theta}^*$ estimates, all their correlations with $\boldsymbol{\theta}_u$ estimates are higher than their corresponding ones with NC subscores, since both $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_u$ are estimated using the models which take the item discrimination and difficulty into consideration. However, for $\boldsymbol{\theta}_u$ estimates, they seem to be more closely related to NC subscores rather than to $\boldsymbol{\theta}^*$ estimates, and the reason may be the assumption of unique proficiency for both estimations.

Table 3.9. Average Correlation Among Construct Estimates, Unidimensional Estimates and NC Subscores

		Design 1			Design 2		
		θ^*	θ_u	NC	θ^*	θ_u	NC
Cluster 1	θ^*	1.0000	0.9847	0.9726	1.0000	0.9783	0.9701
	θ_u	0.9847	1.0000	0.9852	0.9783	1.0000	0.9905
	NC	0.9726	0.9852	1.0000	0.9701	0.9905	1.0000
Cluster 2	θ^*	1.0000	0.9796	0.9729	1.0000	0.9804	0.9638
	θ_u	0.9796	1.0000	0.9911	0.9804	1.0000	0.9801
	NC	0.9729	0.9911	1.0000	0.9638	0.9801	1.0000
Cluster 3	θ^*	1.0000	0.9633	0.9450	1.0000	0.9806	0.9715
	θ_u	0.9633	1.0000	0.9799	0.9806	1.0000	0.9801
	NC	0.9450	0.9799	1.0000	0.9715	0.9801	1.0000

Figures 3.7 and 3.8 show the scatterplots of $\boldsymbol{\theta}^*$ estimates with $\boldsymbol{\theta}_u$ estimates and with NC subscores separately. Clearly there are lower and upper bounds, which are respectively 0 and total number of items in that cluster, for NC subscores. Observed from the figures, the range of $\boldsymbol{\theta}^*$ estimates and $\boldsymbol{\theta}_u$ estimates is restricted by the EAP scoring method, since none of the absolute proficiency estimates exceeds 3. It is also clear that, for the people located at either tail of the proficiency distribution, they are more differentiated with the $\boldsymbol{\theta}^*$ estimates than with the NC subscores, and slightly more than with the $\boldsymbol{\theta}_u$ estimates.

3.2.7 Accuracy of Construct Estimates

$\boldsymbol{\theta}^*$ parameters were created by projecting $\boldsymbol{\theta}$ parameters onto different reference composites for the three clusters. They were regarded as the reference for comparing the

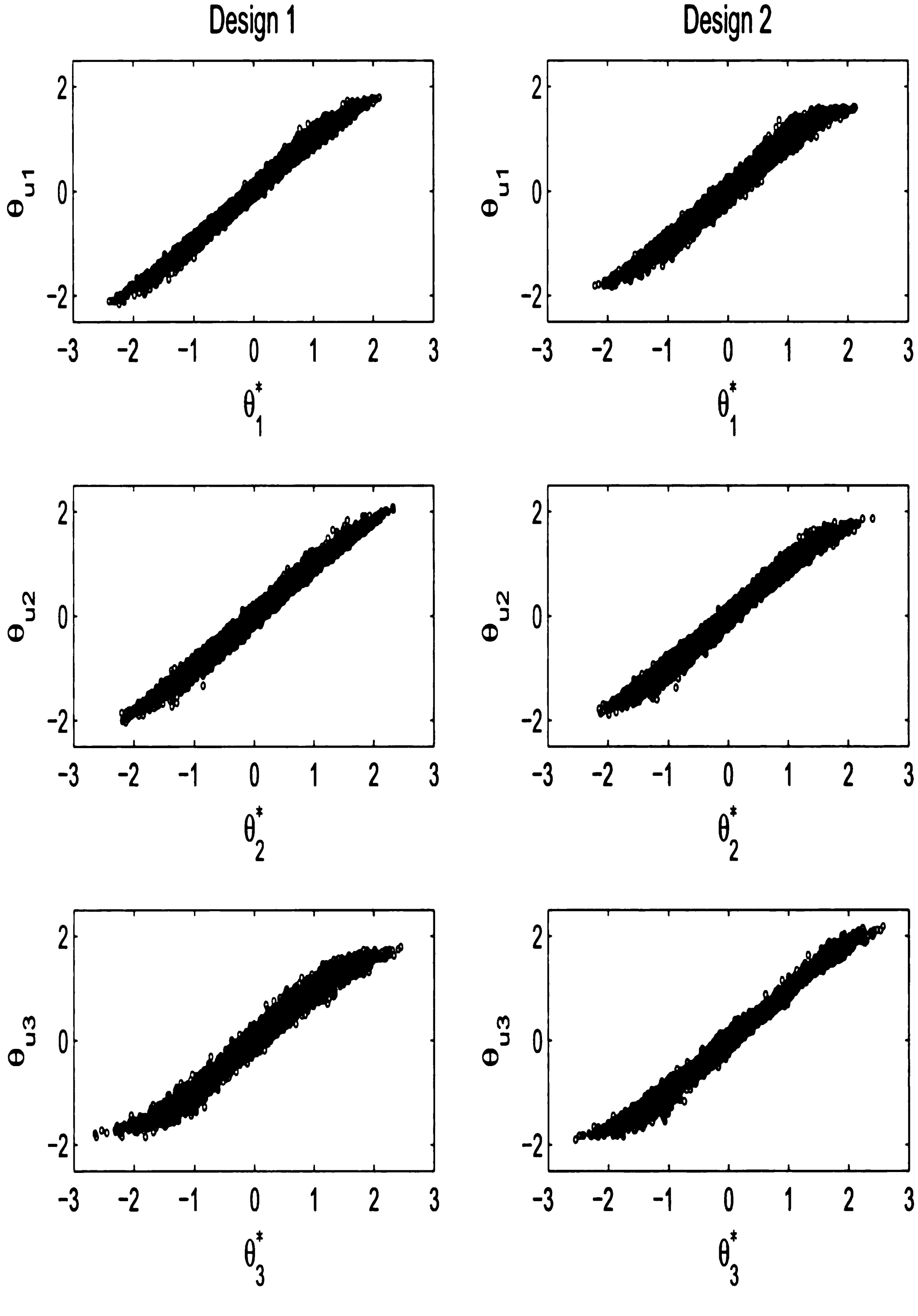

Figure 3.7. Plots of Average Construct Estimates versus Average Unidimensional Estimates across all Replications

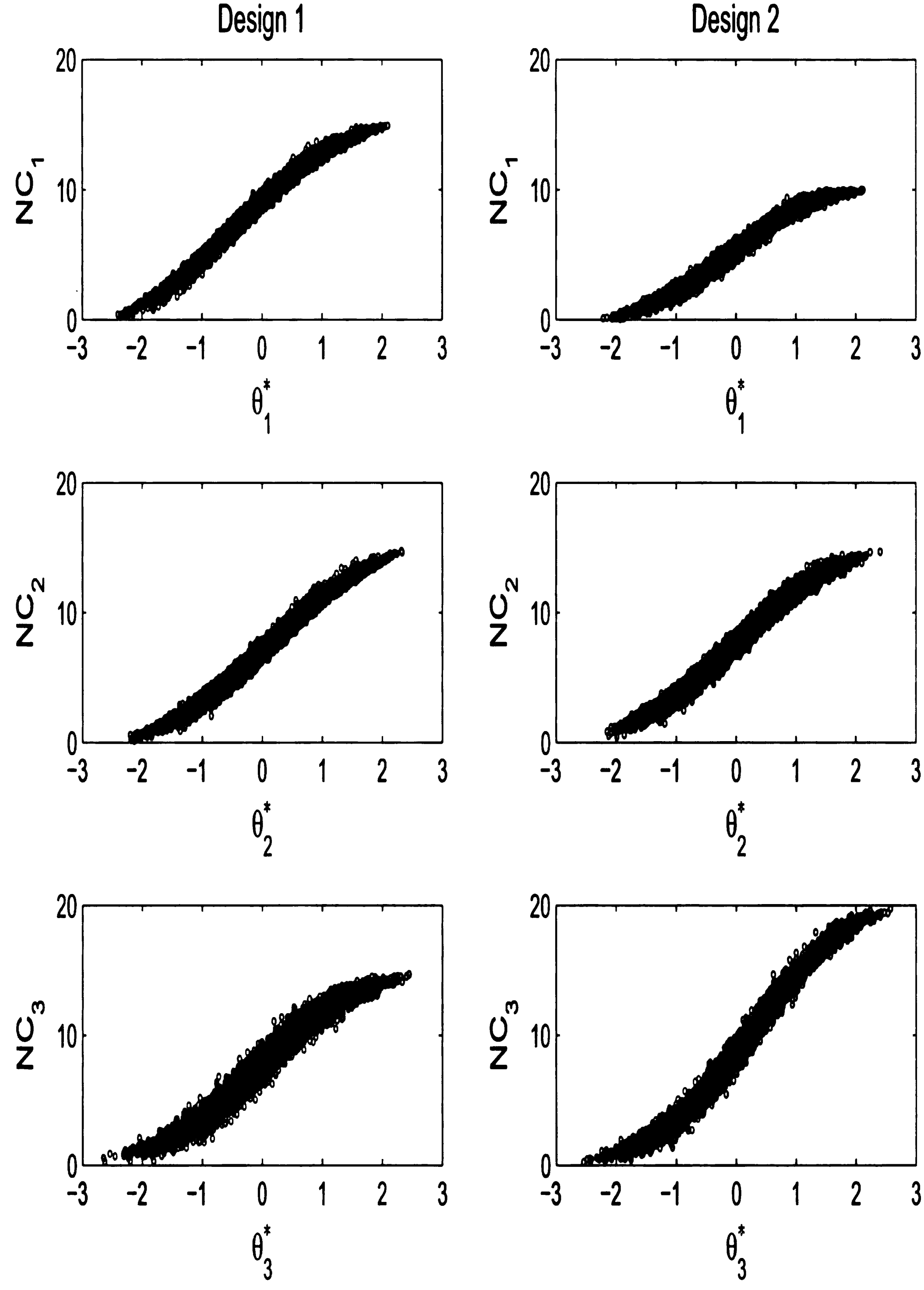

Figure 3.8. Plots of Average Construct Estimates versus Average NC Subscores across all Replications

recovery efficiency by $\boldsymbol{\theta}^*$ estimates with that by unidimensional estimates.

Table 3.10 shows the average correlation across all replications and average Bias/RMSE across all persons for the recovery of $\boldsymbol{\theta}^*$ parameters by the estimated $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_u$. As in previous discussion for the $\boldsymbol{\theta}$ recovery, all the Bias values in this table are negative due to the fact that more people have positive $\boldsymbol{\theta}^*$ values. Also, because of the EAP score shrinkage, all the RMSEs show the large deviations between the estimated and true $\boldsymbol{\theta}^*$.

Table 3.10. Construct Recovery from Unidimensional Estimates and Construct Estimates

		unidimensional $\boldsymbol{\theta}_u$			multidimensional $\boldsymbol{\theta}^*$		
		θ_{u1}	θ_{u2}	θ_{u3}	θ_1^*	θ_2^*	θ_3^*
Design 1	Correlation	0.9121	0.9103	0.9098	0.9243	0.9217	0.9385
	Bias	-0.0192	-0.0394	-0.0446	-0.0312	-0.0347	-0.0483
	RMSE	0.3988	0.4085	0.4041	0.3742	0.3833	0.3421
Design 2	Correlation	0.8984	0.9134	0.9385	0.9150	0.9251	0.9558
	Bias	-0.0169	-0.0394	-0.0438	-0.0130	-0.0373	-0.0362
	RMSE	0.4255	0.3973	0.3381	0.3926	0.3732	0.2909

From the table, all the scale-free correlation indices for the recovery by $\boldsymbol{\theta}^*$ estimates are consistently higher than for the corresponding recovery by $\boldsymbol{\theta}_u$ estimates, and their RMSEs are also lower than those from $\boldsymbol{\theta}_u$ estimates. Both the correlation and the RMSE indicate that $\boldsymbol{\theta}^*$ estimates perform better for the recovery of $\boldsymbol{\theta}^*$ parameters. Compared with previous recovery of $\boldsymbol{\theta}$ parameters in Table 3.5, the recovery of $\boldsymbol{\theta}^*$ is much better by either $\boldsymbol{\theta}^*$ estimates or $\boldsymbol{\theta}_u$ estimates, since their values are not influenced by the orthonormal rotation of the coordinate system.

Table 3.11 gives the hypothesis tests on the difference of correlations for $\boldsymbol{\theta}^*$ parameter recovery by $\boldsymbol{\theta}^*$ estimates and $\boldsymbol{\theta}_u$ estimates. With each replication as one observation, there were 50 observations for each hypothesis test, and the test was conducted separately for different designs and clusters. There is no testing for Bias or RMSE, since they were already averaged across replications. Paired tests were used because the same dataset was calibrated to give both the multidimensional and unidimensional estimates. Also, due to the possible nonnormality, the nonparametric sign test and commonly used Fisher's r-to-z transformation in Equation 3.3 were applied in the test.

$$z = \frac{1}{2} \log \frac{1+r}{1-r} \tag{3.3}$$

Table 3.11. Hypothesis Testing on the Difference of Correlations for Construct Recovery by Construct Estimates and Unidimensional Estimates

Design	Cluster	Scale	Method	Mean	Std Error	Statistics	p-value
1	1	Raw	Paired t	0.0122	0.0002	71.6861	0.000
		Fisher	Paired t	0.0778	0.0011	70.2359	0.000
		Raw	Signtest			6.9296	0.000
	2	Raw	Paired t	0.0114	0.0021	5.4395	0.000
		Fisher	Paired t	0.0762	0.0091	8.3590	0.000
		Raw	Signtest			6.6468	0.000
	3	Raw	Paired t	0.0287	0.0021	13.8690	0.000
		Fisher	Paired t	0.2073	0.0126	16.4260	0.000
		Raw	Signtest			6.0811	0.000
2	1	Raw	Paired t	0.0166	0.0003	47.7248	0.000
		Fisher	Paired t	0.0938	0.0020	47.7994	0.000
		Raw	Signtest			6.9296	0.000
	2	Raw	Paired t	0.0117	0.0025	4.7332	0.000
		Fisher	Paired t	0.0826	0.0106	7.8124	0.000
		Raw	Signtest			6.6468	0.000
	3	Raw	Paired t	0.0173	0.0001	131.6440	0.000
		Fisher	Paired t	0.1694	0.0012	138.7443	0.000
		Raw	Signtest			6.9296	0.000

Therefore, each testing consists of three different tests, which include the raw scale paired t-test, the transformed scale paired t-test, and the raw scale nonparametric sign test. From the table, it is easy to see the p-values in the last column are all 0.000's, which indicates all the tests are highly significant; therefore, the recovery by $\boldsymbol{\theta}^*$ estimates performs significantly better than $\boldsymbol{\theta}_u$ estimates.

Figures 3.9 and 3.10 illustrate the Bias for the $\boldsymbol{\theta}^*$ recovery by the estimated $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_u$ for Design 1 and 2 respectively. As before, parameters with large values are underestimated, and the pattern is the opposite for parameters with small values. The different vertical spread for recovery Bias from both estimates indicates that there is

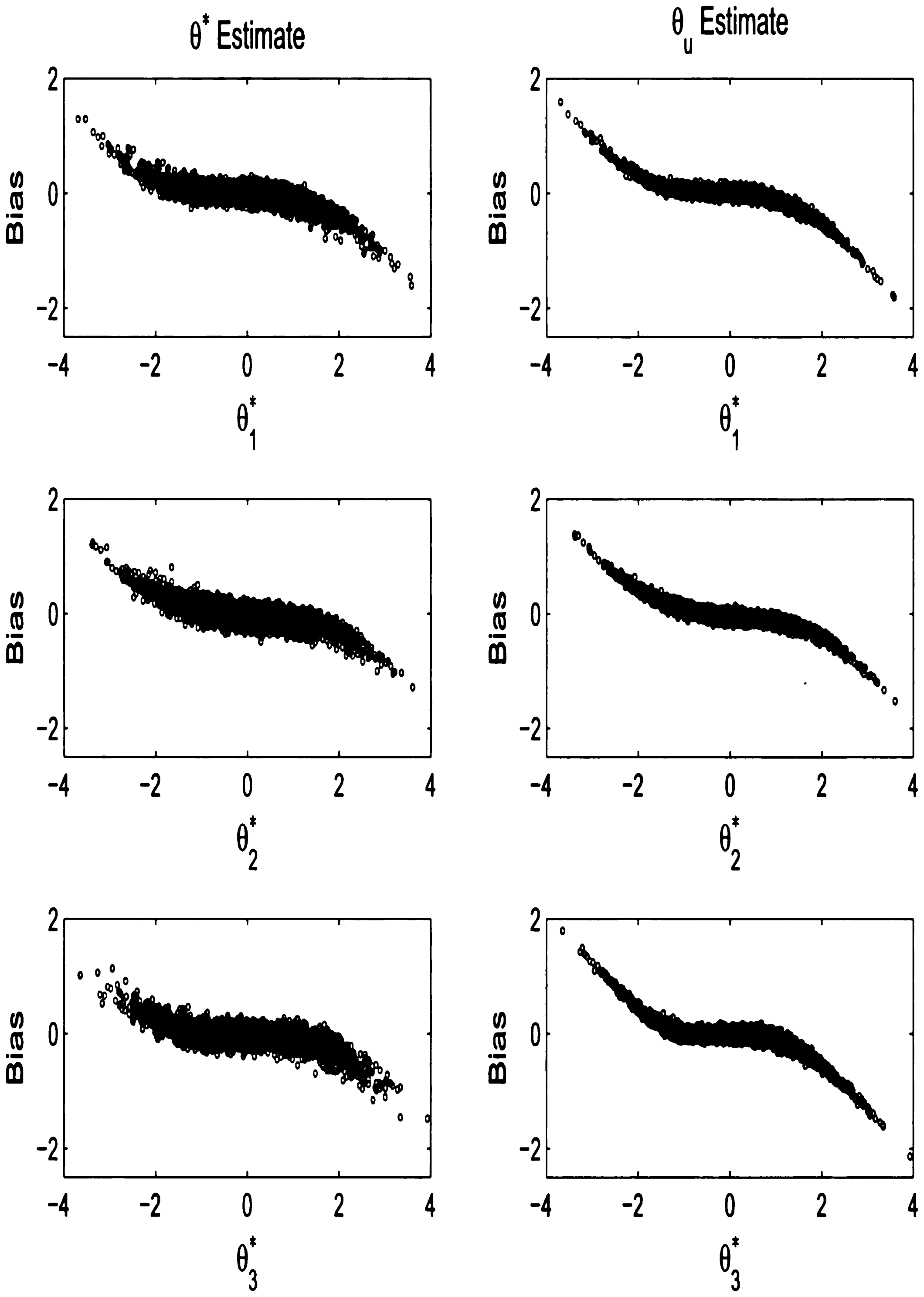

Figure 3.9. Plots of Recovery Bias by Construct Estimates and Unidimensional Estimates in Design 1

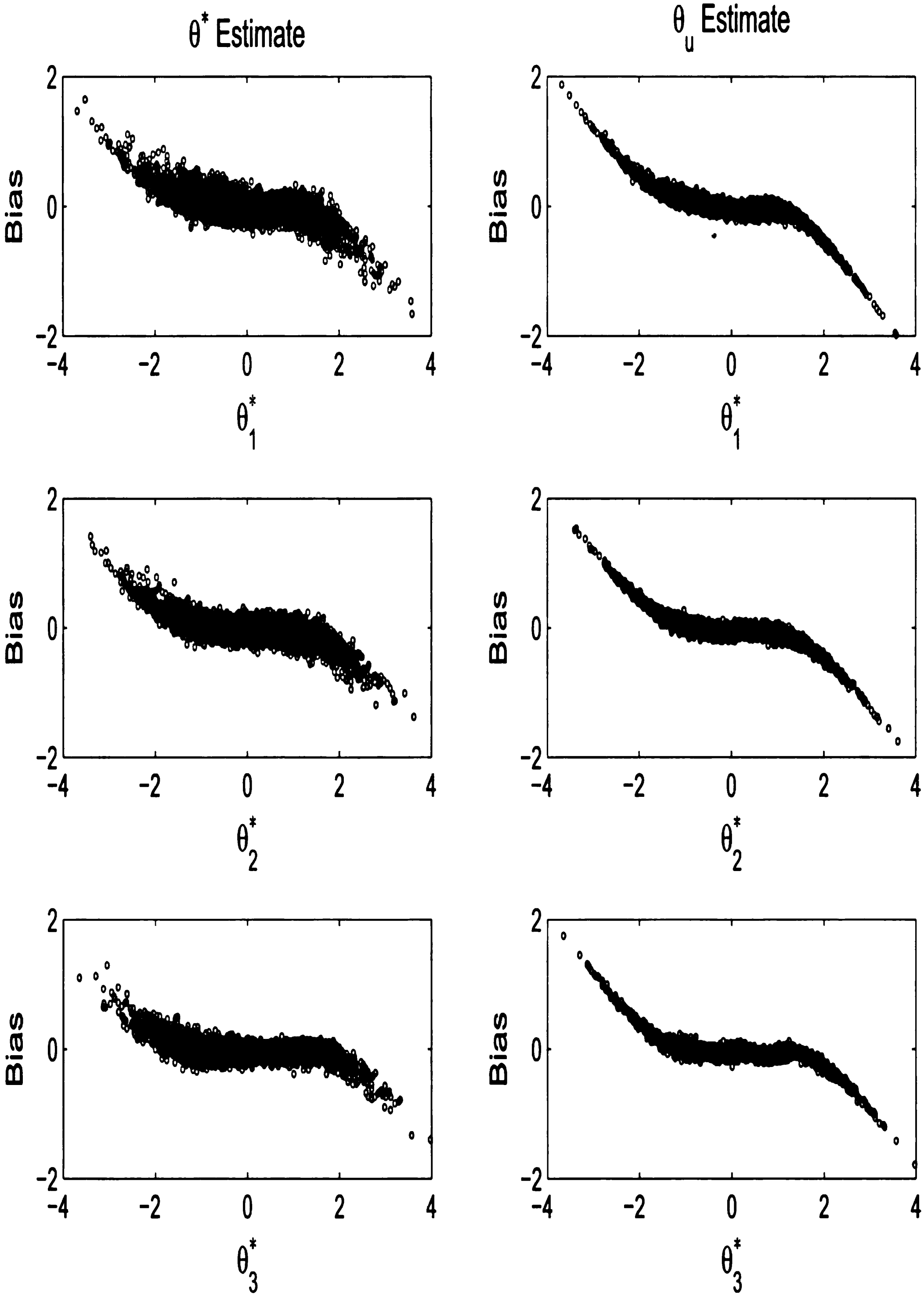

Figure 3.10. Plots of Recovery Bias by Construct Estimates and Unidimensional Estimates in Design 2

more recovery variation for $\boldsymbol{\theta}^*$ estimates than $\boldsymbol{\theta}_u$ estimates, while the absolute Bias values by $\boldsymbol{\theta}^*$ estimates are smaller for the parameters somewhat distant from the middle.

3.2.8 Correlation Recovery

The correlations among $\boldsymbol{\theta}^*$ parameters can be recovered by $\boldsymbol{\theta}^*$ estimates, and in theory, the correlation values depend on how close the reference composite vectors for different item clusters are. Although the TESTFACT software uses the cov$\boldsymbol{\theta} = I_{m \times m}$ assumption for the proficiency estimation, it is likely that there are still small values of correlations among $\boldsymbol{\theta}$ estimates due to the sampling errors and the EAP scoring method, which is already shown in previous Table 3.6.

Table 3.12 shows the correlation matrix recovered from $\boldsymbol{\theta}_u$ estimates, the Promax method and $\boldsymbol{\theta}^*$ estimates, and these correlation matrices were averaged across all replications. The Promax method is based on the result of the factor loadings after the Varimax rotation, whose computation involves the eigenvector calculation in the TESTFACT software. As previously explained, an eigenvector is free to get the sign changed. In order to reduce the effect of this indeterminacy, all correlations from the Promax method were forced to take positive values.

From the table, correlations are underestimated by either unidimensional EAP estimates for $\boldsymbol{\theta}_u$ or the Promax method, while they are overestimated by the EAP estimates for $\boldsymbol{\theta}^*$. It is not surprising that $\boldsymbol{\theta}_u$ estimates give the lowest correlation related to θ_3^*, since the unidimensional calibration does not take it into consideration that even for items in the same cluster, they can most discriminate along slightly different directions and the angular variation in the third cluster is larger than in other clusters. The correlation matrix given by the Promax method is obtained from the rotation matrix, which converts the Varimax transformed factor loadings to be more like a simple structure solution. It is obvious to see that this method still

Table 3.12. Correlation Recovery by Unidimensional Estimates, the Promax Method and Construct Estimates

		Design 1			Design 2		
		θ_1^*	θ_2^*	θ_3^*	θ_1^*	θ_2^*	θ_3^*
True	θ_1^*	1.0000	0.1489	0.6714	1.0000	0.1423	0.6819
	θ_2^*	0.1489	1.0000	0.7173	0.1423	1.0000	0.6659
	θ_3^*	0.6714	0.7173	1.0000	0.6819	0.6659	1.0000
Uni	θ_1^*	1.0000	0.1210	0.5466	1.0000	0.1191	0.5799
	θ_2^*	0.1210	1.0000	0.6006	0.1191	1.0000	0.5685
	θ_3^*	0.5466	0.6006	1.0000	0.5799	0.5685	1.0000
Promax	θ_1^*	1.0000	0.0995	0.5715	1.0000	0.0737	0.6079
	θ_2^*	0.0995	1.0000	0.6300	0.0737	1.0000	0.5960
	θ_3^*	0.5715	0.6300	1.0000	0.6079	0.5960	1.0000
Multi	θ_1^*	1.0000	0.1931	0.7126	1.0000	0.1991	0.7437
	θ_2^*	0.1931	1.0000	0.7676	0.1991	1.0000	0.7206
	θ_3^*	0.7126	0.7676	1.0000	0.7437	0.7206	1.0000

cannot recover the correlations among $\boldsymbol{\theta}^*$ parameters. The slight overestimation of the correlations by $\boldsymbol{\theta}^*$ estimates is expected since there are already small values of correlations among $\boldsymbol{\theta}$ estimates obtained using the EAP scoring method.

CHAPTER 4

Real Data Applications

4.1 Test Information and Analysis Procedure

In this study, the fall 2007 grade 7 mathematics test from the Michigan Educational Assessment Program (MEAP) was used for the real data analysis. To meet the requirement of the federal No Child Left Behind (NCLB) Act, the MEAP is developing tests of English language arts, mathematics, science and social studies, the first two of which are the core subjects used to measure the Adequate Yearly Progress (AYP) of students in the elementary and middle schools. This AYP is used to track year-to-year student achievement, and most importantly, after combined with other indicators, e.g., attendance rate, it is applied to each school and district to evaluate whether their AYP goal is achieved.

The MEAP mathematics test is administered at the beginning of each school year, and it actually covers the contents of the preceding grade level. There are five strands in the mathematics tests, within which multiple domains are measured. The contents of these strands and domains are shown in Table 4.1, and the emphasis for them varies across different grade levels due to different curriculum contents and standard expectations.

Table 4.2 specifies the content classification for each item in this fall 2007 grade 7 mathematics test, and Table 4.3 summarizes the number and percentage of the items within each category. This grade 7 mathematics test actually measures the students' learning in grade 6, and one general scaled score is assigned to each student, which is later categorized into four different achievement levels: Not Proficient, Partially

Table 4.1. Strands and Domains for Michigan Mathematics Content Expectation

Strand	Domain	
Numbers and Operations	ME	Meaning, notation, place value, and comparison
	MR	Number relationships and meaning of operations
	FL	Fluency with operations and estimation
Algebra	PA	Patterns, relations, functions, and change
	RP	Representation
	FO	Formulas, expressions, equations, and inequalities
Measurement	UN	Units and systems of measurement
	TE	Techniques and formulas for measurement
	PS	Problem solving involving measurement
Geometry	GS	Geometric shape, properties and mathematical arguments
	LO	Locations and spatial relationships
	SR	Spatial reasoning and geometric modeling
	TR	Transformation and symmetry
Data and Probability	RE	Data representation
	AN	Data interpretation and analysis
	PR	Probability

Proficient, Proficient and Advanced. Besides this score, the NC subscores are also given according to the five strands specified in Table 4.1. It should be noted that all these MEAP test and item information are retrieved from the web pages of the Michigan Department of Education.

Based on Table 4.3, N and A are the two main strands that are measured in this grade, and both of them cover 80% of the items on the whole test. The other strands M, G and D only contain 3, 6 and 3 items. Since there are only a few items for these three strands, their subscores are theoretically unreliable to be reported if the item responses in each strand are calibrated separately from those in other strands.

This test was designed to measure different constructs related to these content strands or domains; therefore, it can be regarded as a multidimensional test, and analyzed with the MIRT projection method to improve the subscore estimation.

For the real data, the total population size is 124,674 and the number of items for the core test is 60. After listwise deletion for the missing data, the population size reduced to 124,641. The classical p-value for the difficulty of each item was scrutinized, which was calculated as the percentage of people correctly answering this

Table 4.2. Item Content of the Fall 2007 Grade 7 MEAP Mathematics Test

Item	Content Classification	Item	Content Classification
1	N-FL	31	D-PR
2	N-FL	32	D-PR
3	N-FL	33	D-PR
4	N-FL	34	N-MR
5	N-FL	35	N-MR
6	N-FL	36	N-MR
7	N-MR	37	N-FL
8	N-MR	38	N-FL
9	N-MR	39	N-FL
10	N-FL	40	N-FL
11	N-FL	41	N-FL
12	N-FL	42	N-FL
13	N-ME	43	N-ME
14	N-ME	44	N-ME
15	N-ME	45	N-ME
16	A-PA	46	A-FO
17	A-PA	47	A-FO
18	A-PA	48	A-FO
19	A-RP	49	A-FO
20	A-RP	50	A-FO
21	A-RP	51	A-FO
22	A-FO	52	A-FO
23	A-FO	53	A-FO
24	A-FO	54	A-FO
25	A-FO	55	M-UN
26	A-FO	56	M-UN
27	A-FO	57	M-UN
28	G-GS	58	G-TR
29	G-GS	59	G-TR
30	G-GS	60	G-TR

Table 4.3. Strand and Domain Percentage in the Test

Strand	Count	Percentage	Domain	Count	Percentage
N	27	45%	ME	6	10%
			MR	6	10%
			FL	15	25%
A	21	35%	PA	3	5%
			RP	3	5%
			FO	15	25%
M	3	5%	UN	3	5%
G	6	10%	GS	3	5%
			TR	3	5%
D	3	5%	PR	3	5%

item. Item 57 was found to be problematic based on this criterion, because only 8.18 percent of people answered it correctly. The question is as following,

57. What is the total number of square inches in 5 square feet?

A 25

B 60

C 300

D 720

This is one of the only three measurement questions, and the correct answer is D. Clearly, this item was a little difficult for the students. With the BILOG software, the item-test pearson correlation for this item was 0.039, and the classical biserial correlation was 0.071. Both indices were too low, so this item was deleted from the analysis.

Table 4.4 shows the descriptive statistics of the NC total scores and difficulty p-values for the 124,641-by-59 response matrix, which was later used in the subscore estimation.

Table 4.4. Summary Statistics for the Population Data

	Count	Mean	Std	Min	Max
Total Score (Person)	124,641	34.5403	11.6417	0	59
Difficulty p-value (Item)	59	0.5854	0.1518	0.2542	0.8691

In the first analysis for this test data, guessing parameters were estimated from a unidimensional calibration using the BILOG software and these estimates were input to the NOHARM and TESTFACT software for the MIRT calibration. However, with these guessing values, both software gave unusual results, for example, very high value of the item discrimination. Another common method proposed by Lord (1980) was also applied to handle the guessing effect. With this method, all the guessing parameters were fixed to a reasonable value of 0.2 and input into both software. This constant value 0.2 was calculated as $1/(1+n)$, where n was the number of choice options for each item and four was the number for this MEAP data. Unfortunately, this method didn't give any better result. The reason for these bad estimations may be that these guessing parameters were not estimated simultaneously with other item parameters; instead, they were regarded as fixed for the MIRT calibrations. Therefore, although people were supposed to have a chance to guess the answer, at least, it seemed that neither the guessing estimates from the unidimensional model nor the commonly used fixed guessing values fit the MIRT calibration for this MEAP data. Hence, this guessing effect was ignored from the later analysis.

In the following analysis, two datasets were designed to be used. One was the population response matrix, with 124,641 rows and 59 columns. The other was the 5,000-by-59 sample matrix, where people were drawn randomly and without replacement from the population. This sample data were used when the software had difficulty handling the large dataset, and they were also applied for confirming the item grouping structure obtained from the population data.

The big problem for the MIRT calibration is that the number of dimensions is unknown for this real data; therefore, some procedures were applied to determine the dimensionality of the response matrix, or more precisely and conservatively, to find a suitable number of dimensions for the MIRT calibration. First, the DIMTEST software was used on the sample data to test whether the fit of the unidimensional model to the data was rejected or not (Stout et al., 1999). After this hypothesis testing was rejected, the parallel analysis was used to compare the observed eigenvalues with those from the randomly simulated data with the same size and difficulty p-values as the population response matrix (Ledesma & Valero-Mora, 2007). In their study, the number of dimensions was determined by the number of observed eigenvalues from the population data that were larger than the corresponding 95^{th} percentile eigenvalues from a large number of random data samples. In this study, the eigenvalues from all random samples were fairly consistent; therefore, only ten random datasets were simulated and the number of dimensions was determined by the number of observed eigenvalues that were larger than the corresponding ones from all the samples. After the MIRT calibration with this number as the dimensionality, the cluster analysis was performed on the angular dissimilarity matrix from the item discrimination estimates for further analysis on dimensionality.

After the number of dimensions was finally decided for the analysis, the calibration and cluster analysis were gone through one by one for the population data, where the NOHARM software was used for the MIRT calibration due to the huge dataset. Since the item grouping was supposed to be consistent across different samples, the same procedures were applied on the sample data to double check the item clustering.

The raw proficiency coordinates were estimated by implementing the item estimates from the NOHARM software into the TESTFACT software. Then reference composite vectors and $\boldsymbol{\theta}^*$ estimates were calculated. Finally, the summary statistics for $\boldsymbol{\theta}^*$ estimates were reported and the relationship between them and the NC subscores was investigated.

4.2 Dimensionality and Cluster Detection

The DIMTEST software is commonly used for the nonparametric hypothesis testing of whether the response data can be fitted by the unidimensional model (Stout, 1987; Stout et al., 1999; Stout, Froelich, & Gao, 2001).

With this software, all items are divided into a partitioning subtest (PT) and an assessment subtest (AT), where the items are chosen carefully to be dimensionally distinct for these two subtests. The PT test provides the person proficiency estimate, which is used as the conditioning variable for the conditional covariance calculation for items in the AT test. The analysis on the AT test gives a T_L statistic, which is later found to be positively biased. In order to correct the bias, $\overline{T}_G$ is calculated as the average from many simulated data sets that match the observed data. Then the difference between T_L and $\overline{T}_G$ is standardized and compared with the normal distribution. The large value of this test statistic indicates the rejection of unidimensionality. From Table 4.5, this test statistic T is 21.2236, and the p-value is 0.0000. This led to the rejection of unidimensionality for the sample data, which also implicitly indicated the multidimensionality for the population data.

Table 4.5. Results from the DIMTEST Software

T_L	$\overline{T}_G$	T	p-value
26.9574	5.6279	21.2236	0.0000

For the parallel analysis, the TESTFACT software was used to calculate the eigenvalues for the tetrachoric correlation matrix, which is more often used for dichotomous data than the pearson correlation designed for continuous variables. Table 4.6 gives the eigenvalues for the population data and ten random datasets of the same size.

Figure 4.1 shows the scree plot of these eigenvalues, and the line marked with circles is for the population data. Clearly, the eigenvalues for the population data drop dramatically from the first factor to the second one and much slower thereafter. For any random dataset, it seems that all the eigenvalues are roughly the same and lie in a horizontal straight line, especially from the third one to the end. With this plot, different conclusions for dimensionality could be drawn based on different criteria. The commonly used eigenvalue-bigger-than-one criterion gave nine dimensions, and the "elbow" rule may indicate two or four dimensions. With the criterion from the parallel analysis, the preliminary result for the dimensionality was eight.

With eight as the number of dimensions, the NOHARM software was used to calibrate the population data. Then the cluster analysis was conducted on the angular dissimilarity matrix estimated from the Varimax transformed item discrimination matrix. The result of item grouping is shown in Figure 4.2.

From the figure, many small clusters are formed by neighboring items, which in-

Table 4.6. Eigenvalues for the Population Data and Ten Random Datasets

	Eigenvalue									
	1	2	3	4	5	6	7	8	9	10
Population	17.064	2.903	2.286	1.666	1.448	1.332	1.147	1.058	1.040	0.987
1	2.034	1.801	1.049	1.046	1.043	1.039	1.036	1.035	1.032	1.031
2	2.204	1.882	1.052	1.047	1.040	1.037	1.032	1.030	1.026	1.025
3	1.964	1.963	1.055	1.052	1.045	1.039	1.034	1.032	1.031	1.028
4	2.072	1.329	1.062	1.058	1.055	1.049	1.046	1.044	1.042	1.041
5	1.972	1.803	1.140	1.053	1.049	1.041	1.036	1.032	1.031	1.030
6	2.007	1.354	1.060	1.056	1.051	1.048	1.046	1.045	1.043	1.041
7	2.024	1.398	1.062	1.056	1.052	1.048	1.044	1.041	1.038	1.035
8	2.125	1.556	1.059	1.052	1.049	1.046	1.043	1.040	1.034	1.032
9	2.296	1.188	1.066	1.056	1.053	1.048	1.046	1.040	1.037	1.036
10	2.571	1.650	1.245	1.048	1.043	1.036	1.034	1.028	1.023	1.022

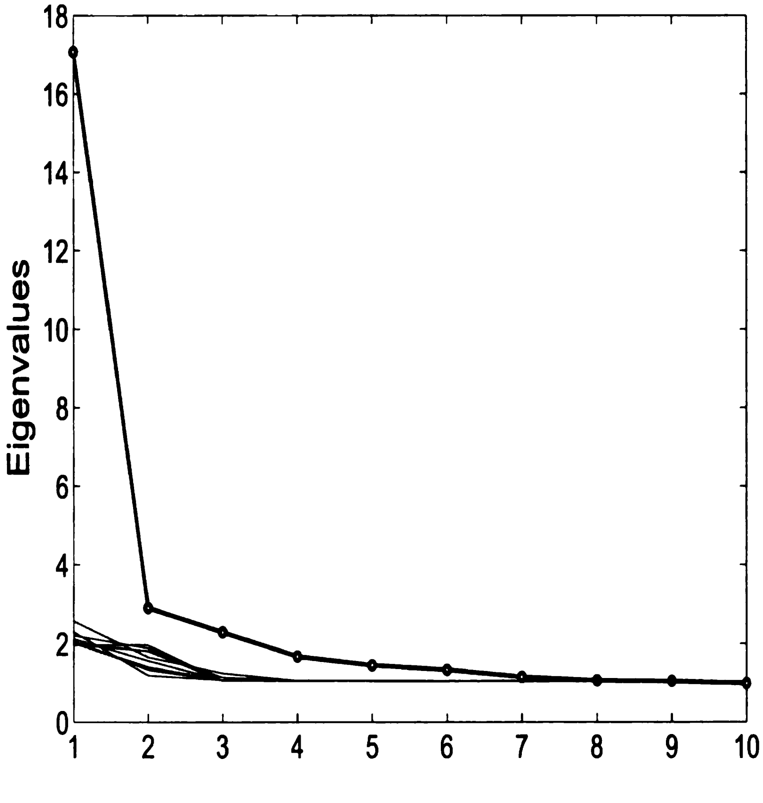

Figure 4.1. Plot of Eigenvalues for the Population Data and Ten Random Datasets

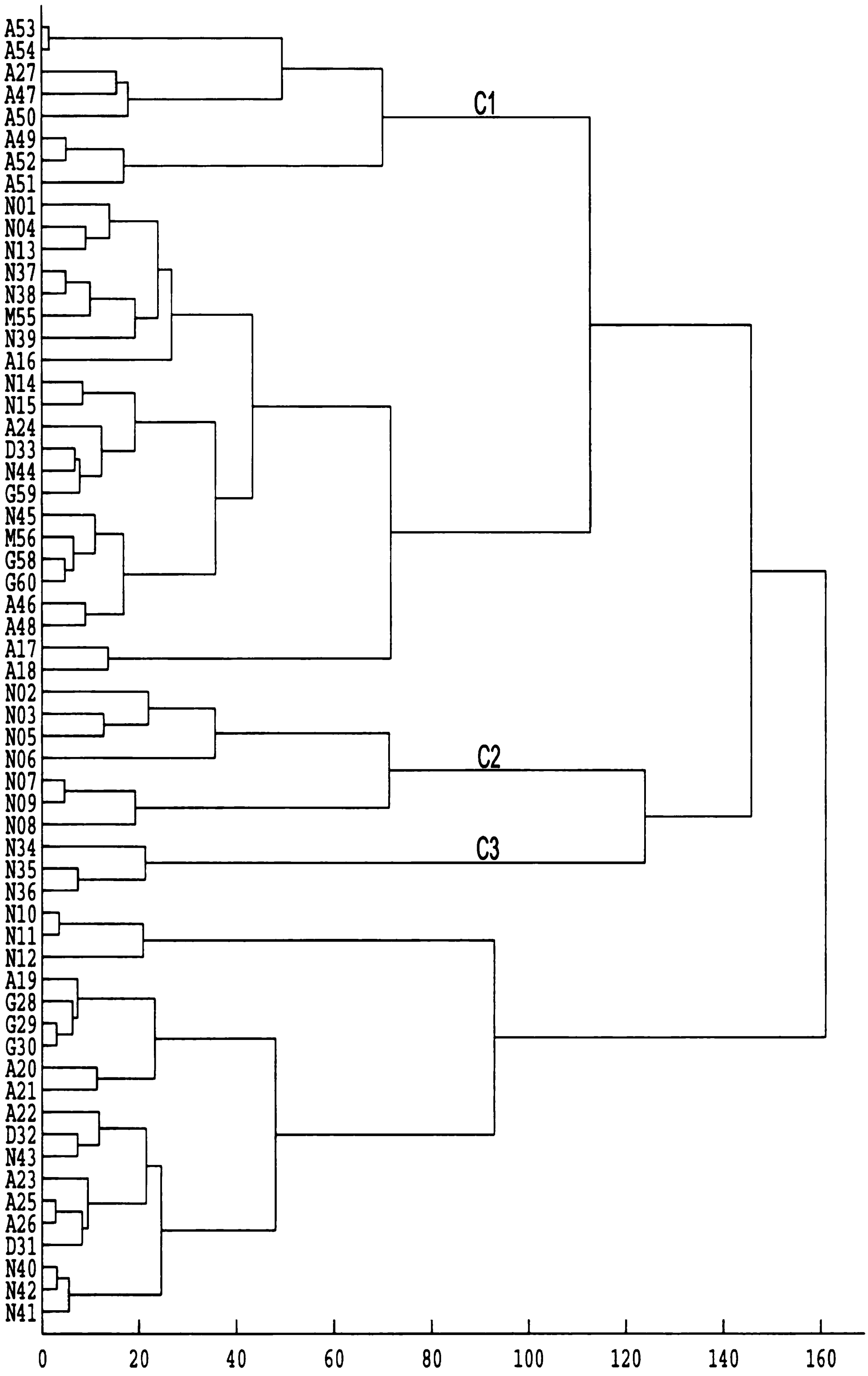

Figure 4.2. Dendrogram for the Population Data Calibrated with 8 Dimensions

dicates that these item vectors point in roughly the same direction and measure the same construct. Especially, items in the cluster labeled as "C1", "C2" and "C3" seem to be more clustered together and separated from other items. As mentioned previously, although overfactoring does not lead to serious consequences, it does need more parameters to be estimated, which may result in more estimation errors. Since the dimensionality could also be inferred from the item clustering, based on this figure and also the "elbow" rule, the dimensionality was reset to four.

In order to check whether four is a reasonable number for the dimensionality, extra calibrations on the population data were conducted using the NOHARM software with the dimensionality assumed from two to five. Three fit indices for each choice of dimensionality are shown in Table 4.7. The first two indices are overall measures of the model-data misfit from the perspective of the residual covariance matrix, and they take smaller values as the model-data fit improves. The last one is the Tanaka index, which indicates a good fit if its value is close to one (Tanaka, 1993).

Table 4.7. Model Fit Indices from the NOHARM Software

	Dimensionality				
	2	3	4	5	8
Sum of squares of residuals	0.0705	0.0490	0.0296	0.0188	0.0103
Root mean square of residuals	0.0064	0.0053	0.0042	0.0033	0.0025
Tanaka index of goodness of fit	0.9659	0.9763	0.9857	0.9909	0.9950

From this table, it is clear that more dimensionality leads to a better fit. Based on all three indices, there seems to be a consistent improvement up to the dimensionality of four, then a slower improvement from the four-dimension solution to the five-dimension solution, and much slower improvement to the eight-dimension solution.

McDonald (1997) suggested that a model is a sufficiently close approximation to the data if there is no more complex model which is identifiable and interpretable, and he also pointed out that the inspection of the residual covariance matrix works the same as the goodness-of-fit indices, especially when the criterion values for these indices are hard to set. Therefore, the estimates from the five-dimension solution were scrutinized and the item grouping from this solution resulted in five item clusters; however, it is hard to give content meanings for the interpretation of two of these clusters. Furthermore, for the residual matrix from the four-dimension solution, the element values are acceptably small and there is no consistent pattern for any block of the residual covariance matrix.

As a result, the dimensionality for this MEAP data was finally determined as four and the item estimates from this MIRT calibration were used in the cluster analysis to determine the item grouping. Figure 4.3 shows the dendrogram based on the estimates from this calibration. The pattern of four item clusters is very clear, and the clusters in the figure are labeled according to the number of items within them.

In order to confirm this was the consistent pattern across different samples, the same procedures were conducted on the 5,000 sample data. The item clustering based on this sample is shown in Figure 4.4. The items in "C3" and "C4" are exactly the same as in the previous analysis. However, "N06", "A18" and "N39" are misclassified into "C2" and "A48" is missing from it. This misclassification is not surprising, since only a small portion of people were included in the analysis. Another cluster analysis was done with the sample size doubled to 10,000. This time, as shown in Figure 4.5, the grouping of all items except item "A48" is exactly the same as that from the population data calibrated with four dimensions. Therefore, except the item "A48", the grouping pattern is very consistent across different samples, especially when the sample size is large enough.

The grouping for item "A48" seems a little inconsistent, since this item is included in cluster "C2" only for the population data calibrated with four dimensions; therefore, anyone seeking to allocate this item into the correct cluster may need suggestions from experts on item contents. For simplicity, this item was classified into cluster "C2" in the later analysis.

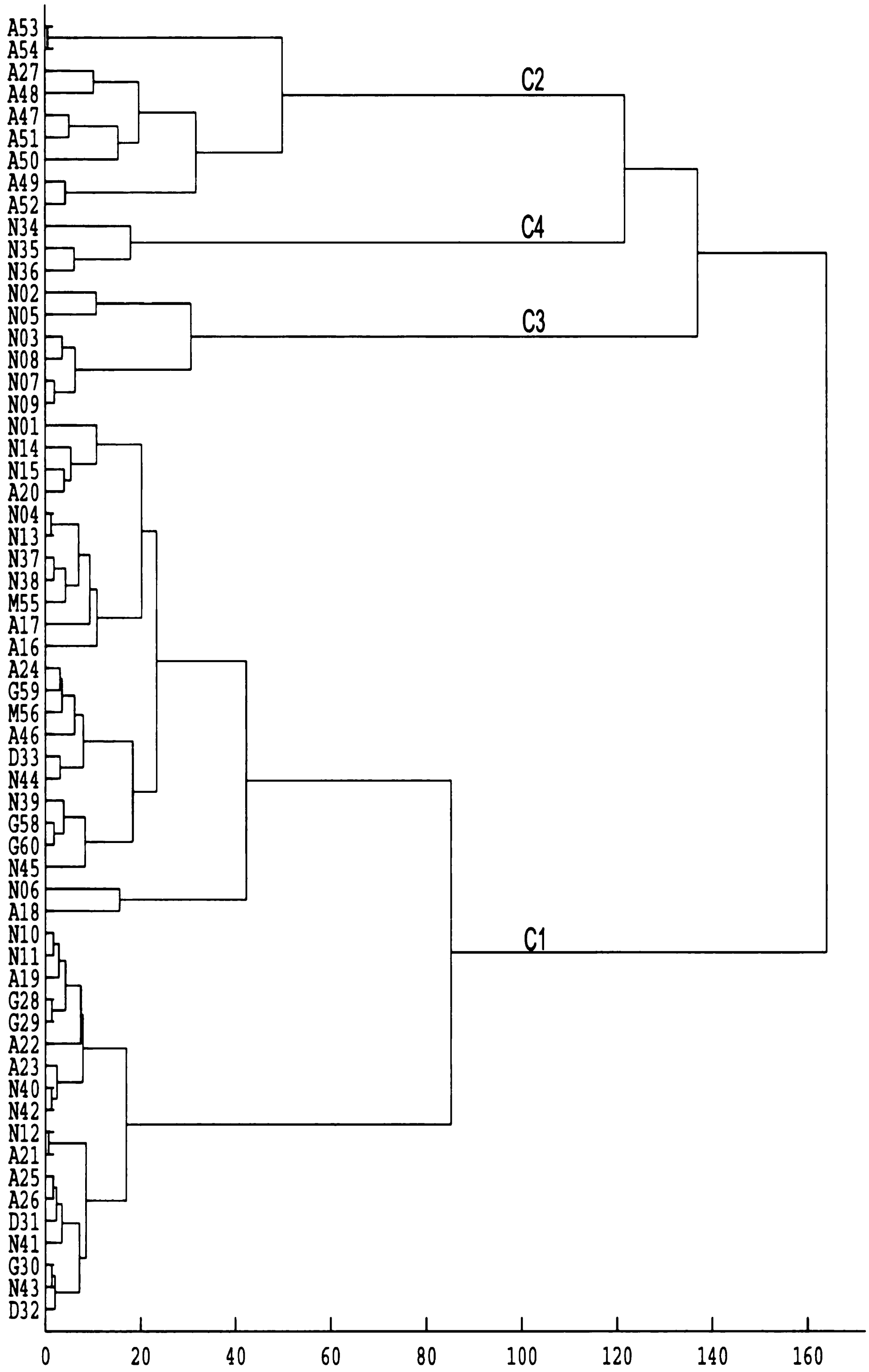

Figure 4.3. Dendrogram for the Population Data Calibrated with 4 Dimensions

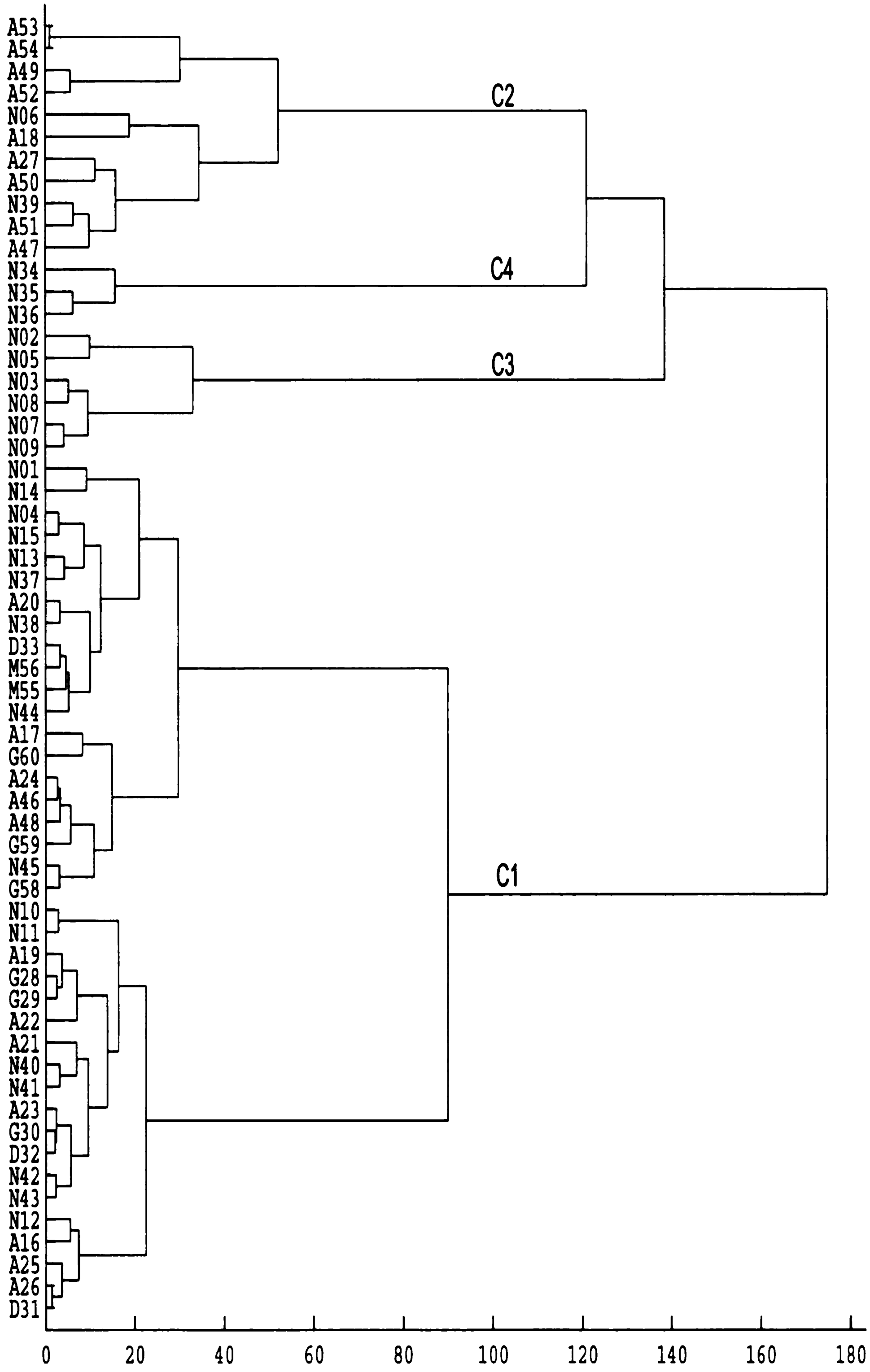

Figure 4.4. Dendrogram for the 5000 Sample Data Calibrated with 4 Dimensions

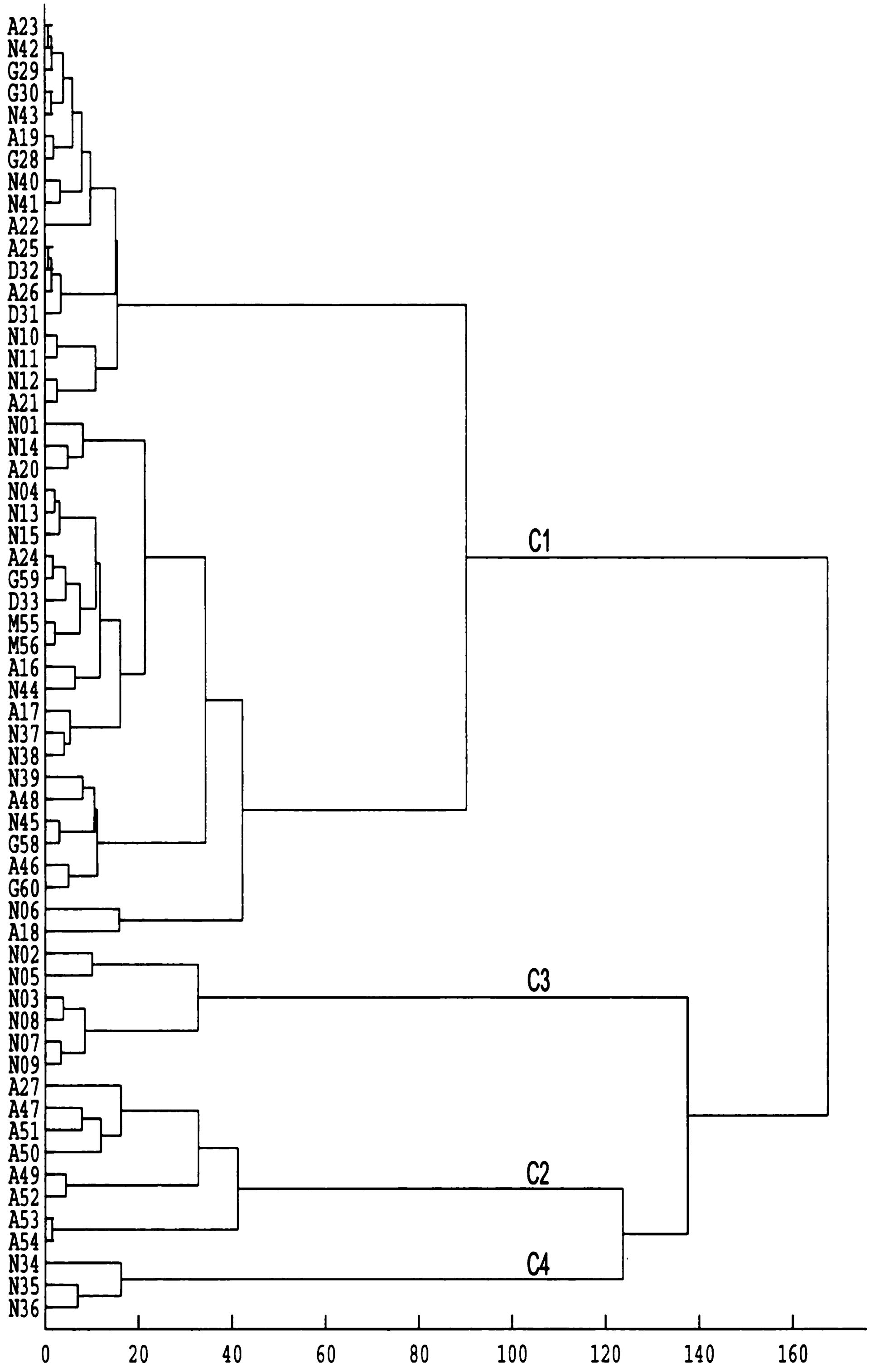

Figure 4.5. Dendrogram for the 10000 Sample Data Calibrated with 4 Dimensions

Based on the analysis for the population data, the number of items within each cluster is 41, 9, 6 and 3. All items for this test can be retrieved from the Michigan Department of Education website, and their contents were scrutinized with distinctions listed in Table 4.8.

Table 4.8. Item Cluster Content of the Fall 2007 Grade 7 MEAP Mathematics Test

Cluster	Strand	Domain	Content
1	N,A,M G,D	FL, ME, PA, RP, FO,GS, PR, UN, TR	mixed contents
2	A	FO	equation representation and calculation
3	N	FL, MR	division with fraction numbers
4	N	MR	percentage in contextual problems

The first observation from the table is that this item clustering is neither determined solely by strands nor by domains. All the clusters except the first one are related to some specific mathematical concept. Since the first cluster includes items from all strands and almost all the domains except "MR", it is difficult to interpret the construct for it. For simplicity, it is labeled as "mixed contents" here. The items in this cluster are assumed to measure the same construct, where no distinction can be detected by the MIRT calibration with four dimensions.

This four-cluster grouping seems to be different than the grouping defined by the five strands that these items were supposed to measure, and this five-strand grouping criterion was adopted for the reporting of subscores. However, it should be noted that item contents can be defined in different ways based on particular details of the complicated cognitive processes and different criteria could lead to different groupings of items; therefore, at least, it can be argued that item subsets can also be determined from domains, specific item descriptions, or a mixture of all these criteria. Most importantly, the grouping criterion on item contents does not take the proficiency correlation into consideration, especially when there are already warnings against the reporting of highly correlated proficiencies (Sinharay et al., 2007).

The cluster analysis determines the item grouping from the empirical data analysis and sorts similar or dissimilar items based on the statistical criteria. This analysis is supposed to define the most statistically varying constructs. An extra analysis was conducted to check the closeness of reference composite vectors based on the item estimates for the population data calibrated with four dimensions. For the four-cluster solution, the values of off-diagonal entries for the $(\boldsymbol{w}_i'\boldsymbol{w}_j)_{ij}$ matrix vary from 0.324 to 0.635; however, for the item grouping based on the five-strand criterion, these values vary from 0.806 to 0.999. If the cut point is defined as 0.8 for an acceptable closeness, the proficiencies resulted from the five-strand criterion are so highly correlated that there is no need to report all of them.

Therefore, although both the expert judgment and the cluster analysis can give evidence for the item grouping, it can be argued that the cluster analysis is preferred because the subscores that are of interest to the users of test results should be statistically distinct rather than defined by the closely related constructs. Furthermore, more reliable item grouping can be obtained if both pieces of evidence are taken into consideration.

4.3 MIRT Calibration and Subscore Reporting

The analysis in the previous section resulted in four dimensions and four item clusters. In this section, in order to obtain more stable item estimates from the NOHARM software, the coordinate system was reconstructed with items measuring different constructs; therefore, item "N02", "N34" and "A53", which were selected from different clusters, were constrained to have special discrimination vectors. Then the item estimates from the NOHARM software were rotated with the Varimax criterion.

Table 4.9 shows the resulting item estimates, generalized discrimination, generalized difficulty and the degree angle with each axis. The descriptive statistics for these estimates are given by Table 4.10. These estimates were regarded as fixed and input into the TESTFACT software for the uncorrelated proficiency estimation.

Table 4.11 shows the directions of four reference composite vectors in the four-dimensional proficiency space defined by the Varimax transformed item discrimination matrix. Almost all the bolded numbers are larger than 0.9, and these numbers are related to different proficiency dimensions. This indicates that all these reference composite vectors are close to different coordinate axes so that any construct estimate is mainly determined by one unique raw proficiency estimate, which is different than those for the other constructs.

Table 4.12 shows all four eigenvalues of the $A_l'A_l$ matrix for each item cluster. For all clusters, the first eigenvalue, which corresponds to the reference composite in Table 4.11, is always the largest and dominant one. The ratio between the first two eigenvalues is also given in the last row. These ratio values are large, which confirms again that the variation of this matrix is mainly determined by the first eigenvector.

Table 4.13 gives all the correlations among NC subscores and construct estimates. The correlations between the NC subscores and corresponding construct estimates for all clusters are highlighted in bold, and almost all of them are higher than 0.9. In addition, the correlation between any two construct estimates is higher than that between corresponding NC subscores. This is expected since construct estimates are calculated under the MIRT model, which simultaneously calibrates the parameters and allows correlations among the estimates. Besides this reason, correlation values among construct estimates are also slightly influenced by the EAP scoring method.

From this table, the means and standard deviations seem good for construct estimates. All the means of construct estimates are slightly above zero, and the reason may be that the mean difficulty of the test is larger than 0.5 based on the classical p-values, or a little below zero with regard to the MIRT difficulty criterion. The standard deviations are smaller than one, due to the EAP scoring method.

Figure 4.6 shows the plots between NC subscores and construct estimates based

Table 4.9. Varimax Transformed NOHARM Item Estimates for the Population Data

Cluster	item	a_1	a_2	a_3	a_4	d	$MDSIC$	B	α_1	α_2	α_3	α_4
1	1	0.24	0.15	0.18	0.40	0.12	0.52	-0.23	63	73	69	40
3	2	0.65	0.17	0.29	0.46	0.10	0.86	-0.11	42	78	70	58
3	3	0.82	0.12	0.19	0.19	-0.10	0.87	0.12	20	82	77	77
1	4	0.26	0.28	0.32	0.68	0.53	0.85	-0.62	72	71	68	36
3	5	0.70	0.20	0.27	0.31	-0.08	0.84	0.10	33	76	71	68
1	6	0.46	0.30	0.34	0.36	-0.67	0.74	0.90	52	66	63	61
3	7	1.20	0.15	0.20	0.28	-0.07	1.26	0.05	17	83	81	77
3	8	1.10	0.20	0.22	0.29	0.01	1.18	-0.01	21	80	79	76
3	9	1.61	0.17	0.23	0.39	0.01	1.68	0.00	17	84	82	77
1	10	0.21	0.07	0.16	0.79	0.90	0.84	-1.07	75	85	79	19
1	11	0.19	0.08	0.15	0.78	0.80	0.82	-0.98	77	84	79	18
1	12	0.10	0.06	0.10	0.33	0.08	0.36	-0.22	74	81	74	25
1	13	0.22	0.21	0.24	0.53	0.33	0.66	-0.50	71	71	68	36
1	14	0.33	0.19	0.27	0.75	0.93	0.88	-1.05	68	78	72	32
1	15	0.27	0.21	0.31	0.67	0.26	0.81	-0.32	71	75	68	35
1	16	0.17	0.27	0.23	0.73	0.57	0.83	-0.68	78	71	74	28
1	17	0.29	0.47	0.32	0.77	-0.02	1.00	0.02	73	62	71	40
1	18	0.31	0.40	0.29	0.44	-0.22	0.73	0.30	65	57	67	53
1	19	0.17	0.07	0.13	0.77	1.41	0.81	-1.75	78	85	80	16
1	20	0.21	0.15	0.23	0.58	0.59	0.67	-0.88	72	77	70	31
1	21	0.13	0.07	0.12	0.40	0.54	0.44	-1.23	73	81	75	25
1	22	0.16	0.06	0.13	1.00	1.61	1.02	-1.57	81	87	83	12
1	23	0.13	0.11	0.15	0.80	0.87	0.84	-1.03	81	82	79	16
1	24	0.17	0.21	0.29	0.55	0.34	0.68	-0.50	75	72	65	36
1	25	0.18	0.15	0.22	0.78	0.71	0.84	-0.85	78	80	75	22
1	26	0.16	0.15	0.24	0.80	0.65	0.86	-0.75	79	80	74	22
2	27	0.16	0.08	0.31	0.32	-0.10	0.48	0.21	71	80	49	49
1	28	0.13	0.07	0.14	0.66	0.75	0.69	-1.08	79	85	78	17
1	29	0.17	0.09	0.19	0.81	1.12	0.86	-1.31	79	84	77	18
1	30	0.13	0.07	0.15	0.55	0.64	0.59	-1.08	77	83	75	21
1	31	0.19	0.20	0.25	0.91	0.84	0.98	-0.86	79	78	75	23
1	32	0.13	0.07	0.17	0.57	0.70	0.61	-1.15	78	84	73	22
1	33	0.17	0.14	0.24	0.50	0.11	0.59	-0.18	73	76	66	33
4	34	0.12	0.65	0.13	0.30	0.32	0.73	-0.43	80	29	80	66
4	35	0.28	1.24	0.27	0.26	0.26	1.32	-0.20	78	21	78	79
4	36	0.40	2.13	0.36	0.24	0.09	2.22	-0.04	79	16	81	84
1	37	0.21	0.28	0.23	0.58	0.27	0.72	-0.38	73	67	71	36
1	38	0.21	0.27	0.24	0.61	0.25	0.74	-0.34	74	68	71	35
1	39	0.10	0.17	0.21	0.31	-0.21	0.42	0.50	76	66	61	43
1	40	0.19	0.16	0.21	0.97	1.27	1.03	-1.24	80	81	78	18
1	41	0.17	0.18	0.19	0.83	1.28	0.89	-1.44	79	78	77	21
1	42	0.21	0.18	0.26	1.17	1.47	1.23	-1.20	80	82	78	18
1	43	0.16	0.11	0.20	0.72	1.33	0.77	-1.72	78	82	75	21
1	44	0.19	0.21	0.30	0.66	0.69	0.78	-0.89	76	75	67	32
1	45	0.19	0.24	0.26	0.39	0.16	0.56	-0.28	70	65	62	46
1	46	0.12	0.17	0.29	0.50	0.10	0.62	-0.16	79	74	62	35
2	47	0.24	0.20	0.46	0.38	0.02	0.67	-0.03	70	72	47	55
2	48	0.15	0.17	0.34	0.46	0.00	0.61	0.00	75	74	56	42
2	49	0.14	0.18	0.35	0.15	-0.72	0.44	1.64	71	66	38	70
2	50	0.23	0.13	0.74	0.44	0.01	0.90	-0.01	75	82	35	61
2	51	0.13	0.16	0.34	0.26	-0.01	0.48	0.02	74	71	44	57
2	52	0.15	0.17	0.40	0.16	-0.45	0.48	0.93	72	69	34	71
2	53	0.31	0.05	2.02	0.46	-0.34	2.09	0.16	82	89	15	77
2	54	0.26	0.05	1.76	0.42	-0.15	1.83	0.08	82	88	16	77
1	55	0.17	0.25	0.25	0.57	-0.20	0.69	0.28	76	69	68	35
1	56	0.11	0.12	0.20	0.32	-0.03	0.41	0.08	74	72	61	39
1	58	0.13	0.18	0.23	0.38	-0.11	0.50	0.22	75	69	62	40
1	59	0.10	0.10	0.16	0.29	-0.21	0.36	0.59	73	74	63	37
1	60	0.07	0.11	0.14	0.22	-0.65	0.29	2.26	76	68	61	41

Table 4.10. Summary Statistics for Item Parameter Estimates

	a_1	a_2	a_3	a_4	d	$MDSIC$	B
Mean	0.28	0.22	0.30	0.53	0.32	0.82	-0.34
Std	0.28	0.31	0.32	0.23	0.55	0.38	0.77

Table 4.11. Reference Composite Vectors from the Population Data Estimation

	C_1	C_2	C_3	C_4
θ_1	0.253	0.183	**0.932**	0.189
θ_2	0.232	0.081	0.142	**0.954**
θ_3	0.294	**0.920**	0.194	0.174
θ_4	**0.892**	0.336	0.273	0.153

Table 4.12. Eigenvalues of $A_l'A_l$ Matrix

Eigenvalue	C_1	C_2	C_3	C_4
1	21.961	9.284	7.863	7.143
2	0.802	1.112	0.147	0.054
3	0.092	0.089	0.009	0.003
4	0.052	0.003	0.001	0.000
1:2	27.369	8.351	53.675	133.257

Table 4.13. Correlations among NC Subscores and Construct Estimates

	NC_1	NC_2	NC_3	NC_4	θ_1^*	θ_2^*	θ_3^*	θ_4^*
NC_1	1.00	0.64	0.50	0.40	**0.98**	0.69	0.62	0.55
NC_2	0.64	1.00	0.42	0.31	0.70	**0.89**	0.52	0.45
NC_3	0.50	0.42	1.00	0.30	0.55	0.47	**0.96**	0.42
NC_4	0.40	0.31	0.30	1.00	0.44	0.32	0.36	**0.92**
θ_1^*	0.98	0.70	0.55	0.44	1.00	0.75	0.67	0.59
θ_2^*	0.69	0.89	0.47	0.32	0.75	1.00	0.58	0.46
θ_3^*	0.62	0.52	0.96	0.36	0.67	0.58	1.00	0.52
θ_4^*	0.55	0.45	0.42	0.92	0.59	0.46	0.52	1.00
Mean	25.92	4.29	2.96	1.68	0.02	0.04	0.04	0.02
Std	8.33	2.24	2.01	1.16	0.96	0.87	0.85	0.83
Mean percent-correct	0.63	0.48	0.49	0.56				

on the analysis of the population data. Clearly, the continuous construct estimate is not one-to-one correspondence with the discrete NC subscore. For the people who have the same NC subscore, their construct estimates may be different. Although not shown here, this construct estimate also gives more score choices than the unidimensional estimate for a limited response pattern with few items in the cluster. For example, there are only three items in cluster "C4". The unidimensional calibration on these items at most gives $2^3 = 8$ different scores; however, the construct estimate almost covers $(-2, 2)$ continuum. Thus, the construct estimate is preferred in multidimensional tests because it not only takes the item characteristics into consideration but also implicitly borrows information from other correlated construct estimates.

Figure 4.7 shows the plots between NC subscores and their dominant raw proficiency estimates. The raw proficiency estimates for the same NC subscore are much more spread out horizontally, compared with the construct estimates. What is more, the difference between the raw proficiency estimates for the neighboring NC subscores is not as clear as that between the construct estimates. It seems that these raw proficiency estimates are unstable with the coordinate system set up by the Varimax transformed item discrimination matrix, which suggests that this rotation may not result in a good coordinate system for the construct interpretation.

In conclusion, the projection method for construct estimates clearly works in this MEAP test, and it not only empirically identified four item clusters, which were assumed to measure different constructs from people, but also gave the subscore estimates under the MIRT framework.

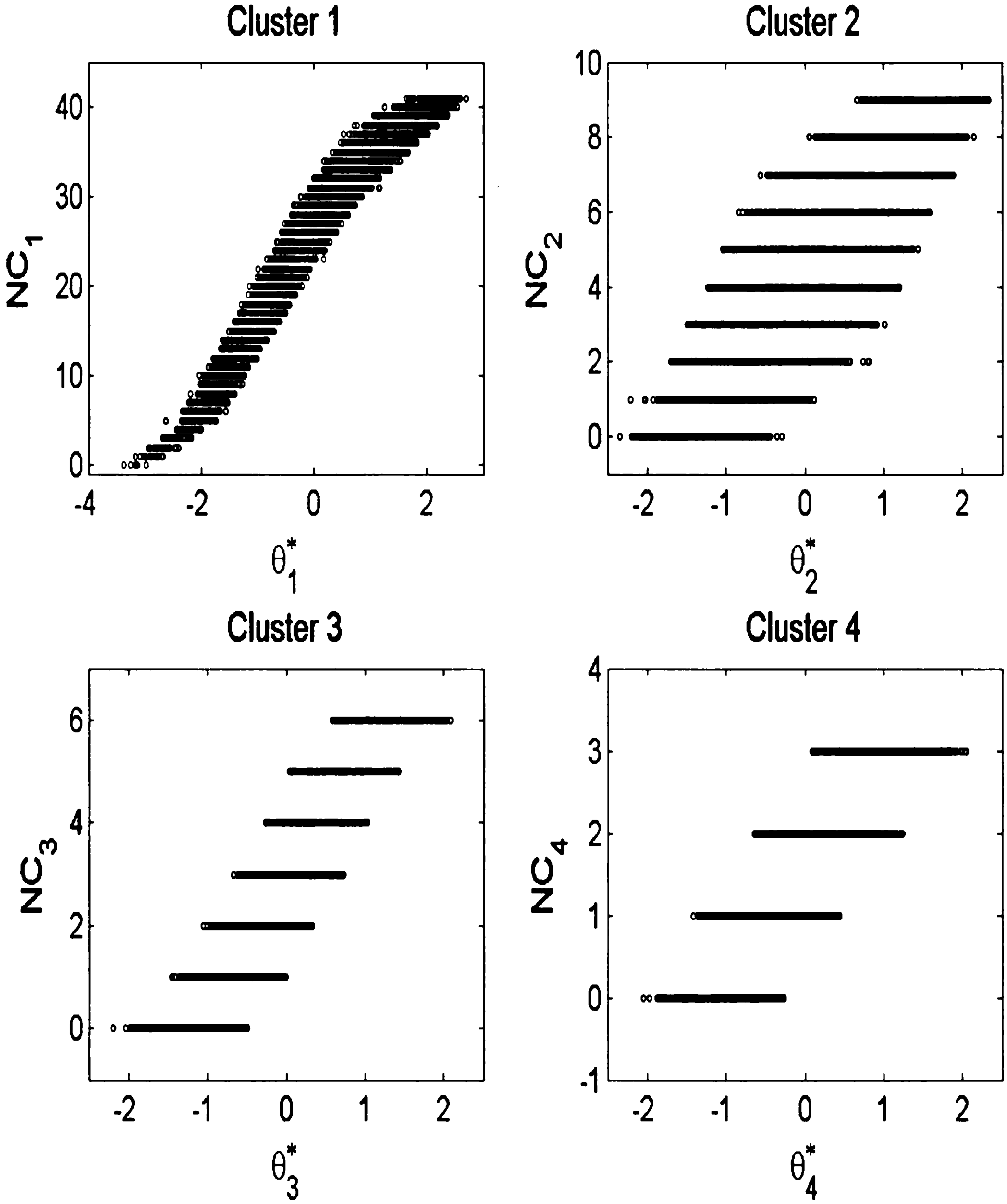

Figure 4.6. Plots of NC Subscores versus Construct Estimates for the Population Data Calibrated with 4 Dimensions

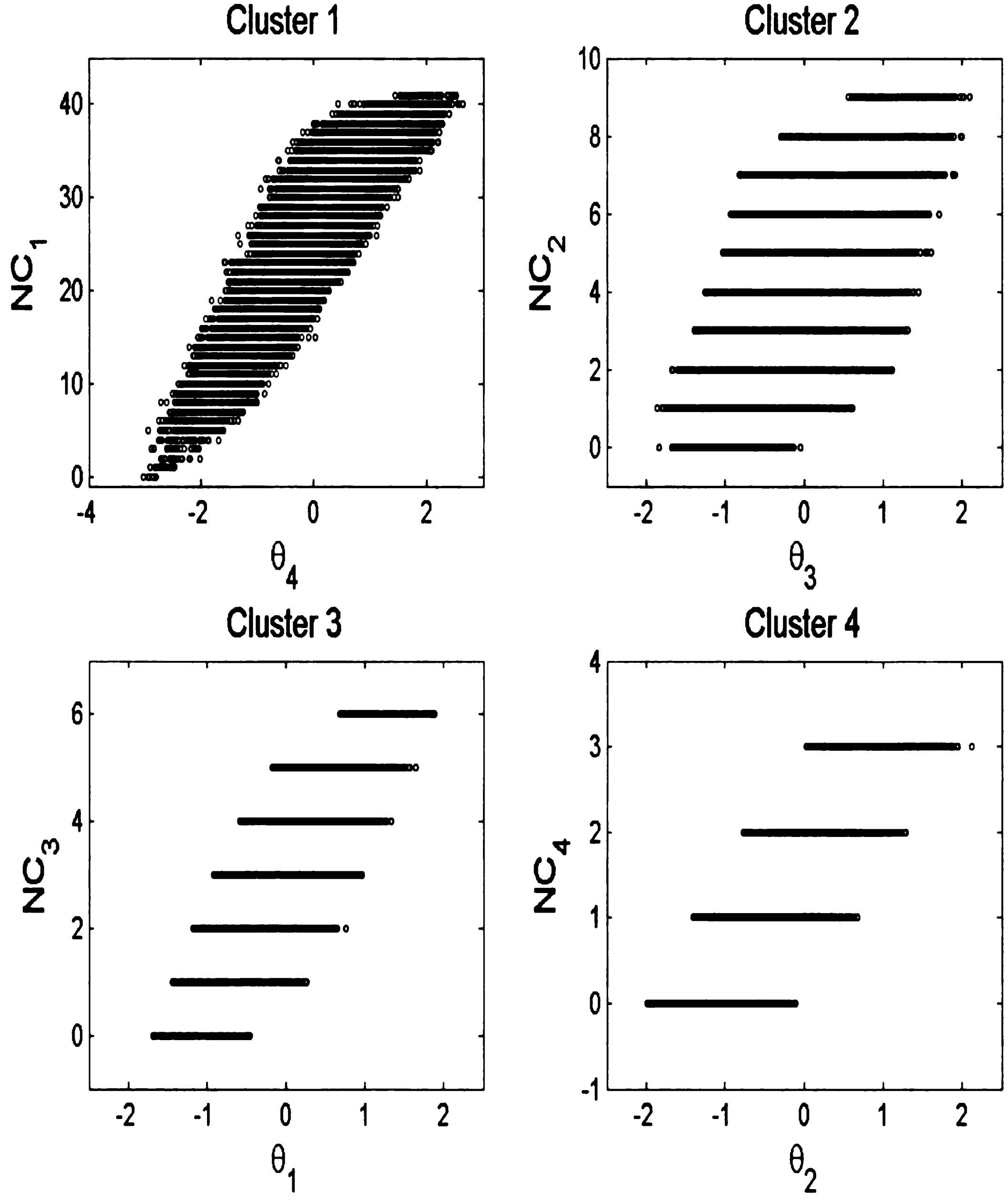

Figure 4.7. Plots of NC Subscores versus Raw Proficiency Estimates for the Population Data Calibrated with 4 Dimensions

CHAPTER 5

Conclusions, Implications, Limitations and Future Research

5.1 Conclusions and Implications

In order to report subscores for multidimensional tests, the NC subscore and unidimensional estimate for items in each cluster may sometimes work well, where the cluster is defined either by items measuring similar contents or by items with similar vector directions, which are estimated using the MIRT models.

However, if the number of items within each cluster is small, the subscore is really suspicious due to lack of sufficient reliability. This problem may be remediated by incorporating information from other clusters, since most proficiencies measured within one test are assumed to be positively correlated. This goal is commonly achieved by some post-hoc adjustments on these subscore estimates; however, this two-stage estimation procedure is not usually preferred, since it does not take into account the measurement error in the first calibration step.

The MIRT model seems to be very useful in multidimensional tests to allow simultaneous estimation of parameters from all dimensions. Currently, because of the coördinate indeterminacy in the MIRT calibration, several popular software just constrains the proficiencies to have zero mean and identity variance-covariance matrix for easy computation. However, it is clear that this zero correlation among proficiencies is not an assumption in the MIRT models.

Precisely speaking, these MIRT calibrations only give one set of possible item and person estimates for a good model-data fit. Because of the constraint discrepancy between the software estimation and actual MIRT models, these estimates may not be ready for interpretation before a suitable coordinate system for person proficiencies is set up.

The commonly used Varimax and Promax methods, which are borrowed from the factor analysis, are mostly used to explain item characteristics. They may be sufficient to identify item clusters, usually by grouping items with the large loading values for the same dimension after the rotation. However, unlike the projection method, they are not designed for proficiency estimation under the MIRT framework.

The simulation study shows that the Varimax rotation within the TESTFACT software is not adept at recovering the generating parameters, partly due to the incorrect alignment of the coordinate axes, especially for the dimension which there is no simple structure item to measure. Commonly the Varimax rotation is conducted on the correlation-adjusted item estimates from the software, and it is most effective when these item vectors are orthogonal to each other. Therefore, it can be imagined that the Varimax rotation is most suitable for situations with simple structure items and uncorrelated proficiencies.

The projection method illustrated in this study focuses on finding an interpretable coordinate system for person proficiency estimates after the MIRT calibration, which leads to its potential usage in subscore reporting, especially when the number of dimensions and number of item clusters are different. In this method, item clusters are obtained from the analysis of the direction cosines of item vectors, instead of the element values of these vectors. The construct score is then calculated as the projection of raw proficiency coordinates onto the most discriminating direction for each item cluster, and its value is invariant with any orthonormal rotation of the coordinate system for the raw proficiency estimates.

Researchers may be concerned with the efficiency of the item grouping obtained from cluster analysis, especially when there are sampling errors in the data matrix and estimation errors for the item discrimination matrix. However, both the simulation study and empirical analysis show that the cluster analysis gives a fairly consistent item grouping pattern and shows robustness to the effects of estimation and sampling errors.

First, the angular distance matrix is invariant to any orthonormal rotation of the coordinate system representing the multidimensional space. However, the method by judging the element values may give inconsistent conclusions because different item discrimination matrices may result from different orthonormal rotations.

Second, from the dendrogram for the first replication of Design 1, the between and within cluster pattern is as clear as expected and it seems that although the estimation errors and the incorrect alignment of the coordinate system could lead to bad recovery of item parameters, the item grouping from the cluster analysis is often "insensitive" to these effects.

Third, in the empirical data analysis, the cluster analysis was conducted on both the large population data and the two sample datasets in order to reduce the effect of sampling errors. The conclusion is that the item grouping pattern for the MEAP data is fairly consistent as long as the sample size is large.

Finally, the grouping pattern was also confirmed based on the results from MIRT calibrations with different dimensionalities. When the population data were calibrated with eight dimensions, all item clusters except the first one were also obvious in this "overfactoring" situation, and as expected, the first cluster was split into different small clusters.

The analysis from the simulation study shows that the construct estimate is highly correlated with the NC subscore and the unidimensional estimate, and the construct recovery from its estimate is better than the unidimensional estimate. Empirical analysis based on the MEAP data identified four item clusters with the last three clearly related to some specific mathematical content, which was not exactly as defined by strands or domains. Although the first cluster contains many items from different content areas, which are supposed to measure different proficiencies, it is clear that all these proficiencies are highly correlated in this analysis. The reason may be either they use the same cognitive process or they are taught at the same time in a school. The clustering of these items also conforms to the rule: when there is high correlation among proficiencies, there is no added value to report all of them.

When the dimensionality and the number of item clusters are the same, in some sense, this projection method is similar to the Promax method, because both of them can be used to find an oblique rotation for the interpretable proficiency solution. They can transform Varimax estimates to the simple structure item discriminations and interpretable person proficiencies, except that the emphasis is slightly different. The projection method focuses on the construct proficiency estimation, which may lead to simple structure item discrimination estimates, while the Promax method searches for the simple structure item discrimination matrix, which results in correlations among proficiency estimates. Additionally, the projection method can use the information from item contents and MIRT estimates to obtain reference composite vectors for the projection purpose, while the Promax method is a pure mathematical criterion to rotate the estimated item discrimination matrix. Finally, the projection method gives construct estimates which are invariant to the orthonormal rotation of the coordinate system for raw proficiency estimates; therefore, it can also use the unrotated estimates as the input for the calculation of construct scores.

This projection method is not only important for defining the coordinate system in one single test administration but also has implications in practice for linking and equating across different forms under the MIRT framework. In all IRT models, the response matrix is assumed to be influenced by person proficiency and item characteristics, both of which are unknown. This inevitably leads to the fact that some constraints or assumptions are required to obtain a suitable set of parameter estimates for interpretation. In the unidimensional IRT, item characteristics can be determined with regard to persons' proficiencies in the reference group, which are commonly set to have zero mean and unit variance. After the data calibration for this group, the item characteristics, such as discrimination and difficulty parameters, can be assumed fixed for the item, and they can be reused for the item pool construction, test construction, and equating.

In the multidimensional IRT, item characteristics can also be set with regard to some reference group, where there may be underlying correlations among proficiencies. During the MIRT calibration, due to the unknown proficiency correlation and indeterminacy of the coordinate system, the calibration software assumes the uncorrelated proficiencies and provides the correlation-adjusted and Varimax transformed $\boldsymbol{a}$ vectors as the item discrimination estimates. It is problematic if these $\boldsymbol{a}$'s are used each time to define the item discrimination power, since the coordinate system for interpreting proficiency estimates is incorrect. The inconsistency can also be easily imagined because these $\boldsymbol{a}$ parameters are not fixed, and they are different for groups with different correlations. This difference could lead to problematic MIRT equating and linking procedures, which assumes the same item discrimination parameters invariant to group characteristics.

Therefore, in addition to the constraints as in the unidimensional IRT, an interpretable base coordinate system for person proficiencies needs to be constructed for the reference group. Ideally, item vectors in this multidimensional space represent the true discrimination power $\boldsymbol{a}^*$, whose estimation is excluded from the effect of the unknown proficiency correlation. One way to define the system is that the data calibration is implemented with the proficiency correlation values from educated guessing, or from the correlations among NC subscores or unidimensional estimates. The second way is to use the Promax method to match the item discrimination matrix to the simple structure. The projection method introduced in this study may provide another promising way to define the coordinate system for person proficiencies and item characteristics.

After the coordinate system for person proficiencies is set up for the reference group, the item characteristics can be obtained and regarded as fixed when these items are administered in different tests, or to different groups of people, even when the proficiency correlation matrix is different from that in the reference group. Since the coordinate system is kept consistent across different test settings, the estimates across different tests can be put into the same scale for comparison and interpretation.

One question now is whether it is legitimate to continue using the commonly used correlation-adjusted $\boldsymbol{a}$ to represent the characteristics of items when they are administered to different groups of people. To ensure the person-independent property of the item characteristics defined here, at least, one assumption is required that the correlation matrix among person proficiencies for different populations is invariant. How possible can this assumption be? If the correlation is inherent within the proficiencies themselves, that may be possible. However, it is commonly accepted that correlation is the characteristic of the population and it can change across different populations. Even if this invariance property is true, these proficiencies may be compared under the same coordinate system; however, they are not compared with regard to the correct interpretation of constructs.

5.2 Limitations and Future Research

The commonly used NC subscores or unidimensional estimates with the item clusters defined by different content areas may lead to unique subscores for each student. These scores can be easily interpreted as proficiencies related to these clusters, although people may not exactly be statistically distinguished by these proficiencies.

For this projection method, it may be a little difficult or subjective to determine the number of item clusters and item grouping in the cluster analysis step, which works best when the within-cluster item vectors are very close while the between-cluster item vectors are far apart. When this step is applied to the complicated real settings, especially when proficiencies are highly correlated, the within or between pattern may not be clear.

The different item grouping may result in different definition, interpretation and calculation for construct estimates, since, theoretically speaking, the raw proficiency estimates can be projected onto any item cluster or even one single item. Thus, in practice, this item grouping pattern needs to be confirmed across different samples and different numbers of dimensions. Furthermore, the item grouping may need the involvement from the expert expectation on test structure, which is also useful for decisions on whether the highly correlated proficiency estimates are reported as one or several scores.

In the real data analysis, the dimensionality for the MIRT calibration is unknown. Although numerous methods for determining dimensionality have been suggested by previous research, due to possible estimation dependency between dimensionality and item clusters, finding ways to achieve a good balance between them should be examined in future research. For example, it may be interesting to see how the cluster analysis and projection method perform with different dimensionality assumed, especially for the overfactoring cases, since it is supposed that overfactoring will not lead to serious bad results.

All in all, it is clear that the solution for construct estimates can be flexible for the analysis of the same data, because different choice of dimensionality, number of item clusters, item grouping or even software calibration will lead to different subscore reporting. More research should be conducted on the magnitude of these differences, and try to find an optimal solution for the estimation.

In this study, the EAP scoring method was adopted for proficiency estimation, due to its easy calculation and possible preference for the reliable estimation in real tests. However, these biased scores caused problems for the calculations of Bias and RMSE, both of which are often used in parameter recovery studies. Large deviation between the estimated EAP score and the true score was expected. As a result, only the relative values of Bias and RMSE were compared across different methods or situations.

The EAP scores in this study were calculated for the uncorrelated proficiency estimates, then they were used in the projection calculation for construct estimates. It may be more desirable that this EAP scoring method is applied directly to construct estimates in order to further reduce the effect of incorrect alignment of the coordinate system chosen by the software. One way to achieve this goal is to use the rotated item estimates based on this method, regard them as the fixed parameter input for the TESTFACT software, and directly obtain the EAP score for the construct estimate.

In this study, it is the uncorrelated proficiency coordinates and mixed structure items that were used in the simulation. Although it can be argued that other situations with different combinations of proficiency correlation and item structure can be easily transformed to this one, future research can definitely confirm the efficiency of this projection method in those situations. For example, the simulation study can focus on correlated proficiencies and simple structure items, and finally it can be extended into more general cases with correlated proficiencies and mixed structure items. It should be noted that, in all cases, this projection transforms the uncorrelated proficiency estimates directly into the construct estimates, where the proficiency correlation effect and item composite effect cannot be distinguished, or it may be argued that there is no need to separate them.

For the empirical data analysis, guessing parameters were omitted, because either the unidimensional estimates or the fixed values led to convergence problems and unusual results for other item parameters in the MIRT calibration. Further analysis of the guessing effect on the MEAP data can be conducted when new software can incorporate its estimation directly in the MIRT calibration.

The article by Field et al. (2006) pointed out that reducing the two-mode data for separate analysis of one-mode characteristics may lose the duality information in the data. From the perspective of social networks, the two-mode data are defined for two sets of social units and contain measurements of a relation from the units in one set to those in the other set; accordingly, the one-mode data are the set of social actors with relations defined only between them (Doreian et al., 2004).

The response matrix can be regarded as the two-mode data, since it contains the interactions between persons and items. The simple aggregation of information in the measurement field is in classical test theory, where person information is represented by the NC total score while item information by the difficulty p-value. However, all IRT calibrations, including the MIRT, use the information from person-item interactions and simultaneously calibrate the two-mode response data into two single-mode scale-dependent item and person characteristics.

While IRT models try to find sufficient dimensions and parameters to insure the local independence assumption between responses, the two-mode analysis in Field et al.'s article directly groups interactive actors and events (rows and columns) into the same block, which is very useful in social network analyses to understand how groups of actors are linked by conducting groups of events. If this two-mode method is applied to the response data, first of all, the blocking conclusion may be limited only to the persons and items in the test. Secondly, the blocking can be due to not only the same or highly correlated proficiencies measured by items, but also the difficulty of items, and both effects cannot be separated. Thirdly, this blocking does not give the continuous proficiency estimate to each person as the IRT models. In the measurement field, it is not enough to only understand that one group of people are similar with regard to one group of items; most importantly, these people need to be put in the same continuum for ordering and comparison regardless of the characteristics of local items. Even in the cluster analysis step for the item grouping, what is needed for this study is roughly the same orientation of item vectors, not how these items are clustered together with regard to the temporary setup of the coordinate system.

All in all, either the submatrix blocking for the two-mode data or the one-mode blocking/continuum provided by the analysis is useful for different problems. In future research, if people need to be categorized into different blocks according to different subsets of items, this two-mode method can be applied as an exploratory way.

Although researchers may agree that tests in practice are really multidimensional, many of them are reluctant to adopt the MIRT to real test settings. This is mostly because of the convergence difficulty in the MIRT calibration and the indeterminacy of the coordinate system for parameter estimates. To solve these problems, it is supposed that the confirmatory MIRT model can be carried out after the exploratory version. With the confirmatory version, not all the parameters need to be freely estimated. For example, some elements in item discrimination vectors can be fixed to zero or other values, or some correlation among proficiencies can be input into the parameter estimation (Yao & Boughton, 2007). Future research can focus on how to extract useful information from the exploratory MIRT calibration for the later confirmatory analysis and whether the confirmatory analysis leads to more reliable results.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127.

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

Bock, D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

Bock, D., Gibbons, R., Schilling, S., Muraki, E., Wilson, D., & Wood, R. (2003). *TESTFACT 4.0 [Computer software and manual]: Test scoring, item statistics, and item factor analysis.* Lincolnwood, IL: Scientific Software International.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

Doreian, P., Batagelj, V., & Ferligoj, A. (2004). Generalized blockmodeling of two-mode network data. *Social Networks, 26*(1), 29-53.

Drasgow, F., & Parsons, C. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Field, S., Frank, K. A., Schiller, K., Riegle-Crumb, C., & Muller, C. (2006). Identifying positions from affiliation networks: Preserving the duality of people and events. *Social Networks, 28*(2), 97-123.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement*, *43*(1), 39-52.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models in latent trait theory.* The University of New England, Center for Behavioral Studies, Armidale, Australia.

Fraser, C., & McDonald, R. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*(2), 167-169.

Gessaroli, M. E., & Champlain, A. F. D. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, *33*(2), 157-179.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, *10*(3), 287-302.

Hendrickson, A. E., & White, P. O. (1964). Promx: a quick method for rotation to oblique simple structure. *The British Journal of Statistical Psychology*, *17*, 65-70.

Hirsch, T. M., & Miller, T. R. (1991, June). *Evaluation of a multidimensional item response theory procedure for investigating test dimensionality.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick, NJ.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187-120.

Kao, S.-C. (2007). *The new goodness-of-fit index for the multidimensional item response model.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Kim, J.-P. (2001). *Proximity measures and cluster analysis in multidimensional item response theory.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in efa: an easy-to-use computer program for carrying out parallel analysis.

Practical Assessment, Research and Evaluation, *12*(2), 1-11.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35*, 42-56.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional irt linking. *Applied Psychological Measurement*, *24*(2), 115-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company, Inc.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, *16*(3), 279-293.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 257-270). New York: Springer.

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education*, *5*(3), 193-211.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of eap scores. *Applied Psychological Measurement*, *9*(4), 417-430.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, *4*(3), 207–230.

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

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 271-286). New York: Springer.

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

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

Reckase, M. D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986, June). *The interpretation of unidimensional irt parameters estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Canada.

Reckase, M. D., & Hirsch, T. M. (1991, April). *Interpretation of number-correct scores when the true numbers of dimensions assessed by a test is greater than two.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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

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

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement*, *15*(2), 109-127.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21-28.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (Tech. Rep. No. ONR 90-8). Iowa City, IA: ACT.

Stoer, J., & Bulirsch, R. (2002). *Introduction to numerical analysis.* New York, NY: Springer-Verlag New York, Inc.

Stone, C. A., & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar examination. *Educational and Psychological Measurement*, *66*(2), 193-214.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stout, W., Douglas, B., Junker, B., & Roussos, L. (1999). *DIMTEST [Computer software]*. The William Stout Institute for Measurement, Champaign, IL.

Stout, W., Froelich, A., & Gao, F. (2001). Using resampling to produce and improved DIMTEST procedure. In A. Boomsma, M. A. J. van Dujin, & T. A. B. Snijders (Eds.), *Essays on item response theory* (p. 357-375). New York: Springer.

Sympson, J. (1978). A model for testing with multidimensional items. In *Weiss DJ (ed) Proceedings of the 1977 Computerized Adaptive Testing Conference.* University of Minnesota, Minneapolis.

Tanaka, J. (1993). Multifaceted Conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Test Structural Equation Models.* Newbury Park, CA: Sage.

Thurstone, L. (1947). *Multiple factor analysis.* Chicago, Illinois: The university of Chicago press.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement, 37*(2), 113-140.

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

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

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. R. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, p. 139-155). Norwood, NJ: Ablex.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada, June, 1-19.

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

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 02956 9583