The Effect of Guessing on Assessing Dimensionality in Multiple-Choice Tests: A Monte Carlo Study with Application

by

Chien-Chi Yeh

B.S., Chung Yuan Christian University, 1988

M.Ed., National Tainan Teachers College, 1998

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Chien-Chi Yeh

It was defended on

February 23rd, 2007

and approved by

Levent Kirisci, Associate Professor, Pharmaceutical Sciences and Psychiatry

Suzanne Lane, Professor, Psychology in Education

Kevin H. Kim, Assistant Professor, Psychology in Education

Dissertation Advisor: Clement A. Stone, Associate Professor, Psychology in Education

The Effect of Guessing on Assessing Dimensionality in Multiple-Choice Tests:

A Monte Carlo Study with Application

Chien-Chi Yeh, PhD

University of Pittsburgh, 2007

The purpose of this study was to investigate the effect of guessing in the assessment of dimensionality in multiple-choice tests using procedures implemented in Mplus and TESTFACT. Levels of item discrimination and the size of correlations between dimensions were also manipulated to explore any interaction between these effects. Four indices based on the proportion of variance, parallel analysis, RMSR reduction and a chi-square difference test were used to estimate dimensionality. The research included two parts, a simulation study using a Monte Carlo approach and an application with TIMSS 2003 data.

The simulation study confirmed the guessing effect. TESTFACT appeared to outperform Mplus for most conditions with data that assumed guessing. The proportion of variance and the RMSR reduction indices more accurately estimated dimensionality in Mplus, whereas the chi-square test and parallel analysis performed best with TESTFACT. A discrimination effect was observed clearly in data that assumed no guessing using the parallel analysis index and in data that assumed guessing using the RMSR index for both methods. Less accurate estimation of dimensionality was observed when using Mplus for tests with either high or low discriminating items, and with TESTFACT for tests with lower discriminating items. Higher correlations between dimensions led to more serious estimation problems. When guessing was not modeled, greater influence from the levels in correlations between dimensions and item discriminations was found. Further, a more pronounced discrimination effect was observed in the high correlation condition.

With regard to the application of TIMSS data, 70% of the items exhibited guessing behaviors and high correlations were observed between scores on the different dimensions (math and science). Based on the simulation study, guessing and correlation effects should thus be considered carefully when choosing a method for assessing dimensionality. Inconsistency in the dimensionality assessment using the four indices with Mplus was observed (1 to 5 dimensions),

whereas TESTFACT consistently estimated 2 dimensions. However, further investigation of the internal structure of the TIMSS assessment did not show any connection to content or cognitive domains.

TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... *XIII*

1.0 INTRODUCTION *1*

1.1 STATEMENT OF THE PROBLEM **1**

1.1.1 The importance of the assessment of dimensionality **1**

1.1.2 The influence of modeling guessing **3**

1.1.3 Methods of assessing dimensionality **5**

1.2 RESEARCH QUESTIONS AND THE DESIGN OF THE STUDY **6**

1.3 SIGNIFICANCE OF THE STUDY **7**

2.0 LITERATURE REVIEW *9*

2.1 INTRODUCTION **9**

2.2 THE ASSESSMENT OF DIMENSIONALITY **11**

2.2.1 Factor analysis (FA) approach **12**

2.2.1.1 Introduction **12**

2.2.1.2 The FA model for dichotomous variables **13**

2.2.1.3 Identification of number of factors/dimensions **15**

2.2.2 Multidimensional item response theory (MIRT) approach **17**

2.2.2.1 Models **17**

2.2.2.2 The relationship between FA and MIRT **20**

2.2.2.3 Methods based on MIRT **21**

2.2.3 Nonparametric methods **27**

2.2.4 Research comparing methods **30**

2.2.4.1 The guessing effect and the assessment of dimensionality **31**

3.0 METHODOLOGY 36

3.1 PHRASE I – SIMULATION STUDY 36

3.1.1 Design of the study 36

3.1.2 Generating data and estimating procedure 40

3.1.3 Validating data generation 43

3.1.4 Outcome measures 47

3.2 PHRASE II – APPLICATION OF THE THIRD INTERNATIONAL MATHEMATICS AND SCIENCE STUDY (TIMSS) 49

3.2.1 Introduction 49

3.2.2 The assessment frame of TIMSS 2003 50

4.0 RESULTS 53

4.1 THE RESULTS OF SIMULATION DATA 53

4.1.1 Number of non-convergent solutions 54

4.1.2 The proportion of correct dimensionality decisions 59

4.1.3 Comparing the number of dimensions 69

4.1.4 Parameter Recovery 82

4.2 THE RESULTS OF THE APPLICATION OF TIMSS 91

4.2.1 A description of the TIMSS sample 91

4.2.2 Guessing and the TIMSS 93

4.2.3 The dimensionality of the TIMSS 96

5.0 DISCUSSION 110

5.1 MAJOR FINDINGS OF THE SIMULATION STUDY 110

5.2 MAJOR FINDINGS OF THE TIMSS APPLICATION 117

5.3 LIMITATIONS 118

5.4 FUTURE RESEARCH DIRECTION 119

APPENDIX A 122

APPENDIX B 126

APPENDIX C *127*

APPENDIX D *128*

APPENDIX E *129*

APPENDIX F *130*

APPENDIX G *131*

BIBLIOGRAPHY *132*

LIST OF TABLES

3.1 The descriptive statistics of different levels of discrimination design under one-to-three dimensional data 38
3.2 Design of a two-dimensional data set with low discrimination 41
3.3 Item parameters of a two-dimensional data set with a low discrimination 42
3.4 The true values of IRT parameters for validation 44
3.5 Comparison of data generation using simple structure for uncorrelated ability dimensions. 45
3.6 Comparison of data generation using approximate simple structure for uncorrelated ability dimensions 46
3.7 Recovery of item parameters under PROMAX rotation 47
3.8 Percentages of content categories of TIMSS 2003 in Grade 8 51
3.9 Percentages of cognitive domains of TIMSS 2003 in Grade 8 51
3.10 Administration and item types of TIMSS 2003 assessment 52
4.1 Cases with convergent solutions in Mplus using WLS under the low correlation condition (r = .3) 56
4.2 Cases with convergent solutions in Mplus using WLS under the high correlation condition (r = .6) 57
4.3 Cases with convergent solutions in TESTFACT under the low correlation condition (r = .3) 58
4.4 Cases with convergent solutions in TESTFACT under the high correlation condition (r = .6) 59
4.5 The mean proportion of correct dimensionality decisions using the proportion of variance index 61
4.6 The mean proportion of correct dimensionality decisions using parallel analysis 62
4.7 The means of the first two eigenvalues in Mplus and TESTFACT (r = .3) 64

4.8 The means of the first two eigenvalues in Mplus and TESTFACT (r = .6) 65
4.9 The mean proportion of correct dimensionality decisions using the reduction of RMSR index 67
4.10 The mean proportion of correct dimensionality decisions using the chi-square test 68
4.11 Valid cases for factor solutions matched the underlying dimensionality in TESTFACT ... 83
4.12 The mean RMSD of parameter recovery in Mplus and TESTFACT 85
4.13 The standard deviation of the mean RMSD of parameter recovery in Mplus and TESTFACT 86
4.14 Descriptive statistics of TIMSS 92
4.15 Descriptive statistics of item discrimination parameters in TIMSS 92
4.16 Average proportion correct (p) for low-ability examinees on items where $p \leqq .7$ 95
4.17 Estimated dimensionality using the four indices in Mplus and TESTFACT 97
4.18 Number of substantial factor loadings for the two-factor solutions in Mplus and TESTFACT 99
4.19 Number of substantial factor loadings for the three-factor solution in Mplus and TESTFACT 99
4.20 Factor loadings of the PROMAX 2-factor solution for Mplus and TESTFACT (Booklet 5) 102
4.21 Factor loadings of the PROMAX 3-factor solution for Mplus and TESTFACT (Booklet 5) 104
4.22 Factor loadings of the PROMAX 2-factor solution for Mplus and TESTFACT (Booklet 11) 106
4.23 Factor loadings of the PROMAX 3-factor solution for Mplus and TESTFACT (Booklet 11) 108

LIST OF FIGURES

2.1 Four kinds of plots for interpreting item parameters of MIRT models 19
4.1 The mean difference of estimated and true dimensionality in Mplus (1D, c = 0) 70
4.2 The mean difference of estimated and true dimensionality in TESTFACT (1D, c = 0)..... 70
4.3 The mean difference of estimated and true dimensionality in Mplus (1D, c = .33) 71
4.4 The mean difference of estimated and true dimensionality in TESTFACT (1D, c = .33)..... 71
4.5 The mean difference of estimated and true dimensionality in Mplus (2D, c = 0, r = .3)..... 73
4.6 The mean difference of estimated and true dimensionality in TESTFACT (2D, c = 0, r = .3) 73
4.7 The mean difference of estimated and true dimensionality in Mplus (2D, c = 0, r = .6)..... 74
4.8 The mean difference of estimated and true dimensionality in TESTFACT (2D, c = 0, r = .6) 74
4.9 The mean difference of estimated and true dimensionality in Mplus (2D, c = .33, r = .3)..... 75
4.10 The mean difference of estimated and true dimensionality in TESTFACT (2D, c=.33, r = .3) 75
4.11 The mean difference of estimated and true dimensionality in Mplus (2D, c=.33, r = .6)..... 76
4.12 The mean difference of estimated and true dimensionality in TESTFACT (2D, c=.33, r = .6) 76
4.13 The mean difference of estimated and true dimensionality in Mplus (3D, c = 0, r = .3)..... 78
4.14 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = 0, r = .3) 78
4.15 The mean difference of estimated and true dimensionality in Mplus (3D, c = 0, r = .6)..... 79
4.16 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = 0, r = .6) 79

4.17 The mean difference of estimated and true dimensionality in Mplus (3D, c = .33, r = .3)... 80
4.18 The mean difference of estimated and true dimensionality in TESTFACT (3D, c=.33, r =.3) 80
4.19 The mean difference of estimated and true dimensionality in Mplus (3D, c = .33, r = .6)... 81
4.20 The mean difference of estimated and true dimensionality in TESTFACT (3D, c=.33, r = .6) 81
4.21 The mean RMSD of parameter recovery in unidimensional cases (Mplus) 88
4.22 The mean RMSD of parameter recovery in Unidimensional cases (TESTFACT) 88
4.23 The mean RMSD of parameter recovery in 2-dimensional cases (Mplus) 89
4.24 The mean RMSD of parameter recovery in 2-dimensional cases (TESTFACT) 89
4.25 The mean RMSD of parameter recovery in 3-dimensional cases (Mplus) 90
4.26 The mean RMSD of parameter recovery in 3-dimensional cases (TESTFACT) 90
4.27 Total score by the proportion correct for two items in Booklet 5 94
4.28 Total score by the proportion correct for two items in Booklet 11 94
4.29 The scree plot for Booklet 5 using Mplus and TESTFACT 97
4.30 The scree plot for Booklet 11 using Mplus and TESTFACT 98

ACKNOWLEDGMENTS

The completion of my dissertation was only accomplished with the assistance of teachers, friends and my family. First among these, I would like to express profound gratitude to my advisor, Dr. Clement A. Stone, for his invaluable support, encouragement, supervision and useful suggestions throughout this research. He was responsible for involving me in the MBE project, which helped me develop my research field. His patience and expert advice have enabled me to complete my work successfully.

Great appreciation is extended to my academic advisor, Dr. Suzanne Lane, who has been responsible for helping me complete our program. She always has confidence in me. This means a lot for me, especially when my spirits were low. Her guidance and encouragement prepared me for this dissertation and her constructive suggestions improved the quality of this dissertation.

I would also like to express my immeasurable appreciation to Dr. Kevin H. Kim. His support and quick responses kept me making progress and made me a better programmer. He brought forward good ideas and gave many insightful comments.

I really appreciate the kindness and all of the help from Dr. Levent Kirisci. What he said at our first meeting stayed with me, "We are here to help you shape the dissertation better," he said.

Sincere appreciation is extended to all my friends in Taiwan and in Pittsburgh for their encouragement and company, which released my stress and kept me moving forward. Their support was important to keep me going. I would also like to say 'thank you' to Kirstin Roehrich for editing this dissertation and helping me learn how to write in English.

Finally, I am, as ever, especially indebted to my parents and my family for their love and support throughout my life. They may not have known what I was doing for my study, but they were always there for me.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

1.1.1 The importance of the assessment of dimensionality

Providing evidence of validity is essential to the use of educational and psychological tests. This evidence is not only for obtaining a meaningful basis of test scores but also for knowing social consequences of score use (Messick, 1995). Several researchers have indicated the importance of providing validity evidence using the assessment of test dimensionality, especially in the development, evaluation, and maintenance of large-scale tests (e.g., Hattie, 1985; Nunnally & Bernstein, 1994; Tate, 2002). For example, the assessment of test dimensionality provides empirical evidence to examine the internal structures of tests underlying the responses to a set of items. This kind of evidence relates to the substantive aspect of validity indicated by Messick (1989). A test is developed for a specific purpose with a theoretical structure. This underlying test structure must be examined and confirmed. Assessing dimensionality helps to identify the construct defined by the test developer, and to examine how well the test measures the underlying structure. In other words, the test developer can use the assessment of dimensionality to identify what domains are measured and the relationships between those domains.

Furthermore, confirmation of the internal structure provided information that can be used to make a decision concerning what scores should be reported or what setting cutscores can be made based on the test structures. This information supplied evidence of the structural aspects of validity. When the dimensions are distinguishable, reporting subscores is appropriate; when there is only one dominant dimension, one total score is reported (Haladyna, 2004). For instance, a mathematics test measures algebra and geometry. If the information of dimensionality is

clearly presented in two dimensions in terms of algebra and geometry, then it is appropriate to report two subscores for algebra and geometry respectively. If the assessment of dimensionality shows only one dimension, reporting a total score for mathematics is preferable. Additionally, any information regarding the internal test structure could be the foundation of either "homogeneous" items in the classical test theory (CTT), or the "unidimensionality assumption" in item response theory (IRT; Tate, 2003). Moreover, for accountability and diagnosis purposes of large-scale assessment programs, the practice of reporting subscores has received more attention (e.g., Goodman & Hambleton, 2004; Martineau, Mapuranga, & Ward, 2006; Skorupski, 2005). For example, the No Child Left Behind Act of 2001 (NCLB, 2001) requests requiring statewide testing programs to provide both total scores and subscores of examinee performance (Goodman & Hambleton, 2004). The assessment of dimensionality can help collect evidence for correctly interpreting subscores and using subscores for instructional purposes.

In terms of the generalizability aspects of validity, the test developer should carefully consider the maintenance of score comparability across groups, settings, and tasks (Messick, 1995). Score comparability means that scores have comparable interpretations for different subgroups or on different occasions (e.g., over time). For example, when scores of large-scale tests are used to describe trends in schools, districts, and state achievement over time, the invariance of the tests' factor structures needs to be examined. When several test forms require an equating procedure for using at different points in time, the changes in test structures can be identified by tracking dimensionality of the tests over time (Tate, 2002, 2003). As indicated by Messick (1995), construct-irrelevant variance is a major bias source for the use and interpretation of test scores. The construct-irrelevant variance can lead to the differential item functioning (DIF) issues of fairness across groups. A test with DIF has items that function differently for different groups. However, a test is supposed to have the same measured function for all intended subgroups in order to use the test scores. One example of this concept can be found in the test scores of a reading comprehension exam between groups of native English speakers and people whose primary language is not English. The ability to speak, read, and write English becomes a key factor for attaining high scores. Consequently, the test is unfair for some subgroups or individuals. In other words, the invariance of test scores across groups provides the foundation for the fairness of test use. This kind of invariance relates to consequential construct

validity. The assessment of test dimensionality can help identify the sources of invalidity related to bias and fairness, such as DIF items in tests.

In summary, the assessment of the test dimensionality is able to identify the internal test structure for the following purposes: (1) to confirm the domains are being measured; (2) to understand the relationship between domains; (3) to examine and maintain the test structure across groups or over time. Furthermore, the assessment of dimensionality provides supporting evidence for validity, including the substantive, structural, generalizability and consequential aspects. In addition, the dimensionality is also useful identifying the major threats of construct validity, construct underrepresentation and construct-irrelevant variance.

1.1.2 The influence of modeling guessing

The issue of guessing is also important to multiple-choice assessments. First, guessing increases measured error since it raises the possibility of correct responses (Rogers, 1999). Also, as indicated by Messick (1995), guessing propensities can be the source of construct-irrelevant variance, which provides a major threat of construct validity. Second, the use of guessing strategies introduces error and attenuated the relationships among items. Therefore, it is reasonable and important to consider guessing in the assessment of dimensionality, especially with regard to multiple-choice tests. Although the guessing parameter is included in the three-parameter models in IRT, most of the methods for factor analysis do not include guessing in their models. In addition, most multidimensional item response theory (MIRT) approaches only allow fixed values of guessing in models (e.g., models implemented in TESTFACT and NOHARM).

Recently, Tate (2003) focused on the comparison of empirical methods in assessing test structure as well as the evaluation of guessing effect. He evaluated the estimated number of dimensions and parameter recovery in unidimensional and multidimensional data using both parametric and nonparametric approaches. Some conditions might be expected to be problematic for some of the methods, such as data with extreme difficulty and discrimination parameters, and one item pair with local dependence. For the evaluation of guessing effect, the results of exploratory factor analysis using Mplus obtained the correct decisions in only 3 out of 14 simulation cases. Without modeling guessing, Mplus did not perform well in the confirmation of correct dimensionality and overestimated dimensionality for all of the cases with guessing.

Additionally, the effect of guessing was reflected in the recovery of the true item parameters. By contrast, TESTFACT and NOHARM, which included guessing parameters in the models, performed well in identifying the correct dimensionality for most unidimensional or multidimensional cases. The results in Tate's study also illustrated the effect of guessing when different methods were used to identify dimensionality. The results also found some problems in assessing dimensionality when there was an interaction between item parameters, such as guessing versus discrimination or guessing versus difficulty. However, due to the large amount of selected methods used in Tate's study, only some specific test conditions were examined to show the differences between the various empirical methods for assessing dimensionality.

Another study about the assessment of dimensionality using real data, the Multistate Bar Examination (MBE), provided more understanding of the relationship between items and the internal structure of a test when guessing was modeled (Stone, & Yeh, 2006). The MBE was a four-option multiple-choice exam with 200 questions. For the 2001 February administration, the examination of the average proportion correct for low-ability examinees showed that more than 50% of items showed that guessing was operating. The results of three methods, Mplus, NOHARM and TESTFACT, demonstrated a similar pattern of dimensionality in conditions that did not model guessing. However, a comparison between NOHARM and TESTFACT showed more solid evidence for higher dimensionality and more indicators in the factors when guessing was modeled. The correction of tetrachoric correlations reflected more realistic relationships between items by considering errors caused by guessing behaviors. In other words, when guessing was operating on multiple-choice items, modeling guessing in the assessment of dimensionality became important. The results of the methods with modeling guessing provided more rich information not only for the assessment of dimensionality or the relationship between items, but also for assessing the internal test structures. Although the results found the influence of guessing in the assessment of dimensionality, the true underlying factor structure remained unknown. Therefore, it was impossible to investigate the effect of assessing dimensionality when guessing was modeled.

The results of the two studies mentioned above demonstrated the effect of guessing in determining the dimensionality or examining the internal test structures. However, due to the limitations of these studies, the effect of guessing has not been investigated in broader or more general conditions. Also, no examination determined the extent of recovery of a true

dimensionality or parameter measures (either factor loading in a factor analysis sense or item parameters in a MIRT sense). These two studies did not fully examine the interaction between guessing and other factors, such as difficulty and discrimination item parameters.

1.1.3 Methods of assessing dimensionality

The most common methods include traditional factor analysis, nonlinear factor analysis (NLFA), and the MIRT approaches. The equivalent of NLFA and MIRT has been discussed (e.g., Knol & Berger, 1991; Takane & Leeuw, 1987). Therefore, three kinds for methods of assessing dimensionality will be discussed here, including traditional factor analysis, the MIRT approach.

As for factor analysis of dichotomous data in multiple-choice tests, Mplus is the most commonly used software (Muthén, 1978). Mplus provides a categorical variable model for either dichotomous or ordered categorical data (Newsom, 2005). In this kind of model, the relationship between the factors and the items is nonlinear. In Mplus, the data for exploratory factor analysis (EFA) can be continuous, categorical, or a combination of both. As for dichotomous data (i.e., categorical data), Mplus provides several options for estimation, including the default option, unweighted least squares (ULS). Mplus allows users to perform EFA and confirmatory factor analysis (CFA) to estimate unidimensional or multidimensional models. Additionally, Mplus provides several statistics to evaluate model fit, such as chi-square fit statistics, root mean square residual (RMSR), and root mean square error of approximation (RMSEA). However, Mplus does not allow users to input guessing parameters.

Given the relationship between factor analysis and MIRT, several programs for assessing dimensionality are based on MIRT. Using different estimation methods, TESTFACT and NOHARM are the most popular programs. TESTFACT is based on the full-information item factor analysis proposed by Bock, Gibbons, and Muraki (1988), whereas NOHARM is based on a polynomial approximation to the normal ogive model developed by McDonald (1967, 1982). The estimates of TESTFACT are based on all of the item response information (i.e., "full-information"). The estimation procedure combines the marginal maximum likelihood (MML) estimation and the expectation-maximum (EM) algorithm (Bock & Aitkin, 1981). TESTFACT provides both EFA and CFA. Unlike Mplus, the fixed values of guessing parameters can be set

in the program. TESTFACT also provides chi-square statistics and residual matrix, but not a residual based fit index.

The ULS estimation method implemented into NOHARM provides a more efficient method than the generalized least square (GLS, used in Mplus) and maximum likelihood (ML, used in TESTFACT) procedures. Therefore, it can be used with a large number of items and/or dimensions. NOHARM also allows users to fit both the exploratory and confirmatory models. The matrix of covariance residuals and RMSR index provides information on the model's lack of fit. This approach does not provide the standard errors for parameter estimates nor tests of the model's goodness of fit (McDonald, 1997). However, Gessaroli and De Champlain (1996) developed an approximate chi-square statistic, which may be used to address this limitation. Several authors found that the performance of NOHARM was similar to that of TESTFACT (Knol & Berger, 1991; Stone & Yeh, 2006).

1.2 RESEARCH QUESTIONS AND THE DESIGN OF THE STUDY

According to the discussion above, the assessment of dimensionality provides not only information concerning the internal test structures, but also validity evidences. As discussed, most empirical methods of dimensionality assessment do not estimate a guessing parameter, and only a few methods allow incorporation of guessing parameters in the analysis of the factor structure. However, there has been no full investigation of the effect of guessing under more general conditions. Thus, the main purpose of this study was to investigate the effect of guessing in the assessment of dimensionality. Through the comparison of the traditional factor analysis and MIRT approaches, the influence of modeling guessing in the assessment of dimensionality can be detected under general conditions. At the same time, manipulation of other item characteristics and factors provided information about the influence of the interaction between these factors. The following research questions were addressed in this study.

1) What is the effect of guessing on assessing dimensionality of multiple-choice tests?
2) How well do different indices perform for estimating the number of dimensions when assessing dimensionality?

3) Does the discrimination level for items affect the assessment of dimensionality?

4) Does the correlation among dimensions affect the assessment of dimensionality?

5) What is the interaction between the guessing effect and the level of discrimination of items and correlations between dimensions?

In order to investigate the effect of guessing, this study used a Monte Carlo approach. Through the known dimensionality, it was possible to examine the differences between estimated and true dimensionality. This study considered the assessment of dimensionality for one, two, and three dimensional data. The main manipulated variables included guessing parameters, item discrimination parameters, and the correlations among dimensions/factors. There were two levels for guessing (0 and .3) whereas there were three levels of discrimination (Low, Medium, and High). The correlations between dimensions were .3 and .6. Two estimation methods as implemented in Mplus and TESTFACT, were used to compare the effect of incorporating a guessing parameter.

The second part of this study included an application with real data. The purpose was to examine if the findings from in the simulation study could be useful for practical applications. The Third International Mathematics and Science Study (TIMSS) in 2003 was selected, and the same procedures used to analyze the data. TIMSS contained two major subjects, mathematics and science.

1.3 SIGNIFICANCE OF THE STUDY

Several studies have been conducted to evaluate different empirical methods used for assessing dimensionality (e.g., Knol & Berger, 1991; Stone & Yeh, 2006; Tate, 2003). Tate (2003) conducted a study that compared the performances of selected methods. Tate's study provided a wealth of information about new procedures for assessing test structures. However, Tate did not consider other issues, such as the probability of empirical power or other factors that might affect the assessment of test dimensionality. In Stone and Yeh's study (2006), the results confirmed the effect of guessing as well. However, this study was conducted using real data. This research extends these previous studies by focusing on the effect of guessing with

simulation data. In order to increase the generalizability of the results of this study, more general test setting conditions were considered for manipulation: three discrimination levels, the use of approximate simple structures, and different correlations between dimensions. In this study, three-dimensional data were included to explore the possible influence of higher dimensionality. A multiple replication design was used to deal with statistical inference issues. This design allows for evaluating the empirical power of the procedures. Moreover, this study focused on the influence of the variation of discrimination and guessing parameters in assessing dimensionality. It was possible to investigate the interaction between the discrimination and guessing parameters. Finally, as a suggestion by the software developers (Wilson, Wood, Gibbons, Schilling, Muraki, & Bock, 2003), estimation of guessing was included instead of incorporating true guessing values in Tate's study.

2.0 LITERATURE REVIEW

2.1 INTRODUCTION

Collecting evidence of validity is essential for test development and maintenance. In the development stage, test developers need to identify constructs based on the purpose of the measurement and its assumed theoretical framework, as well as determine the number of scores to report based on the theoretical framework. The test structure or dimensionality represents the content or process structure that a test is intended to measure. In other words, the information about dimensionality provides structural evidence for the consistency between the internal structure of a test and the structure expected by the known definition of construct domains (Fiske, 2002). Therefore, if any constructs are overrepresented or underrepresented, this situation can be detected by the assessment of dimensionality and the factor structure based on the estimated dimensionality. Moreover, the information of test structure provides the foundation for determining how to report scores, either total scores or subscores. The decision should be determined based on the relationship between factors or ability domains. Reporting subscores is appropriate when the distinguishable factors can be identified or when the test structures were designed at the beginning of developing tests. Otherwise, only a single total score should be reported (Haladyna, 2004).

In the next stage, evaluating the tests, researchers provide evidence of reliability and validity to confirm the consistency of scores, the accuracy of the test used, and the test score interpretations. This kind of evidence includes internal consistency, reliability, information of adequacy and appropriateness of test scores reported and so on (e.g., Haladyna, 2004; Hattie, 1985; Nunnally & Bernstein, 1994; Tate, 2002). Score reliability in classical test theory (CTT) is based on the assumption of unidimensionality. If there is any violation, it is necessary to make adjustments to achieve an accurate estimated reliability (Tate, 2002). On the other hand, if test

developers assume that the structure is multidimensional, the assessment of dimensionality can provide evidence for hypothesized multidimensionality. Furthermore, through assessing the dimensionality, two major threats to test score interpretation, construct underrepresentation and construct-irrelevant variance, can be detected (Messick, 1989). For instance, a mathematics test is designed to measure algebra and geometry knowledge. The results of the test structure or dimensionality can show the test developer if there are only two dimensions (presented algebra and geometry factors) in this test or if any suspicious factors (i.e., construct-irrelevant sources) show up in the test structures. One example of a suspicious factor can be item difficulty or item format. It is possible that the dimensionality and test structures may show that other constructs, such as reading ability, also affect test scores.

Regarding issues of score reporting, as indicated by Haertel (1999), large-scale assessments serve an important role in providing information for accountability, evaluation purpose. In addition, more detail for diagnostic information of test results is required in the NCLB (2001). Not only the total scores but also the subscores are reported for obtaining the information about student achievement and growth over time (Goodman & Hambleton, 2004; Martineau, Mapuranga, & Ward, 2006). Most academic assessments require the use of multiple skills to succeed in proficiency. In general, test constructs include several content domains, especially shown in K-12 state assessment. Therefore, evaluating content validity becomes more important for large-scale assessment programs since this sort of information helps test developers to identify interpretable dimensions (Martineau et al, 2006). In other words, confirming the dimensionality of academic assessments helps define interpretable dimensions and decide what subscores are reported.

The maintenance of score comparability across groups, settings, and tasks should be considered carefully since it is important for the generalizability aspect of validity (Messick, 1995). Score comparability means that scores have comparable meanings for different subgroups or on different occasions (Tate, 2002). For example, when the scores of large-scale tests are used to describe trends in schools, districts, and state achievement over time, it means that the scores reported from different time points have to represent the same meanings. Therefore, it is necessary to examine the invariance of the factor structures of the tests over time. Equating issues also arise when test developers use several test forms at different time points. It is necessary to identify that the test structures in different forms are similar. The changes to the test

structures can be identified by tracking the test dimensionality over time and across forms as well (Tate, 2002, 2003). Moreover, concerning consequential construct validity, it is important to confirm if a test has the same measured function for different subpopulations. As indicated by Helms (2003), if the students differed from the norm group of a test on important dimensions, such as ethnicity, issues of fairness and valid test use will arise. Ethnicity can be a bias source (i.e. construct-irrelevant variance) for the use and interpretation of test scores. Consequentially, the bias source can lead to the different item function (DIF; Tate, 2002). The assessment of the dimensionality helps to identify items with DIF among subgroups.

In conclusion, the assessment of dimensionality provides information about the internal structure and supports evidence of construct validity. In the development stage, the dimensionality helps to identify and confirm the intended structure or the discrepancy between the empirical and expected structures. During the evaluation period, this assessment provides supportive evidence of construct validity, including generalizability as well as structural and consequential aspects. The assessment of dimensionality identifies the sources of the major threats to construct validity, including underrepresentation and construct-irrelevant variance.

2.2 THE ASSESSMENT OF DIMENSIONALITY

Since 1985, methods for the study of dimensionality have evolved due to the development of new methods for dichotomous data and the argument concerning the assumption of the unidimensionality of tests. There are two major approaches for the assessment of test structures, the parametric and nonparametric methods. The parametric approach contains item factor analytic (FA) methods, and methods based on multidimensional item response theory (MIRT). The item factor analytic methods are based on nonlinear factor analysis (NLFA) models, and were developed for dealing with the problems caused by dichotomous variables. It has been proven that the NLFA and MIRT models are mathematically equivalent when the distribution of ability is normal (e.g., Knol & Berger, 19991; Takane & De Leeuw, 1987). Therefore, MIRT can be considered either a special case of FA or an extension of item response theory (IRT). Nonparametric approach receives more attention due to the failure of parametric IRT models and applications with shorter test lengths and smaller sample sizes (Tate, 2003).

The following sections introduce the fundamental concepts of these three approaches, including their models and estimation methods.

2.2.1 Factor analysis (FA) approach

2.2.1.1 Introduction

The common factor model for classical linear factor analysis is defined as the following (Gorsuch, 1983):

$$X_{iv} = w_{v1}F_{1i}+w_{v2}F_{2i} + \ldots + w_{vf}F_{fi} + w_{vu}U_{iv}, \qquad (2.1)$$

where X_{iv} is score on variable v of examinee i, w_{vf} is the weight for variable v on factor f, and F_{1i} to F_{fi} are factor scores of examinee i on the f factors, w_{vu} is the weight for variable v of the unique factor, and U_{iv} is the unique factor scores of examinee i for variable v. In the common factor model, the factors are divided into two groups. The first group consists of the common factors, F_{1i} to F_{fi}. Each of the common factors contributes to two or more variables, which means several variables have these factors in common. The noncommon factor variance for each variable, that includes the uniqueness from each variable, the random error of measurement, and all other sources of error and bias not defined by the model, is summarized in a unique factor. Therefore, sometimes, part of $w_{vu}U_{iv}$ is written as residual, e_i. In classical linear factor analysis, either the observed variables (i.e., item scores) or latent variables (i.e., factors) are assumed to be continuous, even though in some conditions the variables may be dichotomized. Therefore, Pearson or phi correlations are used in traditional linear factor analysis to represent the linear relationships between variables.

FA of dichotomous variables (i.e., NLFA) is an extension of classical linear factor analysis, sometimes called item factor analytic methods (Tate, 2002). However, several problems may arise in the factor analysis of dichotomous variables. Traditional factor analysis methods use correlations based on linear relationships between variables (e.g., Pearson or phi correlations). Since there is a score of only 0 or 1 on dichotomous items, the relationship between item scores and the continuous latent variables is nonlinear (Mislevy, 1986). When phi or Pearson correlations are used in FA of dichotomous data, this situation may lead to the identification of spurious "difficulty" factors. Consequently, it is possible to underestimate the

factor loadings and overestimate the number of dimensions (Bock, Gibbons & Muraki, 1988). Therefore, tetrachoric correlations are used instead of phi correlations in dichotomous cases. Unfortunately, several authors have indicated a number of drawbacks using the tetrachoric correlation matrix (Bock, Gibbons & Muraki, 1988; Knol & Berger, 1991; Pang, 1999). Firstly, the tetrachoric matrix may not be positive definite, therefore, causing a problem for the maximum likelihood (ML) method and the generalized least square (GLS) method, which are common FA estimation methods (Knol & Berger, 1991). Secondly, the tetrachoric correlation matrix is not a ML estimator of the population matrix. Thirdly, the calculation of tetrachoric correlations is unstable when the values approach +1 or –1 (Bock, Gibbons & Muraki, 1988). Generally, the methods for calculating tetrachoric coefficients are accurate, although, a large sample size is necessary for the computational accuracy of the tetrachoric matrix estimates. Additionally, the matrix of sample tetrachoric correlation coefficients may produce Heywood cases with communalities greater than one, which imply one or more unique variances are negative values (Hattie, 1984; Knol & Berger, 1991; Nandakumar, 1991). It is difficult to interpret the results of communalities greater than one. The presence and correction of guessing further undermines the use tetrachoric correlations (Carroll, 1945).

The development of factor analysis methods for categorical variables has been promising in the past decade, including one traditional factor analysis using GLS method (Christoffersson, 1975; Muthén, 1978), and two methods based on MIRT approach, using the ML (Bock & Aitkin, 1981), and the unweighted least-squares (ULS) methods (McDonald, 1967, 1994). These three methods were developed to deal with the problems that may arise in the analysis of dichotomous data.

2.2.1.2 The FA model for dichotomous variables

The FA model for dichotomous data (Christoffersson, 1975; Muthén, 1978) can be defined as

$$\mathbf{Y}=\mathbf{\Lambda\theta}+\mathbf{E}, \qquad (2.1)$$

where $\mathbf{Y}$ is the latent continuous variables (Y_1, …, Y_n), $\mathbf{\Lambda}$ is a matrix of factor loadings of items (λ_{i1}, …, λ_{im}) and $\mathbf{\theta}$ is ability values of examinees (θ_1, …, θ_n). This is a common factor model, but $\mathbf{Y}$ is unobserved. Also, the response variables X_i are defined by the unobserved variables Y_i and threshold variables τ_i as following:

$X_i = 1$, if $Y_i > \tau_i$; otherwise, $X_i = 0$.

Under the assumptions that $\boldsymbol{\theta} \sim$ MVN (0, **I**), $\mathbf{E} \sim$ MVN (0, Ψ^2), both are multivariate normality, where Ψ^2 is a diagonal matrix of residual covariance, and Cov ($\boldsymbol{\theta}$, **E**) = **0**. A GLS estimation procedure has been used to estimate the parameters of the Christoffersson's model; this method minimizes the following fit function:

$$F = (\mathbf{p}-\mathbf{P})' S_e^{-1} (\mathbf{p}-\mathbf{P}), \qquad (2.2)$$

where S_e is a consistent estimator of the residual covariance matrix; **P** is a vector of the expected proportion correct of items P_j and P_{jk}, and **p** is an observed proportion correct of items P_j and P_{jk} (De Champlain, 1999). In Muthén's GLS procedure, the parameters are estimated by minimizing a fit function in the following manner:

$$F = \tfrac{1}{2} (\mathbf{s}-\boldsymbol{\sigma})' \mathbf{W}_\delta^{-1} (\mathbf{s}-\boldsymbol{\sigma}) \qquad (2.3)$$

where $\boldsymbol{\sigma}$ is the population threshold and tetrachoric correlations, **s** is the estimates of threshold and tetrachoric correlations from the sample, and $\mathbf{W}_\delta$ is a weighted matrix. This estimation is equivalent to the Christoffersson's estimator but more efficient in computation. This approach is termed a full-weight matrix approach because a $p^* \times p^*$ weight matrix is used, where p^* is the total number of the elements in the **s** vector (De Champlain, 1999).

Muthén's GLS procedure was incorporated into the computer program Mplus. Generally, Mplus can be used to fit unidimensional or multidimensional models using exploratory and confirmatory approaches. However, the program does not attempt to correct for guessing or to "smooth" the tetrachoric correlations. Although this estimation utilizes the joint proportion correct for items taken one to four at the same time, which are the one-way, two-way, three-way, and four-way margins, it does not use all of the available information as TESTFACT does. Statistical tests of model fit and standard error of estimation are available in this procedure. Computing requirements increase quickly when the test length and number of factors increase. Hence, test length could be limited depending on the capacity of the computer. The newer versions of Mplus provide several options for exploratory factor analysis for categorical variables, including ULS, weighted least squares (WLS), weighted least squares with robust standard errors and mean-adjusted (WLSM), and robust weighted least squares with mean- and variance-adjusted (WLSMV). The default method of estimation is the ULS estimator. ULS estimation needs to meet the multivariate normality assumption. The other three options are all

similar to the asymptotic distribution free function (ADF), but different weight matrixes are chosen. Therefore, WLS, WLSM and WLSMV make no assumptions about the population distribution. However, the WLS estimation needs a relatively large sample size, which may be problematic in most educational settings. Therefore, the last two options are preferable for use with dichotomous data. One option for categorical variables suggested by the user guide of Mplus is WLSMV because the weight matrix chosen (all off-diagonal elements are set to zero) is simpler than the one used in WLS (all off-diagonal elements are estimated). Hence, larger sample sizes are not required as for WLSMV. However, due to the way degrees of freedom are calculated with the WLSMV method, it is impossible to evaluate nested models using the chi-square difference test with WLSMV outcomes.

2.2.1.3 Identification of number of factors/dimensions

Deciding how many factors in a FA reflects the dimensionality of a set of variables. It also is an important issue in assessing test structure because the test structure can be derived by examining the pattern of factor loadings. The decision needs to balance the requirement for a parsimonious test structure and a solution in which there are enough common factors to account adequately for the relationship among measured variables (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Generally speaking, more factors are preferable than fewer factors because a larger number of factors provides more information about the relationship between variables. In addition, empirical research suggests that over-factoring introduces less estimated error than under-factoring does (Fava & Velicer, 1992; Wood, Tataryn & Gorsuch, 1996). However, a meaningful or theoretical explanation is more important than just fitting a model (Cudeck, 2000).

A number of methods for identifying the number of factors have been proposed. First, the most commonly employed method, the K1 rule developed by Kaiser (1960), retains those factors with eigenvalues greater than 1.0. However, the K1 rule is intended to overestimate the number of factors. Secondly, since substantial changes in eigenvalues for consecutive factors indicate the presence of significant factors, the number of factors can be determined by a scree test (plots of eigenvalue number versus eigenvalue scale). According to scree plots, the number of factors is equal to one fewer than the solution corresponding to the "elbow" (Kim & Mueller, 1978) or identified by the last substantial drop in the magnitude of the eigenvalues (Fabrigar et al., 1999). This method has been criticized because there may be no clear way to identify the

"elbow" or "substantial drop". Therefore, it performs well only when strong common factors are present. Third, the amount of variance explained is an index for the substantive importance of factors. A number of factors should be retained until a certain amount of total explained variance is achieved, for instance, 70%. Alternatively, the number of factors can be found by setting a criterion of what should be considered the minimum contribution by a factor to be evaluated as substantively significant. For example, factors that have more than 5% explained variance could be retained. The major advantage is that it is easy to interpret. On the other hand, the major disadvantage of this method is that it uses a subjective criterion (Kim & Mueller, 1978). Finally, parallel analysis is based on the comparison of eigenvalues obtained from sample data and completely random, uncorrelated data, proposed by Horn (1965). The number of factors is obtained by comparing scree plots based on the real data versus random data. As indicated by Horn (1965), the mean eigenvalues of the random data provides a baseline for the comparison between the real data and random data. Recently, several researchers suggested using the desired percentile of the distribution of random data eigenvalues, such as 95% (O'Connor, 2000).

In addition to using eigenvalues as an index for the determination of dimensionality, there are other ways to identify the number of factors. There are several statistics provided by Mplus to evaluate model fit: 1) chi-square fit statistics; 2) root mean square residual (RMSR, since it is a standardized statistic, therefore, also referred as SRMR); 3) root mean square error of approximation (RMSEA) that corrects the chi-square statistics for model complexity. The chi-square fit statistics provide information about the difference between observed and expected correlations defined by models. In other words, the chi-square test evaluates whether the observed data corresponds to the expected data. The RMSR index calculates the standard deviation of the difference between observed and expected correlation matrices. Larger values of RMSR mean a greater difference between the observed data and data expected under the model. RMSEA provides information similar to the chi-square statistics but that corrects the chi-square statistics based on model complexity. Therefore, RMSEA is not affected by sample size. Since the chi-square test is affected by sample size, it is more appropriate to use RMSR and RMSEA to evaluate the model fit under larger samples. The model fit decision can be based on the percentage of reduction of the RMSR (Tate, 2003) or cut-off values (e.g., RMSR < .08, Hu & Bentler, 1999).

2.2.2 Multidimensional item response theory (MIRT) approach

In the late 1970s and early 1980s, researchers began developing practical MIRT models. The link between the normal ogive model and the FA of dichotomous variables greatly contributed to the development of MIRT. Reckase (1972) proposed the multidimensional Rasch model. Later, McKinley and Reckase (1982) considered a greater variety of general Rasch models and settled on the linear logistic form, known as a compensatory model. The development of partially compensatory or noncompensatory models proposed by Sympson (1978) and Whitely (1980) was a separate line of MIRT. These noncompensatory models have become more popular and are commonly used in psychology for ability tests. These two models are presented first. Next, the relationship between NLFA and MIRT is illustrated. Four commonly used methods in FA are introduced at the end of the section, including models, estimation procedure, and the index for determining dimensionality. Note that the informal indices for determining dimensionality illustrated in Section 2.2.1.3 can be used for the methods of the MIRT approach. Only the formal indices will be introduced when presenting each method.

2.2.2.1 Models

The compensatory models in MIRT assume that low ability on any dimension can be compensated for with greater ability on another dimension. For example, a spatial reasoning problem can be solved either as a feature of comparison strategy or as a mental rotation strategy (Bolt & Lall, 2002). Equation 2.4 presents the multidimensional compensatory three-parameter logistic model (MC3PL) (Spray, Davey, Reckase, Ackerman, & Carlson, 1990; Reckase, 1997).

$$P(x_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,d_i,c_i) = c_i + (1-c_i)\frac{1}{1+\exp[-1.7(\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i)]} \quad (2.4)$$

where x_{ij} = response of person j to item i (0 or 1); $\boldsymbol{\theta}_j$ = a vector of latent abilities; $\mathbf{a}_i$ = a vector of discrimination for item i in dimension k; d_i= easiness intercept for item i, related to the difficulty for item i; c_i = the guessing parameter of item i. The potential of person j on item i is reflected by a sum of k weighted traits (i.e., discrimination). Therefore, the model is referred to as an additive model (Reckase, 1997). When c_i sets to zero, this model becomes a multidimensional

compensatory two-parameter logistic model (MC2PL). Moreover, the MC2PL model becomes a multidimensional Rasch model when all of the discrimination parameters are set to 1. The interpretations of item parameters are the same across all compensatory models, and are similar to models in unidimensional IRT. Item discrimination is related to the slope of the item's characteristic surface (ICS, discussed later) in the direction of the corresponding ability axis. The higher the value of discrimination of one dimension indicates a greater importance of that dimension (or trait) in item success (Embretson & Reise, 2000). An overall discrimination index is defined as the maximum discrimination index (MDISC, see Equation 2.5, Reckase & McKinley, 1991). Item difficulty, b_i, is defined in second part of Equation 2.5 (Reckase, 1985) and related to d_i, which is called the easiness intercept by Embretson and Reise (2000). The larger the easiness intercept, the smaller the value of difficulty will be, which means the item is easier. The guessing parameter has the same meaning as in the unidimensional IRT model, representing the probability of a correct response for examinees that are very low in all dimensions.

$$\text{MDISC}_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2} \qquad \text{and} \qquad b_i = \frac{-d_i}{MDISC_i} \tag{2.5}$$

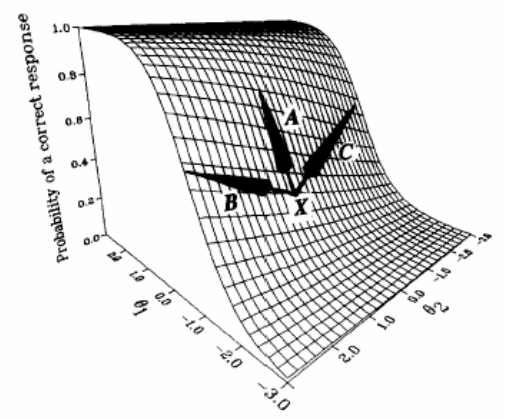
There are four kinds of plots for interpreting the item parameters of MIRT models, shown in Figure 2.1 (cited from Ackerman, 1994). These plots are presented based on two-dimensional data. First, ICS represents the probability of a correct response to an item depending on two or more abilities in MIRT as analog to the item characteristic curve (ICC) in unidimensional IRT. Figure 2.1(a) shows the ICS of an item of a two-dimensional MC2PC model. An ICS leans toward the dimension with a higher MDISC value. The signed distance from the origin to this p=.5 equiprobability line is defined as the difficulty parameter (Ackerman, 1996).

Second, a contour plot of an ICS (see Figure 2.1(b)) is more informative than graphing an ICS. All examinees with the same probability of a correct response lie in a line, called the equiprobability contour. These lines are parallel, but only the contours of compensatory models are straight lines. The higher discrimination (or the steeper the slope of the surface), the closer together the contour lines will be.

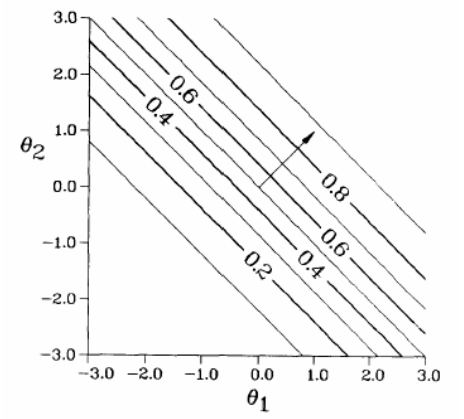
Third, Figure 2.1(c), the item vector plot, depicts each item as a vector in an orthogonal Cartesian coordinate system for a two-dimensional compensatory model. The vectors are only presented in the first and third quadrant because MDISC is always positive (Ackerman, Gierl, &

Walker, 2003). The length of vectors represents the values of MDISC. The representation of item difficulty is as same as in an ICS contour plot. Easy items lie in the third quadrant, while difficult items lie in the first quadrant. The shape of arrows represents the size of guessing parameters.

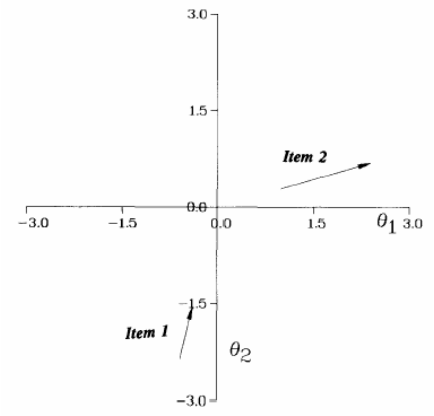
Fourth, referred to Figure 2.1 (d), the contour plot of a test characteristic surface (TCS) is similar to the contour plot of ICS. Examinees with the same true score all lie in a line. The true score, ξ, is the sum of the probability of getting each item correct at θ points (i.e., $\xi=\Sigma P_i$) each. The equi-true-score contours are not linear but rather curvilinear. The closer the contours are, the steeper the TCS surface will be or more precise the measurement precision will be.



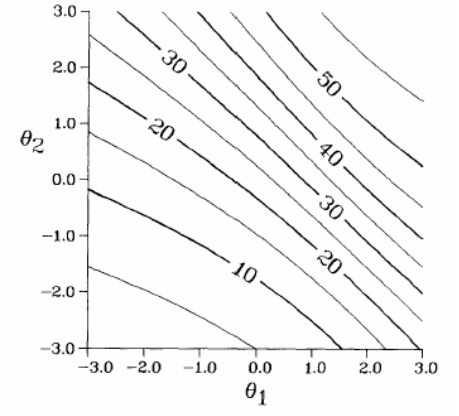
(a) An ICS plot



(b) A contour plot of an ICS



(c) An item vector plot



(d) A contour plot of the TCS

Figure 2.1 Four kinds of plots for interpreting item parameters of MIRT models

The multidimensional noncompensatory three-parameter logistic model (MNC3PL) is defined as the multiplication of the probability of a correct response for each dimension (Sympson, 1978):

$$P(x_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_i,\mathbf{b}_i,c_i) = c_i + (1-c_i)\prod_{k=1}^{m}\frac{1}{1+\exp[-1.7a_{ik}(\theta_{jk}-b_{ik})]} \quad (2.6)$$

All of the parameters in Equation 2.6 are defined as the same as in the MC3PL model (Equation 2.4). The noncompensatory nature derives from the fact that the probability of a correct response can never be greater than the minimum value of the terms in the product (Spray et al., 1990). In this model, because the values of the exponents are fixed, the probability of a correct response decreases when the dimensionality number increases. Hence, the model is referred to as a multiplicative model. In other words, a lower ability in one dimension cannot be compensated for higher abilities in any other dimensions. A noncompensatory model is more appropriate when a multidimensional test requires the simultaneous application of two or more abilities to answer each item correctly (Ansley & Forsyth, 1985; Sympson, 1978). An example is a math word problem test requires reading and mathematic abilities at the same time.

Noncompensatory models have been less commonly used partly due to the increased number of parameters that require estimation. However, it is not always clear when to apply either compensatory or noncompensatory models, because the relationship between abilities required for answering correctly is sometimes unclear. According to Embretson and Reise (2000), compensatory models can be applied to personality and attitude measurement because of the direct interpretability of these constructs. On the other hand, some research in decomposing cognitive processes, such as verbal analogies and abstract reasoning, demonstrates that processing components have noncompensatory impact on item success. In these situations, the noncompensatory models can be applied (e.g., Embretson, 1995; Whitely, 1980).

2.2.2.2 The relationship between FA and MIRT

Several researchers have shown that the compensatory MIRT and the NLFA models are mathematically equivalent (Knol & Berger, 1991; Takane & De Leeuw, 1987). A typical factor analytic model can be expressed as follows (Gorsuch, 1983):

$$Y_i = \lambda_{i1}f_1 + \lambda_{i2}f_2 + \ldots + \lambda_{ik}f_k + \varepsilon_i \quad (2.7)$$

where Y_i are observed scores of item i; f_k are the latent traits or factors (k = 1 , 2,…, m) and λ_{ik} are factor loadings. Given that each dichotomous item has a threshold parameter (γ_i) determining whether a response is correct, the model in Equation 2.7 can be rewritten as a normal distribution function (Φ) for a correct response (McLeod, Swygert & Thissen, 2001):

$$P(u_i = 1 \mid \theta) = \Phi[\frac{\lambda_{i1}\theta_1 + \lambda_{i1}\theta_2 + ... + \lambda_{ik}\theta_k - \gamma_i}{\sigma_i}] \quad (2.8)$$

where u_i are responses of examinees in item i, θ_k denotes the latent traits (e.g., f_k), and σ_i are unique variances (variances of ε_i). Letting

$$a_{ik} = \lambda_{ik} / \sigma_i \quad \text{and} \quad d_i = -\gamma_i / \sigma_i \text{ , where } \sigma_i = \sqrt{1 - \sum \lambda_{ik}^2} \text{ ,} \quad (2.9)$$

the model defined in Equation 2.8 becomes the multidimensional normal ogive model:

$$P(u_i = 1 \mid \theta) = \Phi[a_{i1}\theta_1 + a_{i1}\theta_2 + ... + a_{im}\theta_m + d_i] \quad (2.10)$$

Based on the relationship between FA and MIRT (i.e., Equation 2.8 and Equation 2.10 are equivalent), MIRT parameters can be derived from FA model parameters and vice versa. The discrimination a_{ik} and intercept d_i in MIRT can be calculated using Equation 2.9. On the other hand, given the MIRT parameters, FA parameters can be obtained by:

$$\lambda_{ik} = \frac{a_{ik}}{\sqrt{1 + \sum a_{ik}^2}} \quad \text{and} \quad \gamma_i = \frac{-d_i}{\sqrt{1 + \sum a_{ik}^2}} \quad (2.11)$$

2.2.2.3 Methods based on MIRT

Four methods used in factor analysis based on MIRT were presented in this section. Only the last method, ConQuest is based on Rasch model whereas the other methods were based on three-parameter models.

(1) Full-information item factor analysis by TESTFACT

Bock, Gibbons, and Muraki (1988) developed the exploratory full-information item factor analysis and implemented it in the TESTFACT program. The estimates are based on all of the item response information, so called "full-information," not the partial or summary information used in Mplus and NOHARM. The full-information factor analysis is based on Thurstone's multiple-factor model, and the estimation procedure combines marginal maximum likelihood (MML) estimation and the expectation-maximum (EM) algorithm (Bock & Aitkin,

1981; Dempster, Laird, & Rubin, 1977). This method estimates the nuisance (i.e., subject's ability) and structure parameters (i.e. item parameter) simultaneously. In addition, the posterior distribution (quadrature form) of θ for an examinee is used to estimate θ distribution. An iterative procedure, called the EM algorithm including E- and M- steps, is used to find the maximum likelihood estimation (MLE) for parameters in the presence of the unobserved random ability parameter (θ).

The E-step (Expectation) is used to compute expected values for the distribution of the unobserved ability variable based on observable data. First, the conditional probability, $L_l(X_k)$, of the binary response pattern, $\mathbf{x}_l$, given $\theta = X_k$, and the marginal probability of $\mathbf{x}_l$, $\widetilde{P}_l = \sum_k^q L_i(X_k)A(X_k)$, where X_k is the k-th discrete value (quadrature point) for θ and $A(X_k)$ is the weight of the Gauss-Hermite quadrature (Bock & Aitkin, 1981) are computed. Then the cumulating expected frequencies, $\bar{r}_i$, and the expected number of people with ability **X** normalized to the sample size, $\overline{N}_k$ (Bock, Gibbons, & Muraki, 1988). The M-step (Maximization) is used to compute the values from the E-step to maximize the MML equation (see Equation 2.12) and obtain the item parameters.

$$L_M = P(X) = \frac{N!}{r_1!r_2!...r_s!}\widetilde{P}_1^{r_i}...\widetilde{P}_s^{r_s} \qquad (2.12)$$

Any correlations with extreme proportions are replaced with values calculated by the one-factor version of Thurstone's centroid method. TESTFACT provides a "smooth" option to deal with a non-positive definite tetrachoric correlation matrix (Tate, 2003). The fixed values of guessing parameters can be set in the program. These parameters are used to adjust the computation of tetrachoric correlations and provide the lower asymptote for the solutions (Tate, 2003). If the model includes guessing parameters, Equation 2.13 is used to suppress the artifact effects introduced into the model by calculating a corrected proportion for the four-fold tables, where π_j is measured by the proportion of persons passing item *j*.

$$\pi_j' = 1 - (\frac{1-\pi_j}{1-c_j}) \qquad (2.13)$$

The factor solutions can be rotated orthogonally (VARIMAX) and obliquely (PROMAX). TESTFACT does not provide an RMSR index, but a residual matrix. The dimensionality

decision is based on a test of the difference between chi-square statistics for selections with different numbers of factors. For example, if a second factor is added to the model and the improvement is not significant, then one can conclude that the test is unidimensional. Otherwise, it is necessary to continue adding one factor at a time until the improvement of model fit is not significant. Moreover, the test of exact fit, the likelihood ratio (LR) statistic (Lawley, 1940), is commonly used for assessing model fit in an ML solution. If there is a sufficiently large sample size assuming a normal distribution, the LR statistic approximates a chi-square distribution when the specified number of factors is correct in the population. However, it is highly influenced by sample size. In addition, the hypothesis of perfect fit is not of empirical interest because the goal of FA is to look for a parsimonious solution, not a perfect fit solution (Fabrigar et al., 1999).

The main advantage of the "full-information" method is that it uses all available information in the estimation procedure (i.e., analyzes item response patterns). Additionally, the output contains classical item statistics and factor analytic parameter estimates with standard errors and a likelihood-ratio chi-square test for model fit. Nevertheless, there are some limitations. First, TESTFACT requires no empty cells in the 2^p item vectors (p = the number of items), which restricts the application to practical testing situations (De Champlain, 1999). Several researchers are concerned about the reliability of G^2 likelihood chi-square test statistics, when the expected frequencies are small or near zero (Mislevy, 1986; Knol & Berger, 1991; Wilson, Wood, & Gibbons, 1987). The run time for TESTFACT is an issue with large number of items (Knol & Berger, 1991).

(2) Nonlinear factor analysis by NOHARM

McDonald (1967, 1982) developed the Normal-Ogive Harmonic Analysis Robust Method and implemented it in the NOHARM software. The probability of correct response in multidimensional cases is defined as:

$$P(x_i=1|\boldsymbol{\theta}) = c_i + (1 - c_i)\, N(f_{i0} + \boldsymbol{f_k}\boldsymbol{\theta}) \qquad (2.14)$$

where $\boldsymbol{f_k}$ is an m×1 vector of discrimination parameters on k dimensions. This model uses the estimation of parameters based on an iterative process by minimizing the unweighted least squares (ULS) difference between observed and the expected proportions (Knol & Berger, 1991). The observed proportions present the correct proportions when an examinee successfully answers any two given items whereas the expected proportions are approximated by a

normalized Hermite-Tchebychef third-degree polynomial function (Gosz & Walker, 2002; Fraser & McDonald, 1988), defined as:

$$P(x_i=1|\boldsymbol{\theta}) = b_{i0}h_0(z_i) + b_{i1}h_1(z_i) + b_{i2}h_2(z_i) + b_{i3}h_3(z_i), \quad (2.15)$$

where $z_i = f_i'\theta / d_i$ and $d_i = \sqrt{f_i' P f_i}$ (P is the covariance (correlation) matrix of θ). $h_k()$ is the normalized Hermite-Tchebychef polynomial of degree k, given by $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = (x^2-1)/\sqrt{2}$, $h_3(x) = (x^2-3x)/\sqrt{6}$. Note NOHARM does not estimate the c-parameters; rather they are specified as in TESTFACT.

Underlying this approximation, the distributions of examinee latent traits are normal with a mean of zero and standard deviation of one. Given the pairwise probabilities $\pi_{ij} = P(x_i=1| x_j=1)$, $\hat{\pi}_{ij}$ is approximated by using Equation 2.15. Through an iterative process to minimize the ULS function to estimate parameters:

$$F = \sum\sum [p_{ij} - \hat{\pi}_{ij}]^2 \text{ , where } p_{ij} \text{ are the sample proportion.} \quad (2.16)$$

As indicated by Fraser & McDonald (1988), the ULS function is minimized using either a quasi-Newton or a conjugate gradients minimization algorithm. The algorithm continues iterating until the function value meets the termination criteria set by users.

NOHARM allows users to fit both exploratory and confirmatory models. The specification of the nonzero loadings for a hypothesized model is required in confirmatory analysis (Tate, 2003). The matrix of covariance residuals and RMSR provide information on the lack of fit of the model. Values of RMSR close to four times the reciprocal of the square root of the sample size indicate an acceptable solution (McDonald, 1981).

ULS estimation is efficient compared to the generalized least-squares (GLS) and maximum likelihood (ML) procedures. Therefore, it can be used with a large number of items and/or dimensions. However, there are some limitations with this method. First, only partial information (one-way marginals and two-way marginals) are included. Hence, it is often referred as a "limited" or "bivariate" factor analysis method. However, Knol and Berger (1991) found that the performance in recovering factor analytic parameters of NOHARM was similar to that of TESTFACT. Second, this approach does not provide the standard error for the parameter estimates and a fit statistic for the model (McDonald, 1997). Gessaroli and De Champlain (1996) developed an approximate chi-square statistic that can be used to make up for this limitation.

(3) Approximate Chi-square test of NOHARM solution by CHIDIM

Gessaroli and De Champlain (1996) proposed an approximate chi-square statistic for assessing dimensionality based on McDonald's NLFA model. The approximate chi-square statistic is used to test the null hypothesis if the off-diagonal elements of a residual matrix are equal to zero. This null hypothesis should be true if the correct number of factors is specified to fit an NLFA model (De Champlain & Tang, 1997). This procedure is based on a statistic initially proposed by Bartlett (1950) and outlined by Steiger (1980). There are five steps necessary to calculate this statistic (Gessaroli & De Champlain, 1996). The first two steps is to calculate (1) the proportion correct in either item i or item j as well as both (i.e., $p_i^{(o)}$, $p_j^{(o)}$,and $p_{ij}^{(o)}$), and to calculate (2) the residual joint-proportions (provided by NOHARM, referred as $p_{ij}^{(r)}$). The third step is to calculate the estimated residual correlation ($r_{ij}^{(r)}$) for each pair of items by using the information from previous two steps. Then, it is necessary to transform each one of the estimated residual correlations to a Fisher r to z ($z_{ij}^{(r)}$). The final step is using the following formula to obtain the chi-square statistic:

$$\chi^2 = (N-3)\sum_{i=2}^{n}\sum_{j}^{i-1} z_{ij}^{2(r)} \qquad (2.17)$$

where $z_{ij}^{2(r)}$ is the square of the Fisher r to z, N is the number of sample size, and $i, j = 1, 2, \ldots, n$.

This statistics is approximately distributed as a central chi-square with the degree of freedom equal to $[p(p-1)/2]-t$, where p is the number of items and t is the total number of independent parameters estimated in the model. This test is implemented in the program CHIDIM (De Champlain & Tang, 1997) and requires the product moment matrix and the covariance residual matrix calculated from NOHARM. This statistic can be used for both exploratory and confirmatory analyses. For an exploratory analysis, the search for an appropriate solution is based on increasing the number of factors to an initial unidimensional model until the statistics indicate a good model fit. In a confirmatory application, the hypothesized model is accepted when the null hypothesis is tenable. This chi-square test performs well for unidimensional data, even when the distribution of ability is slightly skewed (De Champlain & Tang, 1993) or when test length is short (e.g. 15) and sample size is small (e.g. 500) (Gessaroli & De Champlain, 1996). In addition, according to the results from Gessaroli and De Champlain (1996), this statistic performed pretty well in identifying unidimensional and two-

dimensional data as well as Type I error control. Compared to Stout's T statistic, the chi-square statistic performs better than Stout's T statistic with a smaller sample size and shorter test lengths.

Gessaroli & De Champlain (1996) pointed out several advantages and two disadvantages for this approximate chi-square test. Starting with the advantages, the first advantage is that the model used for the assessment of dimensionality is related to both NLFA and MIRT. Second, the statistic involves actually testing the hypothesis, therefore, it can be applied to a variety of test settings with greater confidence. Third, the nature of the approximate chi-square statistic, based on the discrepancy function, is consistent with the ULS estimation procedure. Finally, there is no severe limitation on the number of items or dimensions because of the use of the ULS estimation procedure. The two major disadvantages of the approximate chi-square test include: (1) the weak theoretical foundation of this statistic due to the results obtained from ULS estimation; (2) the statistic is affected by a large sample size similar to the other chi-square statistics.

(4) Rasch MIRT model by ConQuest

Another procedure based on the MIRT Rasch modeling approach, called the multidimensional random coefficient multinomial logit model (MRCMLM), was proposed by Adams, Wilson and Wang (1997) and implemented in ConQuest. The MRCLML is an extension of the unidimensional random coefficient multinomial logit model (RCMLM). Given that $\boldsymbol{\theta}$ is a vector of the latent variables in d dimensions, the probability of a response in category k of item i can be found using the following equation (Adams et al., 1997):

$$P(\boldsymbol{X}_{ik} = 1; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}'_{ik}\boldsymbol{\xi})}{\sum_{k=1}^{K_i} \exp(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}'_{ik}\boldsymbol{\xi})}. \qquad (2.18)$$

where $\boldsymbol{A}$ is a design matrix, $\boldsymbol{B}$ is a scoring matrix, $\boldsymbol{b}_{\mathrm{ik}}$ is the score of item i, ξ is the parameter vector of item i. The RCMLM is seen as a generalized Rasch model integrating several existing Rasch models, including the linear logistic latent trait model, the rating scale model, and so on (Adams et al., 1997). This model also provides flexibility in allowing the design of customized models for particular testing occasions (Adams et al., 1997). As an extension of RCMLM, the MRCMLM inherits all of the features and flexibility of the RCMLM. Moreover, this MRCMLM can specify a number of multidimensional models by putting linear constraints on the item parameters (Adams et al., 1997). There are two estimation methods provided in the MRCMLM,

which are the marginal maximum likelihood (MML) method and the condition maximum likelihood (CML) method. In addition, according to the findings of Pfanzagl (1994), only MML estimation is asymptotically efficient in dichotomous Rasch model. According to the description by Hoijtink and Vollema (2003), the model as implemented in ConQuest includes two parts, a measurement model defined in Equation 2.19, and a structural model $\theta \sim N(\mathbf{0}, \Sigma)$:

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_i, \delta_j, \boldsymbol{b}_j) = \frac{\exp(\sum_d b_{jd}(\theta_{id} - \delta_j))}{1 + \exp(\sum_d b_{jd}(\theta_{id} - \delta_j))} \tag{2.19}$$

where θ_{id} is the ability of person i, d = 1, …, D (the number of latent traits); b_{jd} is the discrimination parameter, δ_j is the difficulty parameter of item j; and the covariance matrix Σ of the D latent traits. The model is essentially a confirmatory item factor analysis model and the item parameters are estimated by MML estimation (Hoijtink & Vollema, 2003).

This procedure is appropriate when the discrimination parameter is constant and there is no guessing parameter. The fit of a hypothesized multidimensional factor model is accomplished by using chi-square statistics for testing the difference in the model deviance for a one-factor model and the hypothesized model (Tate, 2003). The previous version of ConQuest assumes that θ has a multivariate normal distribution. However, the later version contains nonparametric distributions for one-dimensional models (Hoijhink, Rooks & Wilmink, 1999). Time consumption is a problem when tests are of the typical length and there is an increasing number of factors (Tate, 2002).

2.2.3 Nonparametric methods

In IRT, dimensionality is defined as the number of abilities required that meet the assumption of "*local independent"* (LI). The original concept of the LI assumption is called strict LI (SLI). Due the difficulty of meeting this strong assumption in the real applications, Stout (1987, 1990) proposed a weak form of LI, called essential local independent (ELI), as the foundation of essential dimensionality. This assumption is defined as:

$$\lim_{N\to\infty} \binom{N}{2}^{-1} \sum_{l \le j} \sum_{k \le N} | Cov(U_j, U_k) \mid \Theta = \theta | \to 0 \tag{2.20}$$

where N is the test length, θ is the latent variable. Under the ELI assumption, a test is assumed unidimensional if $(\mathrm{Cov}(U_j, U_k)|\theta)=0$, $j \neq k$. The methods discussed in this section are basically used as a confirmatory approach and based on the essential unidimensionality assumption (i.e., assuming ELI assumption). Three methods were introduced here, including DIMTEST, HCA/CCPROX, and DETECT.

(1) Test of essential dimensionality by DIMTEST

DIMTEST is a statistical testing procedure developed to test for essential unidimensionality based on asymptotic theory assuming an infinite number of examinees and items (Stout, 1987, 1990). First, a test with N items is split into three subtests: two assessment subtests (AT1 and AT2, each one with M items) and one partition subtest (PT, with $N-2M$ items). AT1 items are selected to reflect item measuring ability in the dominant domain; AT2 items are selected to offset the possible statistical bias found in AT1 subset (due to shorter test length or extreme difficulty levels). Therefore, the difficulty distribution of AT2 items is similar to the one of AT1 subset. PT is used to group examinees into subgroups for calculating T (Hattie, Krakowski, Rogers, & Swaminathan, 1996). Stout, Habing, Douglas, Kim, Roussos, and Zhang (1996) recommended using at least 15 items for PT, and greater than 3, but no more than one third of the number of items in PT for the AT1 and AT2 subsets (Tate, 2003). Stout's T is calculated using following formula:

$$T = \frac{(T_{1a} - T_{1b})}{\sqrt{2}} \tag{2.21}$$

where T_{1a} and T_{1b} are the variance estimated by the AT1 and AT2 item sets. Stout's T tests if one dominant dimension under a set of test items exists (i.e., $d_E=1$) (Nandakumar, 1994). When the value of T is small, the test is essentially unidimensional since the difference between the usual variance estimate (from AT2) and the unidimensional estimate (from AT1) is small. If the T value is greater than z at significance level of .05 (i.e., 1.96), the hypothesis of essential unidimensionality is rejected (Pang, 1999).

For confirmatory applications, selecting an AT1 set depends on prior expectations while an AT2 set is selected using statistical methods, such as HCA/CCPROX and DETECT in exploratory cases (Froelich & Habing, 2001). Nevertheless, even when using AT2 to correct the bias of examinee variability and item difficulty, DIMTEST still exhibits some positive bias (van Abswoude, van der Ark, & Sijtsma, 2004). Nandakumar and Stout (1993) found that

Stoutn's *T* was a poor indicator for testing essential unidimensionality when items of a test have high discrimination (greater than 1.1) and guessing parameters. Due to the selection of AT1 items based the tetrachoric correlations, DIMTEST shares the weakness of those methods using tetrachoric correlations.

(2) Hierarchical cluster analysis of item proximities by HCA/CCPROX

The HCA/CCPROX (Roussos, Stout, & Marden, 1993) is designed to search for homogenous item clusters by an agglomerative hierarchical cluster analysis (HAC) based on item proximity. The proximity measure for two items proposed by Roussos (1995), called CCPROX, is defined as the negative of the conditional covariance for the items plus a constant such that all proximity measures are nonnegative (Tate, 2003). At the initial level, the analysis starts with each individual item of a test treated as a separate cluster, then each item either clusters with other items or remains by itself by using the proximity measures. At the second level of hierarchy, the two clusters having the smallest expected conditional covariance are joined. The process of joining clusters is repeated until all items are collected into one large cluster (Roussos, Stout, & Marden, 1998). The key point of the success of the HCA is the formulation of a proximity measure in the initial level. Three proximity measures are provided, p_{CCOR}, p_{CCOV}, p_{MH} instead of the classical ones (more details see Roussos, Stout, & Marden, 1998). Four HCA methods are provided for the second level proximity measures, including the single link, complete link, and unweighted pair-group methods of average (UPGMA), and the weighted pair-group method of average (WPGMA). Due to the lack of formal criteria to help researchers determine dimensionality, researchers may make the decision based on a *priori* theoretical expectations about the test structure (van Abswoude et al., 2004).

(3) DETECT index of dimensionality using DETECT

The DETECT program estimates the extent of the multidimensional simple structure (Kim, 1994; Kim, Zhang, & Stout, 1995; Zhang & Stout, 1999). The procedure is to find a partition of test items that maximizes the DETECT index, which is defined as the average of all the signed conditional item covariances. In the computation of summing the conditional item covariance for each pair, -1 multiplies a covariance of an item pair in two different partitions. When the partition of items gets closer to the true item cluster representing an approximate simple structure, the value of DETECT index gets larger. A three-step conditional sequence of

decision is used to find the number of dominant dimensions of a test. The first step is to create another DETECT index (D_{ref}) using the second sample, similar to the cross-validated concept. Essential multidimensionality can be concluded when the value of this index is greater than 0.1 (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). The second step is to calculate an R ratio value by using either the original DETECT index or the D_{ref} index calculated in Step 1 divided by the maximum possible DETECT index, which sums the absolute values of the conditional covariance of all item pairs for the optimal partition. If the R ratio value is equal to or greater than 0.8, it indicates that the partition of items is an approximate simple structure. If the first two steps show that the test has essential multidimensionality and approximate simple structure, then the number of clusters in the optimal partition of items indicates the number of dominant abilities. If either the first or the second step fails, then there is no conclusion of the number of dimensionality (Tate, 2003). However, the index shows the magnitude of departure from unidimensionality, but is not an index of the number of dimensions (van Abswoude et al., 2004).

2.2.4 Research comparing methods

According to the summary of Harwell and colleagues (1996), researches on assessing dimensionality prior to 1991 focused on detecting dimensionality (if tests are unidimensional or multidimensional), or the outcome of using unidimensional models. Only two studies compared multidimensional models with multidimensional data. The studies of factor analysis of dichotomous data focused on validating newly developed methods (e.g., Bock, Gibbons, & Muraki, 1988; McDonald, 1982) and comparison of linear factor analysis and nonlinear factor analysis and/or other methods (e.g., residual analysis). For example, Hambleton and Rovinelli (1986) compared four methods, linear and nonlinear factor analysis, residual analysis, and Bejar's method (1980), used five artificial datasets. The results demonstrated the superiority of using nonlinear factor analysis when linear factor analysis overestimated test dimensionality. Regarding the assessment of dimensionality in dichotomous data, the MIRT approach illustrated a different view from classical factor analysis.

In recent decades, most studies of dimensionality using Monte Carlo techniques in IRT have focused on either comparing different methods (e.g., Nandakumar, 1994; Nandakumar &

Yu, 1996; Tate, 2003), or the effect of applying unidimensional item response theory to multidimensional items (e.g., Ackerman, 1989; Way, Ansley, & Forsyth, 1988). Also, due to the equivalence between NLFA and MIRT, there were several studies illustrating the comparison of the methods in either the NLFA or MIRT approach. Concerning the application of those newly developed procedures, more simulation and empirical studies are necessary to verify the applicability and the efficiency of these new procedures to either general testing occasions or some conditions might be problematic for using these procedures.

Among these studies, Mplus, NOHARM and TESTFACT were most commonly chosen by researchers. The study conducted by Knol and Berger (1991) compared the full-information models (i.e., TESTFACT) to the limited information models (such as NOHARM or GLS estimator). The most remarkable finding was that the performance of NOHARM and TESTFACT was highly similar. In other words, the results indicated that full-information models were not superior to the limited-information models. Tate's study (2003) focused on a comparison among a number of empirical methods for assessing test structure. In this study, Mplus performed worse than NOHARM and TESTFACT when the data assumed guessing. A similar conclusion was obtained in Stone and Yeh's study (2006) using real data. In next section, the guessing effect is discussed in more detail.

2.2.4.1 The guessing effect and the assessment of dimensionality

It is essential to consider the influence of guessing on multiple-choice items. Guessing behavior either increases the measured error or can be seen as the source of construct-irrelevant variance (Messick, 1995). It is considered a source of construct-irrelevant since it increases the possibility of a correct response based on an ability other that the ability intended to be assessed (Rogers, 1999). Therefore, it is necessary to model guessing for tests with multiple-choice items. As Stone and Yeh (2006) pointed out, the influence of guessing is dependent on an unknown mechanism or process.

To model guessing behavior, either random guessing or partial-knowledge guessing can be assumed (Waller, 1989). If random guessing is assumed (Lord & Novick, 1968), in multiple-choice items, the probability for individuals answering correctly by chance will be $1/m$, where m is the number of choices (Stone & Yeh, 2006). On the other hand, if the individuals take the exam with partial knowledge, then some of the options can be eliminated. In this case, the

probability of a chance correct response will be greater than 1/*m*. In Waller's study (1989), the three-parameter IRT model not only can remove the effect of random guessing, but also it can make an adjustment for partial-knowledge guessing. However, most traditional factor analysis approaches do not include guessing in the models. Even the factor analysis in the MIRT approach only allows fixed values of guessing to be set in models in some software, such as TESTFACT and NOHARM. It is absolutely necessary in factor analysis to determine dimensionality in educational and psychological test settings using multiple-choice items in order to investigate the influence of guessing. However, it is rare to find a study focusing on the influence of guessing in the decision of dimensionality in either the NLFA or MIRT approaches.

Recently, the guessing effect has been discussed in two studies focusing on the assessment of the dimensionality and test structure. The first study was conducted by Tate (2003), which used a number of empirical methods to assess test structure. Tate examined these assessment methods in either exploratory or confirmatory ways including unidimensional and multidimensional data. The second study proposed by Stone and Yeh (2006) focused on a comparison of three methods (i.e., Mplus, NOHARM and TESTFACT) in the context of determining the dimensionality and test structures. The main situation considered in these two studies was a test with a large number of items and with relative large sample size. Either real data or simulation data was used to evaluate the guessing effects for different approaches for the assessment of dimensionality.

Given using either parametric or nonparametric methods, Tate's study also considered factors likely to affect the assessment of dimensionality: presence of modeled guessing, test item difficulty, test item discrimination, factor correlations, and simple versus complex factor structures. In exploratory approach for assessing dimensionality using simulation data, the methods in four programs were evaluated: Mplus (ULS estimator), NOHARM, CHIDIM, and TESTFACT. Note that only Mplus does not allow for modeling guessing.

Tate (2003) found a low rate of correct dimensionality decisions in cases with modeled guessing using Mplus (3 out 14 multidimensional cases) compared to the other three methods (e.g., 11 out of 14 in NOHARM). Mplus performed fairly well in cases without modeling guessing. However, Mplus tended to underestimate factor loadings and to overestimate dimensionality. Additionally, he found identifying the correct dimensionality with data with extreme item parameters, such as difficulty and discrimination parameters was less effective

when using Mplus. Interestingly, in more than half of the conditions when data assumed guessing for either a simple or a complex structure, Mplus provided supporting evidence for the correct parameter pattern even though it exhibited bias in parameter recovery.

When using NOHARM for exploratory factor analysis, the dimensionality was correctly identified and parameters recovered well for most unidimensional and multidimensional data. However, a tendency to identify a difficulty factor for data with extreme discrimination was observed. For the CHIDIM results (based on a chi-square test), there were 11 out of 18 cases that reached the same conclusion as when using NOHARM (based on the RMSR reduction index). Note that CHIDIM performed better than NOHARM in cases with a greater variation in discrimination or high discrimination. However, with extremely difficult data, CHIDIM was found to obtain more incorrect decisions of dimensionality than NOHARM. Using the RMSR reduction index in TESTFACT, the assessment of dimensionality performed well in general, however, there was only fair recovery with data consisting of extreme difficulty or discrimination item parameters.

The results of Tate's study using Mplus, NOHARM, CHIDIM, and TESTFACT were consistent with the results of most previous studies investigating the performance of these methods (e.g., Knol & Berger, 1991, for Mplus, NOHARM and TESTFACT; Gessoroli & De Champlain, 1996 for CHIDIM). In summary, the effect of guessing was found by comparing the performance of methods both without modeling guessing (i.e., Mplus) and with modeling guessing (i.e., CHIDIM, NOHARM and TESTFACT). Mplus only performed well using data without guessing, whereas the other three methods performed well using the data with or without guessing. In addition, any extreme item parameters in the data affected the identification of dimensionality and parameter recovery. For instance, all four methods performed poor with data consisting of items with extreme difficulty and high discrimination. However, several methodological considerations were evident from Tate's study. It is unknown how successful the conclusions will be when applied to situations that vary in the degree or types of guessing. Tate used true values of guessing which may have biased results. Also, Tate employed a single replication design and therefore could not evaluate Type I error and empirical power rates for the methods.

Stone and Yeh (2006) conducted a comparison of Mplus, NOHARM, and TESTFACT by using real data from an administration of the Multistate Bar Examination (MBE). The MBE is a

multiple-choice test with 200 four-option items. There were six aspects of the content: Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts. Multiple test forms have been developed and administer twice a year. Each test contains two sections consisting of 100 items with equal or comparable content. Examinees take both sections separately in morning and afternoon. For this study, the February 2001 administration was analyzed (N = 20,288). The AM and PM item sets were analyzed separately because of the limitations of Mplus (no more than 100 items may be analyzed). The distribution of item difficulty indices (i.e., proportion correct) for both forms were similar and in quite a broad range (M = .65, SD = .20). The coefficient alpha for each form was around .80. The correlations of item to total score ranged from 0 to .4. In addition, guessing was evaluated for each item by plotting the proportion correct by total scores. Guessing was found to be operating in more than half of the items in each form. The average c-values was around .31 (SD ~ .16) for both forms.

Mplus, NOHARM and TESTFACT were used to assess the dimensionality and internal structures of the MBE. There were two conditions for the factor analyses. Condition 1 provided a comparison of these three methods without modeled guessing, that is no guessing values were incorporated into the analyses. Condition 2 incorporated guessing values into the analyses. As a result, only a comparison between NOHARM and TESTFACT was provided. Note that guessing values estimated by MULTILOG were used in both methods. Also, the WLSMV estimator (not ULS as Tate's study) was used in Mplus. The number of estimated dimensionality was primarily based on eigenvalues, the RMSR statistics (less than .05 for Mplus and TESTFACT, and less than .028 for NOHARM indicated an acceptable factor solution), and the number of substantial loading for factors.

The results for Condition 1 demonstrated similar patterns for Mplus, NOHARM and TESTFACT. Either two or three factors were uncovered, however, more than half the items did not load on any factor. For Condition 2, greater eigenvalues were observed indicating that including guessing resulting in an increase in the proportion of explained variance for the factors. This finding was caused by the correction of the tetrachoric correlation matrix. As indicated by Carroll (1945), tetrachoric correlations correct for the effect of chance success or error and then stronger relationships between items are observed. The average of the tetrachoric correlations under the two conditions confirmed the effect of the correction. For example, the mean correlations for the AM form under the two conditions were .07 (SD = .05) and .11 (SD = .10),

respectively. Moreover, more items with substantial loadings were observed in the solutions under Condition 2. In addition, the correlations between factors under Condition 2 were larger than those under Condition 1. Thus, a more complete understanding of the internal structure was obtained by incorporating guessing into the estimation procedures. Stone and Yeh's study confirmed the findings found in Tate's study. However, the primary limitation was that no systematic examination of the guessing effect could be conducted with a real dataset since the true dimensionality was unknown.

3.0 METHODOLOGY

According to the findings and the limitations of Tate (2003) and Stone and Yeh's (2006) studies, there were two purposes for this study. The major purpose was to investigate the effects of guessing in the context of examining the dimensionality for multiple-choice tests. Only when the true dimensionality is known it is possible to detect the influence of guessing and other factors in assessing dimensionality. However, in most psychological and educational tests, while a specific definition of the test constructs is reported, the true dimensionality is still unknown. Therefore, a Monte Carlo (MC) study was used in this study. Additionally, an MC study can investigate the effects of several factors simultaneously by manipulating the true values of parameters (Harwell, Stone, Hsu, & Kirisci, 1996). The second purpose of this study was to verify if the findings from the simulation study can be applied to the analysis of real data. In order to accomplish this goal, TIMSS 2003 was chosen in this study because of the two subject domains included in this assessment. Therefore, through the analysis of the TIMSS data, it was possible to investigate the dimensionality and factor structures, and verify what was learned in the simulation study.

3.1 PHRASE I – SIMULATION STUDY

3.1.1 Design of the study

Theses factors were manipulated in this study: 1) estimation methods; 2) discrimination parameters of items; 3) guessing parameters of items and 4) correlations among dimensions. The two most popular methods, Mplus and TESTFACT, were selected for this study. Mplus works from the perspective of factor analysis, whereas TESTFACT output in either an FA or MIRT

perspective. Additionally, Mplus has no allowance for modeling guessing whereas TESTFACT can input the guessing parameter for factor analyses. Therefore, the comparison of these two estimation methods can help explore the effect of modeling guessing parameters. NOHARM was not included in this study because the results and performances were similar to those of TESTFACT (e.g., Knol & Berger, 1991; Stone & Yeh, 2006).

The exploratory factor analyses (EFA) were conducted using these two methods for each simulation dataset. The number of factors that was extracted was set to the true dimensionality plus two. For example, one-, two- and three-factor solutions will be conducted by Mplus and TESTFACT for unidimensional data. The convergence criteria of both of the methods was set to .005. The default setting of the number of iterations in Mplus (i.e., 200) was enough for most situations. A smaller number of iterations, 50, was used with TESTFACT. There were two reasons for this setting. Time requirement was the primary concern, since this study used 100 replications in each condition. Additionally, based on the results of a small pilot study of parameter recovery, no more than 15 iterations were necessary when data fit (i.e., the number of extracted factors was equal to the true dimensionality). Therefore, the maximum number of iterations in TESTFACT was set to 50.

Item parameters, discrimination and guessing were also manipulated. Item discrimination was manipulated since these parameters relate to factor loading and reflect the degree to which items relate to factors. The level of discrimination was manipulated by the different proportions of items with three different values (i.e., 0.5, 1.0 and 1.5) of discrimination. For example, a low discrimination condition consisted of 50% of the items with 1.0 and 50% of the items with 0.5. Table 3.1 provides the descriptive statistics for each level of discrimination (i.e., Low (L), Medium (M), and High (H)). As can be seen in Table 3.1, the mean level of the discrimination for L, M, and H levels were 0.75, 1.0, and 1.25, respectively. The intercept coefficients that relate to difficulty were specified in the range of –2.0 to 2.0. The average of the intercept parameters or difficulty level for all the conditions was equal to 0. The standard deviations of item parameters (i.e., discrimination, intercept, and difficulty) in the three different discrimination levels (L, M, and H) were about the same. Note that the range in the intercept, not difficulty, was controlled in the purpose of simulation data. Detail information regarding item parameters for all discrimination levels in one-, two-, and three-dimensional data can be found in Appendix A.

There were two conditions for the guessing parameter, 0 and .33. Since Mplus does not allow the inclusion of the guessing parameter in the calibration procedure, the condition, c = 0, was set as a baseline for comparison and equivalent to a two-parameter model. If a random guessing model (Lord & Novick, 1968) is assumed, then individuals who lack the necessary knowledge would randomly guess. Thus, for multiple-choice items the probability that individuals choose a correct answer by chance is 1 divided by the number of choices (*m*). Values of .33, .25 and .2 correspond to three-, four- and five-option items. The value, .33, was chosen to simplify the design. It is the best to stay with one reasonable and significant value to compare with the baseline value of 0. Moreover, a guessing parameter of .33 was used to explore the effect of test items that reflected a moderate amount of guessing. In Tate's study, the value, .2 was chosen for all conditions. These results of this study provide a useful comparison with Tate's study.

Table 3.1

The descriptive statistics of different levels of discrimination design under one- to-three dimensional data

	Low (L)			Medium (M)			High (H)		
	a	d	b	a	d	b	a	d	b
One-dimension (n = 60)									
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.25	1.15	1.81	0.36	1.15	1.45	0.25	1.15	0.97
Two-dimension (n = 30)									
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.25	1.16	1.72	0.35	1.18	1.43	0.25	1.16	0.97
Three-dimension (n = 20)									
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.26	1.26	1.77	0.36	1.25	1.43	0.26	1.26	1.03

Another variable related to the test structure was the correlation among factors or dimensions. There is a certain amount of correlation among factors (or dimensions) in most educational tests. For this reason, this study incorporated two conditions of correlation among factors, .3 and .6, to represent the possible occasions that occur in most educational tests.

The present study assumes the application of a large-scale testing program with multiple-choice items and administration to a relatively large sample size. Thus, there were several fixed factors in this study. First, most MC studies focus on two-dimensional data because of the increase in complexity where the number of dimensions is greater than two. However, it is possible that realistic tests will have more than two dimensions. Therefore, the inclusion of three-dimensional data provides useful information for practical applications. In this study, one-dimensional data represented unidimensionality as a baseline (unidimensional case) for comparison purposes, while two- and three-dimensional data were designed as multidimensional cases. Second, according to the examination of 17 studies of dimensionality summarized by Harwell et al. (1996), the range of the test length varied from 20 to 60 while the range of the sample size was 1000 to 2000. To adhere to the minimum requirements for the number of items for each dimension, the test should have at least 20 items for each dimension (i.e., in three-dimensional conditions). The total test length should be around 50 to 60 items. Therefore, the total test length was set to 60 items, and the number of items per domain was 60, 30, and 20 for one-to-three dimensional data. Additionally, after considering the requirements for estimation accuracy, the calibration size was set at 2000. According to the recommendations made by Harwell et al. (1996), the number of replications for MC studies should be more than 25. Increasing the number of replications can increase the power; however, if the test length and sample size is large enough, it is unnecessary to increase replications up to 500. Hence, a 100-replication design for each condition was used in the present study.

Since it is hard to find "pure" simple test structures in educational tests, an approximate simple structure was assumed in this study. An approximate simple structure is indicated when the factor loadings for each item loads significantly on only one factor (i.e., the values greater than .3) and the values of factor loading on the other factor(s) are less than .3. The reason for not assuming complex test structures (significant loadings on more than one factor) that most educational tests are designed to measure separate factors or subscales (e.g., MBE). To represent

an approximate simple structure, the minor factor loadings were not set to 0, but a little higher than 0 (.2 was used in this study).

3.1.2 Generating data and estimating procedure

Since the NLFA and MIRT models are basically mathematically equivalent, either the factor analytic or MIRT models can be used to generate multidimensional data. However, the main focus of the present study was to examine the influence of guessing in assessing dimensionality of multiple-choice tests. Thus, it was considered more appropriate to generate multidimensional data by MIRT models, which include guessing in the models. Because the present study was set up to assess dimensionality of multiple-choice tests, the MC3PL model was used to generate data.

In this study, the assumed setting was a large-scale test with multiple-choice items and a large sample size. Hence, as discussed a test length of 60 and a sample size of 2000 was used in all conditions. Also, to obtain a better understanding of the influence of guessing in practical testing applications, an approximate simple structure was assumed. This means that the factor loadings of items belonging to minor factors were set to a relatively of small value. The manipulated variables focused on discrimination, guessing under one-to-three dimensional data. The easiness intercept, related to item difficulty, was specified in the range of –2 to 2. Low and high relationships between factors were specified (i.e., r = .3 and .6).

The first step in generating data was to specify the item parameters (discrimination (a), intercept coefficients (d), and guessing(c)) as per the conditions discussed. For example, in Table 3.2, one condition represented a two-dimensional test with low discrimination. The test assumed no guessing and low correlations between factors (e.g., r = .3). Therefore, there were 30 items for each dimension. Item 1 to Item 30 had higher loadings in dimension one, while Item 31 to Item 60 had higher loadings in dimension two (for more details see Appendix A2). Next item parameters for minor dimensions were specified. The discrimination values for the minor dimension were set to .2. The value was based on the Stone and Yeh (2006) study in which the mean discrimination for items on the minor dimension was approximately .2. To control for any possible effects of related to item difficulty, the range of the easiness intercept was constant across sets of items. It should be noted that the value of discrimination was not set

to a single value, but rather to two values (1.0 and 0.5) to make the situation more realistic. In this case, the average low discrimination was 0.75 and the means of both the easiness intercept (d) and difficulty (b) were 0 (see Table 3.1). Table 3.3 displays the item parameters for a two-dimensional data set with low discrimination.

Based on the item parameters defined for each condition, 100 replications per condition were generated by a SAS program. There were 600 datasets of unidimensional data (3 discrimination conditions (L, H, M) × 2 guessing conditions × 100 replications). There were 2400 datasets of two-dimensional data (6 discrimination conditions (HH, MH, MM, LH, LM, LL) × 2 guessing conditions × 2 correlation conditions * 100 replications). Note that the discrimination conditions were the combination of three discrimination levels. For example, in two-dimensional data, the discrimination conditions were the combination of two discrimination levels. Therefore, the possible combinations were HH, MH, MM, LH, LM and LL. Note that the order of discrimination levels should not affect the results of factor analyses. Therefore, for instance, MH and HM were seen as the same condition. In other words, the MH condition represented two-dimensional data in which the discrimination level for items in one dimension was at a medium level and the discrimination level for items in the other dimension was at a high level. The same characterization applies to generating three-dimensional data. There were 4000 datasets for three-dimensional data (10 discrimination conditions (HHH, HHM, HHL, MMH, MMM, MML, LMH, LLH, LLM, LLL) × 2 guessing conditions × 2 correlation conditions × 100 replications). Therefore, the total number of datasets was 7000 (70 conditions, 100 replications per condition).

Table 3.2

Design of a two-dimensional data set with low discrimination

	Dimension 1	Dimension 2
Subset 1 (Item 1 to 30)	Low discrimination (Mean = .75)	Discrimination = .2
Subset 2 (Item 31 to 60)	Discrimination = .2	Low discrimination (Mean = .75)

Table 3.3

Item parameters of a two-dimensional data set with a low discrimination

Item No.	Dimension 1			Item No.	Dimension 2		
	a	d	c		a	d	c
1	1.0	-2.00	0	31	0.2	-2.00	0
2	1.0	-1.50	0	32	0.2	-1.50	0
3	1.0	-1.25	0	33	0.2	-1.25	0
4	1.0	-1.00	0	34	0.2	-1.00	0
5	1.0	-0.75	0	35	0.2	-0.75	0
6	1.0	-0.50	0	36	0.2	-0.50	0
7	1.0	-0.25	0	37	0.2	-0.25	0
8	1.0	0.00	0	38	0.2	0.00	0
9	1.0	0.25	0	39	0.2	0.25	0
10	1.0	0.50	0	40	0.2	0.50	0
11	1.0	0.75	0	41	0.2	0.75	0
12	1.0	1.00	0	42	0.2	1.00	0
13	1.0	1.25	0	43	0.2	1.25	0
14	1.0	1.50	0	44	0.2	1.50	0
15	1.0	2.00	0	45	0.2	2.00	0
16	0.5	-2.00	0	46	0.2	-2.00	0
17	0.5	-1.50	0	47	0.2	-1.50	0
18	0.5	-1.25	0	48	0.2	-1.25	0
19	0.5	-1.00	0	49	0.2	-1.00	0
20	0.5	-0.75	0	50	0.2	-0.75	0
21	0.5	-0.50	0	51	0.2	-0.50	0
22	0.5	-0.25	0	52	0.2	-0.25	0
23	0.5	0.00	0	53	0.2	0.00	0
24	0.5	0.25	0	54	0.2	0.25	0
25	0.5	0.50	0	55	0.2	0.50	0
26	0.5	0.75	0	56	0.2	0.75	0
27	0.5	1.00	0	57	0.2	1.00	0
28	0.5	1.25	0	58	0.2	1.25	0
29	0.5	1.50	0	59	0.2	1.50	0
30	0.5	2.00	0	60	0.2	2.00	0

Mplus and TESTFACT were used to perform an EFA on each dataset. The number of factors extracted was set to the true values of dimensionality plus two in order to determine if any over factoring occurred. In this way, one-to-three factor solutions were obtained in unidimensional cases, whereas one-to-four factor solutions and one-to-five factor solutions were conducted for two- and three-dimensional data. The criteria for convergence for both methods were set to .005. However, the maximum iterations for both methods were different (200 vs. 50) due to the consideration of practical issues, such as time consumption.

Since TESTFACT may be used to model guessing, but is not jointly estimated with the factor solution, the guessing parameters was fixed at values obtained by other programs, as suggested by the software developers (Wilson et al., 2003). Software developers discussed that the estimation of guessing only depended on the response of low-ability examinees and not on the response functions at all ability levels (Wilson et al., 2003). Hence, a one-dimensional model to estimate guessing is viewed as adequate. In this study, MULTILOG was used to estimate guessing parameters. Additionally, according to Lord's criteria (1980) and the estimation of item parameters resulting from MULTILOG, the values of guessing of those items that did not meet Lord's criterion were set to 0. Lord's criterion is used to fix guessing at 0 for very easy items as well as item with very low discriminations. In these cases, estimation of guessing parameters is difficult to isolate from the difficulty parameters. Since higher values of guessing can lead to higher measured error, the upper boundary was set to control measured error. An upperbound of .5 was imposed as well since guessing parameters greater than .5 may be unrealistic for multiple-choice tests. Therefore, for those datasets assuming guessing, the guessing parameters were fixed to the values estimated from MULTILOG, and also were fixed to zero for some items based on Lord's criterion. An upperbound of .5 was applied in this study if there was c-parameter more than .5 in any item. Note that the c-parameters of data assuming no guessing were set to zero.

3.1.3 Validating data generation

In order to ensure that the SAS program used in this study properly generated data, data generated by this SAS program was compared with data generated by GENMIRT, another program created by Jeffery Kromrey, Cynthia Parshall, Walter Chason and Qing Yi in University

of South Florida (1999). Datasets generated by GENMIRT were based on the normal ogive model. The true values of IRT parameters used to generate data are presented in Table 3.4. The example used a test with 20 items, 2,000 examinees, two-dimensional, no guessing assumed and no correlation between two ability dimensions. The comparison was conducted with no correlation between two dimensions since the GENMIRT program did not allow setting correlations. The validation was performed under two test structures, simple and approximate simple structures. Table 3.5 and Table 3.6 demonstrate the recovery of item parameters estimated by TESTFACT for datasets generated by the SAS and GENMIRT programs.

Table 3.4

The true values of IRT parameters for validation

Item	Simple structure			Approximate simple structure		
No.	a_1	a_2	d	a_1	a_2	d
1	1.0	0.0	1.79	1.0	0.2	1.79
2	1.0	0.0	1.43	1.0	0.2	1.43
3	1.0	0.0	1.07	1.0	0.2	1.07
4	1.0	0.0	.71	1.0	0.2	.71
5	1.0	0.0	.36	1.0	0.2	.36
6	1.0	0.0	-.36	1.0	0.2	-.36
7	1.0	0.0	-.71	1.0	0.2	-.71
8	1.0	0.0	-1.07	1.0	0.2	-1.07
9	1.0	0.0	-1.43	1.0	0.2	-1.43
10	1.0	0.0	-1.79	1.0	0.2	-1.79
11	0.0	1.0	1.79	0.2	1.0	1.79
12	0.0	1.0	1.43	0.2	1.0	1.43
13	0.0	1.0	1.07	0.2	1.0	1.07
14	0.0	1.0	.71	0.2	1.0	.71
15	0.0	1.0	.36	0.2	1.0	.36
16	0.0	1.0	-.36	0.2	1.0	-.36
17	0.0	1.0	-.71	0.2	1.0	-.71
18	0.0	1.0	-1.07	0.2	1.0	-1.07
19	0.0	1.0	-1.43	0.2	1.0	-1.43
20	0.0	1.0	-1.79	0.2	1.0	-1.79

As shown in Table 3.5, for simple structure, the estimated parameters for the two datasets generated using the two programs were very similar and recovered reasonable well. There was slightly underestimation of the discrimination and intercept parameters found when using the SAS program. The correlations between two ability dimensions for the SAS and GENMIRT programs were -.009 and -.044, respectively. As for approximate simple structure, see Table 3.6, the results for the two programs were also similar and also recovered well.

Table 3.5

Comparison of data generation using simple structure for uncorrelated ability dimensions

Item	SAS			GENMIRT		
No.	a_1	a_2	d	a_1	a_2	d
1	0.8	0.0	1.59	0.9	0.1	1.79
2	1.0	0.1	1.44	1.0	0.0	1.46
3	0.9	0.1	0.98	1.0	-0.1	1.08
4	0.9	0.0	0.63	1.0	0.1	0.72
5	1.0	0.0	0.31	1.0	-0.1	0.36
6	0.9	0.0	-0.39	0.9	0.0	-0.31
7	0.9	0.0	-0.77	1.0	0.0	-0.65
8	1.0	0.0	-1.10	1.0	0.0	-1.04
9	0.9	-0.1	-1.41	0.9	-0.0	-1.40
10	0.9	0.0	-1.69	0.8	0.1	-1.68
11	0.0	0.8	1.59	-0.1	1.1	1.92
12	0.0	1.0	1.40	0.0	1.1	1.54
13	0.0	0.9	1.07	0.0	1.0	1.07
14	0.0	1.0	0.72	0.0	0.9	0.69
15	0.0	0.9	0.34	0.0	1.0	0.36
16	0.0	1.1	-0.41	0.0	1.0	-0.38
17	0.1	1.0	-0.67	0.0	1.1	-0.80
18	0.0	1.0	-0.99	-0.1	0.9	-1.01
19	-0.1	0.9	-1.39	0.0	0.9	-1.37
20	0.1	0.9	-1.59	0.1	0.9	-1.71

Table 3.6

Comparison of data generation using approximate simple structure for uncorrelated ability dimensions

Item No.	SAS			GENMIRT		
	a_1	a_2	d	a_1	a_2	d
1	0.9	0.2	1.63	0.9	0.2	1.78
2	0.8	0.2	1.24	1.0	0.2	1.45
3	0.9	0.2	0.97	1.0	0.2	0.99
4	0.9	0.2	0.59	0.9	0.2	0.68
5	0.9	0.2	0.24	1.0	0.2	0.35
6	1.0	0.2	-0.42	0.9	0.2	-0.33
7	1.0	0.3	-0.77	0.9	0.2	-0.73
8	1.0	0.2	-1.12	1.0	0.2	-1.12
9	1.0	0.3	-1.45	1.0	0.2	-1.42
10	0.9	0.2	-1.70	0.9	0.1	-1.65
11	0.1	0.8	1.65	0.2	0.9	1.74
12	0.2	0.9	1.33	0.2	1.0	1.44
13	0.3	0.9	1.05	0.3	1.0	1.08
14	0.2	1.0	0.68	0.3	1.1	0.78
15	0.1	1.1	0.39	0.2	1.0	0.40
16	0.2	1.0	-0.32	0.1	1.0	-0.42
17	0.2	1.0	-0.65	0.1	0.9	-0.74
18	0.2	1.0	-1.14	0.2	1.1	-1.15
19	0.2	0.8	-1.38	0.3	1.0	-1.47
20	0.2	0.8	-1.56	0.1	1.0	-1.92

In order to validate the conditions with correlations among dimensions, recovery of simple structure with a .6 correlation between ability dimensions was evaluated by NOHARM using the data generating by the SAS program. The true values of item parameters (see Table 3.4) were again compared with estimates (see Table 3.7). Close recovery of IRT parameters was observed. The correlation between two dimensions was .623.

Table 3.7

Recovery of item parameters under PROMAX rotation

Item No.	a_1	a_2	d
1	0.8	0.1	1.66
2	0.8	0.0	1.33
3	0.9	0.0	0.94
4	1.0	-0.1	0.64
5	1.0	0.1	0.28
6	1.1	0.0	-0.47
7	1.1	0.0	-0.75
8	0.9	0.0	-1.15
9	1.0	0.0	-1.41
10	0.8	0.1	-1.78
11	0.0	0.9	1.71
12	0.0	1.0	1.38
13	0.0	1.0	1.07
14	0.0	1.0	0.68
15	0.0	1.0	0.35
16	0.0	1.0	-0.36
17	0.0	1.0	-0.72
18	0.1	0.9	-1.04
19	0.0	1.0	-1.39
20	0.0	1.0	-1.75

3.1.4 Outcome measures

The effect of guessing was investigated in the context of the assessment of test dimensionality. Therefore, indices for identifying the number of dimensions were needed in this study. The evaluation of assessing dimensionality included two parts: (1) identifying the number of factors/dimensions; (2) parameter recovery. There were three subjective indices and one objective index used in the first part to identify test dimensionality. Only one index was calculated for evaluation of parameter recovery. Additionally, examination of the parameter recovery focused on factor loadings since factor loadings are only available for Mplus.

The three subjective indices for determining the number of factors were: proportion of variance, parallel analysis, and percentage of RMSR reduction. The proportion of variance was calculated by taking each eigenvalue and dividing it by the sum of all eigenvalues. The amount of variance explained is an index for the substantive importance of factors. The model can retain a number of factors until a certain amount of total variance explained is achieved (e.g., 70%).

Also, the number of factors can be determined by setting a criterion indicating the minimum contribution of each factor. If the percentage of variance less than a certain proportion, this may indicate the presence of nuisance or minor dimensions. In this study, factors were added to the model until the proportion of variance for an eigenvalue was less than 5%. For instance, each proportion of variance for the first three eigenvalues was more than 5% and the fourth eigenvalue was less than 5%, a three-factor solution was obtained.

The second index was based on parallel analysis. Parallel analysis was initially proposed by Horn (1965). The mean eigenvalues from random data was used as a baseline for determining dimensionality. Factors, with greater than the baseline corresponding to eigenvalues based on random data, are retained. However, several authors suggest the use of the 95th percentile for the eigenvalues instead of the mean eigenvalue (e.g., Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995; Turner, 1998). The SAS and SPSS programs for parallel analysis had been developed by O'Connor (2000). The SAS program was used to calculate the 95th percentile of eigenvalues for determining dimensionality in this study.

The last index for determining dimensionality was the percentage of RMSR reduction. This index was also used in Tate's study (2003) for assessing dimensionality in exploratory factor analyses. Factors were added to the model until the percentage of RMSR reduction was less than 10%. All three indices were applied to the factor analysis results obtained from Mplus and TESTFACT. Note that RMSR was not provided in TESTFACT but was calculated by taking the residual matrix provided by TESTFACT.

A statistical test for determining dimensionality is based on the chi-square difference test. A chi-square test statistic was provided by both of Mplus and TESTFACT. The chi-square statistic of TESTFACT may be used to compare nested models (e.g., one factor vs. two factors). However, the chi-square statistic of Mplus using WLSMV estimation may not be used compare nested models since the differences in the degrees of freedom for two nested models will not always be positive values (Wilson et al., 2003). Therefore, the chi-square statistic based on WLS solutions may be used for the chi-square difference test in Mplus. Thus, it was possible to compare the results of Mplus and TESTFACT using chi-square statistics. Note that a large sample was used in the present study since large sample sizes are generally required for WLS methods.

Parameter recovery was another aspect used to evaluate the effects of the factors manipulated in this study. Since the results of Mplus were based on factor models, the recovery of factor loadings between the results in Mplus and TESTFACT were examined. Factor loadings from a PROMAX solution were used given correlations greater than 0. Since data were generated based on a MIRT model (i.e., MC3PL), the true values for factor loading were calculated using Equation 2.11, based on the equivalent relationships between NLFA and MIRT. The root mean square deviation (RMSD) was used for evaluating parameter recovery, and is defined as:

$$\text{RMSD} = \sqrt{\frac{\sum(\hat{p} - p)^2}{n}}, \qquad (3.1)$$

where $\hat{p}$ is the estimated parameter (i.e., factor loading), p is the true parameter and n is the number of items. RMSD was calculated for each replication of experimental conditions.

3.2 PHRASE II – APPLICATION OF THE THIRD INTERNATIONAL MATHEMATICS AND SCIENCE STUDY (TIMSS)

3.2.1 Introduction

The Third International Mathematics and Science Study (TIMSS) has been conducted by the International Association for the Evaluation of Educational Achievement (IEA) since 1995. Every four years, TIMSS implements a study of achievement of students in mathematics and science in three different groups, and collects extensive information about teaching and learning in mathematics and science for students, teachers, and their school principals. More than 40 countries participate in this international comparative study. Most students in Group 1 are 9 years old at the time of testing (i.e., 3rd- and 4th-grade students in most of the countries), whereas students in Group 2 are 13 years old (i.e., 7th- and 8th-grade students). Group 3 consists of students in their final year of secondary education. Students in Group 2 are required to take tests in all countries that participating in the study, but countries are permitted to choose whether or not to administer tests to the students of Group 1 and 3. Each student receives a booklet of

cognitive subject items, including mathematics and science tests, and a questionnaire about their attitudes toward mathematics and science, classroom activities, home background, and out-of-school activities. Mathematics and science teachers and principals also responds to a questionnaire to collect information about the social and cultural contexts of learning, such as teachers' views on the curriculum, school resources, and support for teachers (Gonzalez & Smith, 1997).

The "curriculum" concept in TIMSS is based on three levels: (1) the intended curriculum defined by the educational system and society; (2) the implemented curriculum taught by teachers; and (3) the attained curriculum, what students have learned (Gonzalez & Miles, 2001). The data collected from the assessment administered to students is designed to capture the attained curriculum. In 2003, the cognitive assessment of TIMSS was developed based on two domains: content and cognitive. One feature of the TIMSS assessment is that both multiple-choice and construct-response items were included in TIMSS tests. However, approximately 80% of the items are designed as multiple-choice items. The categories of content domain in mathematics and science of three versions developed in 1995, 1999, and 2003 have been revised slightly over time. For example, there were eight categories of content domain in mathematics in 1995, while there were only six categories in 1999. More details on the assessment framework of TIMSS 2003 are presented in the next section.

3.2.2 The assessment frame of TIMSS 2003

The mathematics and science tests of Grade 8 developed in 2003 were selected for analysis. According to the description of assessment frameworks and specifications of TIMSS 2003 from the International Study Center (Mullis, Martin, Smith, Garden, Gregory, Gonzalez, Chrostorwski, & O'Connor, 2003), there were five content categories included in the mathematics test: Numbers, Algebra, Measurement, Geometry, and Data. In the science test, there were also five content categories: Life Science, Physics, Chemistry, Earth Science, and Environmental Science. Table 3.8 shows the percentage of each content category in the mathematics and science tests in Grade 8 (Mullis et al., 2003). The whole item pool of both subjects included some trend items from TIMSS 1995 or 1999 and some new replacement items developed in TIMSS 2003.

The cognitive domains of mathematics included knowing facts and procedure, using concepts, solving routine problems, and reasoning, whereas the cognitive domains for science included factual knowledge, conceptual understanding and reasoning and analysis (Neidorf & Garden, 2004). These cognitive domains referred to the skills and abilities supposed to be demonstrated by students in their test answers. Table 3.9 presents the percentages of cognitive domains.

Table 3.8

Percentages of content categories of TIMSS 2003 in Grade 8

Math content domain	
Numbers	30%
Algebra	25%
Measurement	15%
Geometry	15%
Data	15%
Science content domain	
Life Science	30%
Physical Science (Chemistry)	15%
Physical Science (Physics)	25%
Earth Science	15%
Environmental Science	15%

Table 3.9

Percentages of cognitive domains of TIMSS 2003 in Grade 8

Math cognitive domains	
Knowing facts and procedures	15%
Using concepts	20%
Solving routine problems	40%
Reasoning	25%
Science cognitive domains	
Factual knowledge	30%
Conceptual understanding	35%
Reasoning and analysis	35%

All the mathematics and science items were combined and divided into 28 blocks (14 blocks for each subject). All 28 blocks were distributed in 12 student booklets. Each booklet contained six blocks, in two forms, either two math blocks and four science blocks or four math

blocks and two science blocks. In other words, there were six booklets consisting of two math and four science blocks and six booklets containing of four math and two science blocks. The block design of TIMSS 2003 ensured that were enough items existed in order to provide reliable measurement in the two subjects for each student. For linking purposes among all booklets, certain amounts of blocks were paired with other blocks. Each block (either math or science) consisted of approximately eight to nine multiple-choice items, three to four short constructed-response items, and zero to one extended constructed-response item. The total number of items in a block ranged from 11 to 16. Therefore, the total score of points in each booklet if all questions were answered correctly ranged from 90 to 97 (M = 94) (Neidorf & Garden, 2004). Each test booklet was randomly assigned to one student; therefore, each item was administered to approximately equal numbers of students. In order to ensure enough students for each item, at least 4500 students were administered the tests. Administration of the TIMSS 2003 assessment contained three timed sections (shown in Table 3.10). Students had two 45-minute sessions to take the test. Note that only the multiple-choice items were analyzed in this study.

Two booklets of Grade 8 were randomly selected, and only multiple-choice responses and subjects administered in the English version were used for the factor analysis. Test lengths were approximately 50 and the sample size was approximately 1500. Sireci and Gonzalez (2003) concluded one dominant factor existed in the TIMSS science item pool using TIMSS 1999 data. However, no study had confirmed the test structure at the booklet level currently. Therefore, this study also provided the evidence of test structures at the booklet level.

Table 3.10

Administration and item types of TIMSS 2003 assessment

Activity	Test time (minutes)	Number of blocks	Item type
Student Booklet - Part 1 (Calculators Not Permitted)	45	3	multiple-choice short constructed-response, extended constructed-response
Student Booklet - Part 2 (Calculators Permitted)	45	3	multiple-choice short constructed-response extended constructed-response
Student Questionnaire	30		

4.0 RESULTS

The results of the simulated data for Mplus and TESTFACT will be presented in the first sections. The results include the proportion of correct dimensionality decisions, the number of dimensions uncovered, and parameter recovery. The second section discusses the examination of the test dimensionality and factor structure of the TIMSS data.

4.1 THE RESULTS OF SIMULATION DATA

Four major indices were used to examine test dimensionality: 1) proportion of variance accounted for, 2) parallel analysis, 3) reduction of RMSR and 4) the chi-square difference test. The first three indices were subjective in nature because essentially the cut points for the decisions were determined by the researcher. The chi-square difference test was a formal test and more likely to provide objective decisions. The proportion of correct dimensionality decisions for these four indices is presented separately for different discrimination conditions. For the decisions of the estimated number of dimensions, the results are primarily presented in figures in order to compare the performances of four indices. For the parameter recovery study, the results are presented to illustrate the influence of guessing, correlations among dimensions, and discrimination in determining dimensionality when using Mplus and TESTFACT. Because the proportion of cases with non-convergent solutions not only affected the results but also demonstrated the estimation problems for using Mplus and TESTFACT, the number of valid cases with convergent solutions is presented first.

4.1.1 Number of non-convergent solutions

Tables 4.1 to 4.4 display the number of convergent solutions (valid cases) when using Mplus and TESTFACT. In this study, the cases were marked as non-convergent cases if the results had not converged after a specified number of iterations or the procedure failed without a factor solution. The specific numbers of iterations for both methods were 200 in Mplus and 50 in TESTFACT. The major reason for setting a smaller number of iterations in TESTFACT was the time consumption for a 100-replication design, and preliminary results indicated that 50 iterations should be adequate under most conditions (see Section 3.1.1). Also, in Mplus, the dimensionality decisions were based on the WLSMV solutions, except the results of the chi-square test index were based on the WLS estimator (more detail see Section 3.1.4). In fact, there were no non-convergent cases in the WLSMV solution, which was more robust to sample size. However, given a test length equal to 60 with 2000 subjects, the WLS estimator might have been problematic in obtaining a convergent solution because of the need for larger sample sizes. Therefore, the information shown in Tables 4.1 and 4.2 is the number of valid cases based on the solutions using the WLS estimator. There were 38 out of 7000 non-convergent cases that totally failed to converge in all factor solutions in Mplus. In TESTFACT, there were a total of 66 out of 7000 cases that had non-convergent problems in all factor solutions.

The effect of a non-convergent case was to stop the process of determining the number of factors. For example, in three-dimensional data, the estimation procedure obtained one- and two-factor convergent solutions but no three- to five-factor solutions. The chi-square test stopped at the difference test for one- and two-factor solutions. Therefore, the test dimensionality would be two if the chi-square test was significant, otherwise, the test dimensionality would be one. Additionally, any convergent factor solution was not counted as a valid case after a non-convergent solution. For instance, if a dataset obtained one-, two-, and four-factor convergent solutions, the possible dimensionality was only one or two dimensions. Note that the situations of having convergent solutions after one non-convergent solution were detected more frequently in TESTFACT than in Mplus, especially for three-dimensional data.

In Tables 4.1 and 4.2, the number of convergent solutions cases for Mplus decreased when the number of extracted factors increased. A decrease in valid cases was observed in data

that assumed guessing and a high correlation. In Tables 4.3 and 4.4, the patterns in the number of valid cases for TESTFACT across different conditions, such as guessing or factor correlations, were similar to the patterns for Mplus (see Tables 4.1 and 4.2). One exception was the number of valid cases decreased more dramatically in TESTFACT when the number of extracted factors increased. There were no valid cases in most conditions for higher factor solutions, such as four- or five-factor solutions. In addition, the number of valid cases was limited for three-dimensional data that assumed guessing particularly for the high correlation condition.

It should be noted that the number of valid cases was treated differently for the indices used to evaluate dimensionality. Since the proportion of variance and parallel analysis indices are based on eigenvalues, which are derived in all runs whether a converged solution or no converged solution, all cases were analyzed. On the other hand, the RMSR reduction index and the chi-square test were based on the output of converged solutions. Thus, only those cases with converged solutions were analyzed. In summary, for the proportion of variance and parallel analysis indices, 100 cases per each condition were analyzed in Mplus and TESTFACT. For the RMSR reduction and the chi-square test indices, only the valid cases shown in Tables 4.1 to 4.4 were analyzed in Mplus and TESTFACT (e.g., 87 cases were used in Mplus for three-dimensional data with c = .33, r = .3 and HHH discrimination condition). Finally, those conditions with less than 25 valid cases were excluded due to the small proportion of valid cases. Due to the presence of a small proportion of valid cases in data with and without guessing in the high correlation condition, these two conditions were excluded for all analyses in following sections. In the parameter recovery section, only those cases whose number of extracted factors matched the true value of the dimensionality were used in the investigation of parameter recovery. For example, the cases of two-factor solutions were used to illustrate how well parameter recovery is for two-dimensional data. Therefore, the number of valid cases used in parameter recovery was less than that used in the proportion of correct dimensionality decisions and the number of dimensions decisions.

Table 4.1

Cases with convergent solutions in Mplus using WLS under the low correlation condition (*r* = .3)

Disc. Condition[a]	c = 0					c = .33				
	1F[b]	2F	3F	4F	5F	1F	2F	3F	4F	5F
Unidimensional data										
H	93	92	89			100	100	80		
M	99	99	99			100	99	90		
L	100	100	100			100	100	88		
Two-dimensional data										
HH	100	100	99	91		100	100	95	50	
MH	98	98	98	97		100	99	97	69	
MM	100	100	100	99		100	100	91	74	
LH	99	99	99	96		100	100	92	69	
LM	100	100	100	99		100	99	92	59	
LL	100	100	100	99		100	100	97	70	
Three-dimensional data										
HHH	100	100	100	95	90	100	100	87	62	41
HHM	99	99	99	95	88	100	99	90	63	36
HHL	99	99	99	97	90	100	100	89	67	37
MMH	100	100	100	98	94	100	99	93	71	41
MMM	100	100	100	100	97	100	100	92	70	56
MML	100	100	100	100	94	100	100	91	70	50
LMH	100	100	99	98	97	100	100	86	69	39
LLH	100	100	100	93	89	100	100	83	52	46
LLM	98	98	97	96	92	100	100	90	57	47
LLL	99	99	99	97	93	100	99	86	54	28

[a] More details about discrimination conditions see Sections 3.1.1 and 3.1.2

[b] "1F" represented one-factor solutions, "2F" represented two-factor solutions and so on.

Table 4.2

Cases with convergent solutions in Mplus using WLS under the high correlation condition (*r* = .6)

Disc.	c = 0					c = .33				
Condition	1F	2F	3F	4F	5F	1F	2F	3F	4F	5F
Two-dimensional data										
HH	100	99	100	92		100	100	86	27	
MH	99	99	99	93		100	98	88	59	
MM	97	97	96	91		100	100	93	53	
LH	99	99	98	88		100	100	80	48	
LM	100	100	98	98		100	99	83	60	
LL	100	100	100	98		100	99	87	54	
Three-dimensional data										
HHH	96	96	91	75	56	100	100	78	20	7
HHM	100	100	95	91	77	99	98	83	20	6
HHL	99	99	95	90	73	100	100	79	35	8
MMH	100	100	96	94	80	100	100	93	41	11
MMM	100	100	100	97	86	100	100	84	49	22
MML	100	100	98	97	94	100	100	91	47	22
LMH	98	98	96	86	82	100	99	88	33	20
LLH	98	98	97	85	81	100	100	83	35	9
LLM	98	98	95	93	87	100	100	84	45	21
LLL	95	94	92	90	87	100	99	77	41	20

Table 4.3

Cases with convergent solutions in TESTFACT under the low correlation condition (*r* = .3)

Disc.	c = 0					c = .33				
Condition	1F	2F	3F	4F	5F	1F	2F	3F	4F	5F
Unidimensional data										
H	100	86	3			99	21	0		
M	100	62	10			96	11	0		
L	100	73	27			95	10	0		
Two-dimensional data										
HH	100	100	91	86		98	95	18	1	
MH	100	100	64	66		97	89	6	0	
MM	100	100	67	68		97	89	4	0	
LH	100	100	70	63		95	85	5	0	
LM	100	100	58	66		95	82	1	0	
LL	100	100	74	71		92	83	1	0	
Three-dimensional data										
HHH	100	100	100	87	63	99	85	67	0	0
HHM	100	100	100	82	55	99	94	56	0	0
HHL	100	100	100	80	49	99	92	51	0	0
MMH	100	100	100	88	62	98	89	46	0	0
MMM	100	100	100	78	56	97	68	34	0	0
MML	100	100	100	86	53	99	83	28	0	0
LMH	100	100	100	81	56	96	87	44	0	0
LLH	100	100	100	82	54	99	83	33	0	0
LLM	100	100	100	80	62	96	81	34	0	0
LLL	100	100	100	82	56	95	59	25	0	0

Table 4.4

Cases with convergent solutions in TESTFACT under the high correlation condition (*r* = .6)

Disc.	c = 0					c = .33				
Condition	1F	2F	3F	4F	5F	1F	2F	3F	4F	5F
Two-dimensional data										
HH	100	100	66	78		99	99	7	0	
MH	100	100	56	60		99	91	4	0	
MM	100	100	74	61		99	87	0	0	
LH	100	100	71	63		99	80	1	0	
LM	100	100	72	55		96	82	0	0	
LL	100	100	60	51		98	72	0	0	
Three-dimensional data										
HHH	100	100	60	37	21	100	94	0	0	0
HHM	100	100	25	16	7	100	94	0	0	0
HHL	100	100	39	20	6	98	86	0	0	0
MMH	100	100	0	0	0	99	84	0	0	0
MMM	100	100	0	0	0	99	71	0	0	0
MML	100	100	0	0	0	100	78	0	0	0
LMH	100	100	5	0	0	99	83	0	0	0
LLH	100	100	16	8	3	98	69	1	0	0
LLM	100	100	0	0	0	99	60	2	0	0
LLL	100	100	0	0	0	98	51	1	0	0

4.1.2 The proportion of correct dimensionality decisions

In this section, the presentation of the results focuses on the extent to which the four indices, the proportion of variance, parallel analysis, the RMSR reduction and the chi-square test, could be used to identify the correct or simulated dimensionality. To facilitate comparisons across conditions, results are organized by the correlation condition, the guessing condition and the number of simulated dimensions. Since any correlation for unidimensional data was not possible, the correlation condition should be ignored for the unidimensional cases.

Tables 4.5 to 4.6 present, respectively, the mean proportion of correct dimensionality decisions based on the first two indices, the proportion of variance and parallel analysis, for determining the number of factors. With regard to the proportion of variance index in Table 4.5, the performances of Mplus and TESTFACT were similar when using data without modeled

guessing. Under data that assumed no guessing, the performance deteriorated when either the dimensionality of data or the correlations among dimensions increased. Additionally, in three-dimensional data, the performance deteriorated when the discrimination parameters decreased. In the lowest discrimination condition (i.e., LLL condition), no correct decisions were found. In data that assumed guessing, the performance of TESTFACT was similar to the performance of Mplus, given unidimensional data and two-dimensional data with higher discrimination, such as HH or MH conditions. With two-dimensional data, TESTFACT appeared superior to Mplus given lower discrimination parameter and r = .3. When r = .6, the correct decision proportions were in the range of 70% to 90% for TESTFACT, whereas no correct decisions were made for Mplus. Given three-dimensional data, both TESTFACT and Mplus performed poorly except for higher discrimination parameter conditions, where TESTFACT was superior. The impact of different factor correlations was greater in data that assumed guessing. Higher correlations decreased the correct decision rates. With three-dimensional data, the performances of both methods were significantly worse than with one- and two-dimensional data. A remarkable finding was that no correct dimensionality decisions occurred in the three-dimensional data with a high correlation condition in both estimation methods (i.e., Mplus and TESTFACT).

The results of parallel analysis in Table 4.6 demonstrated that Mplus performed poorly in all conditions with data that assumed guessing (i.e., no correct decisions), whereas TESTFACT had higher correct decision rates for items with higher discrimination. However, Mplus performed better than TESTFACT with three-dimensional data that assumed no guessing and a high correlation. Similar to the results using the proportion of variance index, the performances of Mplus and TESTFACT in the three-dimensional data that assumed guessing with high correlations were the worst (i.e., all cases had no correct decisions). When discrimination decreased, the correct decision rates decreased, and the discrepancy of correct decision rates between Mplus and TESTFACT increased. With regard to the correlation effect, in data that assumed no guessing, Mplus performed better in the high correlation condition, whereas TESTFACT performed better in the low correlation condition. However, with data that assumed guessing, the correlation effect was not found in both methods, except for three-dimensional data in TESTFACT.

Table 4.5

The mean proportion of correct dimensionality decisions using the proportion of variance index

Disc. Condition	*r* = .3				*r* = .6			
	c = 0		c = .33		c = 0		c = .33	
	Mplus	TSF[a]	Mplus	TSF	Mplus	TSF	Mplus	TSF
Unidimensional data								
H	1.0	1.0	1.0	1.0				
M	1.0	1.0	1.0	1.0				
L	1.0	1.0	1.0	1.0				
Two-dimensional data								
HH	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
MH	1.0	1.0	1.0	0.9	1.0	1.0	0.0	0.9
MM	1.0	1.0	1.0	0.9	1.0	1.0	0.0	0.8
LH	1.0	1.0	0.9	0.9	1.0	1.0	0.0	0.5
LM	1.0	1.0	0.3	0.9	0.4	0.4	0.0	0.2
LL	1.0	1.0	0.0	0.9	0.0	0.0	0.0	0.1
Three-dimensional data								
HHH	1.0	1.0	0.2	0.7	0.0	0.0	0.0	0.0
HHM	1.0	1.0	0.0	0.6	0.0	0.0	0.0	0.0
HHL	1.0	1.0	0.0	0.2	0.0	0.0	0.0	0.0
MMH	1.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0
MMM	1.0	1.0	0.0	0.3	0.0	0.0	0.0	0.0
MML	0.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0
LMH	0.9	0.9	0.0	0.1	0.0	0.0	0.0	0.0
LLH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

[a] **"TSF" represented "TESTFACT".**

Table 4.6

The mean proportion of correct dimensionality decisions using parallel analysis

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	0.7	0.8	0.0	0.8				
M	0.0	0.4	0.0	0.9				
L	0.0	0.7	0.0	0.9				
				Two-dimensional data				
HH	1.0	1.0	0.0	0.8	1.0	1.0	0.0	0.9
MH	0.8	0.9	0.0	0.9	0.9	0.9	0.0	0.9
MM	0.3	0.5	0.0	0.9	0.4	0.7	0.0	0.9
LH	0.5	0.7	0.0	0.8	0.6	0.8	0.0	0.8
LM	0.0	0.4	0.0	0.8	0.1	0.4	0.0	0.9
LL	0.0	0.2	0.0	0.9	0.0	0.3	0.0	0.7
				Three-dimensional data				
HHH	1.0	1.0	0.0	0.7	1.0	0.6	0.0	0.0
HHM	1.0	1.0	0.0	0.6	1.0	0.3	0.0	0.0
HHL	0.9	0.9	0.0	0.5	0.9	0.4	0.0	0.0
MMH	0.9	1.0	0.0	0.5	0.9	0.0	0.0	0.0
MMM	0.7	0.8	0.0	0.4	0.9	0.0	0.0	0.0
MML	0.4	0.6	0.0	0.3	0.7	0.2	0.0	0.0
LMH	0.7	0.9	0.0	0.5	0.8	0.1	0.0	0.0
LLH	0.3	0.5	0.0	0.3	0.7	0.3	0.0	0.0
LLM	0.1	0.3	0.0	0.4	0.4	0.3	0.0	0.0
LLL	0.0	0.1	0.0	0.3	0.3	0.3	0.0	0.0

Because parallel analysis was based on the comparison of estimated eigenvalues from the target data and random data, any decrease or increase of eigenvalues might have an impact on determining dimensionality. A similar impact of the increase or decrease in eigenvalues should be observed in the performance of the proportion of variance index as well. Tables 4.7 and 4.8 provide the first and the second eigenvalues in Mplus and TESTFACT under low and high correlation conditions. The information demonstrated how modeling guessing affected the eigenvalues. In Table 4.7, when the data assumed guessing, the first eigenvalues of Mplus dropped dramatically, compared to the no guessing condition. The eigenvalues were only half the value of the eigenvalues obtained in data that assumed no guessing. Meanwhile, guessing only led to a small decrease, around 10%, in TESTFACT. The pattern of second eigenvalues was similar to the first one. The difference between both methods for the second eigenvalues was smaller than for the first eigenvalues when data assumed guessing. The significant drop in the first and the second eigenvalues for Mplus resulted in an increase in the rest of the eigenvalues. Thus, dimensionality was overestimated because the eigenvalues obtained from random data were relatively close to the eigenvalues of simulation data. Any increase of the eigenvalues, however slight the increase, affected the results in determining dimensionality.

Table 4.8 presents the first two eigenvalues in both methods under the higher correlation condition (r = .6). The pattern and values of the first eigenvalues were similar to the low correlation condition. However, the higher the correlations among dimensions, the larger the first eigenvalues. This again led to a decrease in the rest of eigenvalues which decreased the proportion of overestimation in the high correlation condition. Tests with lower discrimination decreased the first eigenvalues and caused smaller differences between Mplus and TESTFACT in data that assumed guessing.

In summary, the first two eigenvalues of Mplus and TESTFACT were about the same in data that assumed no guessing. However, in data that assumed guessing, the eigenvalues of TESTFACT were relatively higher than those of Mplus. Low discrimination also caused a decrease of the eigenvalues. High correlations increased the first eigenvalues, but decreased the rest of the eigenvalues. Therefore, high correlations among dimensions led to an increase in underestimating dimensionality. In conclusion, TESTFACT did correct the measurement error problem caused by guessing. TESTFACT showed superior results when using parallel analysis and the proportion of variance index with data that modeled guessing.

Table 4.7

The means of the first two eigenvalues in Mplus and TESTFACT (*r* = .3)

	Eigenvalue 1				Eigenvalue 2			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	34.95	34.91	17.50	30.65	1.29	1.26	2.33	2.35
M	27.71	27.67	13.88	23.70	1.46	1.44	1.98	2.63
L	20.57	20.53	10.34	16.93	1.58	1.55	1.73	2.85
				Two-dimensional data				
HH	28.71	28.66	14.55	25.29	8.29	8.28	4.51	6.58
MH	26.65	26.60	13.52	22.95	7.11	7.09	3.90	5.67
MM	24.30	24.24	12.29	20.46	6.20	6.18	3.49	5.22
LH	24.85	24.80	12.58	21.22	5.66	5.65	3.21	4.41
LM	22.18	22.13	11.25	18.40	5.07	5.05	2.94	4.24
LL	19.74	19.70	9.99	16.01	4.21	4.20	2.55	3.71
				Three-dimensional data				
HHH	28.44	28.35	14.27	24.45	5.37	5.34	3.13	4.53
HHM	27.38	27.29	13.74	23.24	5.26	5.23	3.04	4.43
HHL	26.38	26.30	13.25	22.21	5.26	5.23	3.02	4.38
MMH	26.25	26.18	13.19	22.11	4.75	4.74	2.79	4.02
MMM	25.00	24.94	12.54	20.88	4.08	4.07	2.51	3.61
MML	23.95	23.89	11.99	19.91	3.99	3.98	2.44	3.52
LMH	25.21	25.14	12.66	21.04	4.64	4.62	2.71	3.92
LLH	24.20	24.13	12.12	19.95	4.15	4.12	2.51	3.61
LLM	22.80	22.74	11.40	18.65	3.52	3.51	2.21	3.19
LLL	21.60	21.54	10.85	17.38	2.86	2.85	1.97	2.99

Table 4.8

The means of the first two eigenvalues in Mplus and TESTFACT (r = .6)

	Eigenvalue 1				Eigenvalue 2			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Two-dimensional data				
HH	33.51	33.46	16.88	29.55	4.65	4.64	2.77	4.02
MH	30.97	30.91	15.60	26.79	4.04	4.03	2.48	3.52
MM	28.30	28.24	14.29	24.03	3.58	3.57	2.28	3.24
LH	28.51	28.47	14.37	24.34	3.32	3.31	2.17	3.00
LM	25.70	25.66	12.90	21.44	2.98	2.97	2.02	2.89
LL	22.98	22.94	11.56	18.75	2.54	2.54	1.85	2.81
				Three-dimensional data				
HHH	35.22	35.12	17.61	30.60	2.95	2.93	2.21	2.68
HHM	33.83	33.75	16.92	29.19	2.89	2.87	2.16	2.60
HHL	32.48	32.39	16.25	27.82	2.88	2.87	2.11	2.66
MMH	32.44	32.36	16.22	27.77	2.62	2.60	2.10	2.54
MMM	30.97	30.90	15.40	26.29	2.32	2.31	2.02	2.50
MML	29.60	29.53	14.72	24.89	2.27	2.26	2.00	2.57
LMH	31.09	31.01	15.49	26.30	2.55	2.54	2.05	2.58
LLH	29.67	29.59	14.75	24.84	2.37	2.35	2.02	2.63
LLM	28.18	28.11	14.03	23.37	2.04	2.03	1.94	2.65
LLL	26.76	26.70	13.26	21.86	1.76	1.76	1.91	2.72

The assessment of dimensionality based on the RMSR reduction index in Mplus and TESTFACT is shown in Table 4.9. As can be seen the pattern of results was similar to previous results. The performance in data that assumed no guessing was better than in data that assumed guessing. Using both methods, the proportions of correct decisions under high or low correlation conditions were similar in data without guessing, except for three-dimensional data. However, in one- and two-dimensional data that assumed guessing, the performance of TESTFACT was similar or better than Mplus. In all conditions of three-dimensional data, the performance of Mplus was much better than the performance of TESTFACT. In addition, the performance of both methods with the three-dimensional data that assumed guessing with a high correlation was least successful and somewhat unpredictable. Another remarkable finding, the influence of discrimination in Mplus for one- and two- dimensional data, differed for three-dimensional data. In one- and two-dimensional data, better performance using Mplus was observed in lower discrimination conditions, which was opposite to the results of the other indices mentioned above (i.e., the proportion of variance and parallel analysis). However, the patterns of performance in three-dimensional data were not so clear. The influence of factor correlations was observed in data that assumed guessing using Mplus and TESTFACT. Higher factor correlations decreased the rate of correct decisions. Note that results not available due to non-convergence are indicated by N/A in the table. As can be seen these occurred only for TESTFACT in the three-dimension, correlation equal to .6 condition.

Table 4.10 displays the mean proportion of correct dimensionality using the chi-square difference test. In most of the conditions, the performance of TESTFACT was superior to Mplus, except for three-dimensional data that assumed guessing with a high correlation. When using Mplus, there were decision rates in most conditions of no more than 10% correct, except with three-dimensional data that assumed guessing where correct decision rates were in the range of 50% to 80%. A larger discrepancy was found in the results for Mplus between the low and high correlation conditions as well. Interestingly, Mplus performed better with data that assumed guessing than with data that did not assume guessing. In contrast, TESTFACT performed well with one-to-three dimensional data, having decision rates of 100% correct, except for three-dimensional data and the high correlation condition where the number of non-convergent solutions precluded any evaluation. Also, TESTFACT performed well with data that assumed no guessing or guessing.

Table 4.9

The mean proportion of correct dimensionality decisions using the reduction of RMSR index

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	0.6	1.0	0.0	1.0				
M	1.0	1.0	0.0	1.0				
L	1.0	1.0	1.0	1.0				
				Two-dimensional data				
HH	0.9	1.0	0.0	1.0	0.9	1.0	0.0	0.9
MH	1.0	1.0	0.4	1.0	0.9	1.0	0.0	0.4
MM	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.2
LH	1.0	1.0	0.7	0.9	1.0	1.0	0.0	0.0
LM	1.0	1.0	1.0	0.9	1.0	1.0	0.7	0.0
LL	1.0	1.0	1.0	0.5	1.0	1.0	0.0	0.0
				Three-dimensional data				
HHH	1.0	0.9	0.1	0.7	0.8	N/A	0.0	N/A
HHM	1.0	0.6	0.3	0.2	0.8	N/A	0.0	N/A
HHL	1.0	1.0	0.6	0.0	0.9	N/A	1.0	N/A
MMH	1.0	0.9	0.6	0.2	0.9	N/A	0.5	N/A
MMM	1.0	0.8	0.8	0.0	0.9	N/A	0.0	N/A
MML	1.0	1.0	0.9	0.0	0.9	N/A	0.0	N/A
LMH	1.0	1.0	0.8	0.0	0.9	N/A	0.6	N/A
LLH	1.0	1.0	0.3	0.0	0.9	N/A	0.2	N/A
LLM	1.0	1.0	0.3	0.0	0.8	N/A	0.0	N/A
LLL	1.0	1.0	0.0	0.0	0.4	N/A	0.0	N/A

Table 4.10

The mean proportion of correct dimensionality decisions using the chi-square test

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	0.0	1.0	0.0	1.0				
M	0.0	1.0	0.0	1.0				
L	0.0	1.0	0.0	0.9				
				Two-dimensional data				
HH	0.0	0.1	0.1	1.0	0.0	1.0	0.1	1.0
MH	0.0	0.5	0.0	1.0	0.0	1.0	0.1	1.0
MM	0.0	0.4	0.1	1.0	0.0	1.0	0.1	1.0
LH	0.0	0.5	0.1	1.0	0.0	1.0	0.2	1.0
LM	0.0	0.5	0.1	1.0	0.0	1.0	0.2	1.0
LL	0.0	0.4	0.0	1.0	0.0	1.0	0.1	1.0
				Three-dimensional data				
HHH	0.1	1.0	0.4	1.0	0.2	N/A	0.8	N/A
HHM	0.0	1.0	0.4	1.0	0.1	N/A	0.8	N/A
HHL	0.0	1.0	0.3	1.0	0.1	N/A	0.7	N/A
MMH	0.0	1.0	0.3	1.0	0.1	N/A	0.6	N/A
MMM	0.0	1.0	0.3	1.0	0.0	N/A	0.5	N/A
MML	0.0	1.0	0.3	1.0	0.0	N/A	0.5	N/A
LMH	0.0	1.0	0.3	1.0	0.1	N/A	0.7	N/A
LLH	0.1	1.0	0.4	1.0	0.1	N/A	0.6	N/A
LLM	0.0	1.0	0.4	1.0	0.0	N/A	0.5	N/A
LLL	0.0	1.0	0.4	1.0	0.0	N/A	0.6	N/A

In summary, differences in the performance (i.e., proportion of correct dimensionality) of the four indices for determining the number of dimensions were observed. For the conditions of no modeled guessing (c = 0), and factor correlations equal to .3, the RMSR statistic, in general, appeared to yield better performance. For simulated data with one or two dimensions, the proportion of variance index was also effective, and TESTFACT and Mplus were observed to perform similarly. However, for simulated data with three dimensions or factor correlations equals to .6, the performance of both TESTFACT and Mplus deteriorated although the performance of TESTFACT was markedly lower than Mplus.

As for the condition of modeled guessing (c = .33), TESTFACT performed better generally than Mplus. The parallel analysis and the chi-square test indices appeared to perform best in general when using TESTFACT, whereas the proportion of variance and the chi-square test indices performed better in Mplus. As for the case of no modeled guessing, in simulated data with three dimensions and factor correlations equal to .6, decreased performance was observed, particularly with regard to TESTFACT. It should be noted that in many cases there were no valid cases at this condition (see Tables 4.3 and 4.4). In these cases TESTFACT always underestimated true dimensionality.

4.1.3 Comparing the number of dimensions

Based on the proportion of correct decisions, the performance of Mplus and TESTFACT varied, given different underlying dimensionality. This section compares the four indices in terms of the number of estimated dimensions uncovered. The results that are presented are based on difference between the estimated number of dimensions and the true dimensionality (see Appendix B to Appendix E).

Figures 4.1 to 4.4 present the mean differences for unidimensional data for c = 0 and c = .33. Figures 4.1 and 4.2 present the results for the four indices for Mplus and TESTFACT respectively under the no guessing condition. As can be seen, the proportion of variance and the RMSR reduction indices performed well for Mplus and TESTFACT. In addition, the chi-square test and parallel analysis for tests with high discrimination items also performed well. For the condition of assumed guessing (see Figures 4.3 and 4.4), all indices yielded greater dimensionality with Mplus except the proportion of variance index. In contrast, the performance

of these four indices in TESTFACT was very similar (see Figure 4.4). In conclusion, for unidimensional data, TESTFACT performed better than Mplus with data that did and did not assume guessing. The proportion of variance index illustrated the best performance among these four indices in both methods.

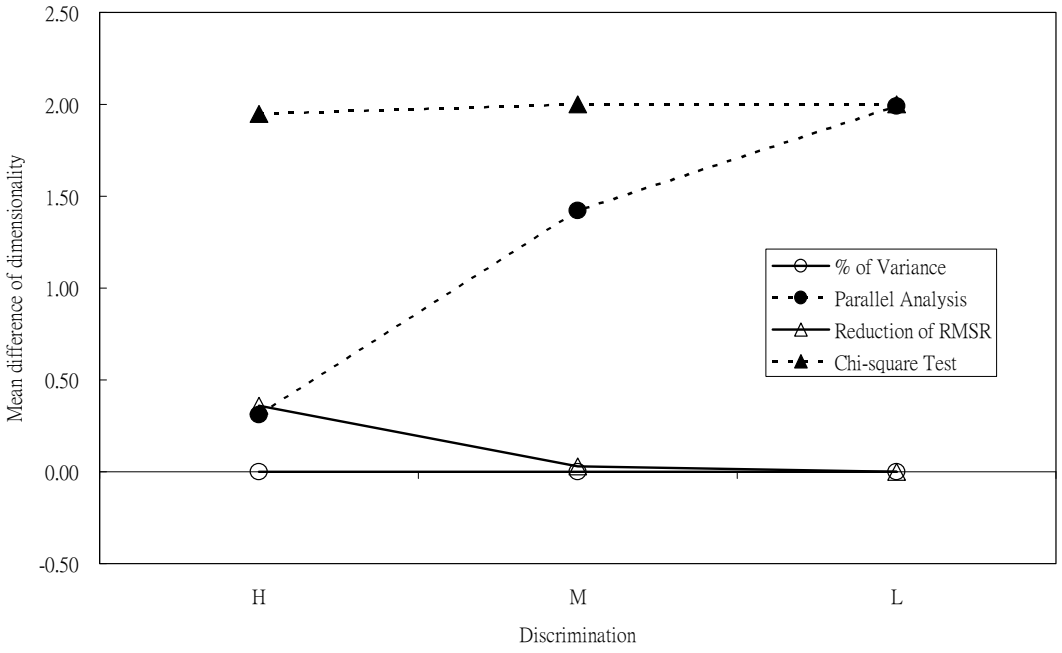


Figure 4.1 The mean difference of estimated and true dimensionality in Mplus (1D, c = 0)

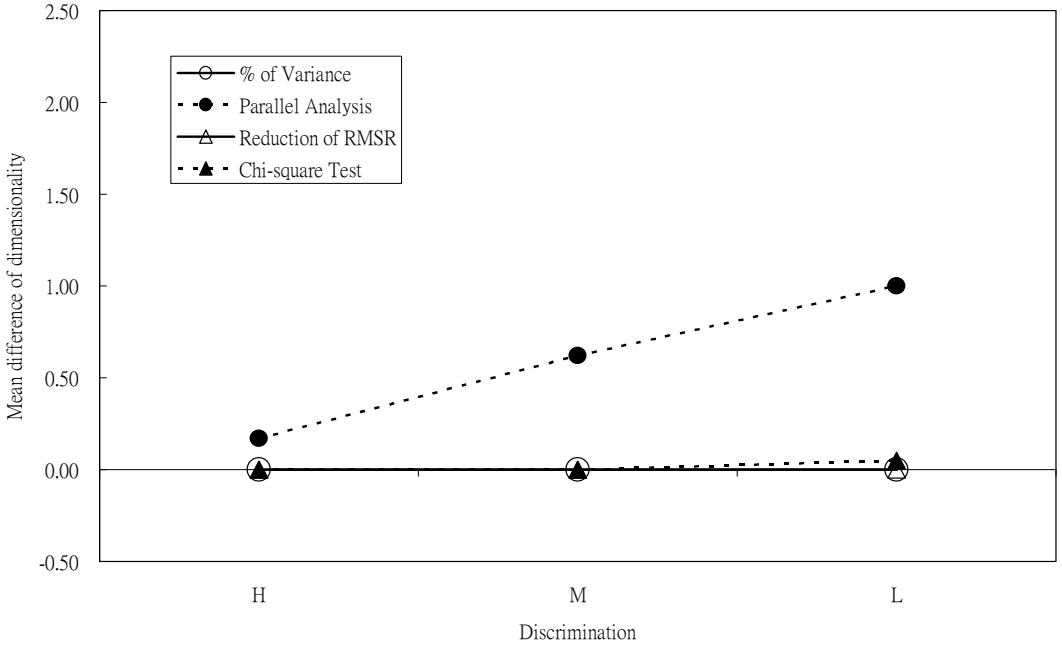


Figure 4.2 The mean difference of estimated and true dimensionality in TESTFACT (1D, c = 0)

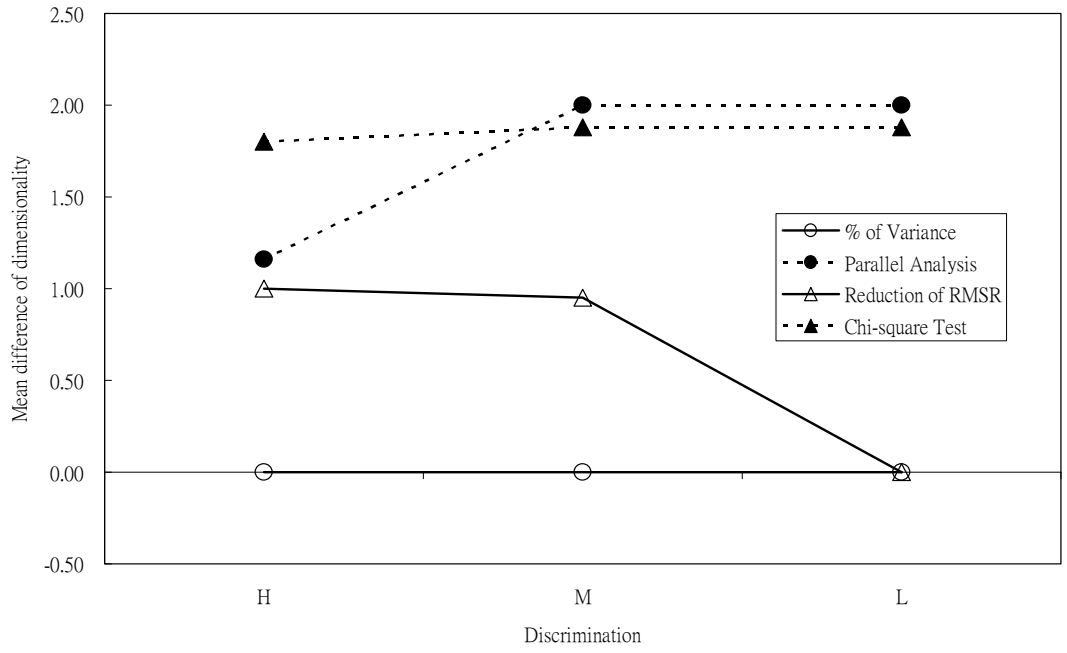


Figure 4.3 The mean difference of estimated and true dimensionality in Mplus (1D, c = .33)

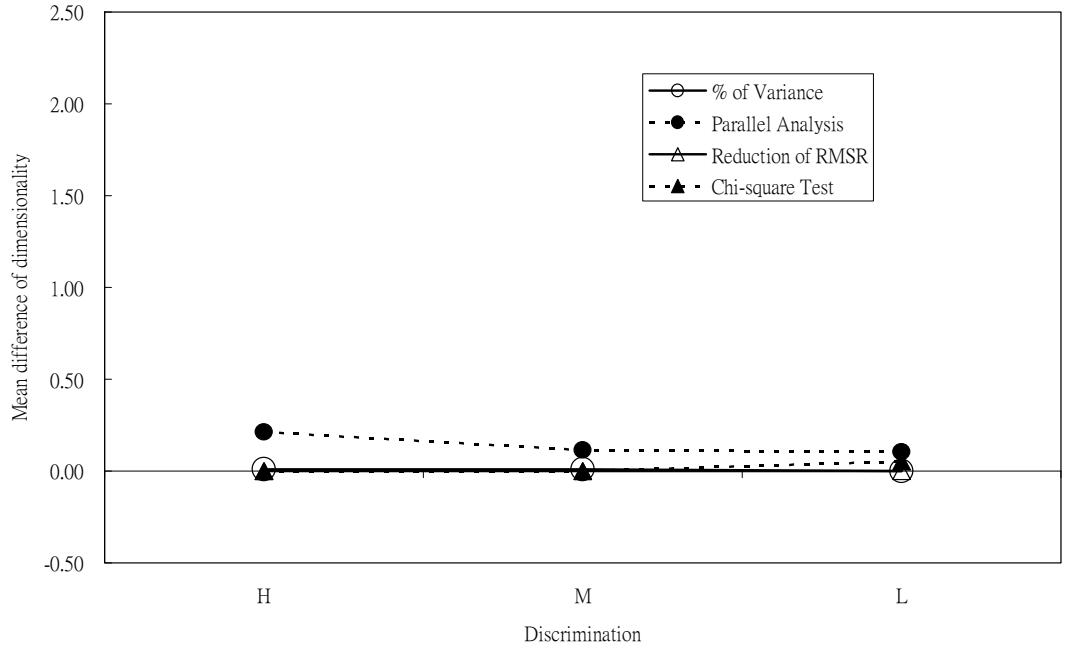


Figure 4.4 The mean difference of estimated and true dimensionality in TESTFACT (1D, c = .33)

Figures 4.5 to 4.12 present the number of estimated dimensions with two-dimensional data using the four indices. In data that assumed no guessing (see Figures 4.5 to 4.8) and either low or high correlation conditions, the performances of Mplus and TESTFACT were similar for all indices, except Mplus had a serious overestimation problem using the chi-square test. Note that the degree of overestimation in using parallel analysis increased as the discrimination decreased in both methods. While dimensionality was overestimated for the most part, the proportion of variance index resulted in some underestimation in tests with low discriminating items and factor correlations equal to .6 for both methods.

In data that assumed guessing (see Figures 4.9 to 4.12), Mplus tended to overestimate dimensionality and TESTFACT underestimated dimensionality when there were differences between TESTFACT and Mplus. For Mplus, the proportion of variance and the RMSR reduction indices performed well in the low correlation condition (see Figure 4.9), but either overestimated or underestimated the true dimensionality in the high correlation condition (see Figure 4.11). In contrast, TESTFACT showed its superiority using the four indices in the low correlation condition (see Figure 4.10). In the high correlation condition (see Figure 4.12), TESTFACT underestimated the true dimensionality slightly using the proportion of variance and the reduction of RMSR indices, whereas the other indices performed fairly well, especially using the chi-square test (the estimated dimensionality was close to the true dimensionality).

In conclusion, in data that assumed no guessing, most indices performed well except the chi-square test in Mplus. With data that assumed guessing, both parallel analysis and the chi-square test consistently overestimated the dimensionality, whereas the other two indices either overestimated or underestimated dimensionality depending on the discrimination and correlation conditions. Using TESTFACT, the four indices performed generally well in all conditions. Finally, the consistency among the four indices in TESTFACT was greater than in Mplus. In Mplus and TESTFACT, the effect of discrimination was observed in the results of most indices except the chi-square test index.

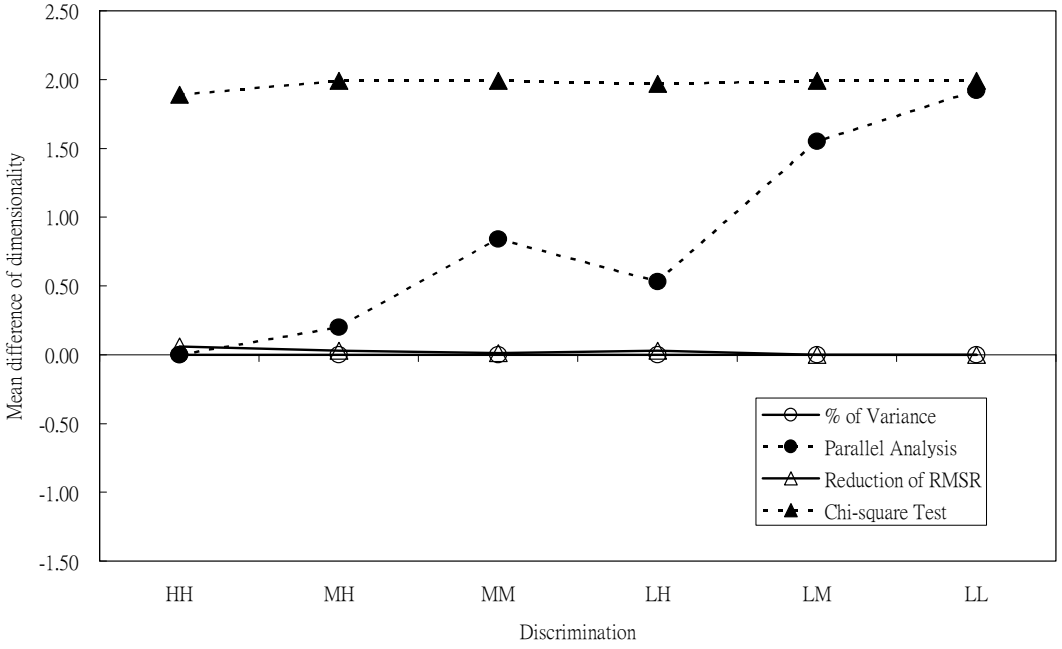


Figure 4.5 The mean difference of estimated and true dimensionality in Mplus (2D, c = 0, r = .3)

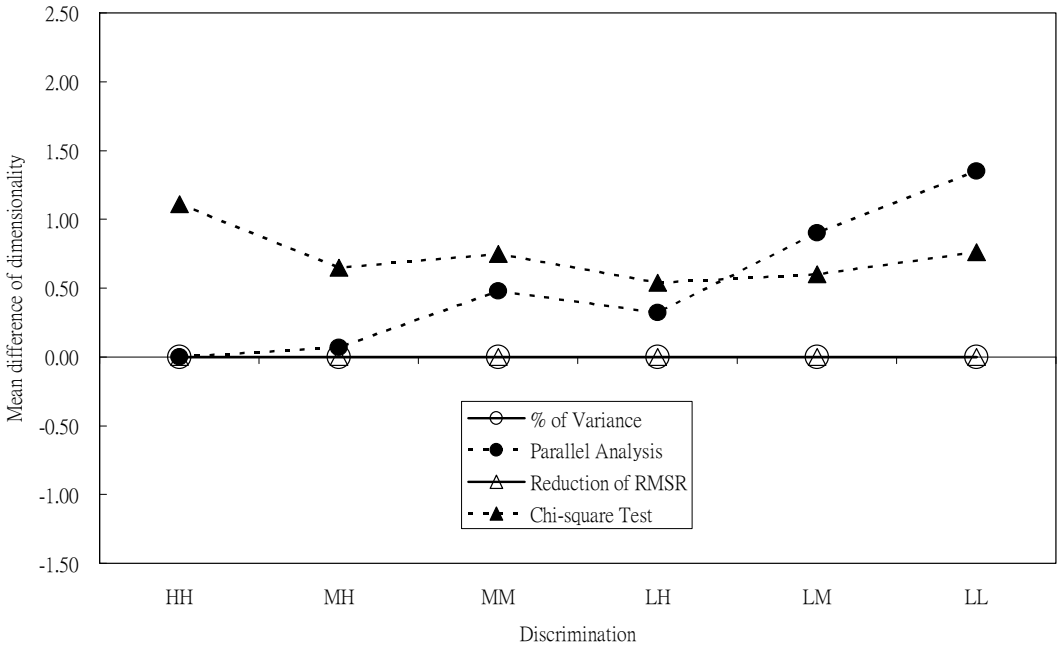


Figure 4.6 The mean difference of estimated and true dimensionality in TESTFACT (2D, c = 0, r = .3)

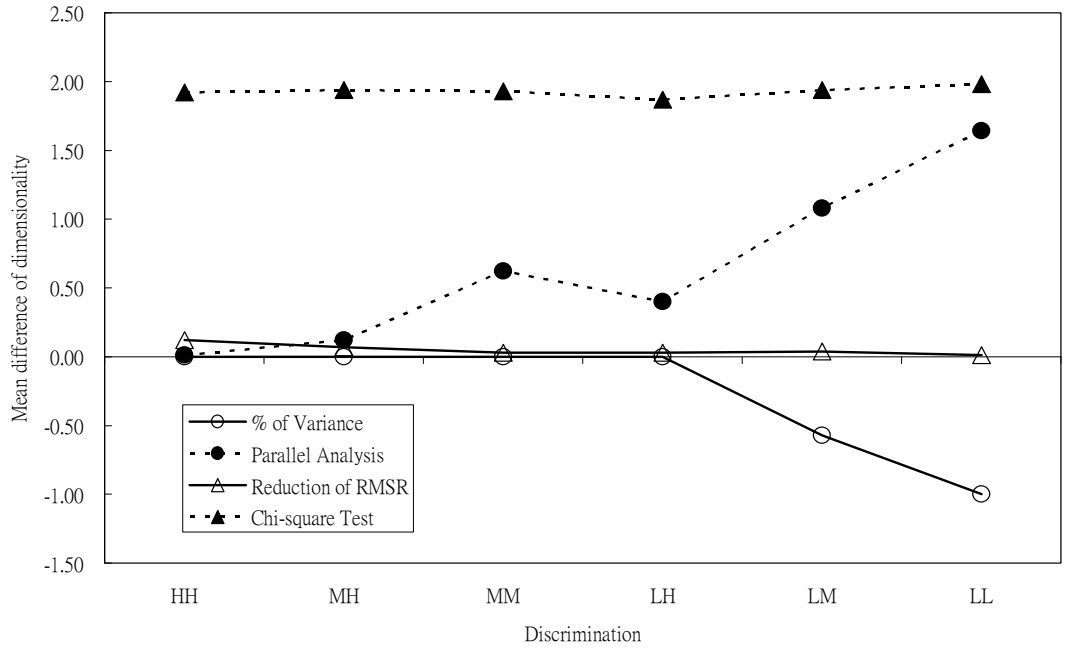


Figure 4.7 The mean difference of estimated and true dimensionality in Mplus (2D, c = 0, r = .6)

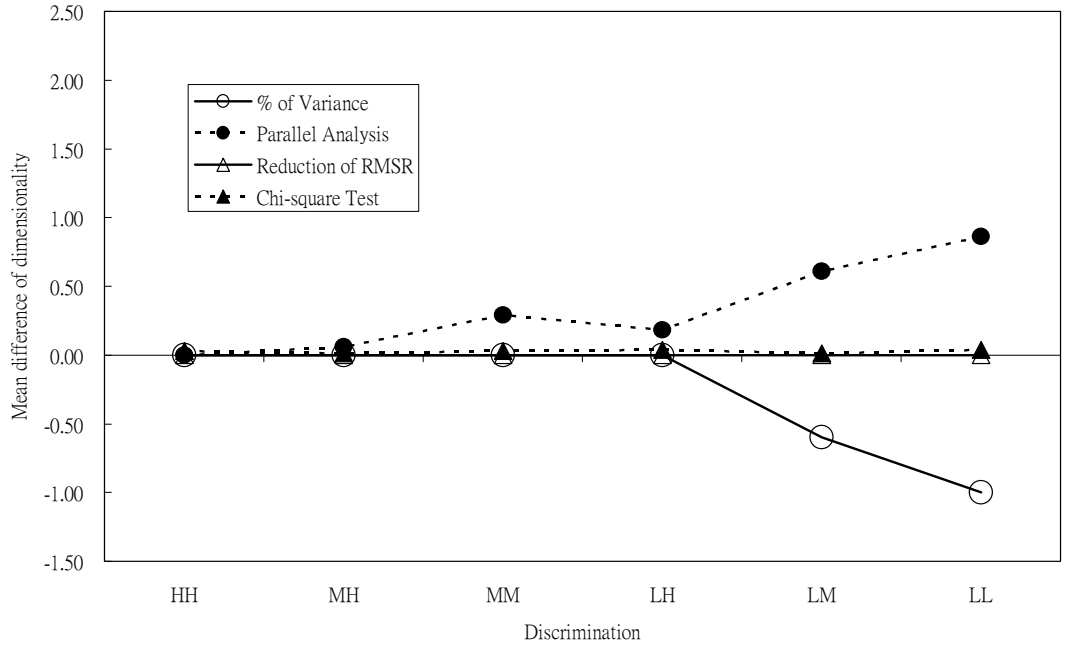


Figure 4.8 The mean difference of estimated and true dimensionality in TESTFACT (2D, c = 0, r = .6)

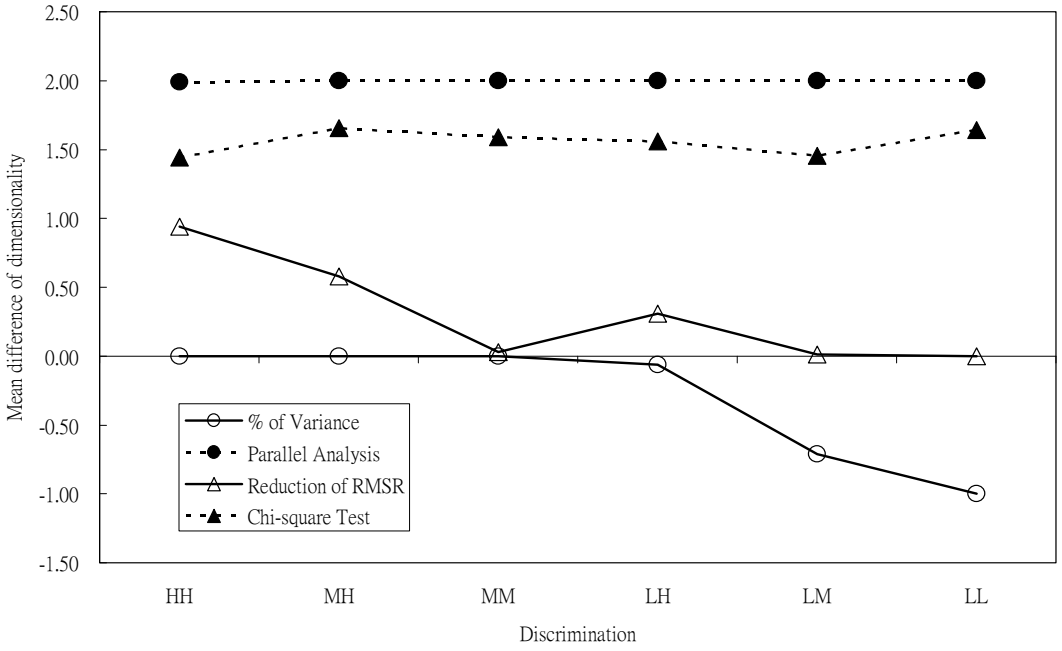


Figure 4.9 The mean difference of estimated and true dimensionality in Mplus (2D, c = .33, r = .3)

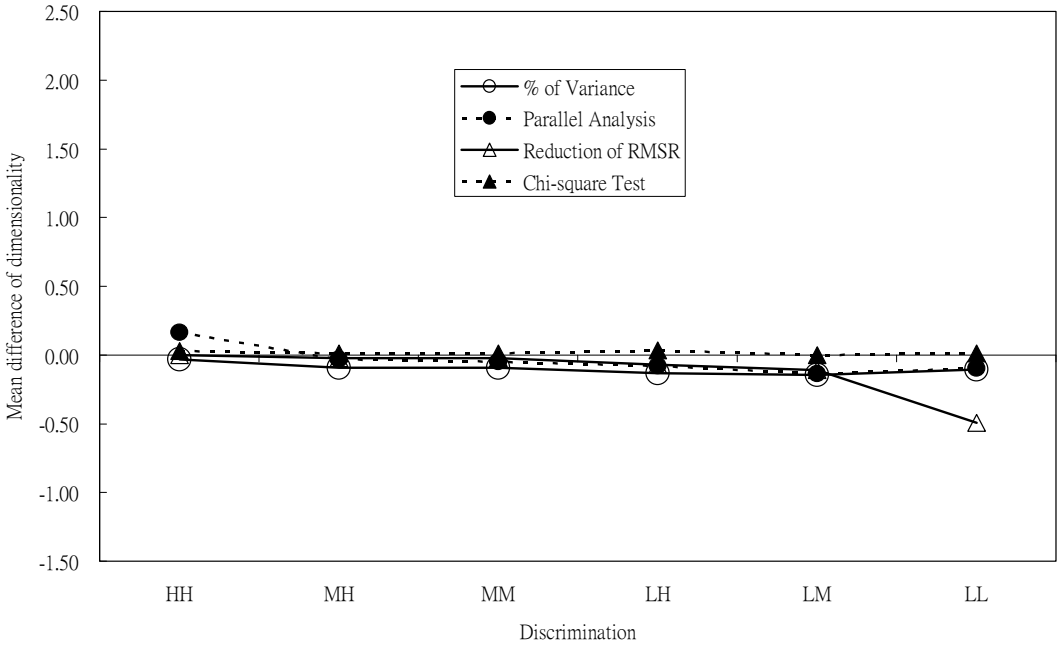


Figure 4.10 The mean difference of estimated and true dimensionality in TESTFACT (2D, c=.33, r = .3)

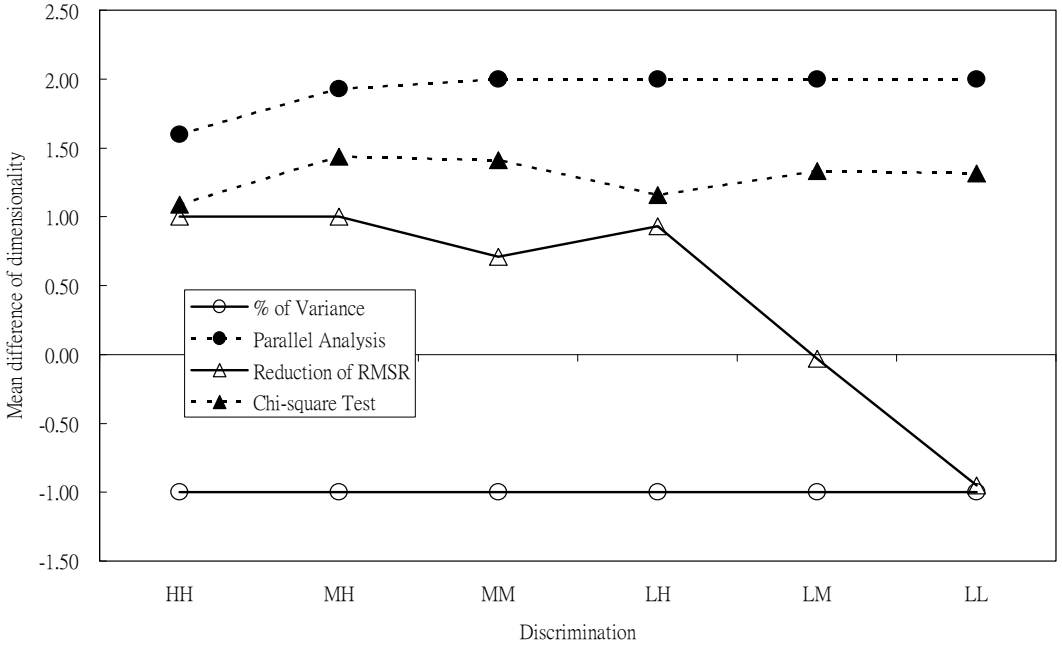


Figure 4.11 The mean difference of estimated and true dimensionality in Mplus (2D, c=.33, r = .6)

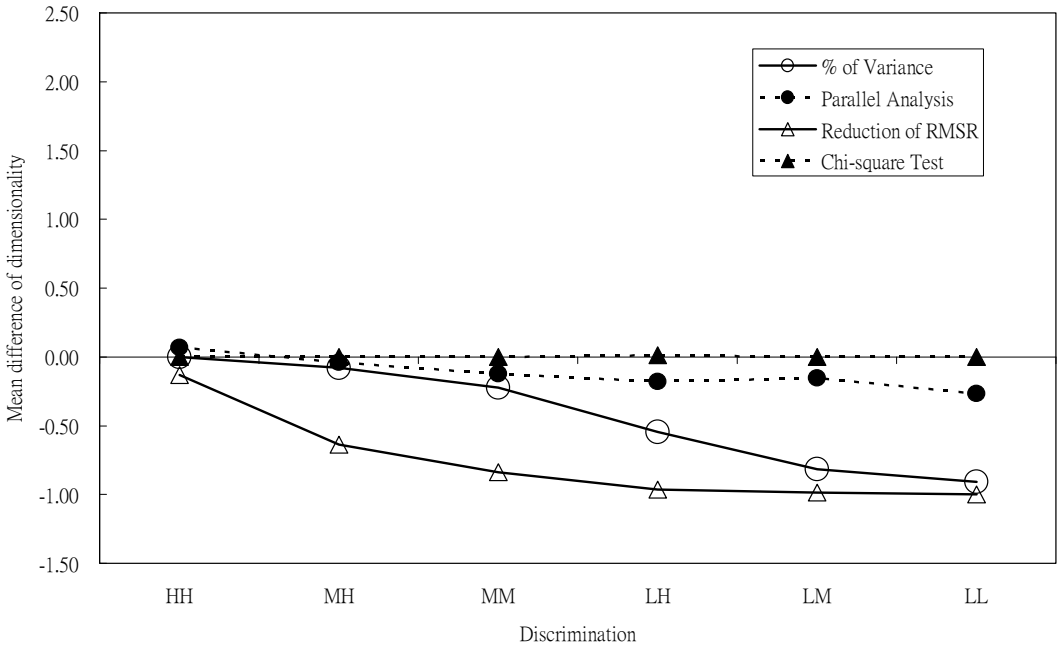


Figure 4.12 The mean difference of estimated and true dimensionality in TESTFACT (2D, c=.33, r = .6)

Figures 4.13 to 4.20 present the mean differences on the number of uncovered dimensions for three-dimensional data. The pattern of three-dimensional data was quite different from that of the two-dimensional data, especially regarding the performance of TESTFACT. For the no modeled guessing conditions (i.e., c = 0, see Figures 4.13 and 4.14), the pattern and the performance of Mplus and TESTFACT were very similar, except the chi-square test consistently overestimated dimensionality with Mplus. When the correlations among dimensions increased (see Figure 4.15 and Figure 4.16), the performances of parallel analysis and the reduction of RMSR index in Mplus were close to the true dimensionality, whereas the same indices with TESTFACT underestimated the true dimensionality. However, the proportion of variance index and the chi-square test either underestimated or overestimated in Mplus, whereas these two indices underestimated dimensionality in TESTFACT. In other words, the superiority of TESTFACT was not evident in three-dimensional data as it was shown to be in one- and two-dimensional data. Note that results based on the RMSR index and chi-square test were not available for TESTFACT under the .6 correlation condition due to the number of non-convergent solutions (see Figures 4.16 and 4.20).

In data that assumed guessing (see Figures 4.17 to 4.20), the performances of all indices in Mplus showed consistent differences from true dimensionality, but there was no consistency in the direction of estimation among indices (the dimensionality was either overestimated or underestimated). Interestingly, the performance of the RMSR reduction index in Mplus was affected by discrimination. Mplus tended to overestimate dimensionality given higher discrimination conditions and to underestimate dimensionality in lower discrimination conditions. However, consistent overestimation was found in Mplus using the chi-square test except in data that assumed guessing with the high correlation condition (only slightly overestimated). In contrast, in TESTFACT, the chi-square test performed very well except the results were unknown in data with a high correlation, again due to problems with non-convergent solutions. The other three indices tended to underestimate dimensionality consistently, but the degree of underestimation was different. The most serious underestimation was found in data that assumed guessing with the high correlation condition. Overall, parallel analysis performed better than the other two indices, especially in data that assumed guessing with the high correlation condition.

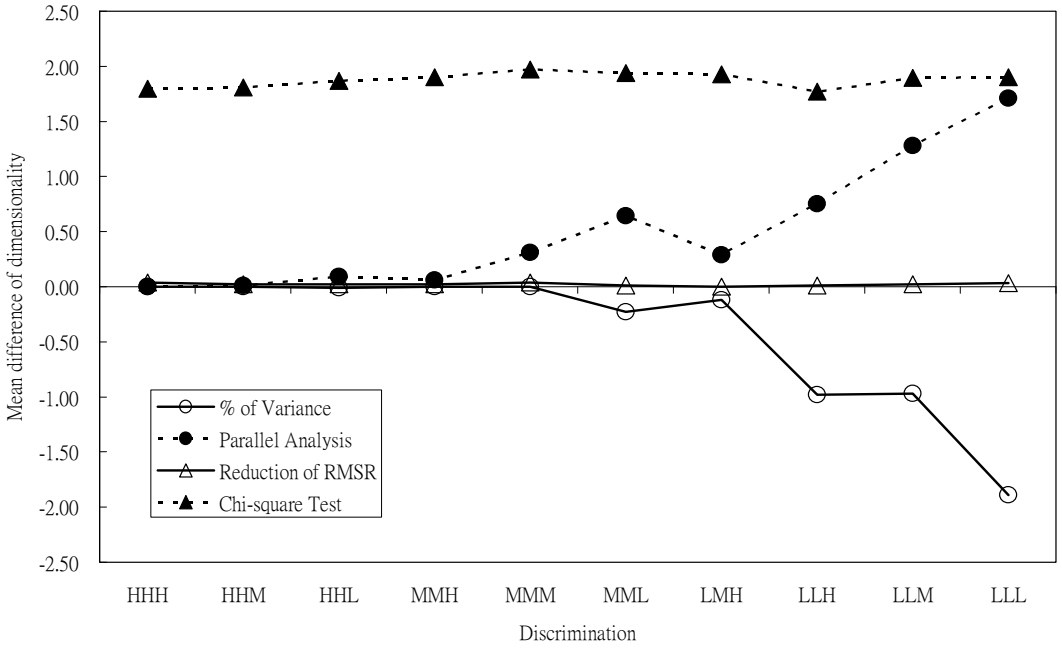


Figure 4.13 The mean difference of estimated and true dimensionality in Mplus (3D, c = 0, r = .3)

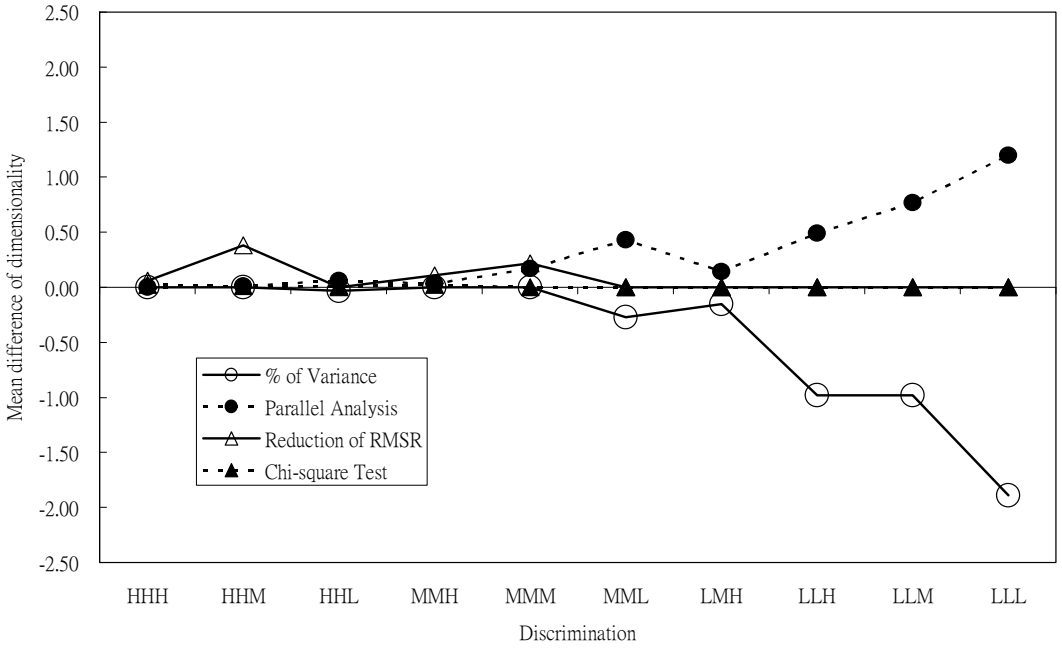


Figure 4.14 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = 0, r = .3)

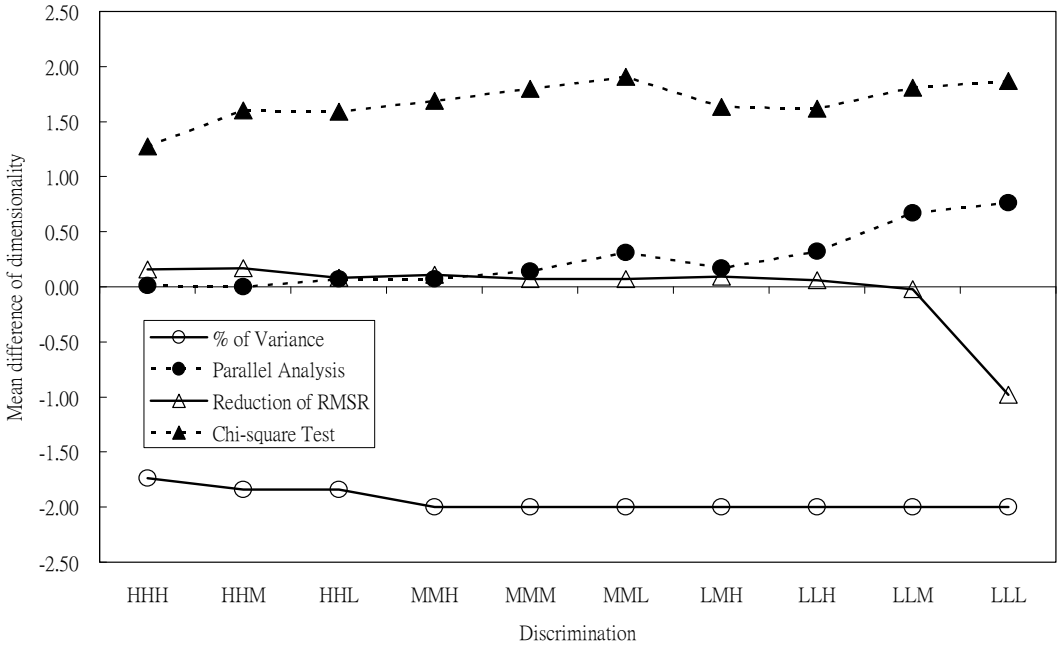


Figure 4.15 The mean difference of estimated and true dimensionality in Mplus (3D, c = 0, r = .6)

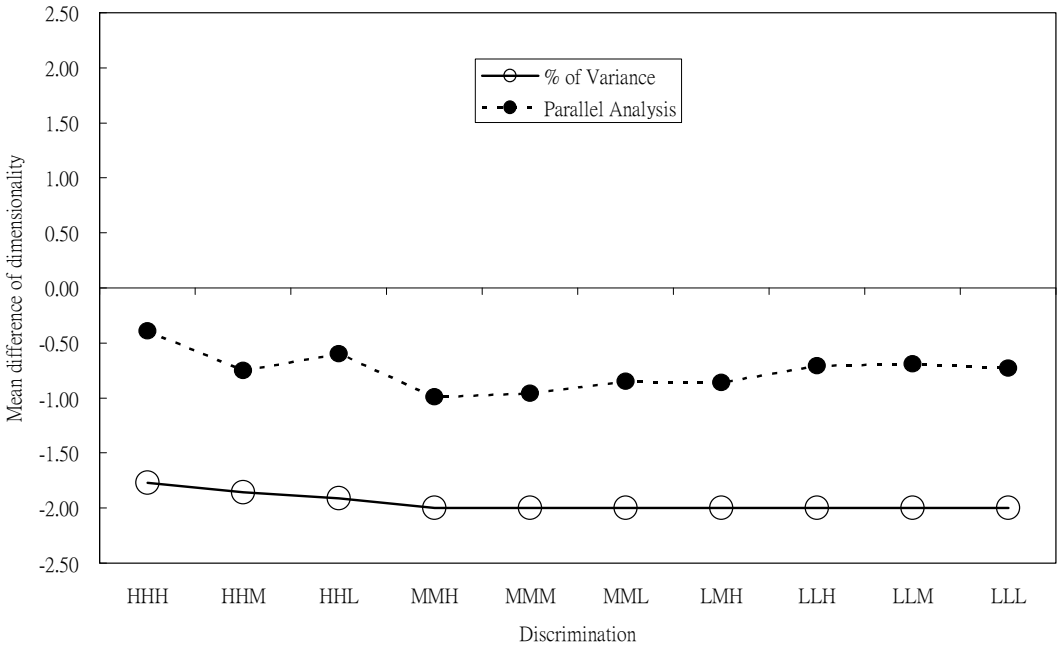


Figure 4.16 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = 0, r = .6)

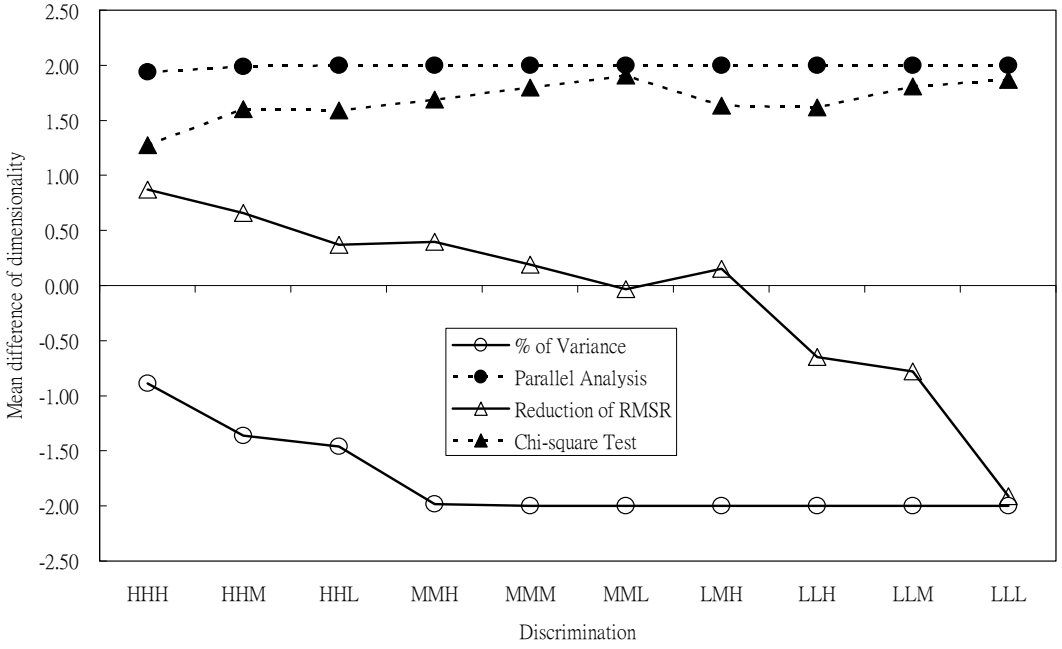


Figure 4.17 The mean difference of estimated and true dimensionality in Mplus (3D, c = .33, r = .3)

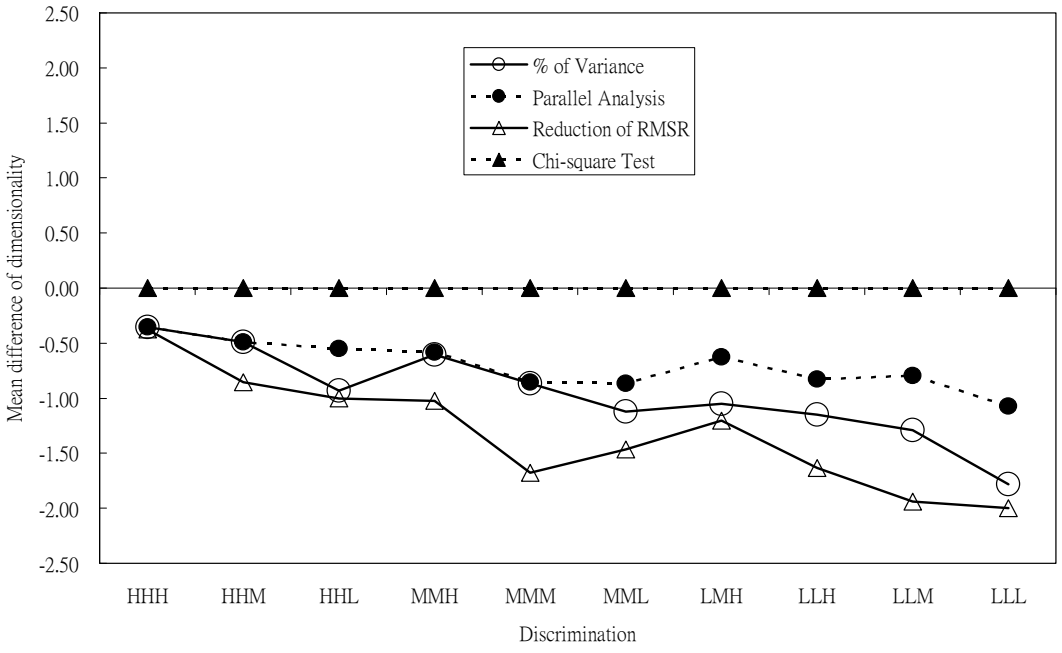


Figure 4.18 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = .33, r = .3)

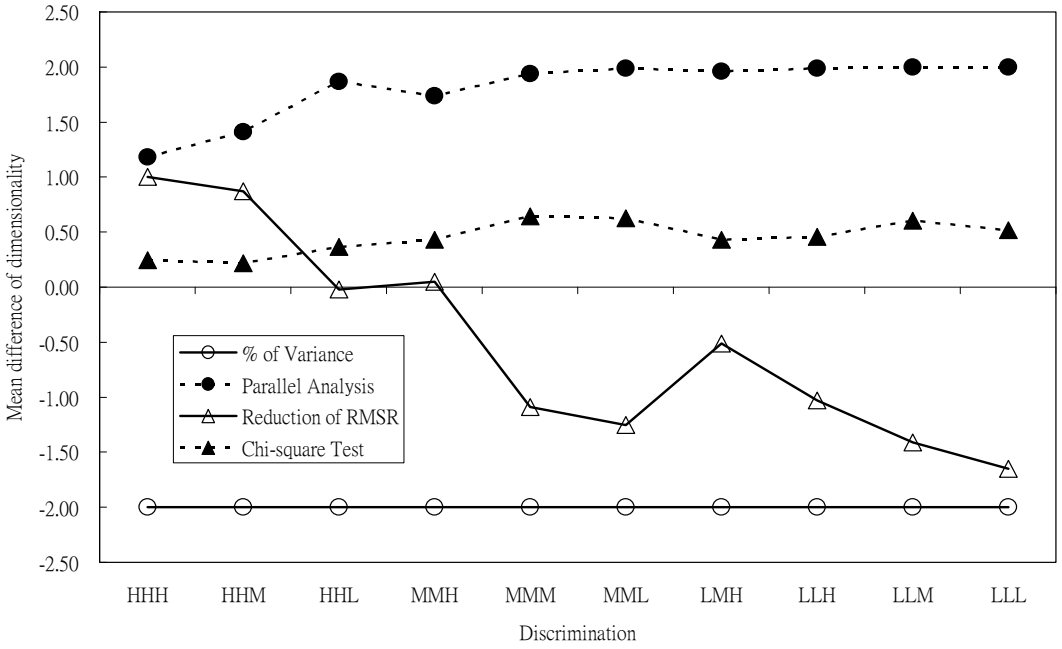


Figure 4.19 The mean difference of estimated and true dimensionality in Mplus (3D, c = .33, r = .6)

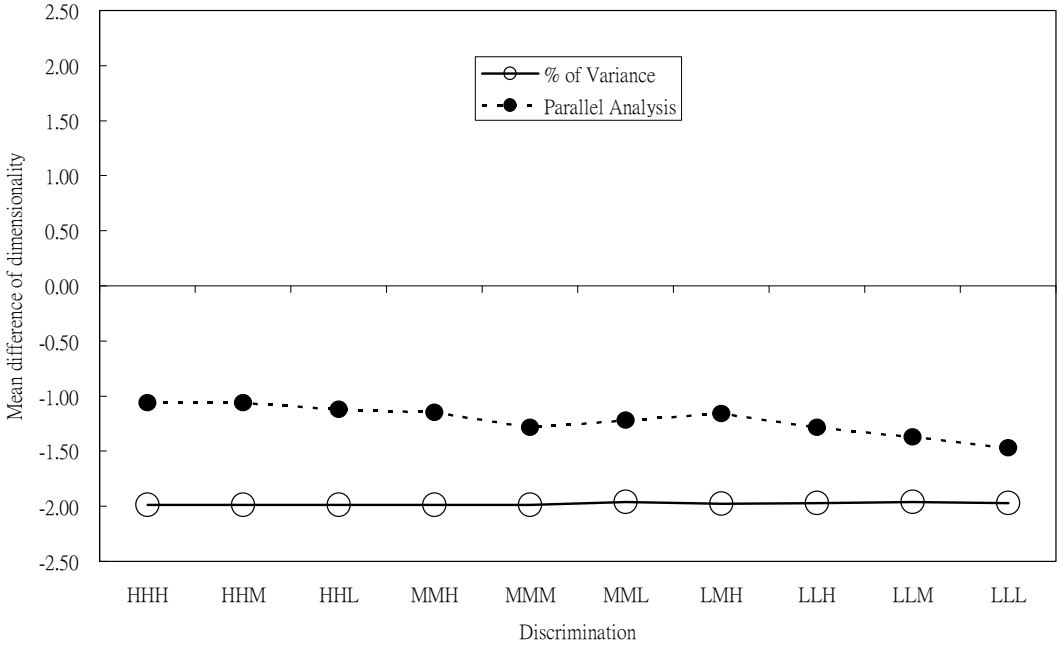


Figure 4.20 The mean difference of estimated and true dimensionality in TESTFACT (3D, c = .33, r = .6)

The results comparing the differences between estimated and true dimensionality illustrated indices that sometimes overestimated dimensionality and indices that sometimes underestimated dimensionality. Because the magnitude of the eigenvalues was the basis of the proportion of variance and parallel analysis indices, the impact of guessing was greater in these two indices since the correction for guessing directly affects the magnitude of the eigenvalues. It is clear that the performance of parallel analysis was significantly worse in Mplus, but not the case in TESTFACT, when the data assumed guessing. In addition, the impact of the guessing correction was greatest in the first eigenvalues. Hence, the proportion of variance index had the best performance in unidimensional data but tended to underestimate dimensionality in higher dimensional data. In contrast, the RMSR reduction index and the chi-square test were not directly affected by the correction of guessing. Instead, the discrimination effect had a greater impact on the performance of the RMSR reduction index, whereas the chi-square test may have been affected by relatively large sample size.

In summary, when comparing the four indices with unidimensional data, the proportion of variance and the RMSR reduction indices performed better in Mplus, whereas most indices performed fairly well in TESTFACT. For two-dimensional data that either assumed guessing or no guessing, the RMSR reduction performed relatively well in Mplus and TESTFACT, compared to other indices, except in the condition that data assumed guessing with a high factor correlation. For three-dimensional data using Mplus, the RMSR reduction index worked well in data that assumed no guessing and the chi-square test performed better in data that assumed guessing. In TESTFACT, parallel analysis and the chi-square test performed either fairly well or better than other indices for most conditions.

4.1.4 Parameter Recovery

In order to evaluate parameter recovery, only factor solutions where the number of factors extracted matched the underlying dimensionality were analyzed. For example, a two-factor solution was used to evaluate parameter recovery given simulated two-dimensional data. The index for this evaluation was RMSD, which compared estimated parameters with true values (see Equation 3.1). RMSD was calculated for each replication for all conditions. An oblique solution (PROMAX) was used in the two- and three-dimensional data because a nonzero

correlation was assumed (.3 and .6). Table 4.11 presents the valid cases for these analyses. Given data that assumed guessing, the number of valid cases was less than those in data that assumed no guessing. However, the impact was greater in TESTFACT. As discussed before, TESTFACT showed serious non-convergent problems with three-dimensional data (see Tables 4.3 and 4.4). Higher factor correlations also decreased the number of valid cases. For example, given three-dimensional data that assumed guessing in TESTFACT, there were 30% to 80% valid cases for three-factor solutions under the low correlation condition, whereas there were almost no valid cases with convergent three-factor solutions in the high correlation condition. Due to the presence of a small proportion of valid cases in data with and without guessing in the high correlation condition, these two conditions were excluded in the analysis of parameter recovery for TESTFACT.

Table 4.11

Valid cases for factor solutions matched the underlying dimensionality in TESTFACT

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
Unidimensional data								
H	93	100	100	99				
M	99	100	100	96				
L	100	100	100	95				
Two-dimensional data								
HH	100	100	100	95	99	100	100	99
MH	98	100	99	89	99	100	98	91
MM	100	100	100	89	97	100	100	87
LH	99	100	100	85	99	100	100	80
LM	100	100	99	82	100	100	99	82
LL	100	100	100	83	100	100	99	72
Three-dimensional data								
HHH	100	100	87	67	91	N/A	78	N/A
HHM	99	100	90	56	95	N/A	83	N/A
HHL	99	100	89	51	95	N/A	79	N/A
MMH	100	100	93	46	96	N/A	93	N/A
MMM	100	100	92	34	100	N/A	84	N/A
MML	100	100	91	28	98	N/A	91	N/A
LMH	99	100	86	44	96	N/A	88	N/A
LLH	100	100	83	33	97	N/A	83	N/A
LLM	99	100	90	34	95	N/A	84	N/A
LLL	99	100	86	25	92	N/A	77	N/A

Table 4.12 displays the mean RMSD of parameter recovery in Mplus and TESTFACT. With data that assumed no guessing and the low correlation condition, the RMSD values obtained for Mplus were slightly smaller than the RMSD for TESTFACT. However, with data that assumed guessing and low factor correlations, the values of RMSD in TESTFACT were significantly smaller than the values of Mplus. These results showed the superiority of TESTFACT when the data assumed guessing. In the high correlation condition, a similar pattern was observed for two-dimensional data. However, larger values of RMSD for three-dimensional data in the high correlation condition were found in Mplus when compared to the low correlation condition.

The table also illustrated an effect due to the size of the discrimination parameters, but this effect depended on the presence of modeled guessing. For c = .33 and the low correlation condition, there did not appear to be an effect using Mplus. As can be seen, all the RMSD values were approximately the same across all three dimensionality conditions. However, for TESTFACT, RMSD values were consistently greater for low discrimination items versus medium high discrimination items.

Table 4.13 presents the standard deviations of the RMSD values. In most conditions, the standard deviations of Mplus were slightly smaller than those for TESTFACT, except in data with a high correlation and lower discrimination condition. In addition, a significantly larger standard deviation in the RMSD statistic across replications was found in lower discrimination conditions as compared to higher discrimination conditions. For example, the standard deviations of the H and L conditions in data that assumed guessing in the low correlation condition were 0.006 versus 0.322 for TESTFACT and 0.006 versus .212 for Mplus. Note that there were significantly larger standard deviations found in unidimensional data with M and L discrimination conditions using TESTFACT (caused by a few extreme RMSD values).

Table 4.12

The mean RMSD of parameter recovery in Mplus and TESTFACT

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	0.02	0.03	0.29	0.06				
M	0.03	0.06	0.27	0.18				
L	0.04	0.10	0.31	0.23				
				Two-dimensional data				
HH	0.03	0.03	0.28	0.08	0.05	0.04	0.29	0.07
MH	0.03	0.04	0.27	0.09	0.05	0.04	0.28	0.09
MM	0.04	0.04	0.25	0.10	0.05	0.05	0.27	0.11
LH	0.04	0.04	0.26	0.11	0.07	0.06	0.29	0.11
LM	0.05	0.05	0.24	0.12	0.06	0.06	0.27	0.12
LL	0.05	0.06	0.23	0.13	0.06	0.07	0.27	0.14
				Three-dimensional data				
HHH	0.04	0.04	0.28	0.08	0.07	N/A	0.42	N/A
HHM	0.04	0.05	0.27	0.09	0.07	N/A	0.42	N/A
HHL	0.05	0.06	0.27	0.10	0.08	N/A	0.40	N/A
MMH	0.05	0.06	0.26	0.10	0.07	N/A	0.41	N/A
MMM	0.06	0.07	0.25	0.11	0.06	N/A	0.39	N/A
MML	0.06	0.08	0.25	0.12	0.08	N/A	0.37	N/A
LMH	0.06	0.07	0.26	0.12	0.08	N/A	0.39	N/A
LLH	0.07	0.08	0.27	0.12	0.09	N/A	0.38	N/A
LLM	0.07	0.09	0.25	0.13	0.09	N/A	0.36	N/A
LLL	0.07	0.09	0.25	0.14	0.09	N/A	0.35	N/A

Table 4.13

The standard deviation of the mean RMSD of parameter recovery in Mplus and TESTFACT

	r = .3				*r* = .6			
Disc.	c = 0		c = .33		c = 0		c = .33	
Condition	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	0.004	0.004	0.006	0.006				
M	0.005	0.187	0.006	0.330				
L	0.005	0.246	0.212	0.322				
				Two-dimensional data				
HH	0.003	0.003	0.006	0.009	0.004	0.004	0.006	0.008
MH	0.004	0.004	0.005	0.014	0.005	0.005	0.006	0.012
MM	0.005	0.005	0.005	0.009	0.007	0.006	0.008	0.014
LH	0.004	0.004	0.005	0.012	0.006	0.005	0.022	0.015
LM	0.007	0.005	0.006	0.012	0.009	0.007	0.019	0.015
LL	0.006	0.005	0.006	0.013	0.012	0.007	0.055	0.013
				Three-dimensional data				
HHH	0.004	0.004	0.007	0.009	0.006	N/A	0.038	N/A
HHM	0.005	0.004	0.007	0.010	0.007	N/A	0.024	N/A
HHL	0.005	0.005	0.007	0.011	0.009	N/A	0.025	N/A
MMH	0.005	0.004	0.007	0.012	0.008	N/A	0.022	N/A
MMM	0.005	0.004	0.006	0.016	0.007	N/A	0.022	N/A
MML	0.006	0.005	0.007	0.014	0.013	N/A	0.023	N/A
LMH	0.005	0.005	0.009	0.013	0.008	N/A	0.025	N/A
LLH	0.007	0.006	0.026	0.016	0.012	N/A	0.028	N/A
LLM	0.007	0.006	0.025	0.017	0.014	N/A	0.023	N/A
LLL	0.008	0.006	0.023	0.020	0.015	N/A	0.024	N/A

Figures 4.21 to 4.26 present the mean RMSD of parameter recovery across discrimination conditions in Mplus and in TESTFACT. These figures are arranged by the simulated dimensionality. As discussed above, the RMSD values for parameters in tests that assumed no guessing were relatively low compared to the data that assumed guessing in both Mplus and TESTFACT. Moreover, the RMSD of the data in the high correlation condition was slightly larger than the RMSD in data under the low correlation condition.

These figures also illustrated the different influence of discrimination on Mplus and TESTFACT. In Mplus, the trend of the data that assumed no guessing was opposite to the one for data that assumed guessing (see Figures 4.23 and 4.25). When the data assumed guessing, the largest RMSD value was found in the high discrimination condition. On the other hand, when data assumed no guessing, the largest value of the RMSD was found in the low discrimination condition. However, in TESTFACT, the patterns across all conditions were similar. The largest RMSD value was found in the lowest discrimination condition. Overall, TESTFACT performed better than Mplus with data that assumed guessing. There were significantly large RMSD values observed in two- and three-dimensional data that assumed guessing in Mplus. The values of RMSD were almost the same across one-to-three dimensional data in the low correlation condition. Furthermore, in two-dimensional data, there was no serious influence of factor correlations shown in the comparison of results in Mplus and TESTFACT. Finally, when data that assumed guessing and a high correlation condition, the values of RMSD in Mplus for three-dimensional data were significantly larger than the values found in two-dimensional data. For example, in HH discrimination condition, it was .29 versus .42 for two- and three-dimensional data. The difference, around .10, was almost 30% more than the value of .29.

In summary, TESTFACT did show superiority in parameter recovery in data that assumed guessing. The guessing effect was associated with increases in the RMSD values in Mplus when data that assumed guessing. A significant correlation effect was also found in three-dimensional data in Mplus. Due to problems of non-convergent solutions with TESTFACT and three-dimensional data, the correlation effect in three-dimensional data using TESTFACT could not be determined.

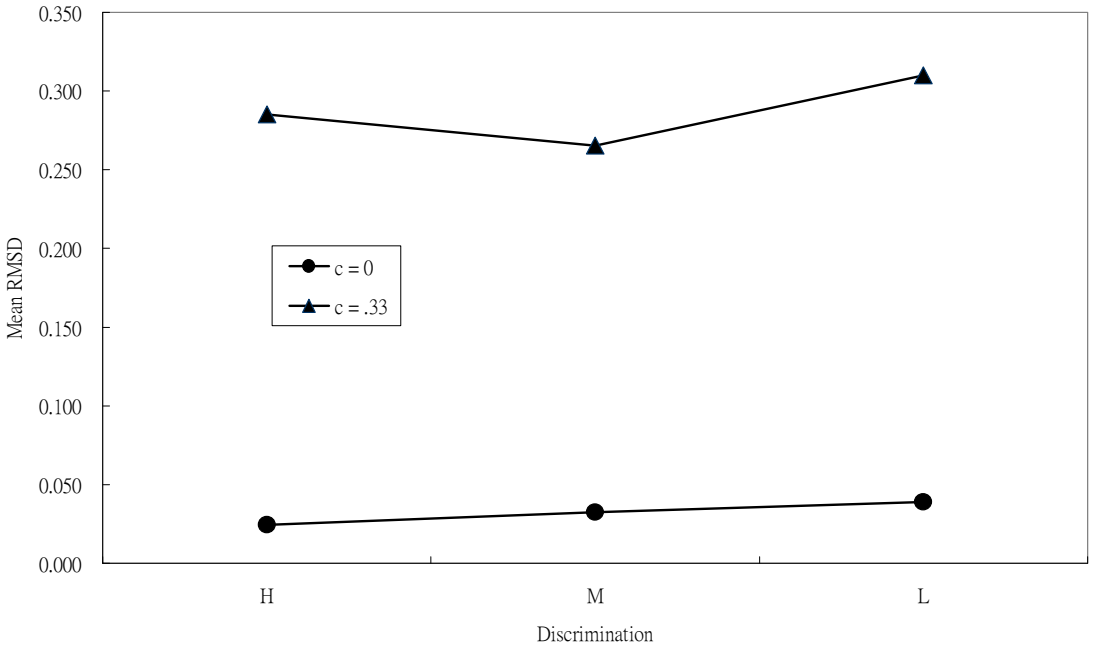


Figure 4.21 The mean RMSD of parameter recovery in unidimensional cases (Mplus)

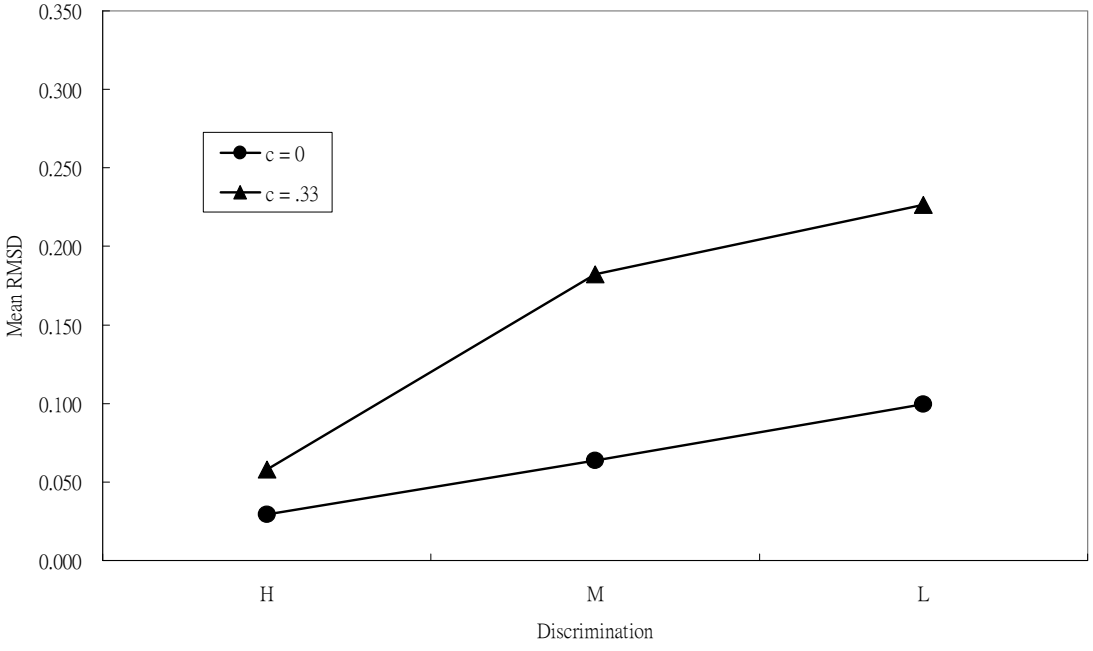


Figure 4.22 The mean RMSD of parameter recovery in Unidimensional cases (TESTFACT)

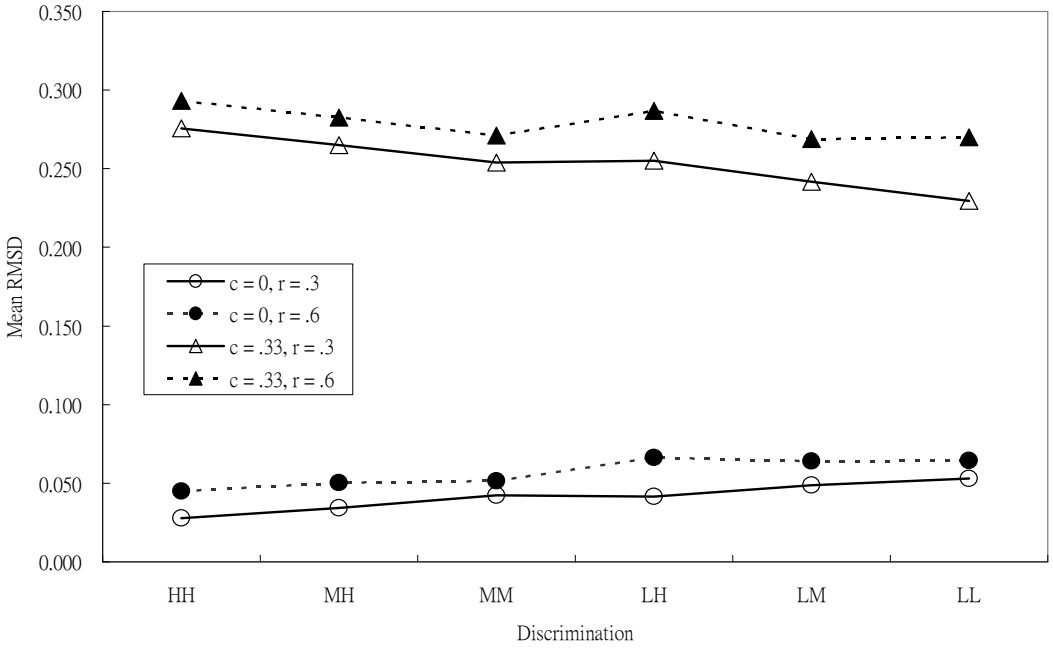


Figure 4.23 The mean RMSD of parameter recovery in 2-dimensional cases (Mplus)

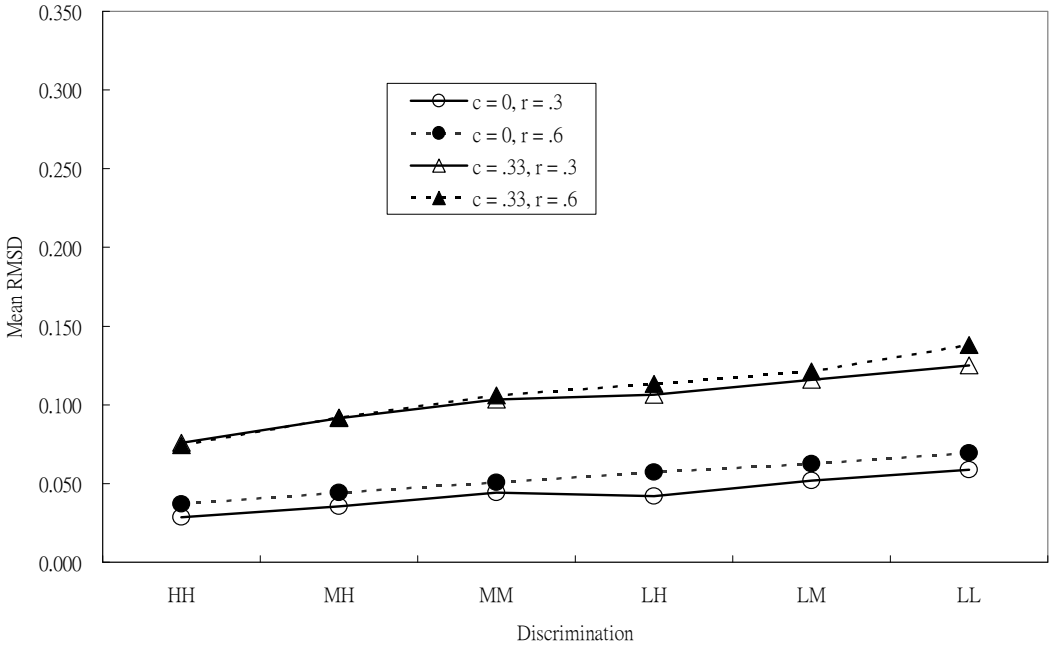


Figure 4.24 The mean RMSD of parameter recovery in 2-dimensional cases (TESTFACT)

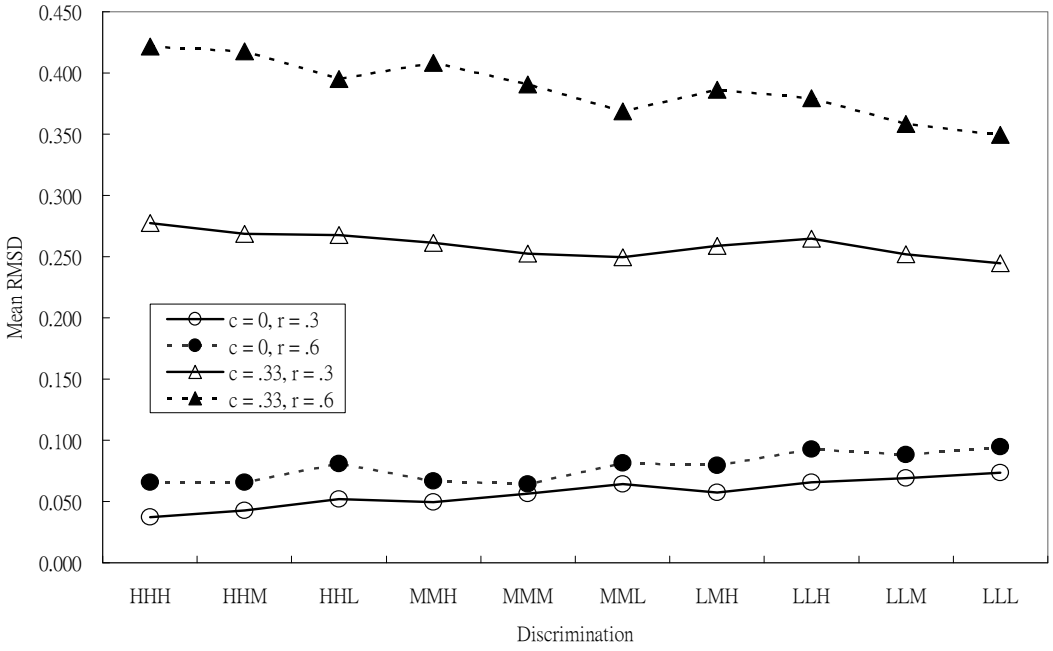


Figure 4.25 The mean RMSD of parameter recovery in 3-dimensional cases (Mplus)

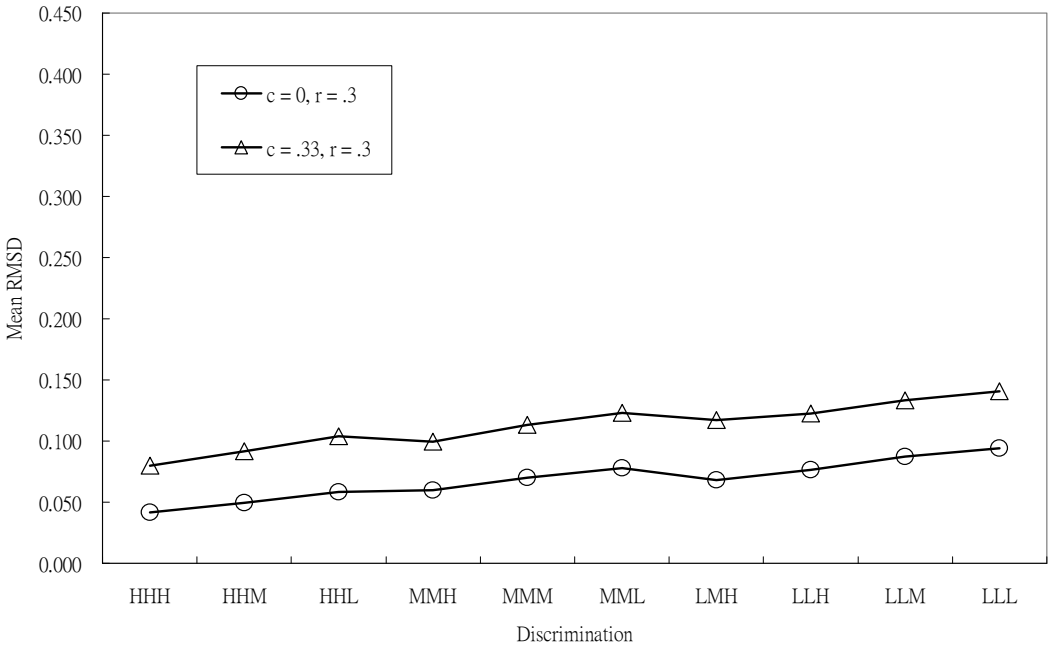


Figure 4.26 The mean RMSD of parameter recovery in 3-dimensional cases (TESTFACT)

4.2 THE RESULTS OF THE APPLICATION OF TIMSS

In this section, the assessment of the test dimensionality and factor structures of two booklets of multiple-choice items selected from the TIMSS assessment are presented. TIMSS data was analyzed to empirically compare Mplus and TESTFACT and investigate the guessing effect with real data. Basic descriptive statistics are first provided. Next the investigation of examinee guessing behavior is presented. Lastly, the results of the dimensionality assessment and the factor structure using Mplus and TESTFACT are compared.

4.2.1 A description of the TIMSS sample

The sample sizes for two booklets were in the range of 1500 to 1800 (N = 1554 for Booklet 5, N = 1847 for Booklet 11). Table 4.14 presents the descriptive statistics for these two booklets. The mean total scores showed that Booklet 5 was easier than Booklet 11 (33.56 vs. 30.64), and the standard deviation for the two tests was about the same (~ 11). The distributions of total scores for the two tests were slightly different in their skewness but similar in kurtosis. The test lengths for the two tests were about the same (55 vs. 54), but the number of mathematics and science items was different due to the block design. The number of mathematics items in Booklet 5 was almost twice the number of science items. In Booklet 11 the number of science items was six more than the number of mathematics items. Finally, the correlations between the mathematics and science subscores in two tests were similar (~ .75).

In order to apply the findings from the simulation study to TIMSS data, the discrimination level of the tests was examined since a discrimination effect was observed in the simulation study. Table 4.15 presents the descriptive statistics for item discrimination parameters and frequencies of low, medium, and high discrimination items. As in the simulation study, .75 and 1.25 were used as cut points to group items into three discrimination levels. For example, those items with discrimination values less than .75 were classified as low discrimination items, whereas those items with discrimination values above 1.25 were classified as high discrimination

items. As can be seen, the mean discrimination values for the two TIMSS booklets were about the same (~ 1.1). The mean discrimination values of three discrimination levels were similar as well. However, there were some small differences in the frequencies of items with these three discrimination levels. According to the statistics shown in Table 4.15, the discrimination level of these two booklets was considered as in the normal range and not extremely high or low. In relation to the simulation study, the frequency of items in different discrimination levels indicated that the may be classified in the LMH discrimination level.

In summary, these two tests were similar in sample size, test length, correlations between two subject domains, the distribution of the total scores and the mean discrimination values. However, the tests also exhibited some small differences in test difficulty, item constitution in the two subject domains and the frequency of low, medium, and high item discrimination values.

Table 4.14

Descriptive statistics of TIMSS

	n	Mean	*SD*	Skewness	Kurtosis	*r*
Booklet 5						
Total	55	33.56	11.60	-0.09	-0.99	
Math	36	21.74	8.38	-0.04	-1.07	.72
Science	19	11.83	4.02	-0.26	-0.82	
Booklet 11						
Total	54	30.64	11.16	0.03	-1.00	
Math	24	12.86	5.69	0.11	-1.00	.75
Science	30	17.77	6.25	-0.09	-0.90	

Table 4.15

Descriptive statistics of item discrimination parameters in TIMSS

	Total		Math		Science	
	n	Mean	*n*	Mean	*n*	Mean
Booklet 5						
Total	55	1.11	36	1.30	19	0.73
Low	15	0.57	5	0.56	10	0.58
Medium	20	0.96	12	1.04	8	0.85
High	20	1.65	19	1.67	1	1.27
Booklet 11						
Total	54	1.06	24	1.19	30	0.96
Low	12	0.58	2	0.59	10	0.58
Medium	27	1.01	14	1.03	13	1.00
High	15	1.53	8	1.61	7	1.43

4.2.2 Guessing and the TIMSS

Because the main purpose of this study was to investigate the guessing effect, it was necessary to examine the extent to which the subjects were guessing in the TIMSS assessment (Stone & Yeh, 2006). As recommended by Hambleton & Swaminathan (1985), the degree of potential guessing behavior for each item was evaluated by plots of the proportion correct by the total scores. If guessing behavior did not exist, the proportion correct should increase from 0 to 1 as the total score increased. If a moderately constant proportion correct (i.e., greater than 0) for low total scores was observed, guessing behavior might be assumed (Stone & Yeh, 2006). Figures 4.27 and 4.28 present plots of proportion correct by total scores for two items in each booklet. Item 22 in Booklet 5 (a four-choice item) and Item 24 in Booklet 11 (a five-choice item), exhibited constant proportions correct around .15 to .25 and .15 to .20, respectively, for total scores between 9 and 26. These relatively constant proportions of correct responses provide evidence that low ability examinees might be using guessing strategies. However, the proportions of correct responses were slightly less than expected values ($1/m$, m is the number of options) under a random guessing model.

On the other hand, Item 20 in Booklet 5 and Item 43 in Booklet 11 in Figures 4.27 and 4.28 illustrated the problem for assuming guessing behaviors in some items. These two items had no constant proportion correct for low total scores. Therefore, it was assumed that there was no guessing behavior operating. This circumstance might happen for very easy items (e.g. Item 20 with p = .79 and a = .51) or for low discriminating items (e.g., Item 43 with p = .71). These items illustrate Lord's criterion (1980), which indicates that an item response theory (IRT) model with a guessing parameter should not be estimated for items when $b - 2/a < -3.5$ (b is the item difficulty parameter and "a" the item discrimination parameter). If an item was very easy (b is negative), or its discrimination parameter (a) was small, the values of $b - 2/a$ would be very small.

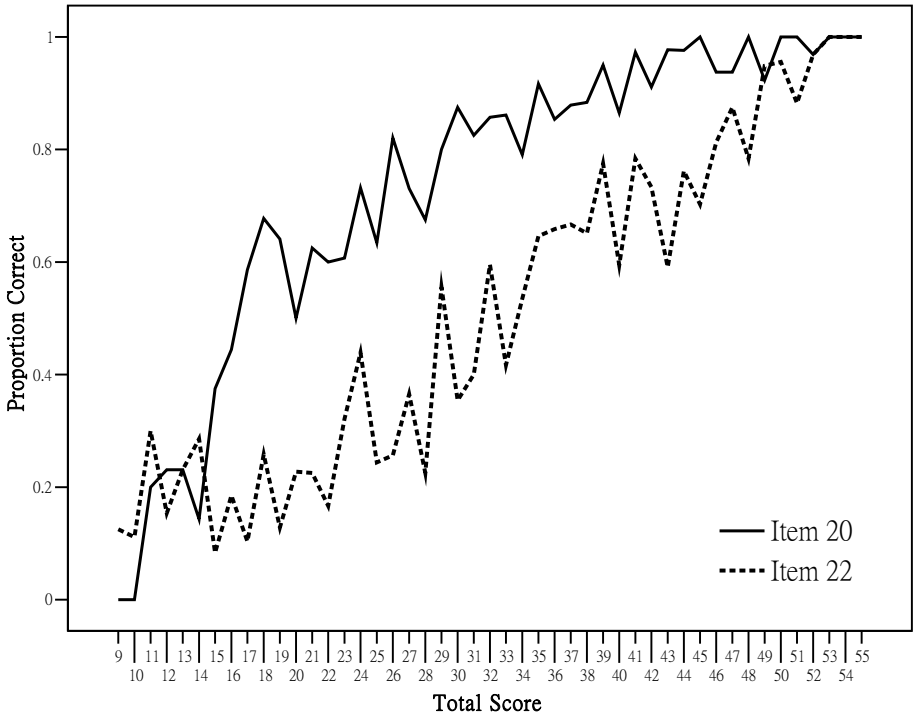


Figure 4.27 Total score by the proportion correct for two items in Booklet 5

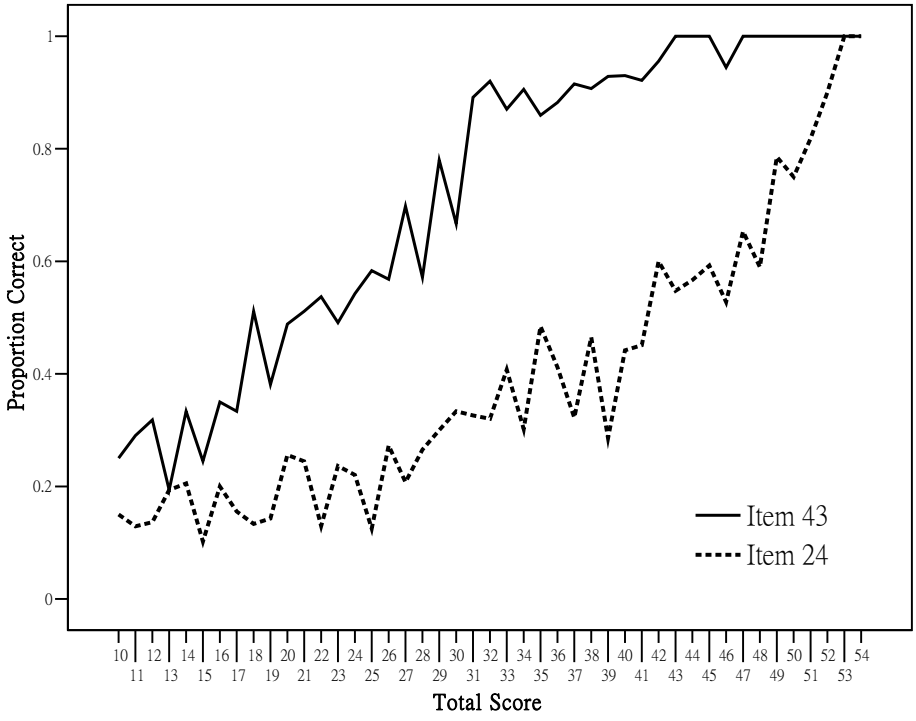


Figure 4.28 Total score by the proportion correct for two items in Booklet 11

Table 4.16 presents the frequency of the average proportion correct for low-ability examinees. Because the average difficulties of the two tests were different, the cutscores for low-ability examinees in two booklets were set to one standard deviation below the mean total scores (22 for Booklet 5 and 20 for Booklet 11). Since easier items could not be used to identify a constant proportion correct, only items with an item difficulty (*p*) less than .7 were included in Table 4.16. Note that most of the items in the two booklets used four-options but some used five-options. As can be seen, overall pattern in the frequency distributions for the two booklets was similar, and the highest frequency for both booklets was in the range of .2 to .3. The results matched expectations under the random guessing model. In addition, the mean of guessing or IRT based c-parameters for the items meeting Lord's criterion was .21 (*SD* = .12) for Booklet 5 and .22 (*SD* = .13) for Booklet 11. These results were also consistent with the random guessing model. In summary, more than 70% of items indicated that guessing was operating. It should be note that there were some items with a below or above average c-parameter (.25). The possible explanations, discussed by Stone and Yeh (2006), might be: (1) increased proportion correct may be caused by the elimination of distractors; (2) use of well-designed distractors may change the guessing strategy of examinees, that is, the random guessing model may not be appropriate.

Table 4.16

Average proportion correct (*p*) for low-ability examinees on items where $p \leqq .7$

p	Booklet 5	Booklet 11
< .1	2	2
.1 - .2	7	11
.2 - .3	20	21
.3 - .4	8	9
> .4	1	2
M	.24	.25
SD	.09	.09
n (items)	38	45
N (examinees)	316	407

4.2.3 The dimensionality of the TIMSS

Based on the assessment design of TIMSS 2003 the assessment consisted of two subject domains, mathematics and science. In addition, the correlations between mathematics and science subscores for these two booklets were high (~ .75, see Table 4.14 in the previous section). Hence, one dominant factor or a two-factor solution might be expected from a factor analysis.

Table 4.17 presents the results of the four indices for estimating the number of factors from an exploratory factor analysis of both booklets using Mplus and TESTFACT. As can be seen in Table 4.17, more consistency among dimensionality decisions using the different indices was found in TESTFACT than in Mplus. The decision based on parallel analysis and the chi-square test indices tended to conclude higher dimensionality in both booklets using Mplus, whereas consistently smaller number of factors were concluded using TESTFACT. This is consistent with the findings shown in the simulation study. Overall, a two- or three-factor structure might be concluded for both booklets. Note that for TESTFACT, factor solutions beyond two factors did not converge and so the RMSR reduction index and the chi-square test could not be used to evaluate higher-order factor models.

In order to consider the results from the simulation study in relation to the TIMSS application, it was necessary to match the conditions of the TIMSS assessment with the conditions evaluated in the simulation study. Given the high correlations between the two subject domains, the level of observed examinee guessing behavior, and the level of item discriminations, the TIMSS assessment appeared to best conform to the following condition from the simulation study: a multi-dimensional test (2 or 3 dimensions) with correlations between dimensions equal to .6 and average item discriminations (LMH). Based on the findings from the simulation study, no index performed well with two-dimensional data using Mplus, whereas most indices except the RMSR index performed fairly well with two-dimensional data using TESTFACT. In other words, the dimensionality decisions based on the results of TESTFACT might be reliable given two-dimensional data. The situation becomes more complex to assess for higher dimensionality (> 2 dimensions). Under this condition, both MPlus and TESTFACT did not reliably estimate true dimensionality. If TESTFACT is assumed to be

more reliable given the observed level of guessing behavior, the results presented in Table 4.17 indicate a two-factor structure in both booklets.

Table 4.17

Estimated dimensionality using the four indices in Mplus and TESTFACT

Form	Method	Proportion of Variance	Parallel Analysis	Reduction of RMSR	Chi-sq. Test
Booklet 5	Mplus	2	5	2	2
	TESTFACT	2	2	2	2
Booklet 11	Mplus	1	4	3	5
	TESTFACT	2	2	2	2

In addition to assessing dimensionality based on the four indices, another common method for determining dimensionality is through the examination of scree plots (plots of eigenvalues). Figures 4.29 and 4.30 present scree plots for both booklets using Mplus and TESTFACT. It was clear that the plots indicated no more than two factors, which is consistent with the TESTFACT analysis.

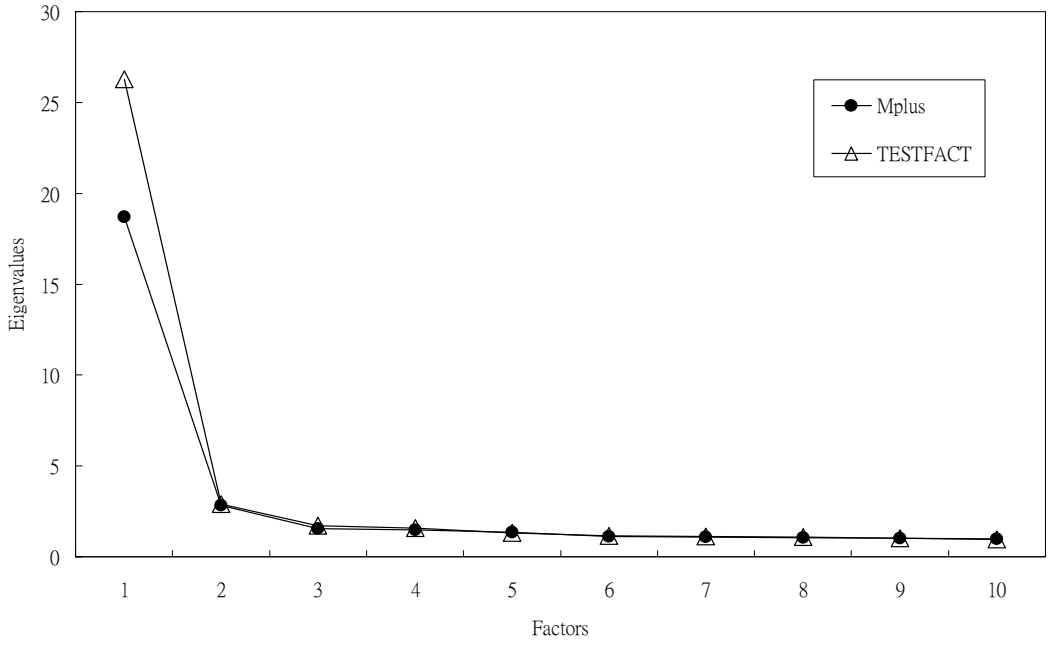


Figure 4.29 The scree plot for Booklet 5 using Mplus and TESTFACT

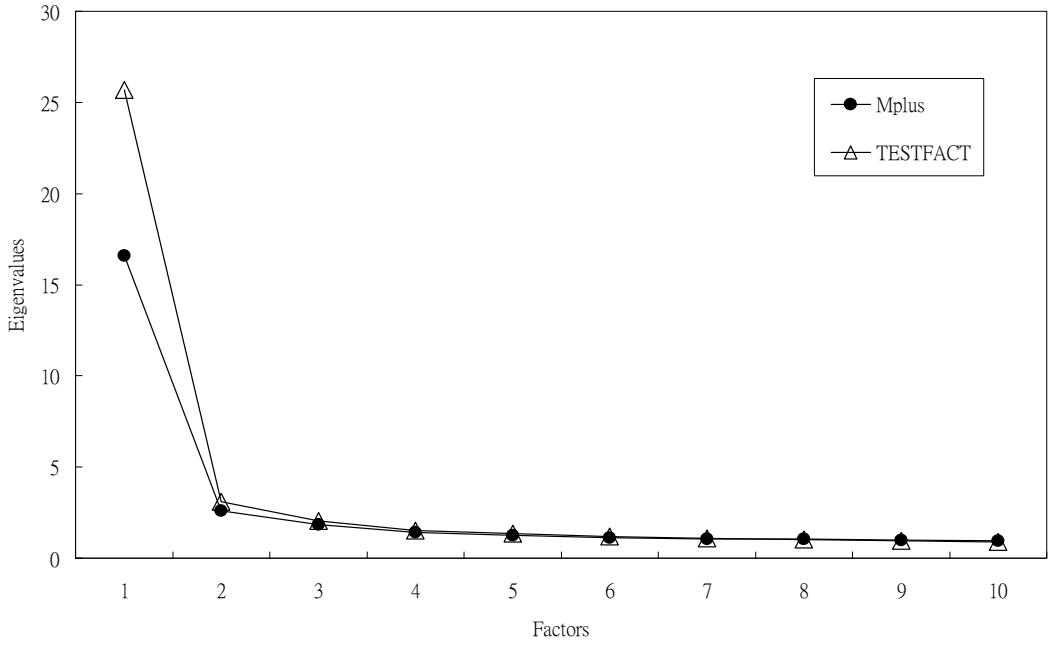


Figure 4.30 The scree plot for Booklet 11 using Mplus and TESTFACT

Because the number of dimensions obtained by Mplus and TESTFACT more commonly reflected no more than three-factor structures, two- and three-factor solutions were further explored. Tables 4.18 and 4.19 present summary information for two- and three-factor solutions. The information included the number of substantial factor loadings (i.e., loadings > .3) and the correlations among factors for these factor solutions. Note that a maximum of 200 iterations was specified in TESTFACT in order to obtain convergent solutions. As can be seen in Table 4.18, the patterns of substantial factor loadings in Mplus and TESTFACT were similar, but more indicators of factors were found in the results of TESTFACT. Moreover, the correlations of Mplus were lower than those of TESTFACT. The correlations of TESTFACT were close to the values calculated by two subscores of mathematics and science items.

Table 4.19 presents the results for the three-factor solutions. Similar to the two-factor solutions, the correlations estimated by TESTFACT were greater than the values estimated by Mplus. There were also more indicators shown in the solutions of TESTFACT. The factor correlations for the three-factor solutions in both methods indicated higher-order factor structures. As suggested by Stone and Yeh (2006), given a moderately high strength in the correlations among factors, it might be useful to evaluate the existence of a second-order factor. However, a

minimum of four first-order factors was required for the evaluation, whereas there were only two or three first-order factors in this case.

Table 4.18

Number of substantial factor loadings for the two-factor solutions in Mplus and TESTFACT

	Mplus		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
Booklet 5				
Math	26	9	30	11
Science	1	17	2	18
Factor correlations				
Factor 2	.65		.76	
Booklet 11				
Math	20	5	23	1(1)
Science	9(1)[a]	20	10(2)	25
Factor correlations				
Factor 2	.63		.76	

[a] the number shown in the parentheses was the number of negative substantial factor loadings.

Table 4.19

Number of substantial factor loadings for the three-factor solution in Mplus and TESTFACT

	Mplus			TESTFACT		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Booklet 5						
Math	25	8	10	23	4	17
Science	1	16	0	1	18	2
Factor correlations						
Factor 2	.63			.68		
Factor 3	.44	.45		.69	.73	
Booklet 11						
Math	18	0	9	21	1(3)	5(1)
Science	5(1)[a]	19	4	4(1)	20	15
Factor correlations						
Factor 2	.60			.69		
Factor 3	.48	.51		.60	.53	

[a] the number shown in the parentheses was the number of negative substantial factor loadings.

In order to further explore the internal structures for the two booklets, more details of the substantial factor loadings for items are presented in Tables 4.20 to 4.23. Mathematics items are indicated by item codes beginning with "M", whereas science items are indicated by item codes beginning with "S". Note the significant factor loadings were estimated using the PROMAX method in two- and three-dimensional data, given the moderately high correlations among factors. Also this analysis was limited since only some items in both booklets have been released to the public and partial information about the items was known (e.g., the content domain and the cognitive domain defined by the TIMSS research groups).

For Booklet 5, there were more items in mathematics than in science (36 vs. 19). In the two-factor solution that appears in Table 4.20, Factor 1 was clearly represented by only mathematics items. However, a combination of mathematics and most of the science items was found for Factor 2. The patterns in Mplus and TESTFACT were similar. Note that most mathematic items with significant factor loadings shown in Factor 2 were easy items. However, the structure of the three-factor solution (see Table 4.21) did not provide a further separation of mathematics and science items. The factor pattern of the three-factor solution was not related to either the content domain or the cognitive domain but to higher *p* values or low c-parameters. There was a slight difference between the results of Mplus and TESTFACT in the three-factor solutions.

As for Booklet 11, the patterns shown in the results of Mplus and TESTFACT were quite different. For two-factor solutions (see Table 4.22), the mathematic and the science items were merged into one factor, Factor 1, in Mplus and TESTFACT. Factor 2 was more likely to be the science factor in TESTFACT, whereas there were more mathematic items merged into Factor 2 in Mplus. Interestingly, the *p* values for most items (either math or science) of Factor 2 using Mplus were relatively high (more than .7). In the three-factor solution in Table 4.23, the result for Mplus illustrated a simple structure compared to the result for TESTFACT. In Mplus, Factor 2 represented a science factor, whereas Factor 1 and Factor 3 contained primarily the mathematics items, but included some science items as well. In TESTFACT, Factor 1 represented mathematics whereas Factor 2 and Factor 3 represented primarily the science domain, but there were a few mathematics items merged into Factor 2 and Factor 3.

Note that the factor loadings in TESTFACT were relatively higher than those in Mplus, especially in the two-factor solution. For example, the factor loadings of Factor 1 in TESTFACT

of the two-factor solution were higher than those in Mplus (the differences were around .2). Also, the values greater than one or negative values were found in the results of TESTFACT. Since the possibility for the presence of a difficulty factor existed, the detailed information about estimated c-parameters and item difficulty (*p* values) for both booklets are provided in Appendix F and Appendix G.

In summary, the TIMSS assessment appeared consistent with a multi-factor test with significant observed guessing behaviors and significant correlations between the factors. Based on TESTFACT, the various indices for determining the number of dimensions in the two TIMSS booklets indicated the presence of at least two dimensions. The inconsistency in the determinations for the four indices with Mplus is consistent with the results of the simulation study. Further exploration of the internal structure did not clarify the internal structure of the item responses to the TIMSS booklets. For one booklet (Booklet 5), the two-factor structures were very similar for Mplus and TESTFACT, whereas some differences were observed for the three-factor solution. For the other booklet (Booklet 11), results were very different for Mplus and TESTFACT. In addition, more substantial factor loadings were found using TESTFACT, and the presence of a difficulty factor was observed using both Mplus and TESTFACT. The differences between the results may be due to the presence of guessing behavior. However, given the performance of Mplus versus TESTFACT under the high correlation condition in the simulation study, more confidence may be placed in the TESTFACT results. Further investigation of the factor solutions did not provide a clearer understanding of the internal test structure of TIMSS.

Table 4.20

Factor loadings of the PROMAX two-factor solution for Mplus and TESTFACT (Booklet 5)

Item ID	Mplus		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
M022043			0.382	
M022049		0.388		0.361
M022050	0.465		0.579	
M022057		0.473		0.445
M022062	0.575		0.671	
M022066	0.888		1.055	
M022097		0.579	0.435	0.390
M022101		0.679		0.611
M022104		0.558		0.421
M022105		0.461		0.483
M022108	0.426		0.550	
M022181		0.477	0.365	0.460
M022257	0.705		0.887	
M032044	0.565		0.760	
M032046	0.772		0.916	
M032079	0.453		0.618	
M032228	0.589		0.753	
M032261	0.511		0.631	
M032271	0.468		0.532	0.341
M032489		0.515		0.436
M032523	0.756		0.921	
M032525	0.583		0.705	
M032533	0.691		0.855	
M032579	0.549		0.714	
M032588	0.350		0.423	0.335
M032678	0.656		0.706	
M032701	0.408		0.719	
M032704	0.638		0.932	
MF32036	0.679		0.901	
MF32447	0.614		0.775	
MF32609	0.457		0.637	
MF32670		0.620	0.352	0.428
MF32690	0.465		0.664	
MF32727	0.639		0.734	
MF32728	0.405		0.678	
MF32732	0.406		0.552	

Table 4.20 (continued)

Item ID	Mplus		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
S032024		0.305		0.533
S032141		0.316		0.781
S032315		0.581		0.591
S032463		0.575		0.704
S032465		0.464		0.524
S032514		0.299		0.428
S032579	0.405		0.538	0.296
SF12001		0.649		0.624
SF12002		0.588		0.559
SF12003		0.790		0.891
SF12004		0.510		0.578
SF12005		0.618		0.743
SF12006		0.402		0.895
SF12013		0.419		0.505
SF12014		0.584		0.558
SF12015		0.501		0.492
SF12016		0.392		0.778
SF12017		0.645		0.709
SF12018			0.341	

Table 4.21

Factor loadings of the PROMAX three-factor solution for Mplus and TESTFACT (Booklet 5)

Item ID	Mplus			TESTFACT		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
M022043			0.302			
M022049						0.334
M022050	0.497			0.587		
M022057		0.394				0.310
M022062	0.392		0.517			0.724
M022066	0.872			0.960		
M022097		0.437	0.390			0.435
M022101		0.603			0.527	
M022104		0.508			0.421	
M022105		0.336	0.361			0.502
M022108			0.502			0.723
M022181		0.427			0.412	
M022257	0.617			0.588		0.431
M032044	0.538			0.480		0.425
M032046	0.748			0.743		
M032079	0.382			0.343		0.436
M032228	0.534			0.555		
M032261	0.487			0.471		
M032271	0.409			0.386		
M032489		0.380	0.380			0.507
M032523	0.717			0.820		
M032525	0.621			0.745		
M032533	0.622			0.591		0.387
M032579	0.392		0.435	0.316		0.664
M032588	0.348			0.366	0.333	
M032678	0.543		0.325	0.373		0.544
M032701	0.441			0.751		
M032704	0.626			0.821		
MF32036	0.639			0.727		
MF32447	0.539			0.458		0.478
MF32609	0.451			0.587		
MF32670		0.486	0.366			0.455
MF32690	0.445			0.482		
MF32727	0.526		0.315	0.381		0.590
MF32728	0.402			0.512		
MF32732	0.351					0.411

Table 4.21 (continued)

Item ID	Mplus			TESTFACT		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
S032024					0.365	0.374
S032141		0.307			0.636	
S032315		0.584			0.604	
S032463		0.547			0.640	
S032465		0.470			0.578	
S032514		0.310			0.409	
S032579	0.441			0.568	0.353	
SF12001		0.659			0.613	
SF12002		0.558			0.546	
SF12003		0.734			0.764	
SF12004		0.571			0.602	
SF12005		0.644			0.728	
SF12006		0.421			0.892	
SF12013		0.393			0.467	
SF12014		0.564			0.494	
SF12015		0.394			0.336	0.363
SF12016		0.451			0.798	
SF12017		0.599			0.597	
SF12018						

Table 4.22

Factor loadings of the PROMAX two-factor solution for Mplus and TESTFACT (Booklet 11)

Item ID	Mplus		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
M032094	0.699		0.958	
M032100		0.616		0.557
M032116	0.622		0.970	
M032132		0.370	0.296	
M032324	0.600		0.715	
M032397	0.427		0.568	
M032402	0.561		0.914	
M032419	0.410		0.609	
M032477	0.542		0.630	
M032662	0.594		0.650	
MF12013	0.537		0.766	
MF12014		0.425	0.476	
MF12015	0.468	0.347	0.591	
MF12016	0.452		0.786	
MF12017	0.465		0.595	
MF22185	0.650		1.073	-0.322
MF22188	0.332		0.512	
MF22189		0.455	0.371	
MF22191	0.440		0.636	
MF22194	0.422		0.608	
MF22196	0.634		0.885	
MF22198	0.620		0.872	
MF22199	0.814		1.084	
MF22251	0.569		0.916	

Table 4.22 (continued)

Item ID	Mplus		TESTFACT	
	Factor 1	Factor 2	Factor 1	Factor 2
S022002	0.321	0.335	0.398	0.300
S022019		0.392		0.590
S022042	0.410	0.377	0.546	
S022054				0.534
S022106		0.428	-0.334	0.565
S022115	0.340		0.411	
S022126	0.387		0.520	
S022150				0.739
S022181		0.370		0.549
S022183	0.418		0.493	0.397
S022208	0.460		0.617	
S022276	0.314	0.320	0.442	0.401
S022290		0.658		0.664
S022294		0.334		
S032008		0.538		0.530
S032035		0.567		0.552
S032055		0.575		0.609
S032083		0.332		0.592
S032150	-0.431	0.967	-0.691	1.242
S032258		0.465		0.401
S032281		0.475		0.665
S032301				0.606
S032385		0.354		0.532
S032446		0.355	0.324	0.416
S032564				0.612
S032607		0.545		0.626
S032683	0.355		0.491	0.338
SF32422		0.413		0.560
SF32574	0.387		0.528	0.393
SF32714		0.547		0.881

Table 4.23

Factor loadings of the PROMAX three-factor solution for Mplus and TESTFACT (Booklet 11)

Item ID	Mplus			TESTFACT		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
M032094	0.630			0.780		
M032100			0.543		0.571	-0.343
M032116	0.576			0.700		0.395
M032132			0.434	0.428		
M032324	0.516			0.713		
M032397	0.338			0.568		
M032402	0.525			0.732		0.350
M032419	0.380			0.561		
M032477	0.440		0.320	0.668		
M032662	0.545			0.796		
MF12013	0.453			0.664		
MF12014			0.562	0.590		
MF12015	0.332		0.486	0.646		
MF12016	0.458			0.652		
MF12017	0.351		0.416	0.685		
MF22185	0.625			0.772	-0.396	0.405
MF22188				0.526		
MF22189			0.609	0.545		
MF22191			0.564	0.655		
MF22194	0.341			0.572		
MF22196	0.525		0.351	0.765		
MF22198	0.551			0.749		
MF22199	0.788			0.844	-0.327	0.372
MF22251	0.525			0.644	-0.327	0.372

Table 4.23 (continued)

Item ID	Mplus			TESTFACT		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
S022002		0.305				0.443
S022019		0.403			0.413	0.469
S022042			0.500	0.566		
S022054		0.310			0.385	0.323
S022106		0.471			0.496	
S022115						
S022126	0.388					0.687
S022150		0.433			0.533	0.453
S022181		0.446			0.398	0.375
S022183	0.428	0.370				0.580
S022208	0.444					0.600
S022276				0.388		
S022290		0.553			0.567	
S022294						0.313
S032008		0.431			0.454	
S032035		0.411	0.310		0.458	
S032055		0.401	0.344		0.462	0.341
S032083		0.416			0.466	
S032150	-0.488	0.870		-0.463	1.088	
S032258			0.362		0.331	
S032281		0.439			0.501	0.349
S032301		0.318			0.465	
S032385		0.444			0.366	0.430
S032446					0.325	
S032564				0.360	0.487	
S032607		0.498			0.508	
S032683	0.324					0.432
SF32422		0.415			0.398	0.400
SF32574	0.363			0.391		0.333
SF32714		0.473			0.760	

5.0 DISCUSSION

5.1 MAJOR FINDINGS OF THE SIMULATION STUDY

The primary purpose of this study was to explore the effect of guessing in determining test dimensionality for multiple-choice tests. The influence of guessing was investigated by comparing the dimensionality decisions and parameter recovery between two factor analysis methods (Mplus and TESTFACT), with data that assumed guessing and not guessing, using the Monte Carlo approach. As will be discussed, some findings observed in this study were consistent with findings from previous studies (i.e., Stone & Yeh, 2006; Tate, 2003). However, this study provided more understanding of issues when assessing dimensionality with multiple-choice assessments. This study not only examined the influence of guessing but also considered factors likely to affect the assessment of dimensionality, such as the level of item discrimination, correlations among dimensions and the number of dimensions.

The second purpose of this study was to see whether what was learned from the simulation study could be applied to real data. Therefore, the dimensionality and factor structure of real data (TIMSS 2003) were also examined. The major findings corresponding to the research questions addressed in the simulation study are presented below. The findings of the application of real data are then presented.

Research Question 1*: What is the effect of guessing on assessing dimensionality of multiple-choice tests?*

The influence of guessing was demonstrated by comparing the results of Mplus and TESTFACT based on the proportion of correct dimensionality decisions, the estimated number of dimensions, and parameter recovery. The results of the correct dimensionality decisions showed that Mplus only performed better in data that assumed no guessing using most indices, whereas TESTFACT performed very well in data that assumed both guessing and no guessing.

For example, in two-dimensional data that assumed no guessing, there was a correct rate of 1.0 observed in both Mplus and TESTFACT. However, when using data that assumed guessing, a significantly smaller correct rate (i.e., 0) was observed in Mplus but not in TESTFACT (e.g., above 50%). With regard to the estimated number of dimensions, TESTFACT generally outperformed Mplus with data that assumed guessing, although underestimation of dimensionality occurred in three-dimensional data with a high correlation condition (r = .6). In Mplus, greater degree of either overestimation or underestimation was observed with data that assumed guessing. However, in TESTFACT, better performance in determining the estimated dimensionality was obtained with data that assumed guessing, even though slight underestimation sometimes occurred. The results of parameter recovery also illustrated the superiority of TESTFACT. Mean RMSD values for TESTFACT in data that assumed guessing were significantly smaller than that those for Mplus, but the RMSD values for both methods were similar in data that assumed no guessing. Finally it should be noted that although results for TESTFACT with three-dimensional data and a high correlation conditions (r = .6) were poor, the results are likely due to the number of iterations that was specified in the study. This issue is discussed in more detail under the Limitations section.

In summary, TESTFACT did show superiority in data that assumed guessing for dimensionality decisions and parameter recovery. The overall trends shown in the results of Mplus and TESTFACT were consistent with previous studies (i.e., Knol & Berger, 1991; Tate, 2003; Stone & Yeh, 2006). The findings substantitated the importance of modeling guessing in the factor analysis method when guessing behavior is found in testing applications. Note that when no guessing behavior was operating in the examinations, either Mplus or TESTFACT provided similar dimensionality assessment. However, Mplus may be preferred because of its ease of use, understandability, and efficiency. Another reason for recommending the use of Mplus is that it provides more diagnostics in detecting the factor structure and more fit indices for assessing model fit (Stone & Yeh, 2006).

Research Question 2*: How well do different indices perform for estimating the number of dimensions when assessing dimensionality?*

With regard to the performance of the four indices for determining dimensionality, differences in performance between the four indices were observed in terms of the proportion of correct dimensionality decisions and the difference between the true and estimated number of

dimensions. Based on the results of correct dimensionality decisions, when using Mplus in data that assumed no guessing, the proportion of variance and the RMSR reduction indices worked fairly well in one- and two-dimensional data (i.e., the correct rate was close to 1.0), whereas the parallel analysis and the chi-square test performed poorly in most conditions (e.g., % of correct decisions was close to 0 when using the chi-square test). Note that only the RMSR reduction index performed very well in higher dimensional data. For data that assumed guessing, only the proportion of variance and the RMSR reduction indices in lower dimensionality conditions were effective with Mplus. In contrast, when using TESTFACT with one- and two-dimensional data, the four indices performed fairly well; that is, a correct decision proportion above 70% was observed for most indices with data that assumed either guessing or no guessing. Underestimation of dimensionality was found in three-dimensional data that assumed guessing using most indices except the chi-square test.

In summary, the proportion of variance and the RMSR reduction indices worked better for Mplus, whereas the chi-square test performed the best when using TESTFACT with data either assumed guessing or no guessing. Different indices interacted with the correlation and the discrimination effects in different ways. More detail is discussed below. This study confirmed the effectiveness of the RMSR reduction index in Mplus and the chi-square test in TESTFACT found by Tate (2003).

***Research Question 3**: Does the discrimination level for items affect the assessment of dimensionality?*

A discrimination effect was observed in tests with either higher or lower discriminating items. Based on the correct dimensionality decisions, the discrimination effect was not detected in data that assumed no guessing for most indices. However, this effect was observed in data that assumed guessing. Given data that assumed guessing, a similar effect was observed in Mplus with the low correlation condition and in TESTFACT with the high correlation condition, when using the proportion of variance and the RMSR reduction indices. For both Mplus and TESTFACT, a lower correct decision rate was observed in tests with low discriminating items. Considering the comparison of the estimated number of dimensions determined by the four indices, the discrimination effect was observed in both Mplus and TESTFACT for all indices except the chi-square test. For data that assumed no guessing, the influence of item discrimination was observed with low discriminating items using parallel analysis in both Mplus

and TESTFACT. For data that assumed guessing, overestimation of dimensionality was observed with high discriminating items and underestimation with low discriminating items when using Mplus and the RMSR reduction index. With TESTFACT, a modest discrimination effect was observed with data that assumed guessing. For example, the degree of underestimation increased when the discrimination level decreased. This occurred in two conditions when using the proportion of variance and the RMSR reduction indices: 1) for two-dimensional data and the high correlation condition; 2) for three-dimensional data and the low correlation condition. Based on the parameter recovery results, a small discrimination effect was observed in data that assumed guessing and no guessing for both methods. Mean RMSD values increased when the level of discrimination parameters decreased in TESTFACT, whereas the values of mean RMSD decreased when discrimination parameters of items decreased in Mplus.

In summary, the discrimination effect can be described as follows: 1) tests with lower discriminating items were associated with decreased correct dimensionality decisions; 2) either overestimation of dimensionality was found with higher discriminating items or underestimation in dimensionality was observed with lower discriminating items; 3) higher factor correlations or dimensionality increased the influence of item discrimination. Considering which method was used, the discrimination effect was more pronounced in Mplus with data that assumed both guessing and no guessing, and in TESTFACT with data that assumed no guessing. Less impact was found when using TESTFACT in data that assumed guessing. With regard to which index may be affected by the discrimination level of items, the RMSR reduction index should be used with caution in data that assumed guessing (either high or low discrimination conditions), as well as parallel analysis in data that assumed no guessing in low discrimination conditions. The present study also confirmed the influence of item discrimination found by Tate (2003). However, item discrimination conditions were varied in this study, whereas only a high discrimination condition was studied by Tate (2003). Therefore, a greater understanding of the item discrimination effect was possible in this study.

Research Question 4*: Does the correlation among dimensions affect the assessment of dimensionality?*

The correlation effect was evaluated by comparing results for the more effective indices under the different correlation conditions. In general, the performances for determining test dimensionality deteriorated under the high correlation condition with both unmodeled and

modeled guessing. For example, a lower correct decision rate was observed in the high correlation condition, when using both methods in data that assumed guessing with the proportion of variance index. According to the results of the estimated number of dimensions, the correlation effect led to overestimation in Mplus and underestimation in TESTFACT. When using TESTFACT, the most serious underestimation was found in data that assumed guessing with the high correlation condition. Only a modest correlation effect was observed in the parameter recovery results when compared to the other outcome measures (i.e., correct dimensionality decisions, and the estimated number of dimensions). The correlation effect was found in data that assumed guessing when using Mplus, especially in three-dimensional data. In addition, a greater impact of this effect occurred in data that assumed guessing for most conditions. However, there were some exceptions. For example, a correlation effect was observed in two-dimensional data that assumed no guessing when using both methods and the parallel analysis index. Note that this effect was opposed to the effect mentioned above, that is, a greater proportion of correct dimensionality decisions was observed in the higher correlation condition. Finally, the correlation effect appeared to be dependent on the level of test item discrimination. For example, with data that assumed guessing, there was a tendency for underestimation of dimensionality in tests with low discriminating items. In other words, the difference between the true and the estimated number of dimensionality was greater in low discrimination conditions.

When comparing the correlation effect in both methods, the effect observed in Mplus and TESTFACT was different. Differences in the proportion of correct dimensionality decisions, the degree of either overestimation or underestimation of dimensionality, and the RMSD values were observed. For example, for two-dimensional data and a guessing rate of .33, TESTFACT estimated the true dimensionality better than Mplus based on the more effective indices. Additionally, the effectiveness of the index used to evaluate dimensionality was affected by the correlation condition. For instance, when using Mplus in data that assumed guessing, better performance was observed in the low correlation condition using the proportion of variance and the RMSR reduction indices, whereas better performance was observed in the high correlation using the chi-square test.

In summary, an overall effect of the correlation between dimensions was observed. An increase in the correlation was associated with: 1) a decrease in the proportion of correct

dimensionality decisions; 2) an increase in the difference between the estimated number of dimensions and the true dimensionality (i.e., the discrimination effect); 3) an increase in RMSD values. When comparing Mplus and TESTFACT, the correlation effect was more pronounced in data that assumed guessing when using Mplus, whereas the effect was similar when using Mplus and TESTFACT in data that assumed no guessing. With regard to the four indices, the correlation effect was observed more frequently when using the proportion of variance and parallel analysis in data that assumed guessing and no guessing, whereas the effect was observed when using the RMSR and the chi-square test only with data that assumed guessing. The correlation effect was more evident with data that assumed guessing and when the number of dimensions increased. The correlation effect was also primarily found in two-dimensional data. However, the correlation effect could not be adequately evaluated in data with higher dimensions since unexpected results were observed in Mplus (e.g., the performance of the chi-square test) and the results for TESTFACT were limited for the high correlation condition. Finally, with regard to previous studies, Tate (2003) evaluated dimensionality decisions under only a correlation condition of .6 and a smaller guessing rate of .2. For this limited case, similar results were observed in the present study.

Research Question 5*: What is the interaction between the guessing effect and the level of discrimination of items and correlations between dimensions?*

The interaction between the guessing and discrimination effects was examined by comparing the discrepancy of the results for the more effective indices between Mplus and TESTFACT with data that assumed guessing under different discrimination conditions. No clear discrimination effect was observed under the no guessing condition (i.e., c = 0). Generally, the guessing effect increased the effect due to the level of test item discrimination. In other words, modeling guessing with data that assumed guessing may decrease the discrimination effect. For example, when using Mplus in data that assumed guessing, a low or zero proportion of correct decisions was observed in low discrimination conditions. This same effect was not observed with TESTFACT. However, when the correlation between factors increased or in higher dimensional data, a discrimination effect was observed when using TESTFACT. For example, when using TESTFACT with two-dimensional data and a high correlation condition, the proportion of correct decisions decreased when the discrimination level decreased. It should be noted that the interaction was observed in most indices except the chi-square test and the pattern

of these indices was different. For example, when using the RMSR reduction index in three-dimensional data with Mplus, the discrimination effect was found for both high and low item discrimination conditions, but the discrimination effect was observed only in low discrimination conditions when using the proportion of variance index. The interaction was also observed in the results based on comparing the estimated number of dimensions with true dimensionality. In addition, the discrimination effect varied depending on the index used to evaluate dimensionality when using both Mplus and TESTFACT in data that assumed guessing. For example, with data that assumed guessing, the discrimination effect was not observed in Mplus and TESTFACT when using parallel analysis (i.e., underestimation in Mplus, and perfectly matched true dimensionality in TESTFACT), whereas the discrimination effect was observed in both methods when using the RMSR reduction index. Finally, based on the parameter recovery results, mean RMSD values in Mplus and TESTFACT changed in opposite directions with data that assumed guessing. Smaller RMSD values were observed in low discrimination conditions when using Mplus, whereas larger RMSD values were found in low discrimination conditions when using TESTFACT. Note that the results of Mplus and TESTFACT were similar with data that assumed no guessing.

With regard to the interaction between the guessing and correlation effects, the pattern in results was similar to that found for the interaction between the guessing and discrimination effects. When comparing results for Mplus and TESTFACT, the magnitude of the correlation effect was greater in Mplus than in TESTFACT. For example, in two-dimensional data and unmodeled guessing (i.e., c = 0), there was essentially no difference in the correct dimensionality decisions for Mplus and TESTFACT as the correlation increased from .3 to .6 (e.g., using the proportion of variance index). However, under modeled guessing (i.e., c = .33), there was a large change in the proportion of correct dimensionality decisions in Mplus as the correlation increased from .3 to .6, whereas there was little change in the proportion of correct dimensionality decisions in TESTFACT as the correlation increased from .3 to .6. Again, a different pattern of the interaction for the different indices was observed. A similar pattern of the results in correct dimensionality decisions or the estimated dimensionality was observed when using the proportion of variance and the RMSR reduction indices. However, this pattern differed under parallel analysis. Based on the parameter recovery results, when guessing was not modeled, higher correlations led to increase mean RMSD values. For example, greater differences

between mean RMSD values under the two correlation conditions were observed (.03 ~ .10) when using Mplus in data that assumed guessing. In contrast, mean RMSD values between the two correlation conditions were about the same when using TESTFACT.

Finally, an interaction between the discrimination and correlation effects was also observed. Generally, higher correlations were associated with an increased discrimination effect. For example, in two-dimensional data that assumed guessing and using the RMSR index, the estimated number of dimensions were overestimated under two high discrimination conditions (i.e., HH and MH) and the low correlation condition, whereas overestimation was found in four high discrimination conditions (i.e., HH, MH, MM, LH) under the high correlation condition. However, for the parameter recovery results, no clear interaction between the discrimination and correlation effects was observed. The only exception occurred when using Mplus in three-dimensional data that assumed guessing. Mean RMSD values decreased when the discrimination level decreased under the high correlation condition, whereas mean RMSD values across all discrimination conditions were about the same under the low correlation condition.

In summary, the three factors, guessing, test item discrimination, and correlation between dimensions interacted with each other to influence the overall impact on the outcome measures. For example, the guessing effect increased the influence of the discrimination effect and the correlation effect increased the degree of the discrimination effect. However, there were inconsistent results observed between two-dimensional versus three-dimensional data, and the effect in three-dimensional data could not be adequately evaluated in TESTFACT due to the non-convergent problems.

5.2 MAJOR FINDINGS OF THE TIMSS APPLICATION

Having discussed the major findings observed in the simulation study, some findings arose from the investigation of real data (i.e., TIMSS data). The descriptive statistics illustrated that more than 70% of the TIMSS items demonstrated guessing behavior. Given the high correlations between dimensions and examinee guessing behavior shown in data, the guessing and the correlation effects should be considered carefully. However, any effect due to the discrimination level of items could be largely ignored since the discrimination parameters were

in medium level, not in extreme high or low conditions. Since TESTFACT did correct problems caused by guessing, this method illustrated stability in the assessment of dimensionality using various indices and obtained a larger number of items with substantial factor loadings. Inconsistency in the dimensionality assessment using the four indices with Mplus was observed (1 to 5 dimensions), whereas TESTFACT consistently estimated two dimensions. The patterns in results for both Mplus and TESTFACT were consistent with the findings in the simulation study and Stone and Yeh's (2006) study. Further investigation of the internal structures did not show any connection to the content or cognitive domains defined by the research groups of TIMSS, in either Mplus or in TESTFACT. Due to the absence of item content (only some items were released to the public), it was impossible to further examine interpretable test structures.

5.3 LIMITATIONS

Due to the design of this study, there may be several limitations for applying the findings observed in the simulation study. First, in order to simplify the comparison and to obtain a clear understanding of the guessing effect, only two guessing conditions (i.e. 0 and .33) were chosen. This leads to a restriction in the generalizability of the results obtained from this study to tests with only very low or very high guessing behavior. It should be noted that even in tests with four- or five-option items, where a .20 or .25 guessing rate may be expected, it is typical that one or two distractors can often be eliminated by examinees. Thus, the guessing rate is often in fact higher than that random model expectation. Although the value of .33 seems a little higher than the values shown in most real applications that have four-options, it did provide a useful comparison. However, more research is needed with lower guessing conditions, such as a .2 guessing rate under the random guessing model.

Second, the item parameters used to generate the simulation data were not based on real tests but defined for the purpose of making comparisons between different levels of item discrimination. For example, the item discrimination parameters were set to three levels: low, medium and high. However, some of the test configurations that were studied may not be realistic. Therefore, the degree and the distribution of discrimination parameters should be investigated in more detail to classify the influence of discrimination parameters.

Third, an approximate simple structure was assumed when generating the simulated data in this study. Although an approximate simple structure may be more realistic than a simple structure, a complex structure may also be observed in real testing applications. Therefore, there is a limitation in the generalizability of these findings to other test settings.

Fourth, any assumptions underlying the methods used in this study were assumed and not manipulated which also may be unrealistic. Therefore, what was learned from this study may not apply to those testing applications that do not meet the assumptions required for using these approaches.

Finally, considering time consumption in a 100-replication design study, the maximum number of iterations in TESTFACT was set to 50. If a non-convergent solution was obtained after 50 cycles, the replication was considered to be an invalid case. However, according to the results shown in this study, TESTFACT tended to underestimate test dimensionality when the number of dimensions increased under this restriction. Based on further examination of these non-convergent cases for 10 randomly selected datasets with non-convergent problems, there were 7 out of 10 conditions that did converge within 200 iterations (the average number of iterations was ~ 120). Note that this problem relates primarily to the three-dimensional data with a high correlation condition ($r = .6$) and only occurred in those factor solutions whose number of factors matched the true dimensions. For example, for three-dimensional data, three-factor solutions could be found that did not converge in 50 iterations but did converge in less than 200 iterations, but all four- and five-factor solutions did not converge at 200 iterations. This may lead to different conclusions for the number of estimated dimensions in the simulation study. Therefore, further investigation should be conducted with a larger specified number of iterations to more fully explore the differences between Mplus and TESTFACT with three- or higher dimensional data.

5.4 FUTURE RESEARCH DIRECTION

According to the discussion of the results from the simulation data, there are several areas for future research that could be considered. First, it may be interesting to further explore the effect of item parameters, including guessing and discrimination parameters. In order to identify

the guessing effect, the conditions with more guessing levels can be investigated. In this study only two guessing levels, 0 and .33, were used. In real data, such as the MBE or TIMSS assessments, lower guessing rates, around .2, have been observed. In addition, the discrimination effect was observed in this study. Therefore, it may be interesting to further explore how different distributions of discrimination parameters affect the dimensionality assessment.

Although most of the indices used in this study were essentially subjective, some of these indices performed well or even better than the objective chi-square test. However, the effectiveness of the different indices for estimating dimensionality varied in this study. As indicated by Tate (2003), it may be of interest to explore whether informal hypothesis testing provides supportive evidence for determining correct dimensionality. In this study, the formal tests (i.e. the chi-square different test) for determining dimensionality only performed well in TESTFACT with one-, two- and three-dimensional data, but only performed well in Mplus with three-dimensional data. Therefore, more empirical research may provide more evidence for the use of these potential indices based on subjective criteria. Also, parallel analysis did not show better performances in determining dimensionality as expected. Parallel analysis compares eigenvalues from simulated items responses that are uncorrelated to eigenvalues from the true item responses. However, the item responses that were simulated for parallel analysis did not assume any guessing behavior. It might be interesting to investigate the performance of parallel analysis if guessing was modeled in the simulation of random data.

As mentioned in Tate's study (2003), the exploratory methods are not sensitive to hierarchical structures (e.g., the second-order or bifactor models) or non-simple structures. In addition, only an approximate simple structure was assumed in this study. Therefore, it may be of interest to investigate how well different methods for the dimensionality assessment perform when using tests with hierarchical or non-simple structures, such as second-order factor models or models with complex factor structures.

As mentioned in the limitation section, the results in TESTFACT obtained in a restricted condition (i.e. all cases, where no convergent solutions were obtained after 50 iteration cycles for estimation, were defined as failures). This restriction led to some unexpected results with three-dimensional data. However, this finding should be interpreted cautiously due to the convergence criteria used in this study. More studies should be conducted for further investigation of the

impact on exploratory factor analysis when using higher dimensionality data. Also, this restriction of 50 iterations may not be realistic for real data, especially in those factor solutions where the number of extracted factors matches the expected or the true dimensionality. Therefore, a larger number of iterations (e.g., 200) can be used to examine whether the number of iterations for estimated procedures in TESTFACT affects the dimensionality assessment.

Due to practical constraints, the simulation only explored a limited number of methods for determining test dimensionality. Different approaches for determining dimensionality not explored in this study may be of interest, including nonparametric methods, or noncompensatory factor models. The requirement of strict assumptions for the parametric approach is often difficult to attain. The potential for using nonparametric methods should be examined. In addition, the assumption of a compensatory model, that a lower ability can be compensated for by using other ability, may not be true in some test settings. Some developments in either parametric or nonparametric methods mentioned in Tate's study (2003) could be options for this research direction. For example, the combination of parametric and nonparametric methods proposed by Douglas and Cohen (2001), or a non-simple structure with nonparametric methods as illustrated by Bolt (2001).

Finally, performance assessments and those tests involving mixtures of multiple-choice and constructed-response items are popular. The performance of those methods in assessing dimensionality in tests with constructed-response items or a combination of multiple-choice and constructed-response items should be explored.

APPENDIX A

A.1 ITEM PARAMETERS OF ONE-DIMENSIONAL DATA

	Low			Medium			High		
No.	*a*	*d*	*b*	*a*	*d*	*b*	*a*	*d*	*b*
1	1.00	-2.00	2.00	1.50	-2.00	1.33	1.50	-2.00	1.33
2	1.00	-1.50	1.50	1.50	-1.50	1.00	1.50	-1.50	1.00
3	1.00	-1.25	1.25	1.50	-1.25	0.83	1.50	-1.25	0.83
4	1.00	-1.00	1.00	1.50	-1.00	0.67	1.50	-1.00	0.67
5	1.00	-0.75	0.75	1.50	-0.75	0.50	1.50	-0.75	0.50
6	1.00	-0.50	0.50	1.50	-0.50	0.33	1.50	-0.50	0.33
7	1.00	-0.25	0.25	1.50	-0.25	0.17	1.50	-0.25	0.17
8	1.00	0.00	0.00	1.50	0.00	0.00	1.50	0.00	0.00
9	1.00	0.25	-0.25	1.50	0.25	-0.17	1.50	0.25	-0.17
10	1.00	0.50	-0.50	1.50	0.50	-0.33	1.50	0.50	-0.33
11	1.00	0.75	-0.75	1.50	0.75	-0.50	1.50	0.75	-0.50
12	1.00	1.00	-1.00	1.50	1.00	-0.67	1.50	1.00	-0.67
13	1.00	1.25	-1.25	1.50	1.25	-0.83	1.50	1.25	-0.83
14	1.00	1.50	-1.50	1.50	1.50	-1.00	1.50	1.50	-1.00
15	1.00	2.00	-2.00	1.50	2.00	-1.33	1.50	2.00	-1.33
16	1.00	-2.00	2.00	1.00	-2.00	2.00	1.50	-2.00	1.33
17	1.00	-1.50	1.50	1.00	-1.50	1.50	1.50	-1.50	1.00
18	1.00	-1.25	1.25	1.00	-1.25	1.25	1.50	-1.25	0.83
19	1.00	-1.00	1.00	1.00	-1.00	1.00	1.50	-1.00	0.67
20	1.00	-0.75	0.75	1.00	-0.75	0.75	1.50	-0.75	0.50
21	1.00	-0.50	0.50	1.00	-0.50	0.50	1.50	-0.50	0.33
22	1.00	-0.25	0.25	1.00	-0.25	0.25	1.50	-0.25	0.17
23	1.00	0.00	0.00	1.00	0.00	0.00	1.50	0.00	0.00
24	1.00	0.25	-0.25	1.00	0.25	-0.25	1.50	0.25	-0.17
25	1.00	0.50	-0.50	1.00	0.50	-0.50	1.50	0.50	-0.33
26	1.00	0.75	-0.75	1.00	0.75	-0.75	1.50	0.75	-0.50
27	1.00	1.00	-1.00	1.00	1.00	-1.00	1.50	1.00	-0.67
28	1.00	1.25	-1.25	1.00	1.25	-1.25	1.50	1.25	-0.83
29	1.00	1.50	-1.50	1.00	1.50	-1.50	1.50	1.50	-1.00
30	1.00	2.00	-2.00	1.00	2.00	-2.00	1.50	2.00	-1.33

	Low			Medium			High		
No.	*a*	*d*	*b*	*a*	*d*	*b*	*a*	*d*	*b*
31	0.50	-2.00	4.00	1.00	-2.00	2.00	1.00	-2.00	2.00
32	0.50	-1.50	3.00	1.00	-1.50	1.50	1.00	-1.50	1.50
33	0.50	-1.25	2.50	1.00	-1.25	1.25	1.00	-1.25	1.25
34	0.50	-1.00	2.00	1.00	-1.00	1.00	1.00	-1.00	1.00
35	0.50	-0.75	1.50	1.00	-0.75	0.75	1.00	-0.75	0.75
36	0.50	-0.50	1.00	1.00	-0.50	0.50	1.00	-0.50	0.50
37	0.50	-0.25	0.50	1.00	-0.25	0.25	1.00	-0.25	0.25
38	0.50	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
39	0.50	0.25	-0.50	1.00	0.25	-0.25	1.00	0.25	-0.25
40	0.50	0.50	-1.00	1.00	0.50	-0.50	1.00	0.50	-0.50
41	0.50	0.75	-1.50	1.00	0.75	-0.75	1.00	0.75	-0.75
42	0.50	1.00	-2.00	1.00	1.00	-1.00	1.00	1.00	-1.00
43	0.50	1.25	-2.50	1.00	1.25	-1.25	1.00	1.25	-1.25
44	0.50	1.50	-3.00	1.00	1.50	-1.50	1.00	1.50	-1.50
45	0.50	2.00	-4.00	1.00	2.00	-2.00	1.00	2.00	-2.00
46	0.50	-2.00	4.00	0.50	-2.00	4.00	1.00	-2.00	2.00
47	0.50	-1.50	3.00	0.50	-1.50	3.00	1.00	-1.50	1.50
48	0.50	-1.25	2.50	0.50	-1.25	2.50	1.00	-1.25	1.25
49	0.50	-1.00	2.00	0.50	-1.00	2.00	1.00	-1.00	1.00
50	0.50	-0.75	1.50	0.50	-0.75	1.50	1.00	-0.75	0.75
51	0.50	-0.50	1.00	0.50	-0.50	1.00	1.00	-0.50	0.50
52	0.50	-0.25	0.50	0.50	-0.25	0.50	1.00	-0.25	0.25
53	0.50	0.00	0.00	0.50	0.00	0.00	1.00	0.00	0.00
54	0.50	0.25	-0.50	0.50	0.25	-0.50	1.00	0.25	-0.25
55	0.50	0.50	-1.00	0.50	0.50	-1.00	1.00	0.50	-0.50
56	0.50	0.75	-1.50	0.50	0.75	-1.50	1.00	0.75	-0.75
57	0.50	1.00	-2.00	0.50	1.00	-2.00	1.00	1.00	-1.00
58	0.50	1.25	-2.50	0.50	1.25	-2.50	1.00	1.25	-1.25
59	0.50	1.50	-3.00	0.50	1.50	-3.00	1.00	1.50	-1.50
60	0.50	2.00	-4.00	0.50	2.00	-4.00	1.00	2.00	-2.00
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.25	1.15	1.81	0.36	1.15	1.45	0.25	1.15	0.97

A.2 ITEM PARAMETERS OF TWO-DIMENSIONAL DATA

	Low			Medium			High		
No.	*a*	*d*	*b*	*a*	*d*	*b*	*a*	*d*	*b*
1	1.00	-2.00	1.96	1.50	-2.00	1.32	1.50	-2.00	1.32
2	1.00	-1.50	1.47	1.50	-1.00	0.66	1.50	-1.50	0.99
3	1.00	-1.25	1.23	1.50	-0.50	0.33	1.50	-1.25	0.83
4	1.00	-1.00	0.98	1.50	0.00	0.00	1.50	-1.00	0.66
5	1.00	-0.75	0.74	1.50	0.50	-0.33	1.50	-0.75	0.50
6	1.00	-0.50	0.49	1.50	1.00	-0.66	1.50	-0.50	0.33
7	1.00	-0.25	0.25	1.50	2.00	-1.32	1.50	-0.25	0.17
8	1.00	0.00	0.00	1.00	-2.00	1.96	1.50	0.00	0.00
9	1.00	0.25	-0.25	1.00	-1.50	1.47	1.50	0.25	-0.17
10	1.00	0.50	-0.49	1.00	-1.25	1.23	1.50	0.50	-0.33
11	1.00	0.75	-0.74	1.00	-1.00	0.98	1.50	0.75	-0.50
12	1.00	1.00	-0.98	1.00	-0.75	0.74	1.50	1.00	-0.66
13	1.00	1.25	-1.23	1.00	-0.50	0.49	1.50	1.25	-0.83
14	1.00	1.50	-1.47	1.00	-0.25	0.25	1.50	1.50	-0.99
15	1.00	2.00	-1.96	1.00	0.00	0.00	1.50	2.00	-1.32
16	0.50	-2.00	3.71	1.00	0.00	0.00	1.00	-2.00	1.96
17	0.50	-1.50	2.79	1.00	0.25	-0.25	1.00	-1.50	1.47
18	0.50	-1.25	2.32	1.00	0.50	-0.49	1.00	-1.25	1.23
19	0.50	-1.00	1.86	1.00	0.75	-0.74	1.00	-1.00	0.98
20	0.50	-0.75	1.39	1.00	1.00	-0.98	1.00	-0.75	0.74
21	0.50	-0.50	0.93	1.00	1.25	-1.23	1.00	-0.50	0.49
22	0.50	-0.25	0.46	1.00	1.50	-1.47	1.00	-0.25	0.25
23	0.50	0.00	0.00	1.00	2.00	-1.96	1.00	0.00	0.00
24	0.50	0.25	-0.46	0.50	-2.00	3.71	1.00	0.25	-0.25
25	0.50	0.50	-0.93	0.50	-1.00	1.86	1.00	0.50	-0.49
26	0.50	0.75	-1.39	0.50	-0.50	0.93	1.00	0.75	-0.74
27	0.50	1.00	-1.86	0.50	0.00	0.00	1.00	1.00	-0.98
28	0.50	1.25	-2.32	0.50	0.50	-0.93	1.00	1.25	-1.23
29	0.50	1.50	-2.79	0.50	1.00	-1.86	1.00	1.50	-1.47
30	0.50	2.00	-3.71	0.50	2.00	-3.71	1.00	2.00	-1.96
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.25	1.16	1.72	0.35	1.18	1.43	0.25	1.16	0.97

A.3 ITEM PARAMETERS OF THREE-DIMENSIONAL DATA

	Low			Medium			High		
No.	*a*	*d*	*b*	*a*	*d*	*b*	*a*	*d*	*b*
1	1.00	-2.00	1.92	1.50	-1.75	1.15	1.50	-2.00	1.31
2	1.00	-1.50	1.44	1.50	-0.75	0.49	1.50	-1.50	0.98
3	1.00	-1.00	0.96	1.50	0.00	0.00	1.50	-1.00	0.66
4	1.00	-0.50	0.48	1.50	0.75	-0.49	1.50	-0.50	0.33
5	1.00	0.00	0.00	1.50	1.75	-1.15	1.50	0.00	0.00
6	1.00	0.00	0.00	1.00	-2.00	1.92	1.50	0.00	0.00
7	1.00	0.50	-0.48	1.00	-1.50	1.44	1.50	0.50	-0.33
8	1.00	1.00	-0.96	1.00	-1.00	0.96	1.50	1.00	-0.66
9	1.00	1.50	-1.44	1.00	-0.50	0.48	1.50	1.50	-0.98
10	1.00	2.00	-1.92	1.00	0.00	0.00	1.50	2.00	-1.31
11	0.50	-2.00	3.48	1.00	0.00	0.00	1.00	-2.00	1.92
12	0.50	-1.50	2.61	1.00	0.50	-0.48	1.00	-1.50	1.44
13	0.50	-1.00	1.74	1.00	1.00	-0.96	1.00	-1.00	0.96
14	0.50	-0.50	0.87	1.00	1.50	-1.44	1.00	-0.50	0.48
15	0.50	0.00	0.00	1.00	2.00	-1.92	1.00	0.00	0.00
16	0.50	0.00	0.00	0.50	-1.75	3.05	1.00	0.00	0.00
17	0.50	0.50	-0.87	0.50	-0.75	1.31	1.00	0.50	-0.48
18	0.50	1.00	-1.74	0.50	0.00	0.00	1.00	1.00	-0.96
19	0.50	1.50	-2.61	0.50	0.75	-1.31	1.00	1.50	-1.44
20	0.50	2.00	-3.48	0.50	1.75	-3.05	1.00	2.00	-1.92
Mean	0.75	0.00	0.00	1.00	0.00	0.00	1.25	0.00	0.00
SD	0.26	1.26	1.77	0.36	1.25	1.43	0.26	1.26	1.03

APPENDIX B

THE MEAN DIFFERENCE OF ESTIMATED AND TRUE DIMENSIONALITY IN MPLUS AND TESTFACT BY THE PROPORTION OF VARIANCE INDEX

	r = .3				*r* = .6			
	c = 0		c = .33		c = 0		c = .33	
	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
Unidimensional data								
H	0.00	0.00	0.00	0.01				
M	0.00	0.00	0.00	0.01				
L	0.00	0.00	0.00	0.00				
Two-dimensional data								
HH	0.00	0.00	0.00	-0.03	0.00	0.00	-1.00	0.00
MH	0.00	0.00	0.00	-0.09	0.00	0.00	-1.00	-0.08
MM	0.00	0.00	0.00	-0.09	0.00	0.00	-1.00	-0.22
LH	0.00	0.00	-0.06	-0.13	0.00	0.00	-1.00	-0.55
LM	0.00	0.00	-0.71	-0.15	-0.57	-0.60	-1.00	-0.81
LL	0.00	0.00	-1.00	-0.11	-1.00	-1.00	-1.00	-0.91
Three-dimensional data								
HHH	0.00	0.00	-0.89	-0.35	-1.74	-1.77	-2.00	-1.99
HHM	0.00	0.00	-1.36	-0.49	-1.84	-1.86	-2.00	-1.99
HHL	-0.01	-0.03	-1.46	-0.93	-1.84	-1.91	-2.00	-1.99
MMH	0.00	0.00	-1.98	-0.60	-2.00	-2.00	-2.00	-1.99
MMM	0.00	0.00	-2.00	-0.87	-2.00	-2.00	-2.00	-1.99
MML	-0.23	-0.27	-2.00	-1.12	-2.00	-2.00	-2.00	-1.96
LMH	-0.12	-0.15	-2.00	-1.05	-2.00	-2.00	-2.00	-1.98
LLH	-0.98	-0.98	-2.00	-1.15	-2.00	-2.00	-2.00	-1.97
LLM	-0.97	-0.98	-2.00	-1.29	-2.00	-2.00	-2.00	-1.96
LLL	-1.89	-1.89	-2.00	-1.78	-2.00	-2.00	-2.00	-1.97

APPENDIX C

THE MEAN DIFFERENCE OF ESTIMATED AND TRUE DIMENSIONALITY IN MPLUS AND TESTFACT BY PARALLEL ANALYSIS

	r = .3				*r* = .6			
	c = 0		c = .33		c = 0		c = .33	
	Mplus	TSF[a]	Mplus	TSF	Mplus	TSF	Mplus	TSF
Unidimensional data								
H	0.31	0.17	1.16	0.21				
M	1.42	0.62	2.00	0.11				
L	1.99	1.00	2.00	0.11				
Two-dimensional data								
HH	0.00	0.00	1.99	0.16	0.01	0.00	1.60	0.07
MH	0.20	0.07	2.00	-0.03	0.12	0.06	1.93	-0.04
MM	0.84	0.48	2.00	-0.05	0.62	0.29	2.00	-0.12
LH	0.53	0.32	2.00	-0.08	0.40	0.18	2.00	-0.18
LM	1.55	0.90	2.00	-0.14	1.08	0.61	2.00	-0.15
LL	1.92	1.35	2.00	-0.10	1.64	0.86	2.00	-0.27
Three-dimensional data								
HHH	0.00	0.00	1.94	-0.35	0.01	-0.39	1.18	-1.06
HHM	0.01	0.01	1.99	-0.49	0.00	-0.75	1.41	-1.06
HHL	0.09	0.06	2.00	-0.55	0.07	-0.60	1.87	-1.12
MMH	0.06	0.03	2.00	-0.58	0.07	-0.99	1.74	-1.15
MMM	0.31	0.17	2.00	-0.86	0.14	-0.96	1.94	-1.28
MML	0.64	0.43	2.00	-0.87	0.31	-0.85	1.99	-1.22
LMH	0.29	0.14	2.00	-0.63	0.17	-0.86	1.96	-1.16
LLH	0.75	0.49	2.00	-0.83	0.32	-0.71	1.99	-1.29
LLM	1.28	0.77	2.00	-0.79	0.67	-0.69	2.00	-1.37
LLL	1.71	1.20	2.00	-1.07	0.76	-0.73	2.00	-1.47

APPENDIX D

THE MEAN DIFFERENCE OF ESTIMATED AND TRUE DIMENSIONALITY IN MPLUS AND TESTFACT BY THE REDUCTION OF RMSR INDEX

	r = .3				*r* = .6			
	c = 0		c = .33		c = 0		c = .33	
	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
Unidimensional data								
H	0.36	0.00	1.00	0.00				
M	0.03	0.00	0.95	0.00				
L	0.00	0.00	0.00	0.00				
Two-dimensional data								
HH	0.06	0.00	0.94	-0.03	0.12	0.00	1.00	-0.13
MH	0.03	0.00	0.58	-0.11	0.07	0.00	1.00	-0.67
MM	0.01	0.00	0.03	-0.11	0.03	0.00	0.71	-0.86
LH	0.03	0.00	0.31	-0.19	0.03	0.00	0.93	-0.97
LM	0.00	0.00	0.01	-0.24	0.04	0.00	-0.03	-0.99
LL	0.00	0.00	0.00	-0.55	0.01	0.00	-0.95	-1.00
Three-dimensional data								
HHH	0.04	0.06	0.87	-0.75	0.16	-0.40	1.00	-2.00
HHM	0.02	0.38	0.66	-1.03	0.17	-0.75	0.87	-2.00
HHL	0.02	0.00	0.37	-1.16	0.08	-0.61	-0.02	-2.00
MMH	0.02	0.11	0.40	-1.32	0.11	-1.00	0.05	-2.00
MMM	0.04	0.22	0.19	-1.79	0.07	-1.00	-1.09	-2.00
MML	0.01	0.00	-0.03	-1.65	0.07	-1.00	-1.25	-2.00
LMH	0.00	0.00	0.15	-1.39	0.09	-0.95	-0.51	-2.00
LLH	0.01	0.00	-0.65	-1.74	0.06	-0.96	-1.03	-2.00
LLM	0.02	0.00	-0.78	-1.97	-0.02	-1.00	-1.41	-2.00
LLL	0.03	0.00	-1.91	-2.00	-0.98	-1.93	-1.65	-2.00

APPENDIX E

THE MEAN DIFFERENCE OF ESTIMATED AND TRUE DIMENSIONALITY IN MPLUS AND TESTFACT BY THE CHI-SQUARE TEST

	r = .3				*r* = .6			
	c = 0		c = .33		c = 0		c = .33	
	Mplus	TSF	Mplus	TSF	Mplus	TSF	Mplus	TSF
				Unidimensional data				
H	1.95	0.00	1.80	0.00				
M	2.00	0.00	1.88	0.00				
L	2.00	0.05	1.88	0.05				
				Two-dimensional data				
HH	1.89	1.11	1.44	0.00	1.89	0.03	1.09	0.00
MH	1.99	0.65	1.63	-0.08	1.94	0.01	1.39	-0.08
MM	1.99	0.75	1.59	-0.08	1.93	0.03	1.41	-0.12
LH	1.97	0.54	1.56	-0.10	1.87	0.04	1.16	-0.18
LM	1.99	0.60	1.43	-0.15	1.94	0.01	1.31	-0.15
LL	1.99	0.76	1.64	-0.10	1.98	0.04	1.29	-0.27
				Three-dimensional data				
HHH	1.80	0.02	0.63	-0.46	1.16	-0.40	-0.03	-1.06
HHM	1.81	0.01	0.68	-0.50	1.47	-0.75	0.01	-1.06
HHL	1.87	0.01	0.72	-0.58	1.48	-0.61	0.08	-1.12
MMH	1.90	0.02	0.87	-0.62	1.58	-1.00	0.33	-1.15
MMM	1.97	0.00	1.01	-0.97	1.80	-1.00	0.38	-1.28
MML	1.94	0.00	0.89	-0.88	1.85	-1.00	0.48	-1.22
LMH	1.90	0.00	0.73	-0.66	1.58	-0.95	0.25	-1.16
LLH	1.77	0.00	0.53	-0.83	1.59	-0.84	0.21	-1.29
LLM	1.87	0.00	0.71	-0.82	1.72	-1.00	0.35	-1.37
LLL	1.90	0.00	0.49	-1.13	1.77	-1.00	0.16	-1.48

APPENDIX F

ESTIMATED ITEM PARAMETERS FOR BOOKLET 5

Var. ID	*c*	Lord's criterion	*P*	Var. ID	*c*	Lord's criterion	*P*
M022043	0.00	-5.10	0.77	MF32036	0.24	-0.96	0.52
M022049	0.00	-5.33	0.66	MF32447	0.24	-0.93	0.54
M022050	0.22	-0.85	0.41	MF32609	0.23	-2.19	0.75
M022057	0.00	-5.42	0.76	MF32670	0.00	-2.90	0.82
M022062	0.10	-1.15	0.38	MF32690	0.22	-1.97	0.52
M022066	0.07	-1.14	0.50	MF32727	0.17	-0.73	0.54
M022097	0.08	-2.48	0.79	MF32728	0.31	-0.71	0.50
M022101	0.00	-4.64	0.75	MF32732	0.29	-1.54	0.58
M022104	0.00	-3.78	0.80	S032024	0.22	-1.84	0.48
M022105	0.00	-2.90	0.46	S032141	0.20	-1.64	0.41
M022108	0.16	-1.97	0.60	S032315	0.02	-3.45	0.63
M022181	0.52	-2.40	0.89	S032463	0.15	-2.45	0.66
M022257	0.20	-1.07	0.57	S032465	0.23	-3.43	0.75
M032044	0.32	-0.83	0.58	S032514	0.19	-1.86	0.47
M032046	0.14	-0.21	0.33	S032579	0.31	-0.68	0.47
M032079	0.22	-0.68	0.40	SF12001	0.00	-4.59	0.66
M032228	0.16	-1.70	0.62	SF12002	0.00	-6.50	0.78
M032261	0.26	-1.25	0.55	SF12003	0.07	-4.15	0.77
M032271	0.25	-1.23	0.59	SF12004	0.24	-3.88	0.75
M032489	0.00	-3.72	0.79	SF12005	0.10	-4.20	0.62
M032523	0.18	-0.16	0.39	SF12006	0.29	-1.65	0.55
M032525	0.17	-1.62	0.53	SF12013	0.00	-6.63	0.26
M032533	0.23	-0.80	0.59	SF12014	0.00	-3.78	0.76
M032579	0.15	-1.65	0.62	SF12015	0.15	-3.31	0.75
M032588	0.28	-2.09	0.67	SF12016	0.51	-2.85	0.77
M032678	0.06	-0.71	0.48	SF12017	0.05	-2.54	0.63
M032701	0.38	-3.17	0.85	SF12018	0.43	-3.03	0.66
M032704	0.32	-1.25	0.64				

APPENDIX G

ESTIMATED ITEM PARAMETERS FOR BOOKLET 11

Var. ID	*c*	Lord's criterion	*P*	Var. ID	*c*	Lord's criterion	*P*
M032094	0.27	-1.30	0.62	S022054	0.29	-1.55	0.57
M032100	0.00	-4.09	0.56	S022106	0.02	-4.36	0.18
M032116	0.28	-1.09	0.52	S022115	0.00	-4.14	0.68
M032132	0.00	-3.34	0.56	S022126	0.37	-2.13	0.66
M032324	0.09	-1.15	0.37	S022150	0.37	-0.85	0.53
M032397	0.22	-1.74	0.59	S022181	0.13	-1.75	0.43
M032402	0.33	-0.13	0.50	S022183	0.17	-0.53	0.39
M032419	0.33	-0.72	0.52	S022208	0.28	-0.97	0.54
M032477	0.17	-1.02	0.54	S022276	0.44	-1.67	0.76
M032662	0.08	-0.23	0.23	S022290	0.00	-3.66	0.66
MF12013	0.30	-1.43	0.64	S022294	0.09	-4.04	0.69
MF12014	0.28	-2.39	0.78	S032008	0.00	-4.11	0.63
MF12015	0.03	-2.15	0.62	S032035	0.00	-3.44	0.68
MF12016	0.46	-0.18	0.57	S032055	0.35	-2.63	0.87
MF12017	0.10	-2.43	0.55	S032083	0.07	-2.50	0.76
MF22185	0.34	-0.95	0.54	S032150	0.00	-5.46	0.71
MF22188	0.24	-1.01	0.43	S032258	0.03	-3.70	0.26
MF22189	0.00	-3.31	0.70	S032281	0.33	-1.41	0.71
MF22191	0.00	-2.73	0.59	S032301	0.26	-1.03	0.48
MF22194	0.22	-2.30	0.58	S032385	0.26	-1.51	0.58
MF22196	0.15	-1.67	0.60	S032446	0.32	-2.24	0.69
MF22198	0.22	-1.05	0.49	S032564	0.16	-0.40	0.32
MF22199	0.21	-0.33	0.40	S032607	0.08	-3.00	0.57
MF22251	0.15	-1.19	0.35	S032683	0.30	-0.92	0.51
S022002	0.16	-2.40	0.63	SF32422	0.22	-1.54	0.57
S022019	0.50	-2.73	0.77	SF32574	0.36	-0.55	0.55
S022042	0.00	-2.41	0.62	SF32714	0.41	-2.97	0.79

BIBLIOGRAPHY

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

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what item and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*(4), 311-329.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77-85.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

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

Bolt, D. M. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25*, 244-257.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlation between items and between tests. *Psychometrika, 26*, 347-372.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5-32.

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolation 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53*, 585-596.

Cudeck, R. (2000). Exploratory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.). *Applied multivariate statistics and mathematical modeling* (pp. 265-297). San Diego: Academic Press.

De Champlain, A. F. (1999). *An overview of nonlinear factor analysis and its relationship to item response theory*: Law School Admission Council.

De Champlain, A. F. & Tang, K. L. (1993). *The effect of nonnormal ability distributions on the assessment of dimensionality.* Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

De Champlain, A. F. & Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57* (1), 174-178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234-243.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence, 20*, 169-190

Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Fava, J. L., & Velicer, W. F. (1992). The effects of over-extraction on factor and component analysis. *Multivariate Behavioral Research, 27*, 387-415.

Fiske, D. W. (2002). Validity for what? In Braun, H. I., Jackson, D. N., & Wiley, D. E. (Eds.). *The role of constructs in psychological and educational measurement* (pp.169-178). Mahwah, NJ: Lawrence Erlbaum.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

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

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*, 377-393.

Gonzalez, E.J., & Smith, T. A. (1997). User guide for the TIMSS international database: *Primary and middle school years*. Chestnut Hill, MA: TIMSS International Study Center.

Gonzalez, E. J., & Miles, J. A. (2001). *TIMSS 1999 user guide for the international database*. Chestnut Hill, MA: TIMSS International Study Center.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum.

Haertel, E. H. (1999). Validating arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items. Mahwah, NJ: Lawrence Erlbaum.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*(1), 1-14.

Helms, J. E. (2003). Fair and valid use of educational testing in grades K-12. In J. E. Wall and G. R. Walz (Eds.), Measure up: Assessment issues for teachers, counselors and administrators (pp. 81-88). Greensboro, NC: CAPS.

Hoijtink, H., Rooks, G., & Wilmink, F. W. (1999). Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. *Psychological Methods, 4*(3), 300-314.

Hoijtink, H., Vollema, M. (2003). Contemporary extensions of the Rasch model. *Quality and Quantity, 37*, 263-276.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structure Equation Modeling, 6*(1), 1-55.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kim , H. R. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation, University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 55-12B*, 5598.

Kim, H. R., Zhang, J.,, & Stout, W. F. (1995). A new index of dimensionality—DIMTEST. Unpublished manuscript.

Kim, J.-O. and C. W. Mueller (1978). *Factor analysis: Statistical methods and practical issues*. Newbury Park, CA: Sage.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*(3), 457-477.

Kromrey, J. D., Parshall, C. G., Chason, W. M., & Yi, Q. (1999). *Generating item responses based on multidimensional item response theory*. Paper posted at the Twenty-Fourth Annual of SAS group international conference (Paper 241), Miami Beach, FL.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinborough, 60A*, 64-82.

Lord, F. N. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. N. & Novick, M. R. (1968). Statistical theories of mental test scores. Rading, MA: Addison-Wesley.

Martineau, J. A., Mapuranga, R., & Ward, K. (2006, April). Confirming content structure in standardized state assessments using multidimensional item response theory. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

McDonald, R. P. (1967). *Nonlinear factor analysis* (Psychometric Monographs, No. 15). The Psychometric Society.

McDonald, R. P. (1981). The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*(4), 379-396.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and Issues* (pp. 63-86). Ottawa: Edumetrics Research Group.

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

McLeod, D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for item scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189-215). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R.L., Linn(Ed.), *Educational Measurement* (3rd ed.), pp. 13-103. New York: American Council on Education/Macmillan.

Messick, S. (1995). Validation of psychological assessment: Validation of inferences form person's responses and performances as scientific inquiry into score meanings. *American Psychologist, 50*, 741-749.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*(1), 3-31.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostorwski, S. J., & O'Connor, K. M. (2003). Assessment frameworks and specifications 2003 2nd Edition. Chestnut Hill, MA: TIMSS International Study Center.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551-560.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41-68.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33*(3), 355-368.

NCLB. No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 stat. 1425, (2001).

Neidorf, T. S., & Garden, R. (2004). Developing the TIMSS 2003 mathematics and science assessment and scoring guides. In Martin, M.O., Mullis, I.V.S., & Chrostowski, S.J. (Eds.), *TIMSS 2003 Technical Report* (pp. 23-65), Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Newsom, (2005). *Alternative estimation methods*. Handouts of Structural Equation Modeling in Winter 2005. URL: http://www.upa.pdx.edu/IOA/newsom/semclass/ho_estimate.doc.

Nunnallly, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32* (3), 396-402.

Pang, X. L. (1999). *Assessing the performance of the approximate chi-square and Stout's T statistics with different test structures*. Unpublished Doctoral dissertation, University of Ottawa.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. H. Fischer & D. Laming, *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 249-263). New York: Springer-Verlag.

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Reckase, M. D. (1985). *Models for multidimensional tests and hierarchically structured training materials* (research report No. ONR85-1).

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* (pp. 271-286). New York: Springer.

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

Rogers, H. J. (1999). Guessing in multiple choice tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 235-243). Amsterdam: Pergamom.

Roussos, L. (1995). A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and constrict validity (Doctoral dissertation, University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 57-04B*, 2956.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis*. Paper presented at the annul meeting of the National Council on Measurement in Education, Atlanta, GA.

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

Sireci, S. G., & Gonzalez, E. J. (2003). *Evaluating the structure equivalence of tests used in international comparisons of educational achievement*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Skorupski, W. P. (2005). *A review of approaches for improving the reliability of objective level scores* (Draft). Washington, DC: the Council of Chief State School Officers.

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* (research report series No. ONR90-8): ACT.

Steiger, J. H.(1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

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 Psychology Measurement, 66*(2), 193-214.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82-89). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Takane, Y. and De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*(3), 159-203.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-Scale Assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). New Jersey: Lawrence Erlbaum.

Turner, N. E. (1998). The effect of common variance and structure on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement, 58*, 541-568.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*(3), 233-243.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.

Wilson, D. T., & Wood, R., & Gibbons, R. D.(1987). *TESTFACT: Test scoring, items statistics, and full-information item factor analysis*. Mooresville, IN: Scientific Software International.

Wilson, D., Wood, R., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). *TESTFACT: Test scoring and full information item factor analysis* (Version 4.0). IL: Scientific Software International.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and over-extraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*, 345-365.

van Abswoude, A. A. H., van de Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Measurement in Education, 28*(1), 3-24.

Yeh, C.-C., & Stone, C. A. (2004). *Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination.* Paper presented at the 2004 Annual Meeting of the National Council on Measurement in Education, San Diego.

Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 231-249.