ASSESSING FIT OF ITEM RESPONSE MODELS FOR PERFORMANCE ASSESSMENTS USING BAYESIAN ANALYSIS

by

Xiaowen Zhu

B.S., Southwest University of Science and Technology, 1996

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

2009

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Xiaowen Zhu

It was defended on

November 20, 2009

and approved by

Clement A. Stone, Professor, Psychology in Education

Suzanne Lane, Professor, Psychology in Education

Feifei Ye, Assistant Professor, Psychology in Education

James E. Bost, Associate Professor, Center for Research on Health Care

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

ASSESSING FIT OF ITEM RESPONSE MODELS FOR PERFORMANCE ASSESSMENTS USING BAYESIAN ANALYSIS

Xiaowen Zhu, PhD

University of Pittsburgh, 2009

Assessing IRT model-fit and comparing different IRT models from a Bayesian perspective is gaining attention. This research evaluated the performance of Bayesian model-fit and model-comparison techniques in assessing the fit of unidimensional Graded Response (GR) models and comparing different GR models for performance assessment applications.

The study explored the general performance of the PPMC method and a variety of discrepancy measures (test-level, item-level, and pair-wise measures) in evaluating different aspects of fit for unidimensional GR models. Previous findings that the PPMC method is conservative were confirmed. In addition, PPMC was found to have adequate power in detecting different aspects of misfit when using appropriate discrepancy measures. Pair-wise measures were found more powerful in detecting violations of unidimensionality and local independence assumptions than test-level and item-level measures. Yen's Q_3 measure appeared to perform best. In addition, the power of PPMC increased as the degree of multidimensionality or local dependence among item responses increased. Two classical item-fit statistics were found effective for detecting the item misfit due to discrepancies from GR model boundary curves.

The study also compared the relative effectiveness of three Bayesian model-comparison indices (DIC, CPO, and PPMC) for model selection. The results showed that these indices appeared to perform equally well in selecting a preferred model for an overall test. However, the advantage of PPMC applications is that they can be used to compare the relative fit of different

models, but also evaluate the absolute fit of each individual model. In contrast, the DIC and CPO indices only compare the relative fit of different models.

This study further applied the Bayesian model-fit and model-comparison methods to three real datasets from the QCAI performance assessment. The results indicated that these datasets were essentially unidimensional and exhibited local independence among items. A 2P GR model provided better fit than a 1P GR model, and a two-dimensional model was also not preferred. These findings were consistent with previous studies, although Stone's fit statistics in the PPMC context identified less misfitting items compared to previous studies. Limitations and future research for Bayesian applications to IRT are discussed.

TABLE OF CONTENTS

PREFACE XIX

1.0 INTRODUCTION 1

1.1 STATEMENT OF THE PROBLEM 1

1.2 SIGNIFICANCE OF THE STUDY 6

1.3 LIMITATIONS OF THE STUDY 8

2.0 REVIEW OF LITERATURE 9

2.1 APPLICATIONS OF IRT TO PERFORMANCE ASSESSMENTS 9

2.1.1 Brief Introduction to Performance Assessments 9

2.1.2 IRT Models for Performance Assessments 11

2.1.2.1 General Description 11

2.1.2.2 Graded Response Model (Samejima, 1969) 13

2.1.3 Main Threats in Applying Unidimensional IRT Models to PAs 17

2.1.3.1 Multidimensionality 17

2.1.3.2 Local Dependence 20

2.2 TRADITIONAL METHODS FOR CHECKING IRT MODEL-FIT 23

2.2.1 Assessing Dimensionality 23

2.2.2 Detecting Local Dependence 25

2.2.3 Evaluating Item-Fit 27

2.2.3.1 Traditional Item-Fit Statistics 28

2.2.3.2 Alternative Item-Fit Statistics 30

2.3 POSTERIOR PREDICTIVE MODEL CHECKING (PPMC) IN A BAYESIAN FRAMEWORK 33

2.3.1 Introduction to Bayesian Inference 33

2.3.2 Posterior Predictive Model Checking (PPMC) 35

2.3.2.1 Description of PPMC Method 35

2.3.2.2 Computation via MCMC Simulation 38

2.3.2.3 Discrepancy Measures 39

2.3.2.4 Advantages of PPMC over Classical Model-Fit Tests 41

2.3.3 Markov Chain Monte Carlo (MCMC) Simulation 41

2.3.3.1 Definition 41

2.3.3.2 Convergence Diagnosis 44

2.4 CHECKING IRT MODEL-FIT USING PPMC 47

2.4.1 Advantages of Using PPMC in IRT 47

2.4.2 Discrepancy Measures Used with Dichotomous IRT Models 49

2.4.2.1 Test-Level Discrepancy Measures 49

2.4.2.2 Item-Level Discrepancy Measures 51

2.4.2.3 Pair-wise Discrepancy Measures 53

2.4.3 Previous Research 58

2.5 MODEL COMPARISON IN A BAYESIAN FRAMEWORK 62

2.5.1 Pseudo-Bayes Factor (PsBF) 63

2.5.2 Deviance Information Criterion (DIC) 66

3.0 METHODOLOGY 68

3.1 SIMULATION STUDY 1 69

3.1.1 Design of Simulation Study 1 69

3.1.2 Generate and Validate Item Response Data 75

3.1.3 Estimate Unidimensional GR Model in WinBUGS 91

3.1.4 Discrepancy Measures Used in Study 1 96

3.1.5 Conduct PPMC 103

3.2 SIMULATION STUDY 2 107

3.2.1 Design of Simulation Study 2 107

3.2.2 Generate Item Response Data 109

3.2.3 Estimate Different Data-Analysis Models in WinBUGS 110

3.2.4 Conduct Model Comparison 126

3.3 REAL DATA APPLICATION 130

4.0 RESULTS 136

4.1 RESULTS FROM SIMULATION STUDY 1 136

4.1.1 Item Parameter Recovery 137

4.1.2 Condition 1 (Ma = Mg = unidimensional GR) 138

4.1.3 Condition 2 (Mg = 2-dim simple-structure GR , Ma = 1-dim GR) 149

4.1.4 Condition 3 (Mg = 2-dim complex-structure GR , Ma = 1-dim GR) 158

4.1.5 Condition 4 (Mg = testlet GR , Ma = 1-dim GR) 167

4.1.6 Condition 5 (Mg = items with improper BCCs , Ma = 1-dim GR) 177

4.2 RESULTS FROM SIMULATION STUDY 2 182

4.2.1 Condition 1 (2P GR vs. 1P GR vs. RS Models) 182

4.2.2 Condition 2 (1-dim GR vs. 2-dim simple-structure GR model) 193

4.2.3 Condition 3 (1-dim GR vs. 2-dim complex-structure GR model) 199

4.2.4 Condition 4 (1-dim GR model vs. GR model for testlet)..... 206

4.3 RESULTS FROM REAL APPLICATION..... 212

4.3.1 QCAI Data 1 – AS91 212

4.3.2 QCAI Data 2 – AS92 222

4.3.3 QCAI Data 3 – BS92..... 228

5.0 DISCUSSION 236

5.1 SUMMARY OF MAJOR FINDINGS 236

5.1.1 Simulation Study 1..... 236

5.1.2 Simulation Study 2..... 244

5.1.3 Real Application 246

5.2 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS 248

APPENDIX A 251

APPENDIX B 254

APPENDIX C 255

APPENDIX D 259

APPENDIX E 261

APPENDIX F 263

BIBLIOGRAPHY 276

LIST OF TABLES

Table 3.1 Design and Conditions in Study 1 69

Table 3.2 Item Parameters of the IRT Models under Conditions 1-5..... 76

Table 3.3 Absolute Differences between Observed and Expected Proportions under GR Model 78

Table 3.4 Item Parameter Recovery under Unidimensional GR Model..... 79

Table 3.5 Factor Analyses of Generated 2-dimensional Simple-Structure Data..... 82

Table 3.6 Local Dependence Test (p-values of Chi-Square Statistics) in IRTFIT – Case 2 85

Table 3.7 Local Dependence Tests (Residual Correlations) in IRTFIT - Case 2..... 86

Table 3.8 Average Absolute Residual Correlations for Different Levels of Dependency 87

Table 3.9 Average Absolute Residual Correlations for Different Testlet Effects..... 89

Table 3.10 Expected and Observed Proportions for Two Misfitting Items..... 91

Table 3.11 Item Parameter Recovery using MCMC Estimation for the GR Model..... 96

Table 3.12 Design and Conditions in Simulation Study 2..... 108

Table 3.13 Item Parameter Recovery for 1P GR Model in WinBUGS..... 113

Table 3.14 Item Parameter Recovery for RS Model in WinBUGS..... 116

Table 3.15 Item Parameter Recovery for 2-dim Simple-Structure GR Model in WinBUGS 119

Table 3.16 Item Parameter Recovery for 2-dim Complex-Structure GR Model in WinBUGS . 123

Table 3.17 Item Parameter Recovery for Testlet GR Model in WinBUGS 126

Table 3.18 Misfitting Items Identified in Stone et al. (1993) and Stone (2000) 135
Table 4.1 RMSD for Item Parameter Recovery in WinBUGS for GR Model 137
Table 4.2 Median PPP-values and Average Proportions of Replications with Extreme PPP-values (< 0.05 or >0.95) when Ma=Mg=unidimensional GR 139
Table 4.3 Median PPP-values and Proportions of Replications with Extreme PPP-values for Item-Level Meausres when Ma=Mg=unidimensional GR 143
Table 4.4 Overall Median PPP-values and Average Proportions of Replications with Extreme PPP-values for all Measures – Condition 2 149
Table 4.5 Overall Median PPP-values and Average Proportion of 20 Replications with Extreme PPP-values for all Measures – Condition 3 158
Table 4.6 Overall Median PPP-values and Average Proportion of 20 Replications with Extreme PPP-values for all Measures – Condition 4 167
Table 4.7 Overall Median PPP-values and Average Proportion of Replications with Extreme PPP-values for all Measures – Condition 5 177
Table 4.8 RMSD for Item Parameter Recovery in WinBUGS for 2P GR Model 183
Table 4.9 Model Selection for Overall Test using DIC and Test-Level CPO – Condition 1 184
Table 4.10 Model Selection for Each Item using Item-Level CPO Index – Condition 1 185
Table 4.11 Number of Items with Extreme PPP-values across 20 Replications (Item-level Measures) 187
Table 4.12 Number of Item-pairs with Extreme PPP-values across 20 Replications (Pair-wise Measures) 188
Table 4.13 Median PPP-values for Each Item-level Measure across 20 Replications 189

Table 4.14 RMSD for Item Parameter Recovery in WinBUGS for 2-dim Simple-Structure Model 193

Table 4.15 Model Selection for Overall Test using Different Indices – Condition 2..... 194

Table 4.16 Model Selection for Each Item using Item-level CPO Index – Condition 2 197

Table 4.17 RMSD for Item Parameter Recovery in WinBUGS for 2-dim Complex-Structure Model 199

Table 4.18 Model Selection for Overall Test using Different Indices – Condition 3..... 200

Table 4.19 Model Selection of Each Item using Item-level CPO Index – Condition 3 202

Table 4.20 RMSD for Item Parameter Recovery in WinBUGS for Testlet GR Model 206

Table 4.21 Model Selection for Overall Test using Different Indices – Condition 4..... 207

Table 4.22 Model Selection for Each Item using Item-level CPO Index – Condition 4 209

Table 4.23 Item Parameter Estimates using WinBUGS and Multilog – AS91 213

Table 4.24 PPP-values for Item-level Measures based on GR and 1P GR Models – AS91 214

Table 4.25 Model Selection Indices for Overall Test – AS91 220

Table 4.26 Item-level CPO Index for Each Item – AS91 220

Table 4.27 Item Parameter Estimates using WinBUGS and Multilog – AS92 222

Table 4.28 PPP-values for Item-level Measures based on GR and 1P GR Models – AS92 224

Table 4.29 Model Selection Indices for Overall Test – AS92 227

Table 4.30 Item Parameter Estimates using WinBUGS and Multilog – BS92 228

Table 4.31 PPP-values for Item-level Measures based on GR and 1P GR Models – BS92..... 229

Table 4.32 Model Selection Indices for Overall Test – BS92 234

LIST OF FIGURES

Figure 2.1 Boundary Category Curves for a 5-category Item under the GR Model 15

Figure 2.2 Category Response Curves for a 5-category Item under the GR Model 16

Figure 2.3 Examples of Graphical Displays in PPMC by using Histograms 37

Figure 2.4 Examples of Graphical Displays in PPMC by using Scatter Plots 38

Figure 2.5 Graphical Description of Implementing the PPMC Method 39

Figure 2.6 History Plots Displaying Evidence of Convergence and Non-Convergence 44

Figure 2.7 Example of Observed and Predictive Test Score Distributions 50

Figure 3.1 Overall Steps in Conducting Simulation Study 1 74

Figure 3.2 Boundary Category Curves (BCCs) for Two Misfitting Items 90

Figure 3.3 Sampling History Plots of Item Parameters Associated with Two Chains - Item 1 93

Figure 3.4 "BGR" Diagrams for the Parameters of Item 1 94

Figure 3.5 Autocorrelation Plots for the Parameters of Item 1 95

Figure 3.6 Example Convergence Diagnostic Plots for Item Parameters under 1P GR Model . 112

Figure 3.7 Example Convergence Diagnostic Plots for Item Parameters under RS Model 115

Figure 3.8 Convergence Diagnostic Plots for Parameters under 2-dim Simple-Structure GR Model 118

Figure 3.9 Convergence Diagnostic Plots for Parameters under 2-dim Complex-Structure GR Model 122

Figure 3.10 Convergence Diagnostic Plots for Parameters under Testlet GR Model 125

Figure 4.1 Distributions of PPP-values for Each Discrepancy Measures under the Null Condition 140

Figure 4.2 Diagnostic Plots based on Test Score Distribution when Ma=Mg=unidimensional GR 142

Figure 4.3 Oberved vs. 90% Posterior Preditive Interval of Item-Total Correlation for Each Item when Ma=Mg=undimensional GR 144

Figure 4.4 Realized vs. Posterior Predictive Values of Item-Level Chi-Square Measure and Yen's Q_1 for Item 1 when Ma=Mg=unidimensional GR 144

Figure 4.5 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) when Ma=Mg= unidimensional GR 146

Figure 4.6 Display of PPP-values (based on a single dataset) for Yen's Q_3 (Left), and Item Covariance Residual (Right) when Ma=Mg= unidimensional GR 147

Figure 4.7 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) when Ma=Mg= unidimensional GR..... 148

Figure 4.8 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 and Item Covariance Residual (for a single data) when Ma=Mg= unidimensional GR..... 148

Figure 4.9 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 2 (ρ=0.6) 152

Figure 4.10 Display of PPP-values (based on a single dataset) for Yen's Q_3 (Left), and Item Covariance Residual (Right) - Condition 2 (ρ=0.6) 153

Figure 4.11 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (top), and Item Covariance Residual (bottom) (for a single data) – Condition 2 / Case 2 (ρ=0.6) 154

Figure 4.12 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) – Condition 2 / Case 2 (ρ=0.6) 155

Figure 4.13 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation (Left) and Histogram of Predicted SDs (for a single replication) for Case 1 (top) and Case 2 (bottom) – Condition 2 156

Figure 4.14 Diagnostic Plots based on Test Score Distribution (for a single data) – Condition2 /Case 1 157

Figure 4.15 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (for a single data) for Case 1 (top) and Case 2 (bottom) – Condition 3 161

Figure 4.16 Scatter plots of Realized vs. Posterior Predictive Values of Item Covariance Residual (for a single data) for Case 1 (top) and Case 2 (bottom) – Condition 3 162

Figure 4.17 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) for Case 1 (top) and Case 2 (bottom) – Condition 3 163

Figure 4.18 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 3/ Case 1 165

Figure 4.19 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 3/ Case 2 166

Figure 4.20 Scatter Plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (for a single data) for Case 1 (top) and Case 3 (bottom) – Condition 4 169
Figure 4.21 Scatter Plots of Realized vs. Posterior Predictive Values of Item Covariance Residual (for a single data) for Case 1 (top) and Case 3 (bottom) – Condition 4 169
Figure 4.22 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 6 with Other Items (for a single replication) for Case 1 (top) and Case 3 (bottom) – Condition 4 171
Figure 4.23 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 4/Case 1 172
Figure 4.24 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 4/Case 3 173
Figure 4.25 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation for Case 1 (top), Case 2 (middle), and Case 3 (bottom) based on a single replication – Condition 4 176
Figure 4.26 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_1 and Stone's X^2 Item-Fit Statistics (for a single data) – Condition 5 180
Figure 4.27 Display of Median PPP-values (left) and Proportion of 20 Replications with Extreme PPP-values (right) for Global OR (row1), Yen's Q_3 (row2), and Item Covariance Residual (row3) – Condition 5 181
Figure 4.28 Box-plots of DIC and Test-Level CPO across 20 Replications – Condition 1 184
Figure 4.29 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation for 2P GR (top), 1P GR (middle), and RS (bottom) Model 191

Figure 4.30 Display of Median PPP-values for Pair-wise Measures when fitting 2P GR (top), 1P GR (middle), and RS(bottom) models to the Data 192

Figure 4.31 Box-plots of Model Comparison Indices across 20 Replications – Condition 2 ... 195

Figure 4.32 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when Fitting 1-dim GR model (top) and 2-dim simple-structure GR model (bottom) to the Data..... 198

Figure 4.33 Box-plots of Model Comparison Indices across 20 Replications – Condition 3 ... 201

Figure 4.34 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when Fitting 1-dim GR Model (top) and 2-dim complex-structure GR Model (bottom) to the Data.. 204

Figure 4.35 Box-plots of Model Comparison Indices across 20 Replications – Condition 4 ... 208

Figure 4.36 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when fitting 1-dim GR Model (top) and testlet GR Model (bottom) to the Data 211

Figure 4.37 Example History and Autocorrelation Plots – AS91..... 213

Figure 4.38 Observed vs. Expected ICCs for Misfitting Items on the AS91 Form..... 217

Figure 4.39 Display of PPP-values for Pair-wise Measures when fitting GR Model (top) and 1P GR Model (bottom) to the data – AS91..... 219

Figure 4.40 Example History and Autocorrelation Plots – AS92..... 223

Figure 4.41 Observed vs. Expected ICCs for Misfitting Items on the AS92 Form..... 225

Figure 4.42 Display of PPP-values for Pair-wise Measures when fitting GR Model (top) and 1P GR Model (bottom) to the Data – AS92..... 226

Figure 4.43 Example History and Autocorrelation Plots – BS92..... 228

Figure 4.44 Observed vs. Expected ICCs for Misfitting Items on the BS92 Form 232

Figure 4.45 Display of PPP-values for Pair-wise Measure when fitting GR Model (top) and 1P GR Model (bottom) to the Data – BS92 233

ABBREVIATION TABLES

1P GR – one-parameter Graded Response

2P GR – two-parameter Graded Response

2-dim – two-dimensional

AIC – Akaike's Information Criterion

AP – Advanced Placement

BCC – Boundary Category Curve

BIC – Bayesian Information Criterion

CPO – Conditional Predictive Ordinate

CTT – Classical Test Theory

DF – Degrees of Freedom

DIC – Deviance Information Criterion

DIF – Differential Item Functioning

GPC – Generalized Partial Credit

GR – Graded Response

ICC – Item Category Curves

IRT – Item Response Theory

LD – Local Dependence

LI – Local Independence

MC – Multiple-Choice

MCMC – Markov Chain Monte Carlo

M-H – Metropolis-Hastings

MH – Mantel Hanzel

MIRT – Multidimensional IRT

ML – Maximum Likelihood

MME – Marginal Maximum Likelihood

NAEP – National Assessment of Educational Progress

NR – Nominal Response

OCC – Operating Characteristic Curve

OR –Odds Ratio

PA – Performance Assessment

PC – Partial Credit

PPMC – Posterior Predictive Model Checking

PPP-value – Posterior Predictive P-value

PsBF – Pseudo-Bayes Factor

RS – Rating Scale

WLS – Weighted Least Square

PREFACE

The following dissertation, while an individual work, benefited from the insights and direction of several people. First, I would like to express my deepest gratitude to Dr. Clement A. Stone, my advisor, for his guidance, caring, and tremendous support throughout my doctoral study. It is he who led me into the IRT model-fit area, sparked my interest in Bayesian methodologies, and introduced this interesting research topic to me. In addition, Dr. Stone provided timely and instructive comments and evaluation at every stage of my dissertation process. I truly appreciate his patient revision of my proposal and this dissertation.

Next, I would like to thank three other members in my dissertation committee, Dr. Suzanne Lane, Dr. Feifei Ye, and Dr. James E. Bost. Their insights and constructive suggestions substantially improved the quality of this project. I would also like to thank all the faculty members in the Research Methodology program for their excellent teaching and instruction. Special thanks go to Dr. Suzanne Lane and Dr. Clement A. Stone for helping me to build up solid foundation in educational measurement, and to Dr. Feifei Ye and Dr. Kevin H. Kim for helping me to improve my knowledge and skills in statistics.

Finally, I must thank my family members. Without their support, the completion of this dissertation would not have been possible. I would like to express my heartfelt appreciation to my parents, Zhen Ma and Shubang Zhu, for their unconditional love and endless support. My love and thanks go to my husband, Zhiwei Shan, and my son, Benjamin Shan. Whenever I felt

frustrated and tired during the course of my dissertation work, they were always there cheering me up and giving me confidence.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

Performance assessments (PAs) require students to perform tasks rather than select an answer from a developed list. They are intended to measure students' learning through emulating the context or conditions in which the intended knowledge or skills are actually applied (AERA, APA, & NCME, 1999). Due to their advantages over multiple-choice (MC) tests, there has been a significant expansion in the use of performance assessments, especially in large-scale assessment and accountability programs (Lane & Stone, 2006).

Item response theory (IRT) has become main-stream for analyzing item response data in educational and psychological measurement including performance assessment data. It consists of a family of statistical models which specify how an examinee's item responses are related to his/her latent traits and item properties (Embretson & Reise, 2000). Compared with classical test theory (CTT), IRT models make a number of strong assumptions such as dimensionality, local independence, and model-data fit. The inferences from applications of IRT models are valid only when the fit between model and data is satisfactory and the underlying assumptions are met. Therefore, it is crucial to check the adequacy of a chosen IRT model in order to validate applications of the model.

Evaluating applications of IRT to performance assessments is critical since in practice, unidimensional polytomous IRT models are commonly used to analyze performance assessment data but the underlying assumptions are more likely to be violated due to the properties of performance tasks. For example, the constructs measured in performance assessments are likely to be multidimensional. Large-scale performance assessments usually cover a broad range of content areas and each item in performance assessments often measures several skills simultaneously. The potential presence of local dependence (LD) may also be a more related issue to performance assessments than multiple-choice (MC) assessments. In MC tests, items are usually carefully designed to be independent of one another. In contrast, a setting or context related to a real life situation is usually used in performance assessments and students are asked several questions related to that setting (Yen, 1993). Thus, a set of items share the same stimulus and might depend on each other. Several potential sources of LD existing in performance assessments have been discussed by Yen (1993).

In many practical applications of IRT, there are several available models that might fit the data, and finding the best model for a particular application is desirable. For example, for a performance assessment which measure examinee's overall math ability across two content subdomains (e.g., algebra and geometry), a simple unidimensional polytomous IRT model and a more complicated 2-dimensional polytomous model might both fit the data. In order to know if a simple unidimensional model is adequate or if a multidimensional IRT model would be preferred for this particular performance assessment application, model comparison techniques should be employed.

In the last ten years, it has become increasingly common to use Bayesian methods for estimating IRT models in educational measurement. Part of this increased use is due to the

development of complex IRT models for different educational testing applications. Using traditional marginal maximum likelihood (MML) estimation method to estimate these complex models is difficult, and Bayesian estimation using Markov Chain Monte Carlo (MCMC) methods offer greater potential for estimating complex IRT models. Since Albert (1992) proposed a full Bayesian method based on Gibbs sampling to estimate 2-parameter normal-ogive IRT model, and Patz and Junker (1999a, 1999b) discussed Metropolis-Hastings (M-H) sampling algorithms to estimate several different IRT models such as 2PL, 3PL and mixed models, full Bayesian methods with MCMC algorithm have become widely used by many researchers to estimate a variety of complex IRT models such as testlet models (Bradlow, Wainer, & Wang, 1999; DeMars 2006; Li, Bolt, & Fu, 2006; Wang, 2002), rater-effect models (Patz & Junker, 2002), and multidimensional IRT models (Béguin & Glas, 2001; Bolt & Lall 2003; Yao & Schwarz, 2006).

In addition to using Bayesian methods to estimate IRT models, Bayesian methods can also be used to evaluate other aspects of IRT applications such as model fit and model comparison. Though a number of classical model-fit and model-comparison methods have been proposed and have been found to be useful in more traditional IRT applications, a similar interest in the assessment of IRT model-fit and IRT model comparisons from a Bayesian perspective is gaining more and more attention.

The Posterior Predictive Model Checking (PPMC) method (Rubin, 1984) is a popular Bayesian model checking tool and has proved useful with IRT models (e.g., Béguin & Glas, 2001; Fu, Bolt, & Li, 2005; Hoijtink, 2001; Levy, 2006; Sinharay, 2005, 2006; Sinharay, Johnson, & Stern, 2006). Conducting PPMC involves simulating data under a presumed model and comparing features of simulated data against observed data using discrepancy measures that

are sensitive to different aspects of misfit. Any systematic differences indicate potential misfit of the model. The rationale underlying PPMC is that if a chosen model fits the data, then observed data should look like replicated data generated from the posterior distributions of model parameters. Differences between observed and predicted data on discrepancy measures in PPMC can be evaluated using graphical displays as well as a numerical summary - Posterior Predictive P-value (PPP-value).

Compared with classical model-fit tests, the advantages of using PPMC for IRT model-fit are threefold: (1) PPMC takes into account uncertainty in parameter estimation by using posterior distributions for model parameters rather than point estimates; (2) PPMC constructs null sampling distributions empirically from MCMC simulations rather than relying on analytically derived distributions; (3) PPMC can be used for assessing the fit of complex IRT models which may be needed in real-world testing applications but can only be estimated using Bayesian methods.

Among a number of Bayesian model comparison indices, Pseudo-Bayes Factor (PsBF; Geisser & Eddy, 1979; Gelfand, Dey & Chang, 1992) and Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin & van der Linde, 2002) are popular indices for model comparisons with MCMC estimation. In IRT modeling, the PsBF index is commonly estimated using the conditional predictive ordinate (CPO). In addition, several researchers recently have found that the PPMC method was also effective for comparing different IRT models when MCMC estimation method was used (Béguin & Glas, 2001; Li et. al, 2006).

The purpose of this study was twofold: (1) to explore the performance of the PPMC method and various discrepancy measures in detecting threats to the use of unidimensional graded response (GR) IRT models to performance assessment applications, and (2) to investigate

the relative effectiveness of three Bayesian model-comparison methods (DIC, CPO, and PPMC) in choosing a preferred model for analyzing performance assessment data. Specifically, the following research questions were addressed:

(1) What is the Type-I error rate for each proposed discrepancy measure used with PPMC in assessing the fit of unidimensional GR model?

(2) What is the empirical power of each proposed discrepancy measure used with PPMC in detecting different aspects of misfit for unidimensional GR model?

(3) Among different types of discrepancy measures (test-level, item-level, and pair-wise measures) proposed in the current study, which measures are most effective in detecting specific misfit?

(4) Do the three Bayesian model comparison criteria (DIC, CPO, and PPMC) perform equally well in selecting the same model as the preferred model for a particular performance assessment data? If not, which criterion performs best?

(5) How do Bayesian model checking and model comparison methods work with data from a real performance assessment?

In order to answer these questions, two Monte Carlo simulation studies were conducted. Study 1 was intended to examine different discrepancy measures used in model checking with the PPMC method. Study 2 was designed to assess the different model comparison methods. In addition, the proposed Bayesian approaches to model-checking and model-comparison were further applied to several QUASAR's performance assessment datasets to examine their use with real data.

1.2 SIGNIFICANCE OF THE STUDY

In recent years, it has become increasingly common to use Bayesian method with MCMC for estimating IRT models, especially for complicated IRT models (e.g., Albert, 1992; Béguin & Glas, 2001; Bolt & Lall, 2003; Bradlow, et al. 1999; Patz & Junker 1999a, 1999b, 2002; Yao & Schwarz, 2006). However, relatively little attention has been given to assessing the fit of IRT models and comparing different IRT models from a Bayesian perspective.

Although PPMC has been previously used to assess IRT model fit (e.g., Béguin & Glas, 2001; Fu, Bolt, & Li, 2005; Hoijtink, 2001; Levy, 2006; Sinharay, 2005, 2006; Sinharay, Johnson, & Stern, 2006), the focus has been on unidimensional IRT models for dichotomous items. The present study was intended to extend previous research to polytomous IRT models and provide a comprehensive application of PPMC in the context of unidimensional GR models. This extension is very important because there has been a significant expansion in the use of performance-based items in educational testing and the unidimensional GR model is commonly used for modeling these items. Since the assumptions under the GR model are very likely to be violated in performance assessment applications, it is critical to check the fit of a GR model to a particular performance assessment data. In addition, many of the discrepancy measures used in the current study reflect polytomous extensions of measures used in previous research for dichotomous IRT models. Thus, it would be useful to assess the extent to which their performance with dichotomous items can be generalized to polytomous items. Finally, though PPMC is useful for simple unidimensional IRT models, its power is that it can be used for assessing the fit of complex IRT models which may only be estimated using Bayesian methods. However, research about applications of PPMC to complex IRT models has been very limited. In this current study, the PPMC method was also used to evaluate the fit of different complex

Bayesian IRT models such as 2-dimensional simple-structure and complex-structure GR models, and GR models for testlet. Thus, this study also extended previous research to the use of PPMC with complex IRT models.

Another objective of this study involved comparing Bayesian model-comparison criteria. Comparing different IRT models and choosing the more appropriate one is important to all testing applications including performance assessments. In practical applications, performance assessments are usually designed to only measure one dominant dimension and thus unidimensional polytomous models are commonly used. However, when the assumptions underlying unidimensional models are violated, more complex polytomous models might be used. Therefore, it is necessary and important to know if a simple or more complex model is more appropriate for a particular performance assessment application.

The research comparing different Bayesian model comparison indices has been limited. Sung and Kang (2006) conducted a study to compare four model selection methods (DIC, PsBF, AIC, and BIC) in terms of their effectiveness. They mainly focused on comparing the different unidimensional polytomous models for Likert-type data. In addition, the PPMC method was not considered in their study. Li et al. (2006) investigated the performance of Bayesian tools (DIC, PsBF, and PPMC) in choosing the true testlet models for dichotomous items. Since the results from these studies indicated the differential performance of these model comparison indices, it is necessary to compare their relative performance in different testing applications. The current study played a significant role in extending the previous research to performance assessment settings that consider different polytomous models which may be more appropriate for performance assessment data including both unidimensional and complex GR models.

1.3 LIMITATIONS OF THE STUDY

This study explored the general performance of the PPMC method in detecting different aspects of misfit for the unidimensional GR models, and also investigated the effectiveness of the different Bayesian model-comparison indices in selecting the true models for performance assessment data using two Monte Carlo simulation studies. Though the conditions were carefully designed and the factors were fixed at realistic values, the results may not generalize to other situations not considered in the current study. For example, this study is limited in terms of the length of tests (15 items), the number of response category (5-category), the polytomous model (GR), and the number of dimensions considered for multidimensional conditions.

Another limitation is that due to computing constraints of the WinBUGS program (Spiegelhalter, Thomas, Best, & Lunn, 2003) and the large number of conditions in this study, only 20 replications were implemented. Though it was smaller than which is typical for other Monte Carlo research, it was typical for previous research involving PPMC and Bayesian model-comparison applications (e.g., a number of researchers used 5 to 30 replications).

In addition, the performance of the PPMC method and the Bayesian model-comparison indices for the GR models requires further study. For example, the effect of factors such as sample size, the number of total items, the number of dimensions, the structure of dimensions, and the inter-dimensional correlation given modeled multidimensionality could be further explored. Other discrepancy measures could be proposed and evaluated. For example, the conditional odds ratios could be used. Other assumptions under the use of IRT models with performance assessments could be also considered in the future such as the normal ability assumption. Finally, the current study did not compare the performance of classical model-fit statistics with the performance of PPMC. Further research could explore this comparison.

2.0 REVIEW OF LITERATURE

This chapter provides the theoretical background for this study which is organized into five sections: 1) applications of IRT to performance assessments, 2) traditional methods for checking IRT model-fit, 3) posterior predictive model checking (PPMC) in Bayesian framework, 4) checking IRT model-fit using PPMC, and 5) model comparison in Bayesian framework.

2.1 APPLICATIONS OF IRT TO PERFORMANCE ASSESSMENTS

2.1.1 Brief Introduction to Performance Assessments

The recent trend in educational testing is moving from exclusively using multiple-choice items to including performance assessment items. Performance assessment (PA) is a form of testing that requires students to perform tasks rather than select an answer from a developed list. It is intended to measure students' learning through emulating the context or conditions in which the intended knowledge or skills are actually applied (AERA, APA, & NCME, 1999). PA is also termed "authentic assessment" since it often provides tasks that are thought to model realistic applications that students will encounter in life. The performance-based items usually have two parts: a clearly defined task and a list of explicit criteria (i.e., rubric) for assessing student performance or product. The responses are constructed by examinees and scored on a response

scale with several levels rather than only as correct or incorrect. PA includes a large range of formats such as constructed-response, essays, experiments, and portfolios.

Lane and Stone (2006) summarized the main advantages of performance assessments: (1) *directness*: they provide a more direct measure of the skills of interest; (2) *meaningfulness:* they are meaningful and thus motivating students because of their relevance to real-life situations; (3) they may influence curriculum and instructional changes in positive ways by encouraging teachers to broaden the focus of their teaching and include reasoning, problem solving, and communication in regular classroom activities. Moreover, performance assessments can measure important skills that cannot be assessed by selected-response item format - for example, assessing dynamic cognitive processes. Therefore, it may be argued that performance assessments provide more valid information about student learning than multiple-choice assessments (Baron, 1991).

Due to the aforementioned benefits, in the last decades there has been a significant expansion in the use of performance assessments, especially in large-scale assessment and accountability programs (Lane & Stone, 2006). Many school districts, state testing programs, and national assessments have incorporated performance assessments into their programs. For example, the National Assessment of Educational Progress (NAEP) is the nationally representative and continuing assessment of what students know and can do in various subject areas. Some NAEP items are performance-based. The Advanced Placement (AP) exams consist of one-section constructed-response items which are used to determine the proficiency attached by high school students in college courses. Besides the national assessments, a number of state assessment programs contain both selected-response items and performance-based items (e.g., Kentucky, Pennsylvania, and Vermont), while others are even entirely performance-based (e.g.,

Maryland). Though performance assessments have been widely used in large-scale assessments for high-stake purposes such as providing school accountability information, evaluating reform efforts, and determining instructional and curriculum changes, they can also be useful for classroom purposes such as diagnosing student's strength and weakness and evaluating the effectiveness of instruction. Lane and Stone (2006) pointed out that classroom performance assessments allow for a direct alignment between assessment and instructional activities and have the potential to simulate the criterion performance better than large-scale assessments.

2.1.2 IRT Models for Performance Assessments

2.1.2.1 General Description

IRT consists of a family of statistical models which are used to analyze item response data. These IRT models can be classified in several ways. One way is by the type of item data. Dichotomous IRT models are used for analyzing dichotomous item data (item response scored in two categories), and polytomous IRT models are used to analyze polytomous item data (item response scored in more than two categories). Another way is by the number of ability dimensions accounting for performance differences among examinees. Unidimensional IRT models assume one underlying dimension, while multidimensional IRT models assume more than one dimension determining examinees' performance. Performance assessment tasks are typically polytomously scored and generally measure one underlying ability dimension, thus unidimensional polytomous IRT models are commonly applied to performance assessments.

There are various unidimensional polytomous IRT models available. The most commonly used polytomous models include (1) the graded response (GR) model (Samejima, 1969); (2) the modified GR model (Muraki, 1990), also called Muraki's rating scale (RS) model;

(3) the partial credit (PC) model (Masters, 1982); (4) the generalized partial credit (GPC) model (Muraki, 1992); (5) the rating scale (RS) model (Andrich, 1978); and (6) the nominal response (NR) model (Bock, 1972). According to the useful taxonomy provided by Thissen and Steinberg (1986) for classifying polytomous models, The GR and Muraki's RS models belong to a class of "difference" models, and the remaining models are classified as "divide-by-total" models. For "difference" models, the probability of responding in a particular category *j* is calculated by taking the difference between cumulative probabilities: for example, the probability of responding at or above *j* and the probability of responding at or above (*j+1*). For "divide-by-total" models, the probability of responding in a given category is obtained by the ratio of the function for that category to the sum of the functions for all the categories (Yen & Fitzpatrick, 2006). Bock's NR model is the most general "divide-by-total" model, and all the other models (PC, GPC, and Andrich's RS) are special cases of the NR model. In addition, the PC and Andrich's RS models are Rasch-based models assuming a constant item discrimination or slope parameter for all items.

For performance assessments, all of the aforementioned models could be used because they are applicable to items with ordered response categories. Nevertheless, the GR, PC, GPC, and NR models are more commonly used because they can be used to analyze a set of polytomous items that differ in the number of score levels. For example, either model could be applied to a test having some items with 5-point rubrics and some with 4-point rubrics. While the two RS models are simplified models, they are only suitable for items associated with the same rating scales and therefore are rarely used with performance assessments. However, Lane and Stone (2006) argued that the rating scale models could be applied to performance assessments if a general rubric is used as the basis for developing specific item rubric since the response scales

and the differences between score levels may be the same across the set of items. They also pointed out that the NR model may not be preferred with performance assessments due to the relatively large number of parameters to be estimated.

2.1.2.2 Graded Response Model (Samejima, 1969)

Samejima's (1969) GR model is the main model applied in this study and is introduced in more detail here. The GR model is an extension of dichotomous 2-paramter logistic (2PL) model and was developed to model items with more than two graded or ordered response categories. Let denote $K_i = (m_i+1)$ to be the number of ordered response categories for item i, with higher response category indicating higher ability, then examinees would receive item scores of x = 0, 1… m_i on this item. Samejima (1969) proposed a two-stage process to obtain the probability that a given examinee with a certain ability level will receive item score x. In the first stage, the response categories of each item are dichotomized into two overall categories: (1) equal to or greater than category score x; and (2) less than category score x. For instance, for a 5-category item, there are 4 types of dichotomies: (1) 0 vs. 1, 2, 3, 4; (2) 0, 1 vs. 2, 3, 4; (3) 0, 1, 2 vs. 3, 4; (4) 0, 1, 2, 3 vs. 4. The probability that an examinee receives a category score x (x = 1, 2… m_i) or higher on item i ($P_{ix}^{*}(\theta)$) can be modeled using the 2PL function:

$$P_{ix}^{*}(\theta) = \frac{\exp\left[Da_i(\theta - b_{ix})\right]}{1+\exp\left[Da_i(\theta - b_{ix})\right]}, \tag{2.1}$$

where

D is the scaling constant (1.7 or 1),

a_i is the discrimination (or slope) parameter of item i,

θ is the ability level, and

b_{ix} is the threshold parameter for category x of item i.

The b_{ix} parameter represents the ability level at which examinees have a .50 probability of receiving item score x or higher on item i. For an item with (m_i+1) categories, one item discrimination parameter (a_i) and m_i threshold parameters (b_{ix}) must be estimated under the GR model. For each threshold parameter, there is one corresponding "operating characteristic curve" (OCC; Embretson & Reise, 2000) or "boundary category curve" (BCC) described by $P_{ix}^{*}(\theta)$.

Once these cumulative probabilities $P_{ix}^{*}(\theta)$ are estimated, the probability of responding to a particular response category $P_{ix}(\theta)$ (x = 0, 1, 2… m_i) can then be computed using the difference between the cumulative probabilities for two adjacent categories: $P_{ix}^{*}(\theta)$ and $P_{i(x+1)}^{*}(\theta)$. $P_{ix}^{*}(\theta)$ is known to be the probability of an examinee obtaining item score equal to or higher than x conditional on ability level, and $P_{i(x+1)}^{*}(\theta)$ represents the probability of that examinee obtaining item score higher than x. The difference is the probability of receiving the actual item score x.

Consider a 5-category item, Equation (2.1) defined the four cumulative probabilities: $P_{i1}^{*}(\theta)$, $P_{i2}^{*}(\theta)$, $P_{i3}^{*}(\theta)$, and $P_{i4}^{*}(\theta)$. In order to calculate the probabilities of obtaining the lowest (0) and highest (4) item scores, two additional definitions should be given: the probability of responding in or above the lowest category score (x = 0) is defined as $P_{i0}^{*}(\theta)=1$, and the probability of responding above the highest category score (x = 4) is $P_{i4}^{*}(\theta)=0$. Thus, the probability of responding in each of the five categories (x =0, 1… 4) can be calculated using:

$$\begin{cases} P_{i0}(\theta)=1-P_{i1}^{*}(\theta) \\ P_{i1}(\theta)=P_{i1}^{*}(\theta)-P_{i2}^{*}(\theta) \\ P_{i2}(\theta)=P_{i2}^{*}(\theta)-P_{i3}^{*}(\theta) \\ P_{i3}(\theta)=P_{i3}^{*}(\theta)-P_{i4}^{*}(\theta) \\ P_{i4}(\theta)=P_{i4}^{*}(\theta)-0 \end{cases} \tag{2.2}$$

The general formula for computing the category response probabilities for an item with (m_i+1) categories (item score x = 0, 1, 2… m_i) is as follows:

$$\begin{cases} P_{i0}(\theta) = 1 - P_{i1}^{*}(\theta) \\ P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i(x+1)}^{*}(\theta) \qquad (x = 1, 2\ldots (m_i\text{-}1)). \\ P_{im_i}(\theta) = P_{im_i}^{*}(\theta) - 0 \end{cases} \tag{2.3}$$

For illustrative purposes, Figure 2.1 displays the four boundary category curves ($P_{ix}^{*}(\theta)$) for a 5-category item (a = 1.7, b_1 = -2, b_2 = -1, b_3 = 0, b_4 =1), and the category response curves ($P_{ix}(\theta)$) for this item are shown in Figure 2.2. Under the GR model, the item parameters determine the shape and location of the boundary category curves and category response curves.

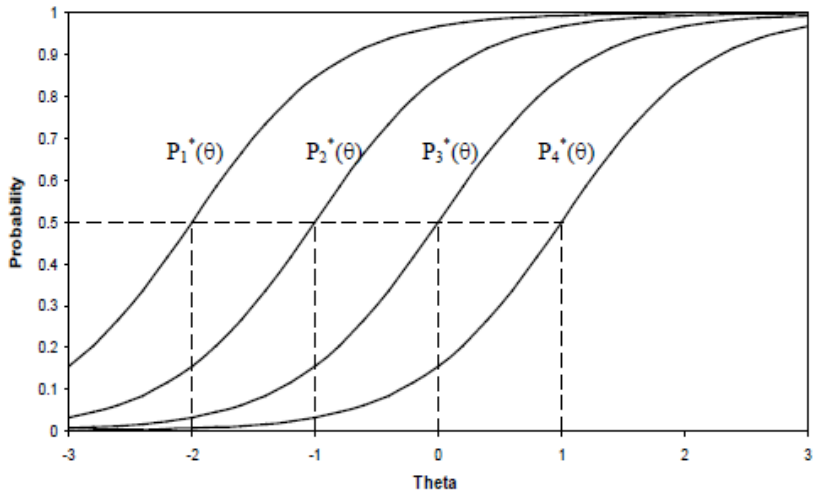


Figure 2.1 Boundary Category Curves for a 5-category Item under the GR Model

For boundary category curves (Figure 2.1), the slope parameter (a_i) determines the steepness of the operating curves: the higher the slope parameter, the steeper the curves. Higher slope indicates that the response categories discriminate or differentiate the examinees at different ability levels fairly well. It should be noted that under the GR model, the slope a_i varies by item i, but within an item, all response categories share the same slope which results in

parallel operating characteristic curves. The constraint of equal slopes within an item prevents negative probabilities for $P_{ix}(\theta)$.

The threshold parameters b_{ix} determine the location of $P_{ix}^{*}(\theta)$. From the intersections of dashed lines in Figure 2.1, it is evident that the threshold represents the ability level at which an examinee has a .50 probability of receiving item score of *x* and higher. For instance, the first threshold for this example item b1 is -2 which means an examinee at ability level of -2 has .50 chance of obtaining a score of 1 or higher on this item. Moreover, the range of threshold values dictates the spread of the boundary category curves. A large range of threshold values results in curves that are more spread out, whereas, a small range of threshold values results in curves that fall closer together. It should be also noted that within an item the threshold parameters are ordered with the constraint $b_{i(x-1)} < b_{ix} < b_{i(x+1)}$. This is a requirement for the GR model, but not for other models such as the PC model.

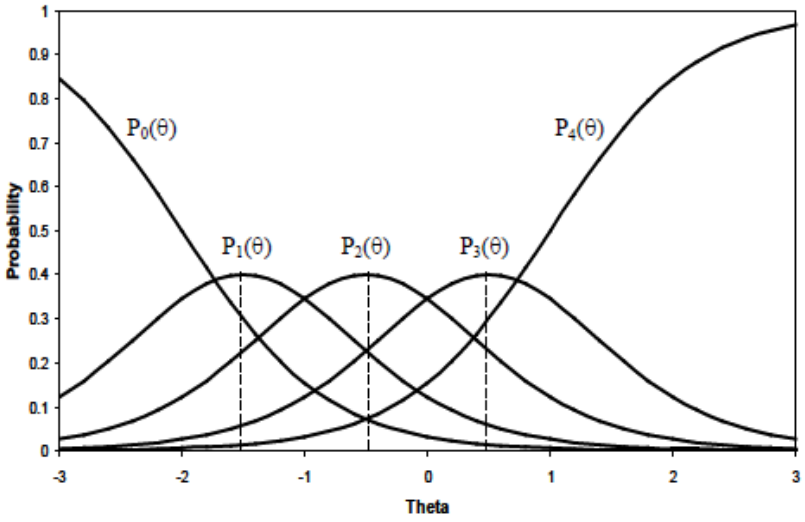


Figure 2.2 Category Response Curves for a 5-category Item under the GR Model

For this 5-categeory item, there are 5 category response curves showed in Figure 2.2. The curve for the lowest category (0) is monotonically decreasing, whereas, the curve for the highest category (4) is monotonically increasing. The curves for the middle three categories (1, 2, and 3)

are bell-shaped. Under the GR model, the slope parameter determines the shape of these curves for the middle categories: the higher the slope parameter, the narrower and more peaked the curves. The threshold parameters determine the locations of the curves for the middle response categories. Specifically, these category response curves peak in the middle of two adjacent threshold parameters. As showed using dashed lines in Figure 2.2, the middle value of two threshold values (b1 = -2 and b2 = -1) is -1.5 which is the mode of the curve for category score 1.

2.1.3 Main Threats in Applying Unidimensional IRT Models to PAs

Due to the well-known advantages of IRT over classical test theory (CTT), IRT has become main-stream for analyzing item response data in educational and psychological measurement. However, IRT is based on many strong assumptions such as dimensionality, local independence, and model-data fit. The inferences about the applications of any IRT model are valid only when all the underlying assumptions for that model are met. Therefore, before any accurate inference is drawn, it is necessary to check the assumptions in order to validate applications of IRT models. It is especially true when unidimensional polytomous IRT models are applied for performance assessments because the assumptions are more likely to be violated due to the properties of performance tasks. This section discusses some main threats in applying unidimensional IRT models to performance assessments.

2.1.3.1 Multidimensionality

Most of the commonly used IRT models assume that one ability dimension determines examinees' performance. However, the constructs measured in performance assessments are very likely to be multidimensional and this multidimensionality is mainly due to the complexity

of performance tasks. Performance tasks are typically developed to measure the complex structure of multiple skills and knowledge needed for solving more realistic problems. Thus, each item in performance assessments usually measures several dimensions simultaneously. For example, a mathematics problem might focus on problem-solving and communication abilities. In order to do well on that problem, students must be able to not only solve the problem, but also communicate their ideas clearly.

Another case is that large-scale performance assessments usually cover a broad range of content areas. For example, a math assessment may measure two content areas: algebra and geometry. Though this test measures student's overall math ability and a unidimensional IRT model is commonly applied, the responses to this test is actually 2-dimensional.

In addition to this planned content structure, many nuisance or construct-irrelevant factors would result in multidimensionality for performance assessments. For example, a performance assessment intended to measure only mathematics ability might also require examinee reading ability. When there is variability on reading ability among the examinees, the reading ability would be viewed as nuisance dimension. Moreover, performance tasks are designed to be contextual or have real-life applications. The degree to which a student is familiar with a specific context would affect his/her performance. If the context effect varies across examinees, it would introduce an additional nuisance dimension. Furthermore, performance tasks often take more time to respond, and if the testing time was inadequate for some examinees, "speededness" would result in another potential construct-irrelevant dimension.

Finally, performance tasks are typically combined with multiple-choice items in order to measure examinees' abilities more accurately. The combination of different item formats would

result in multidimensionality because different formats might measure different level of cognitive processing (Lane & Stone, 2006; Tate, 2002).

In summary, multidimensionality in item responses for performance assessments can be easily caused by various factors such as planed test construct structure, unintended nuisance or construct-irrelevant variances, and mixed item format. Unfortunately, in many practical situations, this multidimensionality is completely ignored and the unidimensional models are often applied to performance assessments. The lack of applications of multidimensional IRT (MIRT) models is due to the difficulties in parameter estimation and the interpretation of the latent ability space, as well as no user-friendly software available for estimating MIRT models (DeAyala, 1994).

When a unidimensional IRT model is used to fit multidimensional data, several problems might arise. Several researchers (Ackerman, 1989; Ansley & Forsyth, 1985, Way, Ansley, & Forsyth, 1988) have investigated the consequences of fitting 2-dimensional dichotomous item data with unidimensional three-parameter logistic (3PL) models and found violations of the unidimensionality assumption clearly affected IRT parameter estimates. DeAyala (1994, 1995) extended the previous work on the influence of multidimensionality on dichotomous model parameter estimation to polytomous models including the GR model and the PC model. For example, it was found that for the GR model, the difficulty parameters were well estimated, the discrimination estimate more accurately estimated the average discrimination than either dimensional a_1 or $a_{2,}$ and the single ability estimate also estimated the average more accurately than either dimensional ability. Using incorrect model parameter estimates would subsequently affect IRT applications such as equating, CAT, as well as the validity of ability score interpretations. Tate (2002) summarized the previous studies and discussed that unidimensional

ability estimates represent a target composite of abilities, and is only robust to violations of the unidimensional assumption when the correlations among ability dimensions are moderate or high. Otherwise, the validity of any inferences from the single ability estimate will be threatened and it may not be appropriate to use unidimensional model (Lane & Stone, 2006).

Reckase (1985) found that difficulty and dimensionality can be confounded in the data and thus the composite of abilities does not remain consistent across the ability scale. For example, if the easy items measured one ability and the hard items measured another ability, low and high scores on the ability scale would not have the same meaning as could be a serious threat to validity of the total score. In addition, Ackerman (1992) demonstrated how the items may display differential item functioning (DIF) if unidimensional model is used to scale multidimensional data. Walker and Beretvas (2001, 2003) found the open-ended items in a large-scale mathematics test functioned differentially in favor of students who were highly capable of communicating their ideas and they further explored the effect of using only a single score on student proficiency classifications in mathematics. Their results indicated that when data believed to be multidimensional are modeled using a unidimensional model, different inferences may be made about student proficiency. Examinees having less mathematics communication ability were more likely to be placed in a lower general mathematics proficiency classification under the unidimensional than multidimensional model.

2.1.3.2 Local Dependence

Local independence (LI) is a fundamental assumption for IRT models which means that there is no relationship between examinees responses to different items after accounting for trait abilities measured by a test. This conditional independence can be expressed mathematically as:

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta). \quad (2.4)$$

It describes that the probability of any pattern of responses to all items (**x**), conditioned on the abilities (θ), is equal to the product of the conditional probabilities of the response to each item. This equation defines the strong form of local independence. A weak form of local independence was proposed by McDonald (1979): the conditional covariances of all pairs of item responses on the abilities are equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on the abilities, is the product of the probabilities of responses to these two items,

$$P(X_i = x_i, X_j = x_j \mid \theta) = P(X_i = x_i \mid \theta) P(X_j = x_j \mid \theta). \quad (2.5)$$

This is a weaker form because higher-order dependencies among items are allowed.

An even weaker form of local independence was proposed by Stout (1987) who called it as "essential item independence" and defined it as "the items in a test can be considered as essentially independent when the average value of the conditional covariances between items approaches zero as test length increases for all ability values". It is a weakest form of local independence since it only requires the average value of covariances rather than all covariances close to 0.

A number of researchers have discussed that the local independence assumption is related to the dimensionality assumption. The strong form indicates that the abilities measured by a test completely explain the difference on examinees' performances. The weak local independence implies that the abilities completely explain the covariance between all item pairs between all item pairs. Finally the essential independence implies that the abilities dominate the difference on examinees' performances (Yen & Fitzpatrick, 2006).

The potential presence of local dependence (LD) may be a more related issue to performance assessments (PA) than multiple-choice (MC) assessments. In MC tests, the items are usually carefully designed to be independent of one another. In contrast, a setting or context related to a real life situation is usually established in PA and students are asked several questions related to that setting (Yen, 1993). Yen (1993) discussed several potential sources of LD in PA such as: external assistance or interference with some items, speededness, fatigue, practice, special item or response format, a shared stimulus or passage, item chaining, items requiring explanation of a previous answer, scoring rubrics or raters, unique content knowledge or abilities, and differential opportunity to learn. Most of these sources reflect an additional nuisance factor (person, item, or rater characteristics) that consistently affects the performance of some students on some items to a great extent, and some sources reflect item interactions such as item chaining and a shared stimulus (Lane & Stone, 2006; Yen, 1993). Several studies (Yen, 1993; Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999) have showed some sources of LD can cause very strong empirical LD.

IRT models are not robust to the violation of local independence assumption. Applying an IRT model to LD response data could cause serious problems. First, the parameter estimates may be biased because the likelihood function for IRT models is based on local independence assumption and the incorrect likelihood would affect the accuracy of parameter estimation. Yen (1993) demonstrated that positive LD would produce higher item discriminations for LD items. Thus, the test information may be overestimated, and the standard errors of test scores would be underestimated. These effects would subsequently affect any application of IRT models. For example, the biased item discrimination estimates would affect item banking, and the underestimated standard errors would cause the premature termination in case of CAT.

In summary, the potential for violations in the assumptions of unidimensionality and local independence may be more likely for performance assessments and the consequences of these violations can not be ignored. Therefore, it is very important to check these two assumptions before a unidimensional IRT model is applied to a performance assessment data.

2.2 TRADITIONAL METHODS FOR CHECKING IRT MODEL-FIT

Assessing the fit of IRT models is a multi-facet procedure that often involves the collection of evidence about different aspects of fit: (1) assessing IRT model assumptions such as unidimensionality and local independence; (2) assessing the goodness-of-fit of IRT models at the item, person, and test levels (Embretson & Reise, 2000). A variety of methods have been proposed for assessing the corresponding different aspects of fit. This section reviews traditional approaches to checking the assumptions of unidimensionality and local independence and evaluating the goodness-of-fit at item level for polytomous IRT models because these three aspects are of the main interest in the present study.

2.2.1 Assessing Dimensionality

Several methods have been developed for assessing the dimensionality of polytomously scored items and most of them are polytomous extensions of methods for dichotomous item response. These methods fall into three categories: (1) factor analytic methods; (2) multidimensional IRT methods; (3) nonparametric methods.

Common linear factor analysis using Pearson product-moment correlations with maximum likelihood (ML) estimation can only be applied when the response scale for polytomous items has a large number of response categories and can be treated as a continuous interval scale. Several factor analytic methods have been proposed specifically for ordinal response data. For example, a weighted least square (WLS) analysis of polychoric correlations has been developed and can be implemented in PRELIS/LISREL (Joreskog & Sorbom, 2006) and Mplus (Muthén & Muthén, 2006). WLS requires a weight matrix which involves the inverse of the covariance matrix of polychoric correlations. The size of the weight matrix is usually substantial and it grows dramatically as the number of items increases. As a result, an adequate estimate of the weight matrix requires a very large sample size. When the sample size is small or moderate, a robust WLS approach (Muthén, duToit, & Spisic, 1997) is considered as the best approach for factor analysis of ordinal variables. The robust WLS approach uses the identity matrix rather than the weight matrix and its estimation does not require extensive computation and enormously large sample sizes. Two robust WLS methods (mean-adjusted WLS and mean- and variance-adjusted WLS) are available in Mplus. In a simulation study, Flora and Curren (2004) showed that WLS performed adequately only at the largest sample size but led to substantial estimation difficulties with smaller samples, whereas, the robust WLS performed well across all simulation conditions.

Compared with factor analytic methods, MIRT approaches use all information in response patterns rather than limited information from correlation matrices. A full-information item factor analysis for polytomous item responses was proposed by Muraki and Carlson (1995) and this method can be implemented in the most recent version of PRELIS/LISREL (Joreskog & Sorbom, 2006). Another is a Rasch MIRT modeling approach proposed by Adams, Wilson, and

Wang (1997) which assumes the slope or discrimination parameter is constant across all items. This method is available in ConQuest (Wu, Adams, & Wilson, 1998).

The unidimensionality assumption indicates that there is a single latent ability measured by a particular test. However, a real-world test will never be strictly unidimensional. Given this fact, Stout (1987) proposed the concept of "essential unidimensionality" for a test which measures a dominant dimension and examinees' performances are unaffected by the presence of minor dimensions. This concept is directly related to "essential local independence" discussed in section 2.1.3.2. To assess whether a test is essential unidimensional for applying a unidimensional IRT model, nonparametric approaches have been developed by Stout and his colleagues (1987, 1990, 1993, & 1996) based on conditional item covariance theory. The simple hypothesis that a test is essentially unidimensional can be examined using DIMTEST software (Nandakumar & Stout, 1993). Poly-DIMTEST (Nandakumar, Yu, Li, & Stout ,1998) is an extension of DIMTEST to accommodate tests that contain polytomous items. DETECT program (Zhang & Stout, 1999) provides more information than DIMTEST by estimating the extent of multidimensional approximate simple structure in a test. Poly-DETECT (Yu & Nandakumar, 2001) is a polytomous extension of DETECT. In addition, HCC/CCPROX program (Roussos, Stout, & Marden, 1998) is used to search dimensionally homogeneous clusters of items using hierarchical cluster analysis technique, and its polytomous version is Poly-CCPROX/HCA (Tay-Lim & Stone, 2000).

2.2.2 Detecting Local Dependence

The IRTNEW software (Chen, 1998) provides five different measures of item local dependence (LD) for dichotomous items. All of them are IRT based and examine LD in the context of

unidimensional IRT models. The first one is Yen's Q_3 statistic (1984, 1993) which measures correlations between pairs of items after accounting for the latent ability. To calculate Q_3, the expected performance of the i^{th} examinee on item j (E_{ij}) is first obtained based on the IRT model:

$$E_{ij} = \sum_{k=1}^{m} (k-1) P_{jk}\left(\hat{\theta}_i\right), \tag{2.6}$$

where m is the total number of response category of item i, and $P_{jk}\left(\hat{\theta}_i\right)$ is the probability of an examinee at ability level $\hat{\theta}_i$ responding in category k. The deviation (d_{ij}) between observed and expected performance is then calculated as:

$$d_{ij} = x_{ij} - E_{ij}. \tag{2.7}$$

For items j and j', the Q_3 is defined as the correlation of deviation scores across all examinees:

$$Q_{3jj'} = r\left(d_j, d_{j'}\right). \tag{2.8}$$

When no local dependence exists, Yen (1984, 1993) suggested the Fisher's Z transformation of the Q_3 index would be approximately distributed normal with a mean of 0 (with a slight negative bias of -1/(n-1)) and a variance of about 1/(n-3), where n is the number of items. However, in a simulation study, Chen and Thissen (1997) found that the empirical distribution of Q_3 did not match this theoretical sampling distribution and produced Type-I error rates that were sufficiently larger than the nominal levels. As a result, it is more common to use Q_3 as a descriptive statistic than a hypothesis testing. The Q_3 values greater than a uniform cutoff value 0.20 generally indicate some degree of local dependence worthy of attention (Chen & Thissen, 1997; Yen & Fitzpatrick, 2006). Yen's Q_3 can be used with either dichotomous or polytomous items.

Four other LD indices in IRTNEW are based on 2 x 2 contingency tables of the observed and expected frequencies of correct and incorrect responses for the item pairs. They include two chi-square indices (the Pearson χ^2 and the likelihood ratio G^2 statistics), as well as two signed measures (the standardized coefficient difference ϕ index and standardized log-odds ratio difference index). An advantage of the signed measures over chi-square indices is that their sign can indicate the direction of association. These four indices were developed by Chen and Thissen (1997) mainly for dichotomous items, but can be extended to accommodate polytomous items based on tests of associations in general m x n contingency tables. Kim, Cohen and Lin (2006) developed a computer program LDIP to detect LD for polytomous items specifically. It provides four indices: Yen's Q_3, the Pearson χ^2, the likelihood ratio G^2, and the Fisher-transformed correlation difference statistic Z_d.

2.2.3 Evaluating Item-Fit

Model-data fit can be evaluated at different levels such as at the test, item or person level. More statistical procedures have been developed to evaluate item fit rather than overall model-fit (Embretson & Reise, 2000). One reason is that a test may include mixed item types and different IRT models need to be used for different type of items. Another reason is that even when the overall model fits the data, some of the items do not function in the intended manner. Inadequacy of model-data fit may have adverse consequences in the applications of IRT models such as biased ability estimates, unfair ranks, and wrongly equated scores (Yen, 1981; Wainer & Thissen, 1987). In addition, item fit analysis can help test constructers to isolate bad items in item pools and retain only items that fit an IRT model. Therefore, assessing model fit at item level is very important.

2.2.3.1 Traditional Item-Fit Statistics

Assessing item-fit involves evaluating the degree to which the model predicts the observed item responses or the degree to which observed item category curves (ICC) agrees with the form predicted by a model. The traditional statistical procedure for assessing item fit for each item is as follows:

(1) Obtain estimates of item and ability parameters for the IRT model;

(2) Specify a number of ability subgroups to approximate the continuous ability scale;

(3) Construct an observed score response distribution by cross-classifying examinees based on their ability estimates and responses and calculating the proportion of examinees responding to each response category.

(4) Construct an expected score response distribution by calculating the probability of response of each subgroup to each response category based on IRT model, item parameter estimates, and an ability estimate representing each group ability.

(5) Evaluate difference between the observed and expected response distribution through chi-square statistics.

Several traditional item-fit statistics that can be used with dichotomous and polytomous items have been proposed based on this procedure. They include Bock's Pearson statistics (1972), Yen's Q_1 statistic (1981), and McKinley and Mills's likelihood ratio statistics G^2 (1985). The following is the simple description of Yen's Q_1 statistic which will be used in the current study. The other statistics are very similar to Yen's Q_1 and only differ in the number of ability subgroups, the methods for constructing the subgroups, or the methods for obtaining the expected proportion.

Yen's Q_1 Statistic (Yen, 1981)

Yen's item-fit statistic is a Pearson chi-square test statistic defined as:

$$\chi^2 = \sum_{j=1}^{10} \sum_{k=1}^{K} N_j \frac{\left(O_{jk} - E_{jk}\right)^2}{E_{jk}}, \tag{2.9}$$

where N_j is the number of examinees within ability subgroup j, O_{jk} and E_{jk} are the observed and predicted proportion of responses to category k for ability subgroup j, respectively. In Yen's statistic, examinees are divided into 10 ability subgroups of approximately equal size after they are rank-ordered by their ability estimates. The expected proportion to a response category for a subgroup is the mean of the probabilities of responses to that category for all the examinees in that subgroup. The distribution of Yen's Q_1 statistic is assumed to have an approximate chi-square distribution with degrees of freedom equal to *10*(K-1)-m*, where *m* is the number of estimated item parameters in the model.

Although the traditional chi-square statistics are useful for detecting item fit in many situations and have been widely used, several problems with them have been identified in the literature. The first problem is whether or not these item-fit statistics follow a known chi-square distribution. IRT-based item-fit statistics are not constructed in the same way as classical goodness-of-fit test statistics in which both variables are known and therefore the observed proportions are based only on observed data. In contrast, for the IRT-based statistics, an IRT model is firstly estimated. Cross-classification of examinees is then based on the ability estimates. The model-dependent observed proportions would cause uncertainty about using a chi-square distribution (Orlando & Thissen, 2000). Further, Stone (2000) pointed out that the model-based expected proportions are also dependent on unknown model parameters. Using estimated values rather than true values of parameters may also affect the chi-square approximation to the distributions of item-fit statistics. Finally, it is not entirely clear what degrees of freedom (DF) should be used for the null chi-square distribution. Though expected

proportions in traditional item-fit statistics depend on both item and ability parameters, the DFs are adjusted for the number of estimated item parameters only (Yen, 1981).

A second problem with the traditional item-fit statistics is that the number of subgroups used to approximate the continuous ability scale and how the subgroups are created is arbitrary. Different choices of subgroups might lead to different values of item statistics, then to different conclusions about item fit.

Finally, classifying examinees into subgroups is based on point estimates rather than true abilities. How accurate ability is estimated would potentially affect this classification, and misclassifications would make the results of item-fit tests questionable, especially for shorter tests like performance assessments (Stone, Mislevy, & Mazzeo, 1994).

2.2.3.2 Alternative Item-Fit Statistics

Given the above disadvantages of the traditional chi-square item-fit statistics, alternative item-fit indices have been proposed. Herein, two widely used statistics (Orlando & Thissen, 2000; Stone, 2000) are introduced.

Item-Fit Statistics based on Total Scores (Orlando & Thissen, 2000)

Orlando and Thissen's method includes forming ability subgroups based on total test scores rather than ability estimates, cross-classifying examinees into the subgroups by their total test scores and item responses, and then comparing expectations and observations using either a Pearson chi-square statistic or a likelihood ratio statistic. The null distributions of these two statistics are approximated by a chi-square distribution with $df = (I\text{-}1)\text{-}m$, where $(I\text{-}1)$ is the number of total score categories and m is the number of estimated item parameters. The effect of sparseness of cell counts may be reduced by collapsing total score groups until all cells have a minimum expected count. This method has advantages over traditional item-fit statistics in two

aspects: the determination of subgroups is not arbitrary – each possible total score defines a group; and the observed proportions are only a function of observed data and no longer model-dependent.

Orlando and Thissen's item-fit statistic was originally developed for dichotomous items, and can be implemented in GOODFIT computer program (Orlando, 1997). Their indices have been directly generalized to accommodate polytomous items (Kang & Chen, 2008), and the generalized indices can be computed through a SAS macro IRTFIT developed by Bjorner, Smith, Stone & Sun (2007).

Item-Fit Statistic Considering Uncertainty in Ability Estimation (Stone, 2000)

Stone, Mislevy, and Mazzeo (1994) and Stone (2000) pointed out that using point estimates for ability rather than true abilities may cause inaccuracy in using a chi-square distribution to approximate sampling distributions of traditional item statistics, particularly for shorter tests such as performance assessments. A simulation study by Stone & Hansen (2000) further showed that the distributions of item-fit statistics for polytomous items were affected by the precision in ability estimation. When the abilities were not estimated precisely, the sampling distribution would differ markedly from the assumed null chi-square distribution. These researchers suggested that this imprecision or uncertainty in ability estimation should be considered when item statistics are used to assess items in a shorter test.

To account for uncertainty in ability estimation, Stone et al. (1994) and Stone (2000) proposed a fit statistic computed based on posterior distribution of ability rather than point estimates of unknown ability parameters. Rather than cross-classifying examinees into only one cell of the item fit table based on his/her item response and point ability estimate, this method assigns each examinee to multiple ability groups based on his/her posterior expectations

(posterior probabilities for each discrete ability level). As ability is estimated less precisely, the posterior expectations would be more spread out across the ability scale and the examinee is classified into more ability levels to account for the uncertainty. For each examinee, calculate his /her posterior expectations and the pseudo-observed count for each cell in item-fit table can then be computed by summing the posterior expectations across all examinees. Treating the pseudocounts as observed counts, a Pearson chi-square or a likelihood ratio statistic can be calculated as for traditional item-fit statistics. This item-fit statistic can be used to evaluate the fit of either dichotomous or polytomous items, and computed through a SAS macro IRTFIT (Bjorner et al, 2007) or a SAS program (IRTFIT RESAMPLE) written by Stone (2000).

Since pseudocounts rather that actual observations are used and the contribution of an examinee's response to the item-fit table is in more than one cell, the independence assumption for goodness-of-fit chi-square test does not hold. Therefore, a null chi-square distribution can not be assumed for this item-fit statistic. Stone, Ankenmann, Lane, and Liu (1993) used Monte Carlo resampling methods to derive an empirical sampling distribution for the item fit statistic and showed that the sampling distribution can be approximated by a scaled chi-square distribution. The procedure takes into the account the uncertainty in the estimation of both item and ability parameters through re-estimating IRT model for each simulated data. However, this method is computationally intensive.

After that, Stone (2000) proposed an alternative resampling method to reduce the computation burden that did not re-estimate the IRT model but instead used the item parameter estimates from the original item responses to calculate item fit statistics across simulated datasets. Based on this distribution of item fit statistics, a scaling factor and degrees of freedom for a scaled chi-square approximation is computed and used for hypothesis tests. For this

method, only uncertainty in ability estimation was considered in generating the empirical sampling distribution. However, uncertainty in item parameter estimation was then considered by adjusting the derived *df* by the number of estimated item parameters. Stone (2000) showed that these two resampling procedures produced comparable results when evaluating the application of IRT to a mathematics performance assessment. Hansen (2004) further proposed using two multilevel equations for predicting scaling corrections based on information (item and sample characteristics) in the observed data, instead of relying on Monte Carlo resampling methods.

2.3 POSTERIOR PREDICTIVE MODEL CHECKING (PPMC) IN A BAYESIAN FRAMEWORK

PPMC is a flexible and powerful method for assessing model-fit in a Bayesian framework. It has several advantages over classical model-fit statistics. Most important, it provides a potential tool for checking the fit of complicated models which can only be estimated using Bayesian analysis. In this section, we first review the basic principle of the Bayesian framework which provides the foundation for the PPMC method followed by detailed discussion of the PPMC technique.

2.3.1 Introduction to Bayesian Inference

Bayesian statistics have received considerable attention over the past decade. In the Bayesian framework, unknown population parameters are treated as random variables that follow a certain distribution. Prior knowledge or beliefs about the possible shape of this distribution are modeled

by specifying a prior distribution on the parameters. The prior distribution will be updated by the data using the likelihood to form a posterior distribution for the parameters. The Bayesian inference about parameters is then drawn based on this posterior distribution.

Mathematically, this is represented as follows. Let **y** denote the data and **ω** denote the vector of unknown parameter(s) in a model. The posterior distribution for **ω** given the data **y** can be obtained through Bayes' theorem:

$$p(\boldsymbol{\omega} \mid \mathbf{y}) = \frac{p(\boldsymbol{\omega}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\boldsymbol{\omega})p(\mathbf{y} \mid \boldsymbol{\omega})}{p(\mathbf{y})} = \frac{p(\boldsymbol{\omega})p(\mathbf{y} \mid \boldsymbol{\omega})}{\int p(\mathbf{y} \mid \boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega}}, \tag{2.10}$$

where $p(\boldsymbol{\omega}, \mathbf{y})$ is a joint probability distribution for $\boldsymbol{\omega}$ and **y;** $p(\boldsymbol{\omega})$is the prior distribution of parameter(s) $\boldsymbol{\omega}$ and it represents researchers' prior information or belief about $\boldsymbol{\omega}$; $p(\mathbf{y} \mid \boldsymbol{\omega})$is the likelihood function of the data given a value of parameter(s), and represents the probability of the data **y** to be observed under a specific value of parameter(s) $\boldsymbol{\omega}$; $p(\mathbf{y})$ is the marginal or unconditional probability of the data across all possible values of $\boldsymbol{\omega}$. Because $p(\mathbf{y})$ is the function of data, it can be considered a constant for a given data. This constant is only used to normalize $p(\boldsymbol{\omega})p(\mathbf{y} \mid \boldsymbol{\omega})$ so that $p(\boldsymbol{\omega} \mid \mathbf{y})$ is a probability distribution. Omitting $p(\mathbf{y})$ will not affect the inferences from posterior distribution, and yields the unnormalized posterior distribution that is proportional to the product of the likelihood and the prior distribution:

$$p(\boldsymbol{\omega} \mid \mathbf{y}) \propto p(\boldsymbol{\omega})p(\mathbf{y} \mid \boldsymbol{\omega}). \tag{2.11}$$

In many situations, it is either infeasible or simply not necessary to compute this normalizing constant and Equation (2.11) is actually applied for Bayesian inferences. The main goal of Bayesian inference is to sample from the posterior distribution $p(\boldsymbol{\omega} \mid \mathbf{y})$ in order to estimate population parameters (e.g., quantiles and moments), to construct credible intervals, and to obtain Bayesian posterior p-values for hypothesis tests (Rupp, Dey, & Zumbo, 2004).

2.3.2 Posterior Predictive Model Checking (PPMC)

2.3.2.1 Description of PPMC Method

PPMC was introduced by Guttman (1976), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984). Gelmen, Carlin, Stern, & Rubin (1996) extended it to allow more direct assessment of the discrepancy between data and presumed model. Conducting PPMC involves simulating data under a presumed model and comparing features of simulated data against observed data using discrepancy measures that are sensitive to different aspects of misfit. Any systematic differences indicate potential misfit of the model. The rationale underlying PPMC is that if a chosen model fits the data, then observed data should look like replicated data generated from the posterior distributions of model parameters. Differences between observed and predicted data on discrepancy measures in PPMC can be evaluated using graphical displays as well as a numerical summary - Posterior Predictive P-value (PPP-value).

Let $\mathbf{y}$ be the observed data and $\mathbf{y}^{rep}$ be the replicated data set that could have been observed if the experiment that produced $\mathbf{y}$ were replicated with the same model and the same values of model parameters $\boldsymbol{\omega}$ that produced the observed data. The PPMC method assesses the fit of a model by examining whether the observed data $\mathbf{y}$ appear extreme with respect to the posterior predictive distribution of replicated data $\mathbf{y}^{rep}$,

$$p(\mathbf{y}^{rep} \mid \mathbf{y}) = \int p(\mathbf{y}^{rep}, \boldsymbol{\omega} \mid \mathbf{y}) d\boldsymbol{\omega} = \int p(\mathbf{y}^{rep} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathbf{y}) d\boldsymbol{\omega} . \quad (2.12)$$

where $p(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior distribution of $\boldsymbol{\omega}$.

A test quantity or discrepancy measure D($\mathbf{y}$, $\boldsymbol{\omega}$) is usually employed to measure the discrepancy between the observed and the predicted data (Gelman et al., 2003). The comparison of observed (realized) and posterior predictive discrepancy measures can be performed using

graphical display as well as a PPP-value. Since the PPMC method should be used as a diagnostic tool for model fit rather than a hypothesis testing, the preferable way to interpret the difference between observed and predicted data in PPMC is to employ graphical plots (Gelman, et al., 2003). However, PPP-values provide a numerical summary measure of the degree to which a model fits the data and are typically used with graphical plots for interpretation. PPP-value is a tail-area probability that predicted data are more extreme than observed data in terms of the values of a discrepancy measure D(**y, ω**):

$$PPP = P\left(D\left(\mathbf{y}^{rep}, \boldsymbol{\omega}\right) \geq D\left(\mathbf{y}, \boldsymbol{\omega}\right) \mid \mathbf{y}\right) = \iint_{D\left(\mathbf{y}^{rep}, \boldsymbol{\omega}\right) \geq D(\mathbf{y}, \boldsymbol{\omega})} p\left(\mathbf{y}^{rep} \mid \boldsymbol{\omega}\right) p\left(\boldsymbol{\omega} \mid \mathbf{y}\right) d\mathbf{y}^{rep} d\boldsymbol{\omega}. \qquad (2.13)$$

It should be noted that the PPP-values can not be interpreted in the same way as traditional hypothesis-testing *p* values. Though both of them are defined as a tail-area probability and share some features, the PPP-values are not necessarily uniformly distributed under the null conditions. In general, they tend to be closer to 0.5 more often than would be expected under a uniform distribution (Levy, 2006; Meng, 1994; Robins, van der Vaart, & Ventura, 2000). As a result, use of PPP-values in a hypothesis testing framework would lead to a conservative test (Bayarri & Berger, 2000; Fu et al., 2005; Levy, 2006; Sinharay, 2005; Sinharay et al., 2006). However, Levy (2006) showed that the distributions of PPP-values were close to uniform for some suitable measures, and their type-I error raters were close to the nominal level in the hypothesis testing framework. PPP-values near 0.5 would indicate that the realized values of discrepancy measures look similar to the posterior predictive values, indicative of data-model fit. Extreme PPP-values near 0 or 1 suggest that the realized discrepancies are inconsistent with the posterior predictive discrepancies and hence are indicative of data-model misfit.

The graphical plots are also commonly used with PPP-values for PPMC to provide graphical evidence of misfit. When the discrepancy measure only depends on the data, the values

of $D(\mathbf{y}^{rep,n})$ (n = 1, 2, …, N, where N is the total number of replications) are plotted in a histogram and the position of observed values of $D(\mathbf{y})$ in this histogram is examined. The observed value of a powerful discrepancy measure $D(\mathbf{y})$ for an inadequate model should be located in the tail area of the histogram. When the discrepancy measure depends on both the data and parameters, pairs of the realized discrepancies $D(\mathbf{y},\boldsymbol{\omega}^n)$ and predictive discrepancies $D(\mathbf{y}^{rep,n},\boldsymbol{\omega})$ are plotted in a scatter plot. Points lying consistently above or below the 45-degree line indicate model misfit.

Figure 2.3 shows two histograms taken from Sinharay el al. (2006). A data was generated based on 3PL model and estimated using 1PL and 3PL models. The discrepancy measure used was the standard deviation of biserial correlation coefficients which was only dependent on the data. As can be seen from the two graphs, when a 1PL model was fitted to the data (left plot), the observed SD was far away from the posterior predictive distribution of the SDs of biserial coefficients. In contrast, when a 3PL model was fitted to the data, the observed SD was very close to the median of the posterior predictive distribution (right plot).

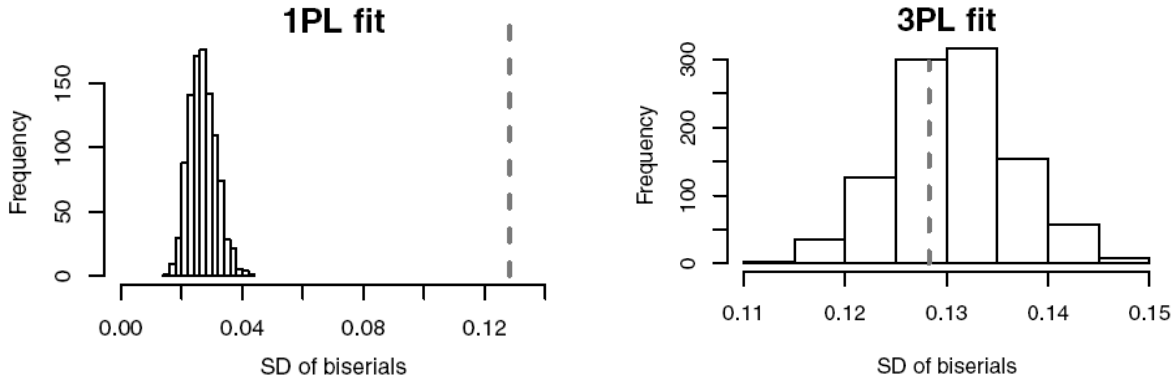


Figure 2.3 Examples of Graphical Displays in PPMC by using Histograms

Figure 2.4 includes two scatter plots also from Sinharay el al. (2006). A data was generated based on 3PL model and estimated using 2PL and 3PL models. The discrepancy

measure used is a chi-square statistic reflecting the difference between the observed and expected test score distribution. This statistic depends on both data and model parameters. The left plot shows the realized discrepancies were consistently larger than the predicted values when 2PL was fitted to the data. The right plot shows 3PL fitted the data well.

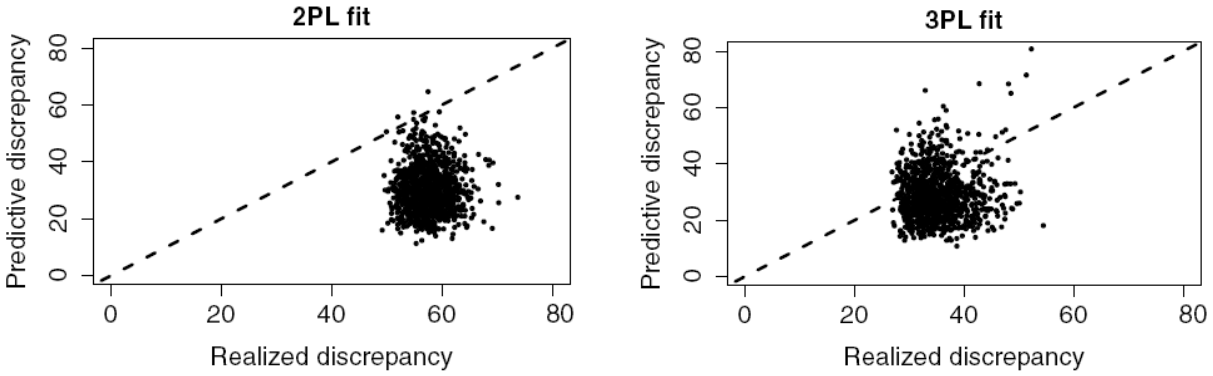


Figure 2.4 Examples of Graphical Displays in PPMC by using Scatter Plots

2.3.2.2 Computation via MCMC Simulation

Computation of posterior predictive distribution of discrepancy measures and PPP-values is typically conducted using MCMC simulation methods. Gelman et al. (1996) pointed out that since the MCMC method is a standard tool for Bayesian analysis with complex models and provides a sample of draws from the posterior distribution $p(\boldsymbol{\omega} \mid \mathbf{y})$, the required computation for PPMC is a byproduct of Bayesian analysis with MCMC simulation.

Given a parameter vector $\boldsymbol{\omega}$, the steps for PPMC via MCMC are as follows: (1) draw *N* parameter estimates $\boldsymbol{\omega}^1$, $\boldsymbol{\omega}^2$, …, $\boldsymbol{\omega}^N$ from the posterior distribution of $\boldsymbol{\omega}$ - $p(\boldsymbol{\omega} \mid \mathbf{y})$ using MCMC algorithm; (2) draw one $\mathbf{y}^{\text{rep}}$ from the predictive distribution $p(\mathbf{y}^{rep} \mid \boldsymbol{\omega})$ for each simulated $\boldsymbol{\omega}$ to produce *N* sets of replicated data, $\mathbf{y}^{\text{rep},1}$, $\mathbf{y}^{\text{rep},2}$, … , $\mathbf{y}^{\text{rep,N}}$ from the joint posterior distribution, $p(\mathbf{y}^{rep}, \boldsymbol{\omega} \mid \mathbf{y})$; (3) compute the realized discrepancies $D(\mathbf{y}, \boldsymbol{\omega}^n)$ and predictive

discrepancies $D\left(\mathbf{y}^{rep,n},\boldsymbol{\omega}^{n}\right)$, n = 1, 2, … N. Through this procedure, the reference distribution of discrepancy is the distribution of $D\left(\mathbf{y}^{rep,n},\boldsymbol{\omega}^{n}\right)$, and the estimated PPP-value is just the proportion of these N replications for which $D\left(\mathbf{y}^{rep,n},\boldsymbol{\omega}^{n}\right)$ equals or exceeds the realized value $D\left(\mathbf{y},\boldsymbol{\omega}^{n}\right)$. Sinharay et al. (2006) described this procedure graphically (see Figure 2.5). The description about MCMC methods will be given in more detail in section 2.3.3.

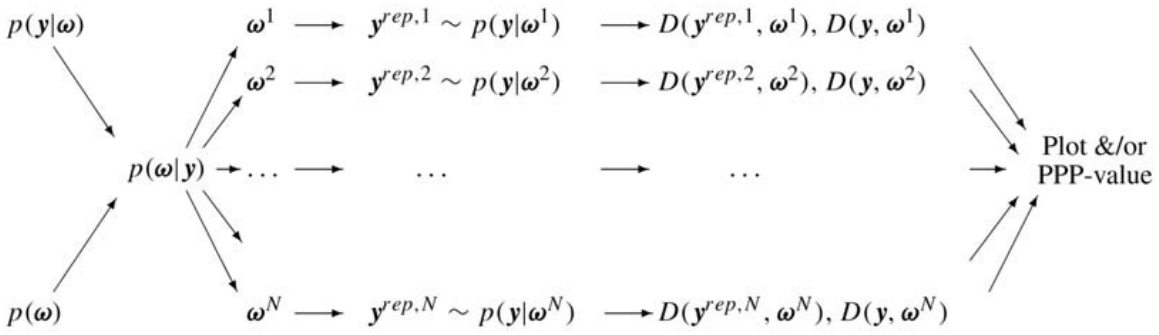


Figure 2.5 Graphical Description of Implementing the PPMC Method

2.3.2.3 Discrepancy Measures

Discrepancy measures play a similar role in Bayesian model-fit checking as test statistics play in classical testing. However, there is much more freedom in choosing discrepancy measures in Bayesian framework because the reference distribution of any measure can be determined through MCMC simulations. They can be any function of the data and/or model parameters. When a discrepancy measure D(**y**) only depends on the data, it is a pivotal quantity similar to classical test statistics. For example, a researcher observed a set of data from an experiment with the mean of 5 and the variance of 10, and wanted to know if the observations follow *N* (5, 10). We know a normal distribution has symmetric short tails. Therefore a useful discrepancy measure for PPMC would be the extreme (minimum or maximum) observation which is only data-dependent. If the observed extreme value is far away from the posterior distribution of

extreme values under the normal model, it indicates the normal model is not adequate for the extreme tails in the observation.

When the pivotal quantity can not capture a particular aspect of a model which reflects inferential interests, an therefore would not have enough power to detect misfit, the discrepancy measure should be chosen to depend on the model parameters $\boldsymbol{\omega}$ as well as the data $\mathbf{y}$: $D(\mathbf{y},\boldsymbol{\omega})$. We continued using the previous example in which the result showed the normal model was inadequate for the extreme tails. If the researcher is interested in whether the distribution of the observation is symmetric, using the extreme observations will not work. A different measure sensitive to asymmetry in the center of the distribution should be used. For example, the measure $D(y,\mu)=\left|y_{(10th)}-u\right|-\left|y_{(90th)}-u\right|$ should be useful, where u is the mean of the normal distribution, $y_{(10th)}$ and $y_{(90th)}$ are the 10^{th} and 90^{th} percentile of the observed data, respectively. This measure relies not only on data but also on the model parameter u.

The choice of discrepancy measures is a key issue in an application of the PPMC method for assessing the fit of a model. They should be chosen to reflect aspects of model misfit of greatest concerns, but not directly addressed by the model (Gelman et al., 2003). For example, Sinharay and Johnson (2003) showed the biserial correlation coefficients were not powerful discrepancy measures for 2PL/3PL model, but they were useful for Rasch model, because 2PL/3PL models have slope parameters to address the biserial but Rasch has not. Identifying an appropriate discrepancy measures is a challenge to researchers in applying PPMC. A useful strategy is to think about what is the main concern in application of a model to a specific dataset and develop a discrepancy most related to this concern. If there are no prior concerns, it is recommended to employ a number of different discrepancies for assessing different aspects of model fit (Sinharay, Johnson, & Stern, 2006).

2.3.2.4 Advantages of PPMC over Classical Model-Fit Tests

The PPMC method has several advantages over classical model-fit tests. Firstly, it incorporates uncertainty in the estimation of parameters into the sampling distributions of discrepancy measures by using posterior distributions of parameters rather than point estimates. Modeling sources of uncertainty is a major advantage of Bayesian framework.

Secondly, for most models, the exact theoretical sampling distributions of classical model-fit statistics are difficult to derive, and the null sampling distributions used are only asymptotically justified. The departure of the approximate distribution from the true sampling distribution may affect the performance of model-fit statistic and thus cause an incorrect decision about the fit of a model. In contrast, the PPMC method forms posterior predictive distributions of discrepancy measures empirically from MCMC simulations. These empirical distributions reflect exact null sampling distributions.

Lastly, the recent rapid development of Bayesian computation allows us to fit more realistic and sophisticated models than previously possible. However, classical model-fit tests are not applicable for assessing the fit of these complicated models. PPMC may be the only general model-checking method for them.

2.3.3 Markov Chain Monte Carlo (MCMC) Simulation

2.3.3.1 Definition

In many situations, the joint posterior distributions can not be obtained analytically and thus direct sampling from them is not possible. MCMC simulation provides a flexible way to draw samples or values from any posterior distribution. MCMC methods are widely considered as the most important development in statistical computing in recent history and their occurrence

makes Bayesian methodology more attractive and popular in many disciplines. MCMC methods include a class of algorithms for sampling (drawing values of parameters) from probability distributions based on constructing a Markov chain whose stationary distribution is the target probability distribution. The key to MCMC is to create a Markov chain and run the simulation long enough so that the distribution of the draws beyond some point of time reflects this target distribution. The expectations of relevant functions of parameters are then approximated using Monte Carlo integration. In a Bayesian framework, the target distribution is the posterior distribution.

Different MCMC methods are distinguished by the sampling algorithms used in simulating the Markov chains. Three well-known MCMC algorithms are the Gibbs sampler, the Metropolis algorithm, and the Metropolis-Hastings (M-H) algorithm. The Gibbs sampler is also called alternating conditional sampling (Gelman et. al, 2003), and it is used to create a Markov chain by successively sampling from a set of "complete conditional" distributions which will eventually approximate the joint posterior distribution. For example, the steps of sampling p unknown parameters $\theta_1, \theta_2 \ldots \theta_p$ is as follows (Ruff et. al, 2004):

Step 1: Specify the joint posterior distribution:

$$p(\theta_1, \theta_2 \ldots \theta_p \mid Y) \propto p(X \mid \theta_1, \theta_2 \ldots \theta_p) p(\theta_1) \ldots p(\theta_p)$$

Step 2: Identify the complete set of conditional distributions:

$$p(\theta_1 \mid \theta_2 \ldots \theta_p, Y), p(\theta_2 \mid \theta_1, \theta_3 \ldots \theta_p, Y), \ldots p(\theta_p \mid \theta_1, \theta_2 \ldots \theta_{p-1}, Y)$$

Step 3: Provide initial values for $\theta_{1,0}, \theta_{2,0} \ldots \theta_{p,0}$ (at iteration 0) using direct specification or sampling from appropriate distributions.

Step 4: Generate new values at iteration i as follows.

$$\theta_{1,i} \sim p\left(\theta_1 \mid \theta_{2,i-1}, \theta_{3,i-1}, \ldots, \theta_{p,i-1}\right)$$
$$\theta_{2,i} \sim p\left(\theta_2 \mid \theta_{1,i}, \theta_{2,i-1}, \theta_{3,i-1}, \ldots, \theta_{p,i-1}\right)$$
$$\vdots$$
$$\theta_{p,i} \sim p\left(\theta_p \mid \theta_{1,i}, \theta_{2,i}, \theta_{3,i}, \ldots, \theta_{p-1,i}\right)$$

Step 5: Repeat step 4 until the Markov chain is convergent.

The Gibbs sampler is conceptually the simplest of Markov chain sampling methods. It works best when all of the complete conditional distributions can be obtained in closed form.

When the conditional distributions are not of a known distributional form, M-H sampling methods are needed. The M-H algorithm, also known as rejection sampling algorithm, samples a proposal θ^* value from any convenient proposal distribution (jumping distribution) at time t, $q\left(\theta^* \mid \theta^{t-1}\right)$ which depends on the previous state θ^{t-1}. This proposal θ^* is accepted ($\theta^t = \theta^*$) with probability α where $\alpha = \min\{r, 1\}$ and

$$r = \frac{p\left(\theta^* \mid y\right) q\left(\theta^{t-1} \mid \theta^*\right)}{p\left(\theta^{t-1} \mid y\right) q\left(\theta^* \mid \theta^{t-1}\right)} \tag{2.14}$$

If the proposal is not accepted, the old value of the parameter will be kept ($\theta^t = \theta^{t-1}$). The difference between the Metropolis and M-H algorithms is on the proposal distribution $q\left(\theta^* \mid \theta^{t-1}\right)$. The Metropolis algorithm requires $q\left(\theta^* \mid \theta^{t-1}\right)$ to be symmetric, satisfying the condition that $q_t\left(\theta_a \mid \theta_b\right) = q_t\left(\theta_b \mid \theta_a\right)$ for all θ_a, θ_b, and t. A symmetric proposal distribution simplifies calculations of the ratio r since when $q\left(\theta^{t-1} \mid \theta^*\right) = q\left(\theta^* \mid \theta^{t-1}\right)$, the terms cancel out. The M-H algorithm generalizes the Metropolis algorithm using asymmetric proposal distributions. Allowing asymmetric proposal distributions can be useful in increasing the speed of the random walk or convergence to a stationary posterior distribution (Gelmen et. al, 2003).

2.3.3.2 Convergence Diagnosis

The key to MCMC success is that the Markov Chain has converged to the target posterior distribution. If the chain does not converge, the simulated draws from this chain would not represent the posterior distribution of parameters of interest. Thus, the inference about parameters based on the distribution of these draws would be invalid. Therefore, it is very important to assess convergence of Markov chains before any Bayesian inferences are made.

A number of convergence diagnostics have been developed. Cowles and Carlin (1996), and Brooks and Robert (1998) provide an excellent review. The most popular diagnostics are time-series plots, autocorrelation plots, and the Gelman-Rubin statistic *R*. A time-series plot, also called a "history plot", is a scatter plot showing the generated values of a parameter at each iteration number in a chain of sample values. Clear trends in the plot indicate that successive simulated values of parameters are highly correlated and a chain has not converged. Time-series plots provide a simple way to check the stability of simulated parameter values. Figures 2.6 provide example illustrations of two chains. The left plot shows a high likelihood of convergence, but the right plot demonstrates the non-convergence since there is a clear trend: the sampled values decreased as the iteration number increased.

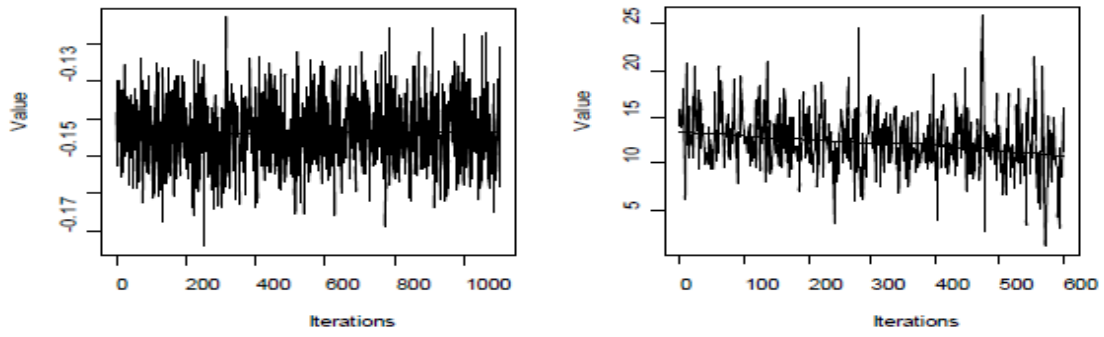


Figure 2.6 History Plots Displaying Evidence of Convergence and Non-Convergence

An autocorrelation plot is a plot of the correlation between sequential draws of a parameter in Markov chain. It is a commonly-used tool for checking randomness (independence) in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. Autocorrelation plots are not strictly a convergence diagnostic tool, but they help indirectly to assess convergence. A MCMC algorithm generating highly correlated parameter values will need a large number of iterations to converge to the appropriate posterior distribution. In other words, such autocorrelation can cause inefficient MCMC simulation. Solution to high autocorrelation is to "thin" the chains by keeping every k^{th} simulation draw from each sequence and discarding the rest.

Gelman and Rubin (1992) suggest monitoring convergence based on multiple chains with different or over-dispersed starting points. The motivation for this statistic is that "even if an iterative simulation appears to converge, it still may actually be far from convergence if important areas of the target distribution were not captured by the starting distribution and are not easily reachable by the simulation algorithm" (Gelman et al., 2003, p297). This statistic is computed through comparing between-chain (B) variance and within-chain (W) variance for each parameter, and defined as the ratio of the estimated marginal posterior variance V(θ|y) to the within-chain variance W:

$$\hat{\mathrm{R}} = \sqrt{\frac{\hat{V}(\theta)}{W}}\ , \qquad \text{where } \hat{V}(\theta \mid y) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B. \tag{2.15}$$

Before convergence, W underestimates total posterior variance in θ because it has not fully explored the target distribution. V(θ|y) on the other hand overestimates variance in θ because the starting points are over-dispersed relative to the target. Once convergence is reached, W and V(θ) should be almost equivalent because variation within the chains and variations

between the chains should coincide, so $\hat{R}$ should approximately equal one. $\hat{R}$ near 1 for all parameters of interest means the MCMC algorithm has converged. In addition, the convergence of multiple chains to the same stationary distribution is often reflected by a large overlap in their sampling histories.

Although the Gelman-Rubin statistic is a useful diagnostic tool, one drawback is that its value depends on the choice of initial values. Since there is no single definitive convergence diagnostic tool, the use of multiple tools is recommended in order to increase the chance of correctly assessing convergence (Sinharay, 2004).

Given convergence of MCMC algorithm, inferences about parameters of interest should be based on the draws after the point of convergence. Several additional issues to be considered in this process were discussed by Kim and Bolt (2007). The first concerns the number of early iterations of the simulation which should be discarded in order to diminish the effect of the starting distribution. The practice of discarding early iterations in Markov chain simulation is referred to as "burn-in". Raftery and Lewis (1992) recommended the length of burn-in should be at least as large as the distance between samples needed to achieve an autocorrelation of 0. Gelman et al (2003) suggested discarding the first half of iterations as a conservative choice. However, Kim and Bolt (2007, p43) pointed out that "because the actual burn-in usually involves a relatively small number of iterations, the effect of some inaccuracy is generally of minimal significance".

A second consideration involves thinning the chain to reduce substantial autocorrelations in the chain by taking every n^{th} draw. However, when large number of parameters are involved, computer storage is always a problem for saving too much draws from the chains. In this

situation, Gelman et al. (2003) suggested to thin the chain so that the total number of iterations saved is no more than 1000.

The final concern is that how large of a posterior sample is necessary for obtaining precise posterior inferences. It is important to recognize that the error in posterior estimation can be attributed not only to the standard error of the point estimates as reflected by the posterior sample standard deviation, but also to sampling error, referred to as Monte Carlo error (Kim & Bolt, 2007). As a rule of thumb, the simulation should be run until the MC error for each parameter of interest is less than about 5% of the sample standard deviation (Spiegelhalter et al., 2003). The MC error can always be reduced by lengthening the chain.

2.4 CHECKING IRT MODEL-FIT USING PPMC

2.4.1 Advantages of Using PPMC in IRT

The common advantages of using PPMC over classical model-fit statistics were summarized in Section 2.3.2.4. This section discusses the advantages of PPMC for assessing IRT model-fit.

Even though numerous classical approaches have been proposed to assessing different aspects of model fit in IRT, many model-fit indexes have well-known shortcomings and none of them is entirely satisfactory. One common issue with classical model-fit indices involves the use of point estimates of IRT model parameters (item and ability) which do not take into account the uncertainty in parameter estimation. In contrast, the PPMC method takes into account this uncertainty by using the entire posterior distributions of model parameters rather than point estimates.

Another common problem with existing model-fit indices is that their sampling distributions only asymptotically approximate a null chi-square distribution, and as discussed previously, it is not entirely clear what degree of freedom should be used. The discrepancy of true sampling distributions from assumed chi-square distributions would result in high type-I error rate and high false alarm rate of some fit statistics (e.g., Orlando & Thissen, 2000; Sinharay, 2006; Stone, 2000). Compared with classical model-fit statistics, the PPMC method is free from the sampling distribution issue because it is constructed empirically from MCMC simulation.

In the last ten years, the family of IRT models has expanded tremendously and complex IRT models have been developed in response to different educational testing applications. When IRT models become more complex, estimation of the models becomes more difficult using traditional marginal maximum likelihood (MML) estimation methods. Bayesian estimation using MCMC methods offer much potential for estimation of complex IRT models. Since Albert (1992) proposed a full Bayesian method based on Gibbs sampling to estimate 2-parameter normal-ogive IRT model, and Patz and Junker (1999a, 1999b) developed M-H sampling algorithms to estimate several different IRT models such as 2PL, 3PL and mixed models, full Bayesian methods with MCMC algorithm have become widely used by many researchers to estimate a variety of complex IRT models such as testlet models (e.g., Bradlow et al., 1999; DeMars, 2006; Li et al., 2006; Wang, 2002), rater-effect models (e.g., Patz & Junker, 2002), and multidimensional IRT models (e.g., Béguin & Glas, 2001; Bolt & Lall 2003; Yao & Schwarz, 2006). However, as for any IRT models, the application of those complex IRT models are valid only if the modes fit the data. Unfortunately, though Bayesian estimation of complex IRT models have received intensive attention, relatively little attention has been given to assessing the fit of

these models from a Bayesian perspective and further research is needed. The PPMC method, as a popular and flexible Bayesian model diagnostic tool, may address this issue.

2.4.2 Discrepancy Measures Used with Dichotomous IRT Models

Previous research using the PPMC method with IRT models has focused on unidimensional IRT models for dichotomous items. Before the review of the previous research, the discrepancy measures examined in those studies are firstly reviewed in this section. These discrepancy measures were designed to assess model-fit at three levels: test level, item level, and item-pair (or pair-wise) level. It should be noted that these measures were mainly proposed for dichotomous IRT models. In the present study, the polytomous models are of interest. The extensions of these measures to account for polytomous IRT models will be introduced in Chapter 3.

2.4.2.1 Test-Level Discrepancy Measures

One simple measure at test level is "observed test score distribution" (number of examinees with each total test score). The overall model fit can be examined through comparing the observed and posterior predictive test score distributions. The credible interval for the posterior predictive score distribution across multiple predicted response data sets and observed score distribution can be shown in a same graph. If the observed score distribution falls within the credible interval, there is no evidence of model misfit at the test level. This measure can be directly used for polytomous IRT models.

Figure 2.7 illustrates the observed frequency, the posterior mean frequency and their central 95% posterior interval (between 2.5% and 97.5%) for a polytomous response data

generated based on a unidimensional GR model and estimated using the same model. As can be seen from this figure, the observed distribution was well within the posterior interval, indicating that the unidimensional GR model fitted the observed data reasonably well regarding the test score distribution.

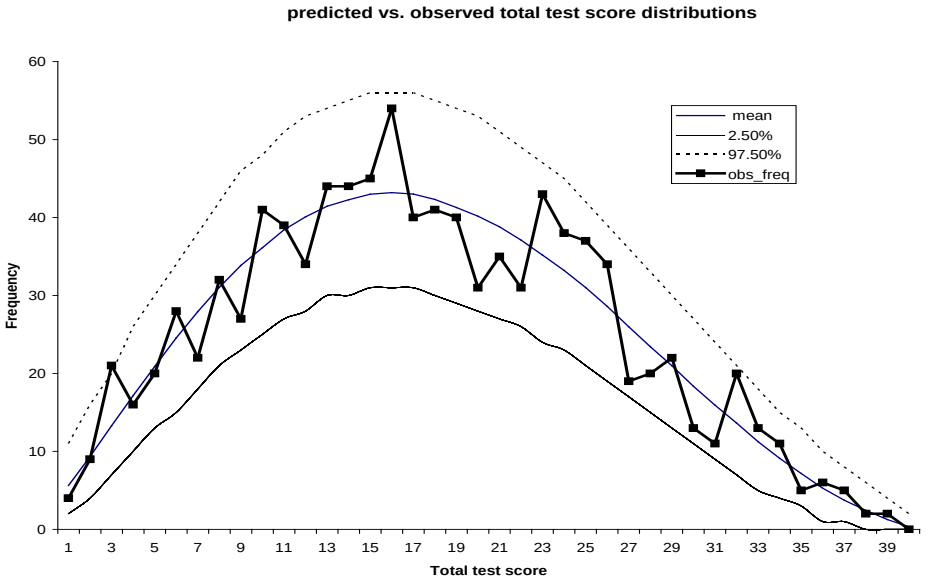


Figure 2.7 Example of Observed and Predictive Test Score Distributions

In addition, Béguin and Glas (2001) suggested using Pearson's chi-square statistic (χ_T^2) to summarize the difference between the observed and expected frequencies of test scores:

$$\chi_T^2 = \sum_t^T \frac{\left[N_t - E(N_t)\right]^2}{E(N_t)}, \tag{2.16}$$

where T is the maximum total test score, N_t is the observed number of examinees with total score t, and $E(N_t)$ is the expected number of examinees with total score t based on the model. For a test score point t, $E(N_t)$ can be calculated as:

$$E(N_t) = N \sum_{\mathbf{y}|t} \int p(\mathbf{y} \mid \theta) g(\theta) d\theta, \tag{2.17}$$

where N is the total number of examinees, $\mathbf{y}|t$ represents the set of all possible response patterns resulting in a score t, $p(\mathbf{y}|\theta)$ is the probability of response pattern $\mathbf{y}$ given ability θ and the item parameters, and $g(\theta)$ is the assumed density of the ability distribution. For dichotomous items, $E(N_t)$ may be computed using a recursive algorithm proposed by Lord and Wingersky (1984). For polytomous items, a generalized recursive algorithm (Thissen, Pommerich, Billeaud, & Williams, 1995) has been discussed.

Béguin and Glas (2001) pointed out that though the statistic χ_T^2 does not follow a chi-square distribution, it can be used with the PPMC method since PPMC constructs null sampling distributions empirically from MCMC simulations. Previous research also found that this measure can be used to detect certain types of misfit for dichotomous IRT models (Béguin & Glas, 2001; Sinharay et al., 2006).

2.4.2.2 Item-Level Discrepancy Measures

(1) *Bayesian χ^2 Statistic*

The Bayesian χ^2 measure is the Bayesian version of unweighted mean square item fit statistic (Masters & Wright, 1997) which is the summation of unweighted standardized squared residuals across the examinees. The *Bayesian χ^2 statistic* for item j is defined as:

$$\chi_j^2 = \sum_{i=1}^{N} \frac{\left[y_{ij} - E\left(y_{ij}\right)\right]^2}{Var\left(y_{ij}\right)}, \tag{2.18}$$

where N is the total number of examinees, y_{ij} and $E(y_{ij})$ are the observed and expected response of examinee i to item j respectively, and $Var(y_{ij})$ is the variances of the response y_{ij} . For the j^{th} polytomous item with the total of response categories (M_j+1), $E(y_{ij})$ and $Var(y_{ij})$ can be calculated using:

$$\begin{cases} E(y_{ij}) = \sum_{k=0}^{M_j} kP_{ijk}(\theta_i) \\ Var(y_{ij}) = \sum_{k=0}^{M_j} (k - E(y_{ij}))^2 P_{ijk}(\theta_i) \end{cases} \tag{2.19}$$

Although it is an intuitive and useful statistic for evaluating overall fit for many statistical models, several researchers (Li et al., 2006; Sinharay et al., 2006) found that it was not useful for IRT model checking as it failed to detect problems with inadequate models.

(2) *Item Score Distribution*

Item score distribution represents the number of examinees responding to each response category for each item. Similar to test score distribution, the difference between observed and posterior predictive item score distributions can be summarized using a goodness-of-fit statistic ($\chi_j^2\_Fit$). For dichotomous items, it is defined as:

$$\chi_j^2\_Fit = \sum_{k=0}^{1} \frac{[O_{jk} - E_{jk}]^2}{E_{jk}}, \tag{2.20}$$

where O_{jk} (E_{jk}) is the observed (predicted) number of examinees scoring in response category *k* on item *j*. E_{jk} can be calculated by summing the probabilities of responding to category *k* on item *j* across all *N* examinees:

$$E_{jk} = \sum_{i=1}^{N} p_{ijk}(\theta_i) \qquad k = 0, 1. \tag{2.21}$$

Joreskog and Moustaki (2001) used this statistic to measure the model fit and further discussed that the null distributions of this measure does not follow a chi-square distribution. They suggested fit values larger than 4.0 indicate poor fit. This measure was also used with the PPMC method in detecting misfit of dichotomous IRT model (Levy, 2006) and polytomous

fusion model (Fu et al, 2005). The PPMC method is free from sampling distribution issue and therefore can avoid using the arbitrary guidelines.

(3) *Classical Item-fit Statistics*

Classical item-fit statistics can also be used as discrepancy measures with PPMC in order to check IRT item-fit. As mentioned previously, for classical item-fit statistics, their sampling distributions only asymptotically approximate assumed chi-square distributions and the uncertainty in the estimation of model parameters (item and ability) is not taken into account. These issues might affect the performance of these classical item-fit statistics especially for shorter tests such as performance assessments. Since the PPMC method is free from these issues, it is important to explore the performance of these classical item fit measures in a Bayesian framework.

Though any classical item-fit statistics can be used with PPMC, only Orlando and Thissen's item-fit statistics were used by Sinharay (2006) to assess the fit of dichotomous items. The results showed these item-fit statistics performed better than that in frequentist framework.

(4) *Biserial Correlation Coefficient*

The biserial correlation coefficient estimates the correlation between examinee total test scores and binary outcomes on a particular dichotomous item which also reflects item discrimination. Sinharay and Johnson (2003) found that the standard deviation of the biserial correlations was powerful in detecting misfit of Rasch models when data were generated from a 2PL or 3PL model.

2.4.2.3 Pair-wise Discrepancy Measures

Pair-wise measures reflect the association between the responses to item pairs, and they as a whole have been found to be more powerful than test- and item-level measures in detecting

misfit for unidimensional IRT models. It is because there are no parameters in unidimensional IRT models to address association/interaction among items and pairwise measures that capture the association among items therefore have the potential to detect possible misfit such as multidimensionality and local dependence. Several pairwise measures were used in previous studies.

(1) *Yen's Q_3 Statistic*

Yen's Q_3 is the most popular statistic used to measure local independence. The definition is given in section 2.2.2. As discussed before, Chen and Thissen (1997) showed the empirical distribution of Q_3 did not match this theoretical sampling distribution that produced the Type-I error rates sufficiently larger than the nominal levels. As a result, it is more common to use Q_3 as a descriptive statistic than a hypothesis testing in the classical (frequentist) framework. Using this statistic in the PPMC context can avoid the sampling distribution issue. Levy (2006) used this measure with PPMC to detect the local dependence among responses to dichotomous items, and found that Q_3 was most effective among several measures.

(2) *Chen & Thissen's Chi-Square LD Index*

The LD indices proposed by Chen and Thissen (1997) are based on 2x2 contingency tables. For each pair of dichotomous items j and j^*, the following contingency table can be constructed:

		Item j	
Item j^*			
		n_{00} ($E(n_{00})$)	n_{01} ($E(n_{01})$)
		n_{10} ($E(n_{10})$)	n_{11} ($E(n_{11})$)

In this table, n_{pq} and $E(n_{pq})$ are the observed and expected number of examinees having response *p* on item *j* and response *q* on item j^*, where 1 and 0 represent the correct and incorrect responses, respectively. A Pearson's χ^2 index is then defined as:

$$\chi^2_{jj^*} = \sum_{j=0}^{1}\sum_{j^*=0}^{1}\frac{\left[n_{jj^*} - E(n_{jj^*})\right]^2}{E(n_{jj^*})}, \tag{2.22}$$

and the corresponding likelihood ratio G^2 statistic is given by:

$$G^2_{jj^*} = 2\sum_{j=0}^{1}\sum_{j^*=0}^{1} n_{jj^*}\log\frac{E(n_{jj^*})}{n_{jj^*}}. \tag{2.23}$$

These two LD indices are assumed to follow chi-square distribution with degree of freedom of 1 when the assumption of local dependence is hold. However, Chen and Thissen (1997) found the empirical sampling distributions of these two indices are very nearly as a χ^2 distribution with degree of freedom slightly less than one in the null conditions. Using the assumed χ^2 distribution as the null sampling distribution would cause conservative results. Levy (2006) used these measures with PPMC to detect the local dependence among responses to dichotomous items, and found that though they were more effective than item-level measures, but less effective than other pair-wise measures used in his study.

(3) *Odds Ratio (OR)*

Odds ratio (OR) is one of Chen & Thissen's LD indices which were developed for dichotomous items based on 2 x 2 contingency tables. Following the denotations for the two chi-square LD indices above, the OR for a pair of dichotomous items (*j* and j^*) is:

$$OR = \frac{n_{00}n_{11}}{n_{01}n_{10}}. \tag{2.24}$$

OR can be used in detecting the violation of local independence assumption for unidimensional dichotomous models. If local independence is met, a unidimensional model can

fit the observed OR. Otherwise, the observed OR will be larger than what is expected under a unidimensional IRT model for within-cluster items, and smaller than expected for between-cluster items. Chen and Thissen (1997) found that the standardized *log(OR)* difference does not follow a standard normal distribution and hence is not a useful diagnostic in a frequentist framework. However, many studies (Sinharay et al, 2005, 2006; Li et al, 2006; Levy, 2006) showed OR measure is a useful discrepancy measures for checking several aspects of model fit in the PPMC context.

(4) *Mantel-Haenszel (MH) statistic*

An odds ratio for one item pair conditional on rest score (i.e., the raw score on the test excluding the two items) *r* can be defined as:

$$OR_r = \frac{n_{00r} n_{11r}}{n_{01r} n_{10r}} \quad , \tag{2.25}$$

where n_{kk^*r} is the number of examinees with rest score "*r*" obtaining a score *k* on one item and a score k^* on the other, *k, k** = 0, 1. The *MH* statistic is a pooled conditional odds ratio across all possible rest scores as:

$$MH = \frac{\sum_r n_{11r} n_{00r} / n_r}{\sum_r n_{10r} n_{01r} / n_r} \quad , \tag{2.26}$$

where n_r is the total number of examinees obtaining a total rest score "*r*".

Like the OR index, the *MH* statistic is also useful in detecting local dependence for unidimensional dichotomous IRT models. If the local independence (LI) holds, the conditional covariance between the scores on the two items is close to zero, and the *MH* statistic should be near 1; if the LI is violated, the conditional covariance is positive for within-cluster items which causes the MH statistic more than 1, and negative for between-cluster items which causes the

MH statistic less than 1. As a result, when the LD is not met, the observed *MH* statistics are likely to be higher (lower) than what is expected for within-cluster (between-cluster) item pairs (Sinharay, Johnson, & Stern, 2006). This statistic has been showed to be very effective with PPMC in detecting local dependence for dichotomous items (Sinharay et al, 2005, 2006).

(5) *Absolute Item Covariance Residual*

For any pair of the items j and j^*, the absolute item covariance residual $RESID\left(j, j^*\right)$ is defined as the difference between the sample item covariance $S^2\left(j, j^*\right)$ and model-based item covariance $\sigma^2\left(j, j^*\right)$:

$$RESID\left(j, j^*\right) = \left|S^2\left(j, j^*\right) - \sigma^2\left(j, j^*\right)\right|, \tag{2.27}$$

where:

$$S^2\left(j, j^*\right) = \frac{\sum_{i=1}^{N}\left(y_{ij} - \bar{y}_{ij}\right)\left(y_{ij^*} - \bar{y}_{ij^*}\right)}{N} \tag{2.28}$$

$$\sigma^2\left(j, j^*\right) = \frac{\sum_{i=1}^{N}\left(E\left(y_{ij}\right) - \bar{E}\left(y_{ij}\right)\right)\left(E\left(y_{ij^*}\right) - \bar{E}\left(y_{ij^*}\right)\right)}{N}, \tag{2.29}$$

where N is the total number of examinees, y_{ij} and $E\left(y_{ij}\right)$ represent the observed and expected score of i^{th} examinee on item j, respectively, and $\bar{y}_{ij}$ and $\bar{E}\left(y_{ij}\right)$ denote the mean observed and expected item score across N examinees, respectively. For two dichotomous item pairs, this residual is simplified as (Levy, 2006):

$$RESID\left(j, j^*\right) = \left|\frac{n_{11}n_{00} - n_{10}n_{01}}{N^2} - \frac{E(n_{11})E(n_{00}) - E(n_{10})E(n_{01})}{E(N^2)}\right| \tag{2.30}$$

This measure has been shown to be relatively effective for detecting the departure of response data from unidimensionality in a frequentist framework (Hattie, 1984, 1985; McDonald

& Mok, 1995). Recently, Fu et al. (2005) and Levy (2006) found that it is also effective in detecting certain types of misfit with PPMC.

(6) *Hoijtink's LI Indices*

Hoijtink (2001) developed two fit statistics based on conditional item covariances and demonstrated that they were effective for checking conditional independence (CI) for 2PL models using the PPMC method. These fit statistics can also be used for checking item independence for polytomous models. The proposed item-level statistic is:

$$CI_j = \sum_{j^* \neq j} \sum_{R^{jj^*}} \sqrt{n_{R^{jj^*}}} Cov^2\left(y_j, y_{j^*} \mid R^{jj^*}\right), \tag{2.31}$$

where y_j and y_{j^*} are the responses to item j and j^*, respectively, R^{jj^*} is an examinee's rest score if two items j and j^* are deleted, and n is the number of examinees with rest score R^{jj^*}. This fit statistic weights each conditional covariance with the number of examinees in the rest score groups in order to ensure the larger groups have more influence on the outcome. It should be noted that although this statistic is at item level, it is grouped into "pairwise measures" since it is based on conditional item covariances.

2.4.3 Previous Research

Previous research using PPMC methods with IRT models has focused on unidimensional dichotomous models. Sinharay (2005) applied the PPMC method to a number of real applications of unidimensional dichotomous IRT models. The first application was to assess which model, a simple 3PL model data or a more complicated hierarchical model, fits an operational CAT data better. The discrepancy measure used is *standard deviation (SD) of the proportion corrects* of the 10 items. Through comparing the observed and predicted SD, the

results showed that the hierarchical model explained the SDs satisfactorily. Another application is to examine the speededness in a basic skill test using two pairwise discrepancy measures (*OR* and *MH*) with PPMC. The last example used the PPMC method to check if a 3PL model can be a good fit to a real data from NAEP. Several measures were employed to evaluate different aspects of misfit including *observed score distribution*, *biserial correlation, OR*, and *MH*. The results suggested the 3PL model performs extremely well. Overall, through using several real applications, this study shows the PPMC method provides a straightforward way to evaluate different aspects of model misfit.

As follows, Sinharay, Johnson, and Stern (2006) conducted several simulation studies to show the ability of PPMC to detect a range of misfitting conditions using similar discrepancy measures as in Sinharay (2005). They included *observed score distribution, biserial correlation coefficient*, *OR*, and *MH* statistics. The results showed that the *biserial correlations* and *OR* measures can be used to detect inadequacy of Rasch models when the data are generated under 2PL/3PL model, and the *observed score distribution* measure can identify the lack of fit of a 2PL model to a 3PL data. Moreover, the *OR* and *MH* statistics were found to successfully detect misfit whenever there is a violation of the local independence assumptions (e.g., for a multidimensional or a speeded test), and the *observed score distribution* was very useful to detect misfit when the assumed ability distribution was not correct. In this study, the authors used graphical displays to present the PPMC results, providing graphical evidence about misfit.

Sinharay (2006) also used PPMC to assess item-fit of simulated and real data by using item-fit plots and the discrepancy measures based on *Orlando and Thissen (2000)'s item-fit statistics* S-X^2 and S-G^2. These Bayesian item-fit measures have reasonable Type-I error rates, false alarm rates, and acceptable power, even for a short test and/or small sample size.

Hoijtink (2001) developed two fit statistics for evaluating conditional independence (CI) and differential item functioning (DIF), then applied PPMC to evaluate the effectiveness these fit statistics. The results showed the PPMC method with these fit statistics were powerful in detecting CI and DIF for 2PL models.

Fu, Bolt, and Li (2005) used PPMC to evaluate item fit for a polytomous fusion model using a number of univariate and bivariate discrepancy measures. The univariate measures check item fit through responses to a single item which is named as "item-level" discrepancy measures in section 2.4.2. They included *Orlando and Thissen (2000)'s item-fit statistics* and *item score distribution*. Bivariate measures are based on the joint responses to an item pair which is called as "pairwise measures" in the present study. Two bivariate measures were included in their study: "*absolute item residual covariance"* and "*bivariate item response discrepancy"* which is a polytomous extension of Chen and Thissen (1997)'s chi-square LD index. It was found that bivariate item test statistics had more power in detecting misfit items than univariate statistics and moreover the absolute item covariance discrepancy measure performed best.

In the context of person-fit, type-I error rates of most statistics for 2PL and 3PL models are not consistent with empirical rates due to the use of estimated abilities rather than true abilities. Since PPMC takes into the account the uncertainty of the estimation of model parameters, Glas and Meijer (2003) applied it for assessing person fit of 3PL models using several discrepancy measures. They found that this Bayesian analysis of person fit produced reasonable Type-I error rates, even for a short test and small sample size.

Levy (2006) conducted a simulation study to explore the effectiveness of PPMC for dimensionality assessment of responses to dichotomous items. In his study, several factors that would influence dimensionality such as correlations between dimensions, data-generating model,

proportion of multidimensional items, strength of dependence, and sample size were systematically manipulated. A number of univariate (item-level) and bivariate (pairwise) discrepancy measures were investigated. The univariate measures included *proportion correct,* and *item score distribution*. The bivariate measures included *Chen and Thissen's chi-square LD index*, *Yen's Q_3 statistic*, *model-based item covariance*, *absolute item covariance residual*, *log(OR)*, and *standardized log(OR)*. It was found that the univariate measures were wholly ineffective for detecting the multidimensionality and the most effective measures were two bivariate measures: *model-based covariance* and Q_3. Furthermore, all discrepancy measures showed empirical proportion of extreme PPP-values below nominal levels, but the *model-based covariance* and Q_3 had PPP-values quite close to nominal levels. The performance of these discrepancies was also found to be related to the manipulated factors.

The studies presented so far have focused on using PPMC to check the fit of a single model. Some researchers have also used PPMC for model comparison. For example, Béguin and Glas (2001) compared the fits of one- and two-dimensional 3PL models by comparing observed and posterior predictive score distributions to a data, and found that the two models were comparable with regard to the reproduction of the observed score distribution. Li, Bolt, and Fu (2006) applied several Bayesian model comparison methods including PPMC to compare different testlet models. PPMC using the *OR* measure was found to be effective in choosing the data-generating testlet model as the best model.

2.5 MODEL COMPARISON IN A BAYESIAN FRAMEWORK

The purpose of model checking is to determine if a chosen model is appropriate. It is useful and necessary when researchers or practitioners may already have a model before collecting the data just based on their preferences, practical concerns or available software, and they want to know if this model is adequate for the data. However, when there are several available models that might fit the data, finding the best model for a particular data is always desirable. For example, for a performance assessment which measure examinee's overall math ability across two content subdomains, algebra and geometry, a simple unidimensional polytomous IRT model and a more complicated 2-dimensional polytomous model might both fit the data. In order to know if a simple unidimensional model is good enough or if a MIRT model is needed for this particular data, model comparison techniques should be employed.

There are several methods for model comparisons: (1) the likelihood ratio G^2 test statistic; (2) Akaike's Information Criterion (*AIC*; Akaike, 1974); (3) Schwarz's Bayesian Information Criterion (*BIC*; Schwarz, 1978); (4) Pseudo-Bayes Factor (*PsBF*; Geisser & Eddy, 1979; Gelfand, Dey & Chang, 1992); (5) Deviance Information Criterion (*DIC*; Spiegelhalter, Best, Carlin & van der Linde, 2002); and (6) PPMC method. Among them, the likelihood ratio G^2 statistic is only appropriate for comparing nested models (e.g., RSM, PCM, and GPCM), and the other four criteria can be used to compare either nested or non-nested models. The difference between the G^2 statistics for two nested models is distributed as a chi-square with the degrees of freedom equal to the difference between the numbers of estimated model parameters. A significant G^2 indicates the more complex model fits the data better. The *AIC* and *BIC* are information-based criteria and are often used when maximum likelihood estimates (MLE) of model parameters are obtained. For some complex IRT models, the MLE may not always be

available and thus the *AIC* and *BIC* would be not appropriate. The *DIC* and *PsBF* are two Bayesian criteria for model comparisons with MCMC estimation. In addition, as mentioned in section 2.4.3, several researchers have found that the PPMC method was also effective for model comparison when MCMC estimation methods were used (Béguin & Glas, 2001; Li et. al, 2006).

In the current study, Bayesian estimation methods with MCMC would be used for the estimation of several polytomous IRT models. As a result, Bayesian criteria (*DIC*, *PsBF*, and PPMC) would be adopted for model comparisons. Since the PPMC method has been described in previous section, only *DIC* and *PsBF* are discussed herein.

2.5.1 Pseudo-Bayes Factor (PsBF)

A common Bayesian approach to comparing the fit of two models is to compute the Bayes factor (*BF*). Consider two models (M_1 and M_2), the *BF* is defined as the posterior odds of Model 1 (M_1) to Model 2 (M_2) divided by the prior odds of M_1 to M_2. By using Bayes theorem, the *BF* further reduces to the ratio of marginal likelihoods of the data under each model:

$$BF = \frac{P(M_1 \mid y)/P(M_2 \mid y)}{P(M_1)/P(M_2)} = \frac{P(y \mid M_1)}{P(y \mid M_2)}. \tag{2.32}$$

A *BF* larger than 1 supports selection of M_1 and a value less than 1 supports selection of M_2. The relative magnitude of *BF* also can be used in evaluating the relative weight of evidence in favor of either model. For example, a value of *BF* between 1 and 3 is considered as minimal evidence in favor of M_1 , between 3 and 12 as positive evidence in favor of M_1, between 12 and 150 as strong evidence, and larger than 150 as very strong evidence in support of M_1 (Raftery, 1996).

There are several issues with the *BF* criteria. For instance, it is often difficult to calculate because the estimation of marginal likelihoods in equation (2.32) is difficult especially for

complex models. In addition, the prior has effects on the estimation of the *BF*. If the prior is improper, the *BF* is not well defined. In order to overcome these problems, an alternative criterion called Pseudo-Bayes factor (*PsBF*; Geisser & Eddy, 1979; Gelfand, Dey & Chang, 1992) has been proposed and commonly used to approximate the *BF*.

The *PsBF* method requires the calculation of cross-validation predictive densities. Let $\mathbf{y}_{(r),obs}$ denote the set of observations $\mathbf{y}_{obs}$ with the r^{th} observation omitted, and let $\boldsymbol{\eta}$ denote all the parameters under the assumed model. The cross-validation predictive density is defined as:

$$f\left(y_r \mid \mathbf{y}_{(r)}\right) = \int f\left(y_r \mid \boldsymbol{\eta}, \mathbf{y}_{(r)}\right) f\left(\boldsymbol{\eta} \mid \mathbf{y}_{(r)}\right) d\boldsymbol{\eta}. \tag{2.33}$$

The density $f\left(y_r \mid \mathbf{y}_{(r)}\right)$ indicates the values of y_r that are likely when the model is fitted to all observations except y_r. This density is also known as the conditional predictive ordinate (*CPO*). In the context of item response data, y_r represents a single examinee's response to an individual item. The product of the *CPOs* across all observations can be used as an estimate of the marginal likelihood in Equation (2.32). Thus, the PsBF can be defined as:

$$PsBF = \frac{\prod_{r=1}^{R} f\left(y_{r,obs} \mid \mathbf{y}_{(r),obs}, M_1\right)}{\prod_{r=1}^{R} f\left(y_{r,obs} \mid \mathbf{y}_{(r),obs}, M_2\right)} = \frac{\prod_{r=1}^{R} \left(CPO_r \mid M_1\right)}{\prod_{r=1}^{R} \left(CPO_r \mid M_2\right)}, \tag{2.34}$$

where *R* denotes the total number of item responses from all examinees. When comparing the models at the item level, *R* equals the number of examinees *N*. When compare the models at the test level, *R* equals the number of the responses of all examinees to all items (i.e., $R = N \times I$ where *I* is the total number of items).

In the context of IRT estimation with MCMC methods, to compute the *PsBF* index, the *CPOs* are first estimated at the level of an individual item response using the inverse likelihood of each observation for *T* draws when the chain is convergent after a sufficient burn-in period:

$$CPO_{ij} = \left(\frac{1}{T} \sum_{1}^{T} \frac{1}{f\left(y_{ij} \mid \eta_t\right)} \right)^{-1}, \tag{2.35}$$

where y_{ij} is the response of an examinee *i* on a particular item *j*, and $f\left(y_{ij} \mid \eta_t\right)$ is the likelihood of the observed item response y_{ij} based on the sampled parameter values at draw *t*. In the WinBUGS program (Spiegelhalter, Thomas, Best, & Lunn, 2003), the computation of the CPO_{ij} is very straightforward since it only requires tracing the inverse probability of each observed item response over the *T* draws from the convergent chain. The CPO_{ij} is the average of these inverse probabilities across the *T* draws which is given in the Summary Statistics in WinBUGS.

A *CPO* index for each item can be summarized by taking the log of the product of the values of CPO_{ij} across all examinees, that is:

$$CPO_j = \log\left(\prod_{i=1}^{N} CPO_{ij} \right), \tag{2.36}$$

where *N* is the total number of examinees. The preferred model for item *j* is the one returning the higher CPO_j. The corresponding *PsBF* is:

$$PsBF_j = \frac{\prod_{i=1}^{N} \left(CPO_{ij} \mid M_1\right)}{\prod_{i=1}^{N} \left(CPO_{ij} \mid M_2\right)}, \tag{2.37}$$

and larger values of the *PsBF* provides the evidence supporting M_1 for item *j*. Thus, the same conclusion can be obtained by using either CPO_j or $PsBF_j$ index. In addition, a *CPO* index for the overall test can be easily computed by taking the log of the product of the item-level CPO_j across all the items. In the current study, the two levels of *CPO* index were used: the test-level *CPO* was used to compare the models for the overall test, and the item-level *CPOs* were used to choose a better model for each items. The larger the value of *CPO*, the better the model is.

2.5.2 Deviance Information Criterion (DIC)

Another popular Bayesian model comparison criterion is *DIC* (Spiegelhalter et. al, 2002). The *DIC* is similar to AIC and BIC in that they all consider the penalty on model complexity in identifying the preferred model. AIC and BIC can be expressed as:

$$AIC = -2\log p(\mathbf{y} \mid \eta) + 2n, \tag{2.38}$$

$$BIC = -2\log p(\mathbf{y} \mid \eta) + n\log(N), \tag{2.39}$$

where $p(\mathbf{y} \mid \eta)$ is the maximum likelihood function, n denotes the total number of model parameters, and N is the total number of observations. The first component of AIC and BIC $-2\log p(\mathbf{y} \mid \eta)$ is often called the "deviance between data and model". The smaller the deviance for a model, the better the model fits the data. The second component in both indices is penalty functions for model complexity. As can be seen from the equations, the penalty function in AIC takes into account the number of model parameters, whereas, the penalty function in BIC considers the effects of both sample size and the number of parameters. As a result, BIC gives higher penalty to the number of parameters when the sample size is larger and thus tends to choose a less complex model than that selected based on AIC.

The AIC and BIC indices are often used under maximum likelihood estimation. When the model is estimated using Bayesian estimation with MCMC methods, the DIC index is widely used to compare different models. Similar to the AIC and BIC, the DIC is also composed of two terms: deviance and penalty function and defined as:

$$DIC = \overline{D}(\eta) + p_D, \tag{2.40}$$

where $\overline{D}(\eta)$, a posterior mean of the deviance between data and model, is a Bayesian measure of fit, and is computed based on the posterior distribution of the deviance:

$$\overline{D}(\eta) = E_{\eta|y}[D(\eta)] = E_{\eta|y}[-2\log p(y|\eta)]. \quad (2.41)$$

The second term, p_D, measuring model complexity, represents the effective number of parameters in the model. It is defined as the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters:

$$p_D = \overline{D}(\eta) - D(\hat{\eta}), \quad (2.42)$$

where $\hat{\eta}$ is the posterior mean of model parameters. As for AIC and BIC, the smaller the value of DIC, the better the fit of a model is. However, any difference in DICs less than 5 units for two models does not provide sufficient evidence in favor of one model over another (Spiegelhalter et al., 2003). The WinBUGS program provides the DIC index.

3.0 METHODOLOGY

The purpose of this study was twofold: (1) to explore the performance of the PPMC method in detecting aspects of lack of fit for unidimensional graded IRT models using different discrepancy measures; (2) to investigate the effectiveness of Bayesian model-comparison methods (DIC, CPO, and PPMC) for comparing different polytomous IRT models. In order to accomplish these two goals, two Monte Carlo simulation studies were conducted. In addition, the proposed Bayesian approaches to model-checking and model-comparison were further applied to real performance assessment data.

This chapter presents the methodology of this study which is organized in three sections. The first section describes simulation Study 1, including the design of the study, generation and validation of item response data, estimation of unidimensional GR models using MCMC, description of the discrepancy measures used in the study, and implementation of the PPMC method. The second section discusses simulation Study 2, including the design of the study, estimation of different types of polytomous IRT models using MCMC, and computation of different Bayesian model-comparison criteria. The last section introduces the application of the methodology to real data.

3.1 SIMULATION STUDY 1

Study 1 was intended to extend previous research on evaluating IRT model fit to polytomous IRT models and explore the performance of PPMC in checking different aspects of fit for unidimensional GR models. A variety of discrepancy measures were considered and the usefulness of these discrepancy measures in detecting different types of misfit was compared.

3.1.1 Design of Simulation Study 1

Table 3.1 Design and Conditions in Study 1

Data-Analysis Model (Ma)	Data-Generating Model (Mg)	Condition Number	Violated Assumption
Unidimensional GR	Unidimensional GR	1	None
	2-dimension simple-structure GR • *Case1:* correlation (dim1, dim2) = 0.3 • *Case2:* correlation (dim1, dim2) = 0.6	2	Unidimensionality
	2-dimension complex-structure GR (one dominant dimension) • *Case1:* mild dependence a2/a1 = 0.5 • *Case2:* large dependence a2/a1 = 1.0	3	Local Independence
	Testlet GR • *Case1:* mild dependence $\sigma^2_{d(i)} = 0.5$ • *Case2:* large dependence $\sigma^2_{d(i)} = 1.0$ • *Case2:* extreme dependence $\sigma^2_{d(i)} = 2.0$	4	
	Some items with improper BCCs	5	Item-Fit

In order to explore the performance of the PPMC method in evaluating different types of fit of a data to the unidimensional GR model, a number of response datasets were generated based on

different IRT models. All simulated data were then estimated using a unidimensional GR model. Let "Mg" denote a "data-generating model" and "Ma" denote the "data-analysis model", Table 3.1 presents the design and specific conditions used in this study.

Condition 1 represents the null condition in which the generating model and analysis model were the same, and Type-I error rates of PPMC were investigated. In Conditions 2 to 5, different types of misfit were simulated which address the main threats to the applications of unidimensional IRT models to performance assessments as reviewed in Chapter 2. In these four conditions, Mg was different from Ma, and thus the empirical power rates of the PPMC method were examined.

In Condition 2, responses were simulated based on a simple-structure 2-dimensional GR model, reflecting a violation of the assumption of unidimensionality. Large-scale performance assessments may assess a broad range of content areas and/or cognitive skills. For example, a math assessment may measure two content areas: algebra and geometry; or, a test may measure computation and high-order thinking skills. Though the tests measure student's overall ability, the responses to the tests may reflect 2 dimensions. Two levels of inter-dimensional correlation (0.3 and 0.6) were considered to reflect a high and moderate degree of multidimensionality, respectively.

In Conditions 3 and 4, two typical locally-dependent data situations in performance assessments were simulated. Condition 3 simulated responses to a test which mainly measures a dominant ability (e.g., math), but a subset of items also measure a nuisance or construct-irrelevant ability (e.g., reading). This nuisance factor may cause local dependence among the subset of items. In this condition, two levels of local dependence were considered which was represented by the ratio of the nuisance dimension slope (a_2) to the dominant dimension slope (a_1

). Condition 4 was designed to simulate responses to a test with a testlet. The items in the same testlet (e.g., a shared stimulus or passage) would be locally dependent. Three levels of testlet effect variance $\sigma^2_{d(i)}$ reflected mild, large, and extreme dependence among the testlet items.

Condition 5 was intended to evaluate the effectiveness of PPMC in assessing item fit. The responses to the misfitting items were simulated based on a function that differed from the logistic function for boundary category curves (BCCs) underlying the unidimensional GR model.

Overall, in simulation Study 1, the manipulated independent factors were the type of "data-generating model" (e.g., 2-dimensional simple-structure GR model), and the type of discrepancy measure. The dependent variable was Type-I error rates and empirical power rates for the proposed discrepancy measures. The discrepancy measures used in this study are described later.

For each condition, test length, sample size, and item parameters were fixed at typical values encountered in performance assessment applications (e.g., NAEP). Specifically, test length was fixed at 15 items and the number of response categories was fixed at 5. Since the focus of this study was on the effectiveness of different discrepancy measures with PPMC, a large sample size should be used to ensure that the model parameters can be estimated precisely and the PPMC results would not be affected by any inaccurate estimation of model parameters,. Reise and Yu (1990) examined the effect of sample size on parameter recovery for GR models and found that sample size had little effect on the recovery of ability parameters, but had an effect on item parameter recovery. They concluded that a sample size of at least 500 examinees was needed to obtain acceptable parameter estimates for the 25 items with 5-category used in their study, and sample sizes between 1000 and 2000 would be needed for more accurate estimation of item parameters. Ankenman and Stone (1992) also found that a size of 500 was the

minimum for accurate and stable parameter estimates when the ability distribution was normal. When the ability distribution was not normal, however, more than 1000 observations would be needed. De Ayala (1994) suggested a ratio of 5:1 (examinees to item parameter estimates) could provide reasonable item parameter estimation when the distribution of item responses is not extreme. Yen and Fitzpatrick (2006) further discussed that large sample sizes were needed when tests contain polytomous items that are extremely hard or extremely easy. Based on these research conclusions, the sample size was fixed at 2000 for all conditions in this study in order to ensure accurate and stable parameter estimates. For a specific condition, additional fixed factors will be described in more detail in section 3.1.2.

In order to investigate Type-I error rates and empirical power rates, multiple responses datasets were generated and the PPMC method was implemented for each generated data for each condition. In general, at least 100 replications is required for exploring Type-I error rates or empirical power in Monte Carlo simulations. However, in this study, the number of replications was set to 20 due to computing constraints of the WinBUGS program and the large number of experimental conditions. It should be noted that 20 replications may be defensible based on previous research. Previous studies have used a small number of replications especially when the WinBUGS program was used to implement MCMC estimation. For example, Bolt and Lall (2003) used only 5 replications for each condition in order to evaluate parameter recovery of multidimensional IRT models using the MCMC estimation in WinBUGS. Fu et al (2005) used 30 replications in order to examine the performance of several item-fit measures with PPMC for polytomous fusion model which was estimated in WinBUGS. Sung and Kang (2006) used 10 replications to compare the relative performance of several Bayesian model-comparison criteria using WinBUGS. In order to evaluate the MIRT approach to subscore estimation, Yao and

Boughton (2007) used 20 replications though the BMIRT rather than WinBUGS program was used in their study. Even Levy (2006) with a much more efficient computer program C++ to implement MCMC estimation and PPMC for dichotomous IRT models used 50 replications in his study.

Note that in the current study, misfit was indicated if PPP-values represented extreme values either below 0.05 or above 0.95, corresponding to a two-tailed test with α=0.10 in a hypothesis testing framework.

For each replication, the overall steps to conducting Study 1 are shown in Figure 3.1 and described as follows:

(1) Using defined item parameters and simulated ability values, generate one set of item responses **y** using Mg;

(2) Estimate Ma with the generated data using MCMC estimation in WinBUGS;

(3) Obtain the posterior distributions of model parameters $p(\boldsymbol{\omega}|\mathbf{y})$ and posterior predictive distributions of item responses $p(\mathbf{y}^{rep}|\boldsymbol{\omega})$ in WinBUGS;

(4) Save T draws of model parameters (person and item) estimates for $\boldsymbol{\omega}$ ($\boldsymbol{\omega}^n, n = 1...T$) from $p(\boldsymbol{\omega}|\mathbf{y})$ after the Markov chain has converged;

(5) Save T draws of predictive (replicated) responses $\mathbf{y}^{rep,n}$ (n=1,…,T) from the likelihood distribution $p(\mathbf{y}^{rep}|\boldsymbol{\omega}^n)$;

(6) Compute the realized discrepancy measure $D(\mathbf{y}, \boldsymbol{\omega}^n)$ for each draw of $\boldsymbol{\omega}$ based on observed response to get the realized distribution of discrepancy measure;

(7) Compute the predictive discrepancy measure $D(\mathbf{y}^{rep,n}, \boldsymbol{\omega}^n)$ for each draw of $\boldsymbol{\omega}$ based on replicated data to get the posterior predictive distribution of discrepancy measure;

(8) Estimate PPP-values using the proportion of *T* draws for which $D(\mathbf{y}^{rep,n}, \boldsymbol{\omega}^{n})$ exceeds $D(\mathbf{y}, \boldsymbol{\omega}^{n})$. Extreme PPP-values (either < 0.05 or > 0.95) indicate model misfit;

(9) Repeat Steps 1) to 8) 20 times to obtain the estimates of Type-I error rates or empirical power at a significant level of 0.10 (e.g., α=0.10).

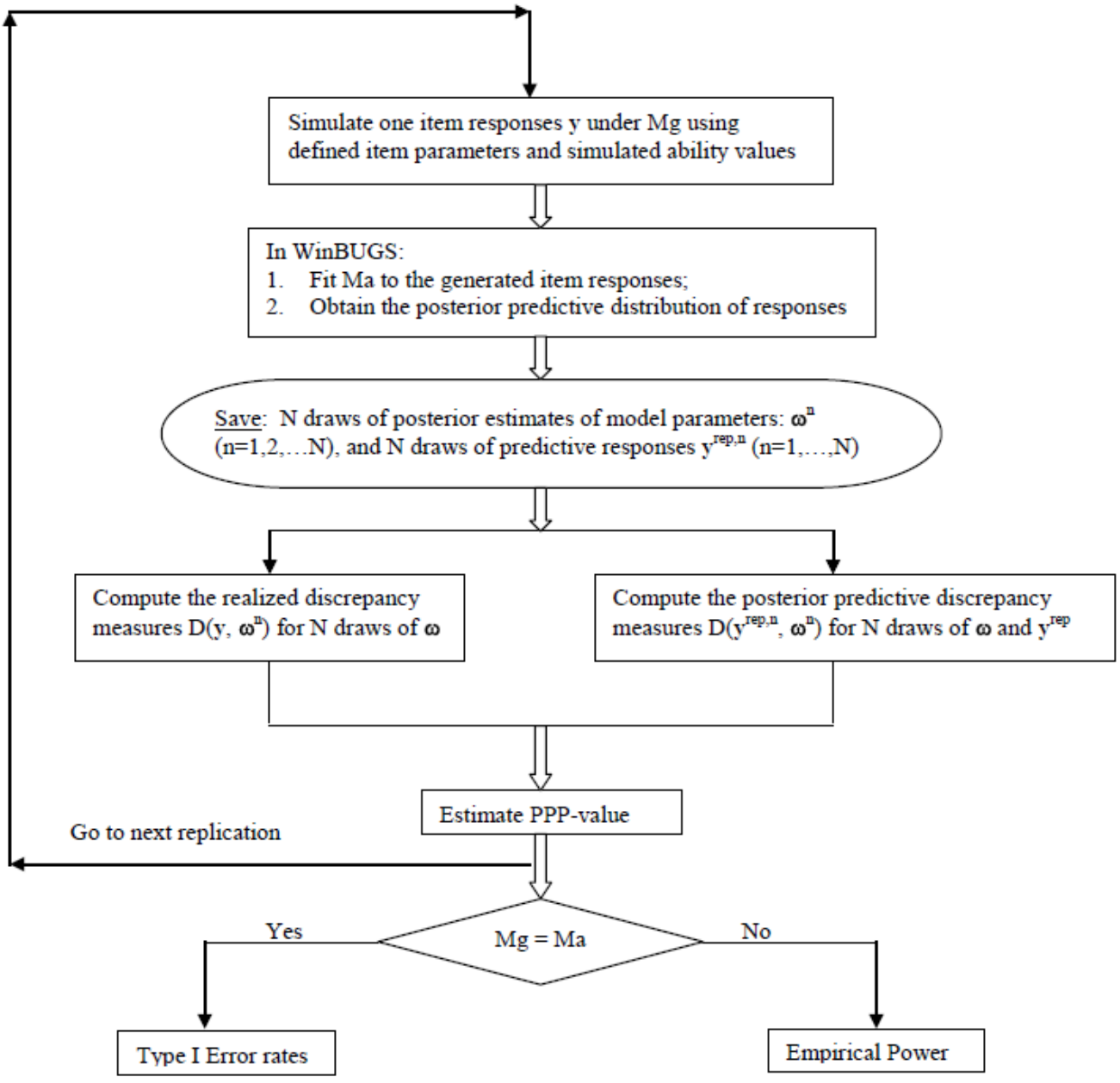


Figure 3.1 Overall Steps in Conducting Simulation Study 1

3.1.2 Generate and Validate Item Response Data

For each condition, 20 item response data sets were generated using Mg with each dataset containing responses for 2000 simulated examinees to 15 polytomous items with 5 response categories. Table 3.2 presents the item parameters used for each condition. The details about the configurations of the item parameters are discussed in each condition.

Unidimensional Graded Response Data (Condition 1)

For Condition 1, the unidimensional data were generated under Samejima's (1969) GR model (Equation 2.1). The scaling constant D was set to 1. The configuration of item parameters for the GR model (see Table 3.2) involved a combination of 3 levels for the slope parameter: 1.0, 1.7, and 2.4 (low, average and high discrimination) and 5 levels of threshold parameters reflecting varying levels of item difficulty: (1) -2.0, -1.0, 0.0, 1.0; (2) -1.5, -0.5, 0.5, 1.5; (3) -1.0, 0.0, 1.0, 2.0; (4) -3.0, -1.5, -0.5, 1.0; (5) -1.0, 0.5, 1.5, 3.0. This configuration was intended to reflect a wide range of items. The last column (b) in Table 3.2 provides the average item difficulty for each item. For items with more than two response categories, the average item difficulty is the ability level at which the expected item score on the item divided by the maximum item score is equal to 0.5, that is, the ability value at which examinees are most likely to receive half of the possible score points. As can be seen in this column, the average difficulty values covered five different levels: -1.0, -0.5, 0, 0.5, and 1.0. In addition, ability parameters for simulated examinees were randomly selected from a N (0,1) distribution.

Table 3.2 Item Parameters of the IRT Models under Conditions 1-5

	Conditions 1, 4 and 5						Condition 2						Condition 3					
Item	**a**	**b1**	**b2**	**b3**	**b4**	**b**	**a1**	**a2**	**b1**	**b2**	**b3**	**b4**	**a1**	**a2**	**b1**	**b2**	**b3**	**b4**
1	1.0	-2.0	-1.0	0.0	1.0	-0.5	1.0	0	-2.0	-1.0	0.0	1.0	1.0	1.0 (0.5)	-2.0	-1.0	0.0	1.0
2	1.0	-1.5	-0.5	0.5	1.5	0.0	1.7	0	-1.5	-0.5	0.5	1.5	1.0	1.0 (0.5)	-1.5	-0.5	0.5	1.5
3	1.0	-1.0	0.0	1.0	2.0	0.5	2.4	0	-1.0	0.0	1.0	2.0	1.0	1.0 (0.5)	-1.0	0.0	1.0	2.0
4	1.0	-3.0	-1.5	-0.5	1.0	-1.0	1.0	0	-3.0	-1.5	-0.5	1.0	1.0	1.0 (0.5)	-3.0	-1.5	-0.5	1.0
5	1.0	-1.0	0.5	1.5	3.0	1.0	1.7	0	-1.0	0.5	1.5	3.0	1.0	1.0 (0.5)	-1.0	0.5	1.5	3.0
6	1.7	-2.0	-1.0	0.0	1.0	-0.5	2.4	0	-2.0	-1.0	0.0	1.0	1.7	0	-2.0	-1.0	0.0	1.0
7	1.7	-1.5	-0.5	0.5	1.5	0.0	1.0	0	-1.5	-0.5	0.5	1.5	1.7	0	-1.5	-0.5	0.5	1.5
8	1.7	-1.0	0.0	1.0	2.0	0.5	1.7	0	-1.0	0.0	1.0	2.0	1.7	0	-1.0	0.0	1.0	2.0
9	1.7	-3.0	-1.5	-0.5	1.0	-1.0	0	2.4	-3.0	-1.5	-0.5	1.0	1.7	0	-3.0	-1.5	-0.5	1.0
10	1.7	-1.0	0.5	1.5	3.0	1.0	0	1.0	-1.0	0.5	1.5	3.0	1.7	0	-1.0	0.5	1.5	3.0
11	2.4	-2.0	-1.0	0.0	1.0	-0.5	0	1.7	-2.0	-1.0	0.0	1.0	2.4	0	-2.0	-1.0	0.0	1.0
12	2.4	-1.5	-0.5	0.5	1.5	0.0	0	2.4	-1.5	-0.5	0.5	1.5	2.4	0	-1.5	-0.5	0.5	1.5
13	2.4	-1.0	0.0	1.0	2.0	0.5	0	1.0	-1.0	0.0	1.0	2.0	2.4	0	-1.0	0.0	1.0	2.0
14	2.4	-3.0	-1.5	-0.5	1.0	-1.0	0	1.7	-3.0	-1.5	-0.5	1.0	2.4	0	-3.0	-1.5	-0.5	1.0
15	2.4	-1.0	0.5	1.5	3.0	1.0	0	2.4	-1.0	0.5	1.5	3.0	2.4	0	-1.0	0.5	1.5	3.0

Using these item and ability parameters, the probability of each examinee responding to each item response category was calculated based on the unidimensional GR model, and the cumulative probabilities were then obtained for each category. A random error component was incorporated into each response by selecting a random number from a uniform distribution *U* (0, 1) and comparing it to the cumulative probabilities for each response category. The ordinal position of the first cumulative probability that was greater than the random number was taken as the examinee's response to the item. The SAS code used to generate unidimensional GR data is attached in Appendix A.

Two methods were used to validate the data generation process in Condition 1. The first method involved comparing observed and model-based expected proportions of examinees responding in each category for each item. An extreme sample size of 20,000 examinees were simulated with the same item parameters, but all examinees had a fixed ability value (θ=0), and the observed and expected proportions were compared. If the data generation procedure was valid, the differences between the two proportions would be small. Table 3.3 presents the absolute differences between the observed and expected proportions at each response category level for each item. As can be seen from this table, the largest absolute difference was 0.006 and the average absolute difference across all categories and all items was 0.002. The small differences provided evidence for the validation of the data generation process.

Table 3.3 Absolute Differences between Observed and Expected Proportions under GR Model

Item	Cat1	Cat2	Cat3	Cat4	Cat5
1	0.005	0.003	0.000	0.001	0.001
2	0.003	0.002	0.002	0.001	0.003
3	0.001	0.002	0.002	0.001	0.000
4	0.001	0.001	0.003	0.000	0.003
5	0.002	0.004	0.001	0.001	0.000
6	0.002	0.003	0.000	0.001	0.004
7	0.003	0.001	0.004	0.004	0.003
8	0.001	0.002	0.004	0.002	0.001
9	0.001	0.001	0.003	0.006	0.003
10	0.002	0.001	0.000	0.000	0.000
11	0.001	0.001	0.002	0.004	0.001
12	0.000	0.002	0.003	0.002	0.000
13	0.001	0.002	0.002	0.000	0.000
14	0.000	0.001	0.001	0.001	0.002
15	0.004	0.002	0.002	0.000	0.000

The second method used to validate the generation of unidimensional GR data involved checking item parameter recovery. If the data were properly simulated, item parameter estimates should be close to true values. For this check, a large set of responses (10,000 examinees) were generated and the GR model was estimated using MULTILOG (Thissen, 1991). Table 3.4 provides the true and estimated parameters. As can be seen from this table, item parameters were recovered very well, again providing support for the validation of the data generation process.

Table 3.4 Item Parameter Recovery under Unidimensional GR Model

Item	True					Estimates				
	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	1.0	-2.0	-1.0	0.0	1.0	1.01	-2.01	-0.99	0.01	1.02
2	1.0	-1.5	-0.5	0.5	1.5	1.00	-1.50	-0.51	0.52	1.49
3	1.0	-1.0	0.0	1.0	2.0	1.00	-0.95	0.03	1.01	1.98
4	1.0	-3.0	-1.5	-0.5	1.0	0.97	-3.02	-1.51	-0.48	1.05
5	1.0	-1.0	0.5	1.5	3.0	0.98	-1.02	0.54	1.54	3.09
6	1.7	-2.0	-1.0	0.0	1.0	1.71	-1.99	-1.01	0.01	1.01
7	1.7	-1.5	-0.5	0.5	1.5	1.68	-1.50	-0.48	0.51	1.52
8	1.7	-1.0	0.0	1.0	2.0	1.70	-0.97	0.01	0.99	1.97
9	1.7	-3.0	-1.5	-0.5	1.0	1.67	-3.03	-1.54	-0.48	1.01
10	1.7	-1.0	0.5	1.5	3.0	1.70	-1.00	0.51	1.52	3.06
11	2.4	-2.0	-1.0	0.0	1.0	2.39	-1.97	-0.99	0.01	1.03
12	2.4	-1.5	-0.5	0.5	1.5	2.40	-1.47	-0.49	0.52	1.53
13	2.4	-1.0	0.0	1.0	2.0	2.42	-0.99	0.02	1.01	1.97
14	2.4	-3.0	-1.5	-0.5	1.0	2.41	-2.95	-1.49	-0.48	1.00
15	2.4	-1.0	0.5	1.5	3.0	2.40	-0.99	0.51	1.53	3.02

Two-dimensional Simple-Structure Graded Response Data (Condition 2)

For Condition 2, a multidimensional extension of GR model discussed by De Ayala (1994) was used to generate 2-dimensional simple-structure data. Based on this extended model, the probability of an examinee with ability Θ receiving a category score x (x = 1, 2… m_i) or higher on item i ($P_{ix}^{*}(\Theta)$) is defined as:

$$P_{ix}^{*}(\Theta) = \frac{\exp\left[D\sum_{h} a_{ih}(\theta_h - b_{ix}) \right]}{1+\exp\left[D\sum_{h} a_{ih}(\theta_h - b_{ix}) \right]} \tag{3.1}$$

where

D is the scaling constant (1.7 or 1),

a_{ih} is the discrimination (slope) parameter of item i on dimension h,

θ_h is the ability level on dimension *h*, and

b_{ix} is the threshold parameter for category *x* of item *i*.

Similar to the GR model, $P_{ix}^{*}(\Theta)$ is the cumulative probability, and the probability of responding in a particular category, $P_{ix}(\Theta)$, equals the difference between the cumulative probabilities for adjacent categories.

In this condition, the total test measured two dimensions but each item measured only one dimension (simple-structure condition). The first 8 items were designed to measure the first dimension, and the remaining 7 items were designed to measure a second dimension. As can be seen from Table 3.2, the threshold parameters were the same as for Condition 1, but the configuration of the slope parameters was different from Condition 1 though there were still three levels 1.0, 1.7, and 2.4. This configuration was intended to ensure that the two dimensions had items with similar discrimination.

The correlation between two dimensions was fixed at one of two levels: 0.3 or 0.6. The 0.6 case represents typical correlations among different dimensions for many large-scale operational tests which cover a wide range of content domains. The rationale for the 0.3 case was based on Levy (2006)'s study in which the effect of the inter-dimensional correlations on PPMC was investigated. In his study, the correlation was varied across four levels: 0, 0.3, 0.7 and 0.9, and the general finding was that the power of PPMC in detecting misfit increased as the inter-correlation decreased. Specifically, PPMC performed better for two lower level correlations (0 and 0.3) but worse for two higher level correlations (0.7 and 0.9). However, the performance of PPMC did not increase when the correlation decreased from 0.3 to 0 or from 0.9 to 0.7. As a result, a value of 0.3 was used in the current study to reflect a low correlation between dimensions and this level provided a useful comparison with the more moderate correlation case

(0.6). Ability parameters for two dimensions were randomly selected from a bivariate normal distribution (0, 1) with the specified correlation (0.3 or 0.6).

Using the above model parameters and following the same logic for generating unidimensional GR data under Condition 1, 2-dimensional simple-structure item responses were generated by computing the cumulative probabilities for each response category based on Equation (3.1), and then comparing the probability to a random number from a uniform distribution U (0, 1).

In order to validate the data generation process for 2-dimensional simple-structure, exploratory factor analyses of two generated datasets were conducted using the robust WLS estimation approach in Mplus (Muthén & Muthén, 2006). Table 3.5 provides the Root Mean Square Residual (RMSR) fit indices for the one-factor and two-factor solutions as well as promax rotated factor loadings from the analyses. For each of the two cases (correlation 0.3 and 0.6), the two-factor model fit the data significantly better than a one-factor model since the RMSR values were much less for the two-factor model and also below the recommended critical value of 0.08. In addition, as seen from the promax rotation loading pattern, the first 8 items loaded on the first factor and the remaining 7 items loaded on the second factor. The estimated correlations between two dimensions were 0.27 and 0.53 and close to the true correlations (0.3 and 0.6, respectively). All these results indicated the data were properly generated.

Table 3.5 Factor Analyses of Generated 2-dimensional Simple-Structure Data

	Correlation = 0.3			Correlation = 0.6		
Item	F1	F2		F1	F2	
1	**0.51**	0.00		**0.52**	-0.02	
2	**0.69**	0.01		**0.68**	0.02	
3	**0.80**	0.00	Promax factor correlation: **0.27**	**0.78**	0.06	Promax factor correlation: **0.53**
4	**0.47**	0.04		**0.52**	0.00	
5	**0.69**	-0.01		**0.69**	0.03	
6	**0.79**	0.00		**0.78**	0.06	
7	**0.53**	0.02		**0.48**	0.01	
8	**0.68**	0.00		**0.71**	-0.00	
9	0.04	**0.80**	One-factor RMSR: *0.17*	-0.01	**0.80**	One-factor RMSR: *0.10*
10	-0.01	**0.49**		0.01	**0.49**	
11	0.01	**0.69**	Two-factor RMSR: *0.01*	0.03	**0.68**	Two-factor RMSR: *0.01*
12	-0.01	**0.79**		0.06	**0.76**	
13	-0.05	**0.53**		0.07	**0.45**	
14	0.00	**0.68**		0.01	**0.70**	
15	0.01	**0.79**		0.05	**0.76**	

Local Dependent Data due to a Nuisance Factor (Condition 3)

In Condition 3, a 2-dimensional GR model (Equation 3.1) was used to simulate responses to a test in which all items measure a dominant ability (e.g., math), but a subset of items also measure a nuisance or construct-irrelevant dimension (e.g., reading). This nuisance factor may cause local dependence among the subset of items. Compared with the simple-structure data in Condition 2, the data here reflects a complex-structure since some items were designed to measure two dimensions (i.e., dominant and nuisance) at the same time.

As shown in Table 3.2, all 15 items measured a dominant dimension, but the first 5 items also measured a nuisance dimension. The number of items reflecting two dimensions (5 items) accounted for 1/3 of the total number of test items (15). The slope parameters for the dominant dimension and the threshold parameters had the same basic configuration as in Condition 1. Ackerman (1996) pointed out that if all the multidimensional items were easy or hard items, any

pattern of misfit may be attributable to item difficulty rather than dimensionality. As a result, the five 2-dimensional items in the current study were intended to cover all five levels of threshold parameters in order to not confound threshold and dimensionality.

The degree to which simulated examinees' performance on an item was determined by the nuisance factor was captured by the ratio of the slope parameter (a_1) for the dominant dimension to the slope parameter (a_2) for the nuisance dimension. Levy (2006) varied the ratio of a_2 to a_1 from 0.25 to 0.5 to 0.75 to 1.0 in order to vary the strength of dependence on the nuisance factor going from weak to strong. He found that the performance of PPMC improved as the strength of dependence on auxiliary dimension increased, and for the lowest ratio of a_2 to a_1 (0.25), it was hard to detect the misfit of a unidimensional model. This may be reasonable since item performance was determined primarily by the dominant dimension. It was also found that when the dependence was strong (0.75 and 1.0), the PPMC method performed almost equally well. Based on his findings, the ratio of a_2 to a_1 for the first 5 items was set to two levels (0.5 and 1.0) in the current study. These values reflected mild to large dependence between the dominant and nuisance dimensions. As can be seen from Table 3.2, the same slope values for the dominant dimension (a_1 = 1.0) were used for the first 5 items, and the second slope parameters (a_2) were all 1.0 when the ratio of a_2 to a_1 was 1.0, and all 0.5 when the ratio of a_2 to a_1 was 0.5.

The correlation between the dominant dimension and the nuisance dimension was fixed at a low level 0.3 because the test was designed to measure one dominant ability dimension. Ability parameters for two dimensions were randomly selected from a bivariate normal distribution (0, 1) with the specified correlation (0.3 or 0.6) for each case.

Based on the above specified model parameters, the 2-dimensional complex-structure item responses were generated using a SAS program. As for Condition 2, cumulative

probabilities for each of the response categories were computed based on Equation (3.1) and then compared to a random number from a uniform distribution U (0, 1).

For this condition, it was expected that the underlying factor structure of the generated response data would have only one main dimension since all items were designed to measure a dominant dimension. On the other hand, it was also expected that the generated responses to the first five items would be dependent because of a nuisance dimension also being measured. In order to validate the data generation, two evaluations were conducted. One was to test if the data were essentially unidimensional, and the other was to test if the first five items were locally dependent.

The factor structures of the generated data for two cases (a_2 = 1.0 and a_2 = 0.5) were examined using the robust WLS estimation approach implemented in Mplus (Muthén & Muthén, 2006). For the dataset with a_2 = 1.0, the largest eigenvalue was 7.789 and the second largest was 1.098, and all other eigenvalues were less than 1. For the dataset with a_2 = 0.5, the largest eigenvalue was 7.632, and all other eigenvalues were less than 1. These factor analysis results provided evidence that the generated response data were essentially unidimensional.

The IRTFIT macro (Bjorner et al., 2007) was used to evaluate any local item dependencies in the generated responses. This macro compared the observed and expected counts in a cross-tabulation table for each item pair, and two local dependence statistics were calculated: (1) a chi-square fit statistics which is a polytomous extension of Chen and Thissen's (1997) chi-square local dependence index for dichotomous items; (2) a residual correlation based on the difference between predicted and observed polychoric correlations. Before using the IRTFIT macro to conduct local dependence test, the generated responses were calibrated in

MULTILOG using unidimensional GR models and the item parameter estimates were then used with the IRTFIT macro.

Table 3.6 Local Dependence Test (p-values of Chi-Square Statistics) in IRTFIT – Case 2

Chi-square (p-value)	*Y1*	*Y2*	*Y3*	*Y4*	*Y5*	*Y6*	*Y7*	*Y8*	*Y9*	*Y10*	*Y11*	*Y12*	*Y13*	*Y14*
Y2	**0.00**	.	.	.	.	.	.	.	.	.	.	.	.	.
Y3	**0.02**	**0.00**	.	.	.	.	.	.	.	.	.	.	.	.
Y4	**0.00**	**0.00**	**0.00**	.	.	.	.	.	.	.	.	.	.	.
Y5	**0.01**	**0.00**	**0.00**	**0.00**	.	.	.	.	.	.	.	.	.	.
Y6	0.27	0.19	0.19	0.04	0.47	.	.	.	.	.	.	.	.	.
Y7	0.46	0.31	0.32	0.98	0.22	*0.55*	.	.	.	.	.	.	.	.
Y8	0.69	0.20	0.08	0.04	0.02	*0.72*	*0.95*	.	.	.	.	.	.	.
Y9	0.27	0.15	0.14	0.00	0.19	*0.56*	*0.63*	*0.54*	.	.	.	.	.	.
Y10	0.87	0.51	0.29	0.31	0.37	*1.00*	*0.68*	*0.08*	*0.27*	.	.	.	.	.
Y11	0.61	0.51	0.47	0.04	0.04	*0.42*	*0.57*	*0.88*	*0.91*	*0.50*	.	.	.	.
Y12	0.78	0.40	0.05	0.46	0.65	*0.35*	*0.82*	*0.33*	*0.86*	*0.34*	*0.80*	.	.	.
Y13	0.82	0.70	0.34	0.12	0.22	*0.19*	*0.04*	*0.76*	*0.92*	*0.77*	*0.63*	*0.45*	.	.
Y14	0.22	0.57	0.25	0.04	0.06	*0.55*	*0.53*	*0.80*	*0.91*	*0.38*	*0.98*	*0.37*	*0.16*	.
Y15	0.06	0.32	0.47	0.78	0.20	*0.21*	*0.30*	*0.47*	*0.23*	*0.64*	*0.36*	*0.71*	*0.52*	*0.14*

Table 3.6 provides the p-values for the chi-square tests for all item pairs based on the generated response data for Case 2 (a_2/a_1 = 1.0). The elements in bold represent p-values for the pairs of the simulated dependent items (i.e., Items 1-5), the italicized elements represent the p-values for the pairs of the simulated independent items (i.e., Items 6-15), and the remaining elements in the table reflect the item pairs between the independent and dependent items. As can be seen, the chi-square tests were significant for the dependent item pairs, indicating that the null hypothesis of local item independence was rejected for item pairs for the first 5 items. However, for item pairs for the 10 independent items or the pairs reflecting independent and dependent items, most of the p-values were not significant, suggesting there was no sufficient evidence to reject the local independence assumption. Table 3.7 presents the residual correlations for all the

item pairs. It can be seen that the residual correlations for the dependent item pairs (elements in bold) were quite large relative to the other item pairs.

Table 3.7 Local Dependence Tests (Residual Correlations) in IRTFIT - Case 2

Residual Correlation	*Y1*	*Y2*	*Y3*	*Y4*	*Y5*	*Y6*	*Y7*	*Y8*	*Y9*	*Y10*	*Y11*	*Y12*	*Y13*	*Y14*
Y2	**0.12**	**.**	**.**	**.**	.	.	.	.	.	.	.	.	.	.
Y3	**0.09**	**0.12**	**.**	**.**	.	.	.	.	.	.	.	.	.	.
Y4	**0.11**	**0.11**	**0.12**	**.**	.	.	.	.	.	.	.	.	.	.
Y5	**0.09**	**0.10**	**0.11**	**0.09**	.	.	.	.	.	.	.	.	.	.
Y6	-0.04	-0.02	-0.03	-0.03	-0.03	.	.	.	.	.	.	.	.	.
Y7	-0.04	-0.01	-0.03	-0.02	-0.01	*-0.01*	.	.	.	.	.	.	.	.
Y8	-0.01	-0.03	-0.02	-0.01	-0.05	*0.00*	*0.01*	.	.	.	.	.	.	.
Y9	-0.03	-0.03	-0.04	-0.05	-0.03	*0.02*	*0.00*	*-0.02*	.	.	.	.	.	.
Y10	-0.01	-0.03	-0.02	-0.04	0.01	*-0.01*	*0.00*	*-0.01*	*0.00*	.	.	.	.	.
Y11	-0.03	-0.03	-0.03	-0.06	-0.01	*0.00*	*0.01*	*0.01*	*0.01*	*0.01*	.	.	.	.
Y12	-0.02	-0.02	-0.03	0.00	-0.01	*0.00*	*-0.02*	*0.03*	*-0.02*	*-0.01*	*-0.00*	.	.	.
Y13	-0.00	-0.03	-0.04	-0.03	-0.03	*-0.00*	*-0.00*	*0.01*	*-0.00*	*-0.01*	*0.00*	*0.02*	.	.
Y14	-0.03	-0.01	-0.04	-0.02	-0.02	*0.01*	*-0.00*	*0.00*	*0.02*	*0.00*	*-0.00*	*-0.01*	*0.00*	.
Y15	-0.04	-0.04	-0.01	-0.02	-0.04	*0.00*	*-0.00*	*-0.00*	*0.02*	*0.01*	*0.02*	*0.01*	*0.01*	*-0.01*

The results for Case 1 (a_2 = 0.5) are not shown here. For this case, the chi-square test statistic was not powerful enough to detect any dependence among the first 5 items. Only 1 of 10 p-values was lower than 0.05. However, the residual correlations results clearly indicated dependency among item pairs for Items 1 to 5.

Table 3.8 summarizes the average absolute residual correlations for different types of item pairs for two levels of dependency. The no dependence condition (a_2 = 0) was also included as a baseline for comparisons. As can be seen in this table, the average absolute residual correlation across the dependent item pairs increased as the amount of dependency increased. However, the average absolute residual correlations across the independent item pairs for two dependence cases were similar to the values in the baseline condition. In summary, results from

the factor analysis and local dependence tests imply that the locally dependent response data in Condition 3 were generated properly.

Table 3.8 Average Absolute Residual Correlations for Different Levels of Dependency

Item Pairs	Amount of Dependency		
	None (a2=0)	Mild (a2/a1=0.5)	Large (a2/a1=1)
All item pairs	0.010		
Independent		0.009	0.008
Dependent		0.044	0.106
Between		0.013	0.024

Local Dependent Data due to Testlet Effect (Condition 4)

A common circumstance in which the assumption of local independence is not likely to be true is when a test is constructed of "testlets". A testlet is defined as an aggregation of items based on a single stimulus. For example, a testlet including 3 or 4 items might be constructed based on a common reading passage. Performance assessments often contain testlets that include a more complex stimulus and a set of items paired with each stimulus (Lane & Stone, 2007). As a result, for performance assessments with testlet(s), the assumption of local independence is more likely to be violated. The item responses to the items within a testlet would be more highly related than predicted by the overall latent ability for the entire test.

For Condition 4, the locally dependent data were generated under a modified GR model for testlets proposed by Wang, Bradlow and Wainer (2002). According to their model, the probability of an examinee *j* receiving a category score *x* (*x* = 1, 2… m_i) or higher on item *i* within a testlet *d(i)* is defined as:

$$P^{*}_{jix}(\theta) = \frac{\exp\left[Da_i\left(\theta_j - b_{ix} - \gamma_{jd(i)}\right)\right]}{1+\exp\left[Da_i\left(\theta_j - b_{ix} - \gamma_{jd(i)}\right)\right]}. \qquad (3.2)$$

In this equation, a random testlet effect ($r_{jd(i)}$) is introduced to reflect an interaction for person *j* with testlet *d(i)*. Ability θ_j is typically assumed to have a *N* (0, 1) distribution, and $\gamma_{jd(i)}$ is assumed to be distributed as *N* (0, $\sigma^2_{d(i)}$). The values of $\gamma_{jd(i)}$ are specified to be constant for examinee *j* over all items within a given testlet with the constraint that $\sum_j \gamma_{jd(i)} = 0$ (Bradlow, Wainer, & Wang, 1999). For any independent item, $\gamma_{jd(i)}$ is set to be 0. The variances of the testlet effects, $\sigma^2_{d(i)}$, are testlet specific, allowing the testlet effect to vary across different testlets. As the variance increases, the amount of local dependence increases. When $\sigma^2_{d(i)}$=0, the items within the testlet can be treated conditionally independent.

In this condition, item parameters (see Table 3.2) were the same as for GR model in Condition 1 except one testlet was considered (Items 6, 7 and 8). The variance of the testlet effect was varied in order to simulate varying degrees of dependence. Based on previous research (Bradlow et. al, 1999; Dresher, 2002; Li et. al, 2006; Wainer, Bradlow, & Du, 2000; Wang et. al, 2002), the variance $\sigma^2_{d(i)}$ was specified at three levels: 0.5 (mild dependence), 1.0 (large dependence), and 2.0 (extreme dependence). Note that all values of the variance were relative to 1 which is the variance of person abilities, $\theta_j \sim N$ (0, 1), and is commonly used to identify the model. Ability parameters were randomly selected from *N* (0, 1), and the testlet effect $\gamma_{jd(i)}$ was randomly selected from a *N* (0, $\sigma^2_{d(i)}$) for the items in the testlet. Given the specified model parameters, the responses were generated using a SAS program.

In order to validate the data generation procedure, the IRTFIT macro (Bjorner et al., 2007) was again used to identify any local item dependencies in the generated response data. For three generated datasets (one for each case - $\sigma^2_{d(i)}$ = 0.5, 1.0, and 2.0, respectively), the p-values

of the chi-square local independence statistics for all the item pairs within the testlet were significant, in contrast, the tests for the independent pairs and the pairs between independent and testlet items were not significant. The average absolute residual correlations for different type of item pairs for three levels of testlet variance are provided in Table 3.9. The no testlet effect condition ($\sigma^2_{d(i)} = 0$) was also included as a baseline for comparisons. As can be seen, the average absolute residual correlations were much higher for the item pairs within the testlet than for the non-testlet item pairs or for between testlet and non-testlet item pairs. Moreover, the average residual correlations across the testlet item pairs increased as the amount of dependency increased. The results indicate the responses to a test with a testlet were generated as desired.

Table 3.9 Average Absolute Residual Correlations for Different Testlet Effects

	Variance of Testlet Effects			
Item Pairs	None $\sigma^2_{d(i)}=0$	Mild $\sigma^2_{d(i)}=0.5$	Large $\sigma^2_{d(i)}=1.0$	Extreme $\sigma^2_{d(i)}=2.0$
All item pairs	0.010			
Non-testlet		0.013	0.012	0.011
Testlet		0.130	0.273	0.443
Between		0.020	0.030	0.034

Item Responses to the Items with Improper BCCs (Condition 5)

In this condition, responses to 12 of the 15 items (Items 1-6 and Items 9-15) were generated based on the GR model where the boundary category curves (BCCs) are modeled by 2PL functions (see Equation (2.1)). However, responses to two items (Items 7 and 8) were simulated with BCCs which did not follow the common logistic functions under the GR model. The BCCs for item 7 were based on a cubic form logit function (Douglas & Cohen, 2001) rather than regular logit function:

$$P^{*}_{jix}(\theta) = \frac{\exp\left[a_i\left(\theta_j - b_{ix}\right) + c_i\theta_j^3\right]}{1 + \exp\left[a_i\left(\theta_j - b_{ix}\right) + c_i\theta_j^3\right]} \ . \tag{3.3}$$

As in Douglas & Cohen (2001), the coefficient for the cubic term c_i was set to 0.75. The functions of BCCs for Item 8 were defined using the two-step Guttman functions as in Kang and Chen (2008). Figure 3.2 illustrates the BCCs of these two items.

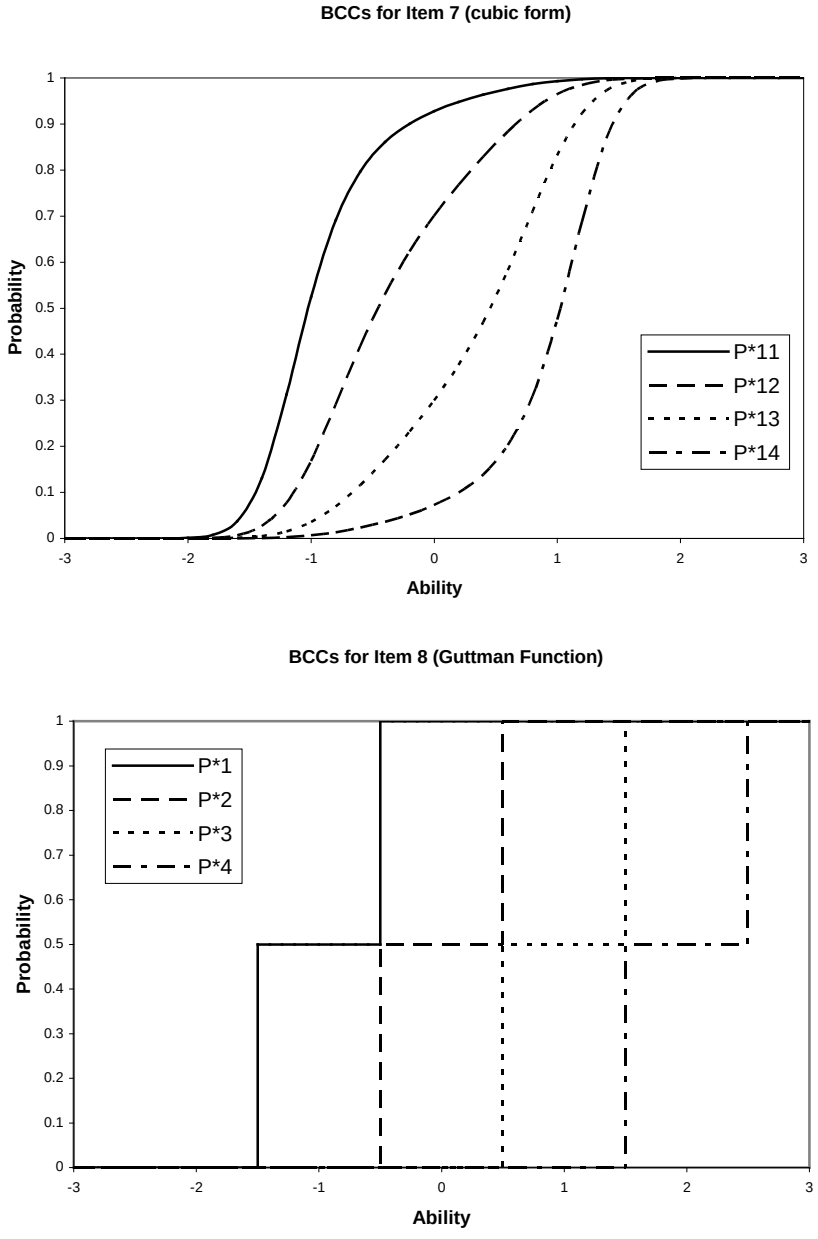


Figure 3.2 Boundary Category Curves (BCCs) for Two Misfitting Items

Item parameters for this condition had the same configuration of slope and threshold parameters as Condition 1 (see Table 3.2). Ability parameters for simulated examinees were randomly selected from the N (0, 1) distribution.

The generation of the responses to these two misfit items (Items 7 and 8) was evaluated using simulated responses for 20000 examinees at a fixed ability level (-1). The expected and observed proportions of examinees responding to each response category on Item 7 and Item 8 are reported in Table 3.10. The close match between two proportions indicates that the responses to the two misfit items were properly generated.

Table 3.10 Expected and Observed Proportions for Two Misfitting Items

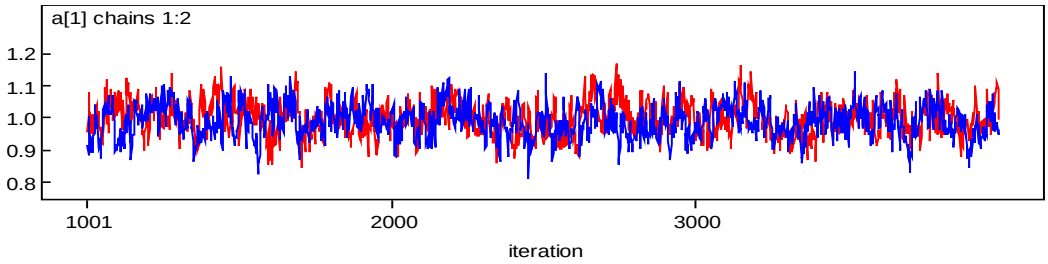
Item	Category1		Category2		Category3		Category4		Category5	
	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs
7	0.475	0.475	0.357	0.363	0.132	0.129	0.029	0.027	0.007	0.006
8	0.500	0.507	0.500	0.493	0	0	0	0	0	0

3.1.3 Estimate Unidimensional GR Model in WinBUGS

For each condition in Study 1, the data-analysis model (*Ma*) was the unidimensional GR model. In order to evaluate the fit of this model using PPMC, for each of 20 data sets generated in each condition, a unidimensional GR model was first estimated using MCMC estimation and WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003). For the GR model, the following priors were used: $\theta_i \sim Normal(0,1)$ for all persons i, and $a_j \sim Lognormal(0,1)$, $b_{j1} \sim Normal(0,0.25)$ and $b_{j(k+1)} \sim Normal(0,0.25)I(b_{jk})$ for all items j, where the notation $I(b_{jk})$ indicates that $b_{j(k+1)}$ was always sampled to be larger than b_{jk} which is a requirement

under the GR model. It should be noted precision parameters rather than variance parameters are used in these prior distributions. The WinBUGS code used for estimation of the GR model is given in Appendix B.

As reviewed in Chapter 2, convergence of parameter posterior distribution to a stationary distribution is crucial to MCMC estimation. A preliminary study was conducted to determine how long the chain should run to achieve convergence and how many iterations were needed after convergence to estimate parameters of the unidimensional GR model. One unidimensional GR dataset was generated containing responses for 2000 simulated examinees to 15 polytomous items with 5 response categories. Using WinBUGS, two chains of 4000 iterations were run. The first 1000 iterations were discarded as part of the burn-in phase, and the remaining 3000 iterations in each chain were thinned by taking every other iteration to reduce any autocorrelation among the draws. Convergence was examined through visual inspection of several convergence diagnostic plots available in WinBUGS. The first plot is a "sampling history plot" for each parameter. Figure 3.3 illustrates the histories for the slope and four threshold parameters of item 1. These sampling histories show that each chain displayed relatively quick convergence to a stationary distribution and an overlap of the sampling histories for the two chains further indicated convergence. Similar results were observed for the other items.



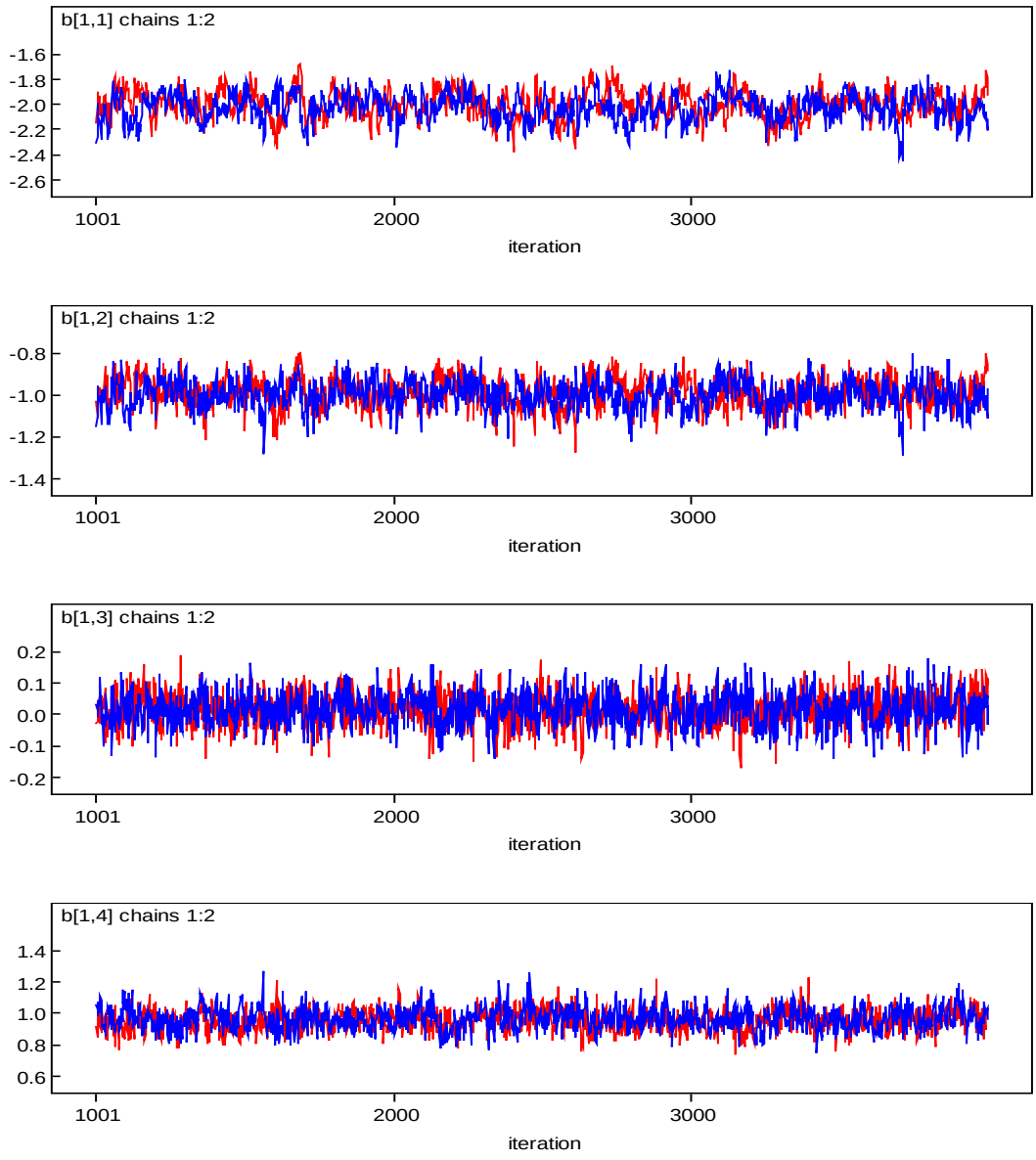


Figure 3.3 Sampling History Plots of Item Parameters Associated with Two Chains - Item 1

In WinBUGS, a "BGR diagram" is often used to show the Gelman-Rubin convergence statistic for multiple chains. It includes three lines in different colors. The green (G) and blue (B) lines reflect the pooled and within-chain posterior variances, respectively. The ratio of these two variances, that is, the Gelman-Rubin statistic, is represented by the red (R) line. Figure 3.4 includes the "BGR diagrams" for the slope and four threshold parameters of Item 1. As can been seen, the red line (Gelman-Rubin statistic) converged to 1, indicating equality between the

pooled and within-chain variances. Thus, these plots demonstrate the convergence of the two chains with 4000 iterations was attained for all the parameters of Item 1. Similar results were obtained for the other model parameters.

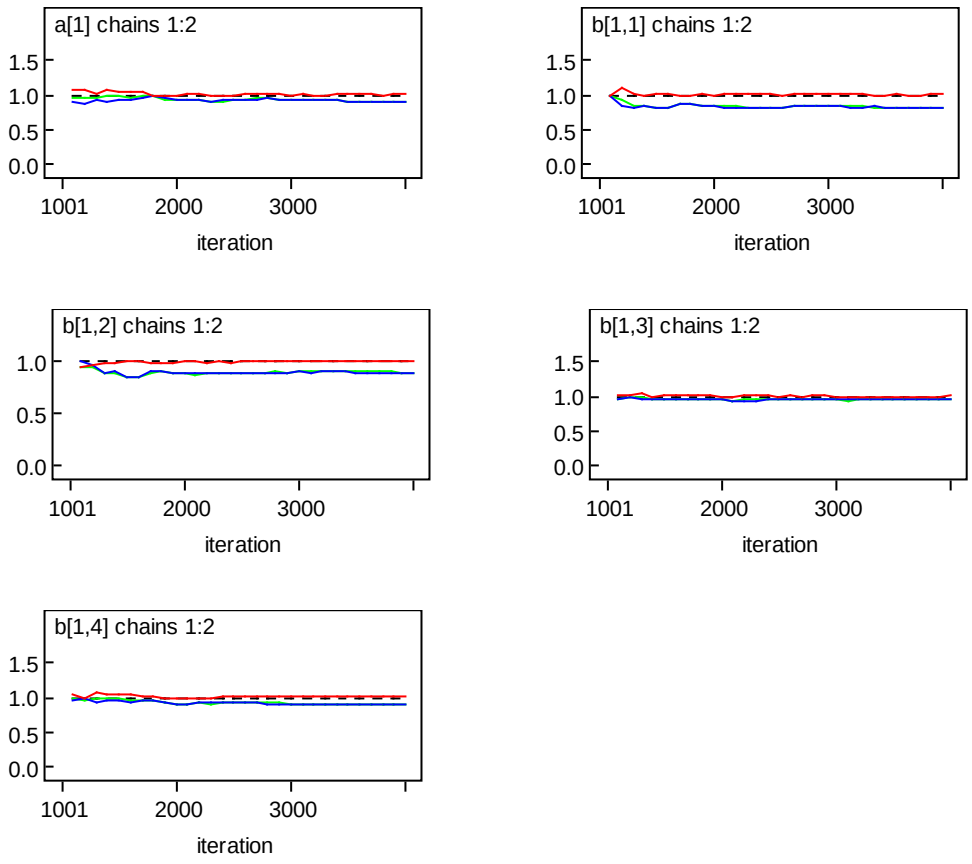


Figure 3.4 "BGR" Diagrams for the Parameters of Item 1

As reviewed in Chapter 2, autocorrelation plots are also helpful in evaluating convergence. High correlations between adjacent states imply a slow rate of convergence, thus requiring more iterations to achieve stationary posterior distributions for the model parameters. Figure 3.5 provides the autocorrelation plots for the parameters of Item 1. As can been seen, the correlations among the successive draws were reduced to 0 very quickly, indicating the length of 4000 iterations was sufficient to ensure convergence. Similar autocorrelation plots were found for other item parameters.

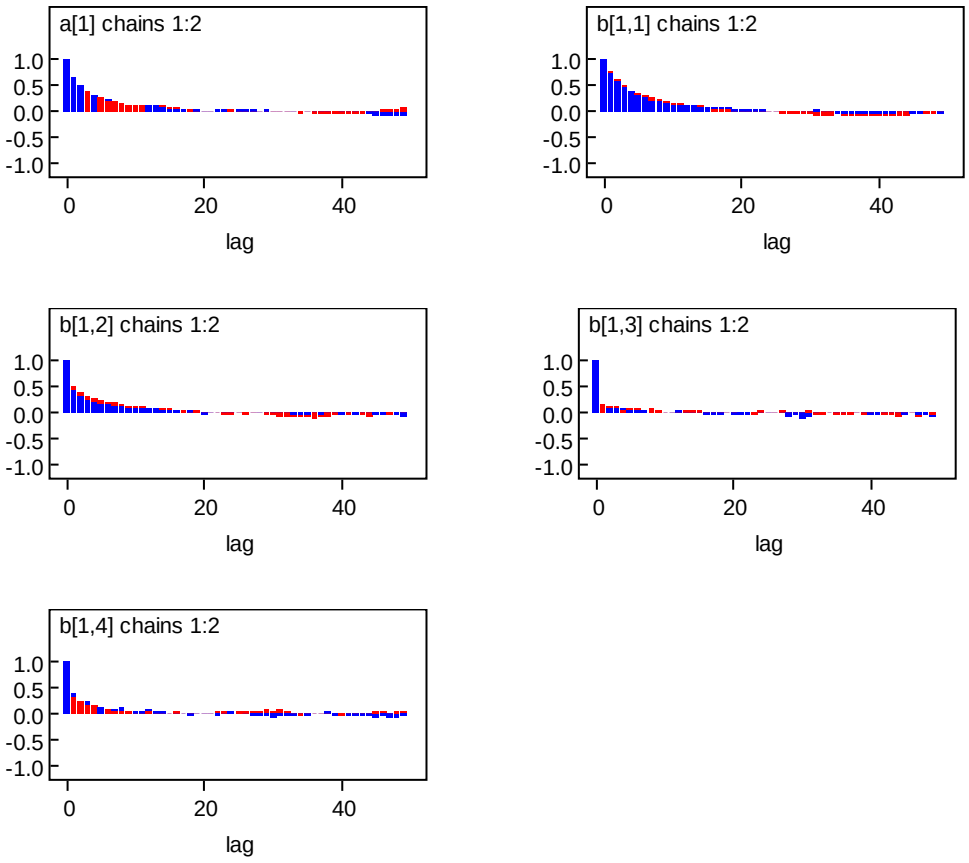


Figure 3.5 Autocorrelation Plots for the Parameters of Item 1

After the burn-in iterations were discarded and the chains were thinned, posterior estimation of model parameters was conducted based on the remaining 3000 iterations and the recovery of item parameters was examined. The degree of parameter recovery using the MCMC method is an important factor in determining whether PPMC could be implemented successfully since PPMC is based on posterior estimation of model parameters. In this preliminary study, parameter recovery was evaluated by computing the difference between the estimated and true parameter values (i.e., bias). Table 3.11 includes the generating item parameters and their corresponding estimates in WinBUGS. The average absolute bias in the estimation of slope parameters across all items was 0.061. The average absolute bias for all the threshold estimates

across all items was 0.047. The results indicated that the item parameters were recovered well using MCMC estimation in WinBUGS with two chains of 4000 iterations.

Table 3.11 Item Parameter Recovery using MCMC Estimation for the GR Model

Item	True					Estimates				
	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	1.0	-2.0	-1.0	0.0	1.0	0.99	-2.02	-1.00	0.01	0.96
2	1.0	-1.5	-0.5	0.5	1.5	1.02	-1.61	-0.51	0.50	1.50
3	1.0	-1.0	0.0	1.0	2.0	1.01	-0.99	0.03	1.02	1.99
4	1.0	-3.0	-1.5	-0.5	1.0	1.04	-2.96	-1.45	-0.50	0.94
5	1.0	-1.0	0.5	1.5	3.0	0.96	-1.04	0.45	1.54	3.08
6	1.7	-2.0	-1.0	0.0	1.0	1.69	-2.12	-1.12	-0.05	0.95
7	1.7	-1.5	-0.5	0.5	1.5	1.70	-1.53	-0.48	0.53	1.47
8	1.7	-1.0	0.0	1.0	2.0	1.68	-1.03	-0.01	1.05	2.04
9	1.7	-3.0	-1.5	-0.5	1.0	1.81	-3.17	-1.47	-0.52	0.96
10	1.7	-1.0	0.5	1.5	3.0	1.71	-1.05	0.49	1.51	2.90
11	2.4	-2.0	-1.0	0.0	1.0	2.42	-2.00	-1.04	-0.06	1.00
12	2.4	-1.5	-0.5	0.5	1.5	2.22	-1.51	-0.51	0.49	1.56
13	2.4	-1.0	0.0	1.0	2.0	2.31	-1.04	-0.05	0.96	1.92
14	2.4	-3.0	-1.5	-0.5	1.0	2.23	-3.25	-1.65	-0.56	1.03
15	2.4	-1.0	0.5	1.5	3.0	2.21	-1.04	0.49	1.54	3.19

3.1.4 Discrepancy Measures Used in Study 1

As discussed in Chapter 2, the choice of discrepancy measures is a key issue in an application of the PPMC method. It was argued that the measures should be chosen to reflect relevant threats to model fit for a specific testing application. The purpose of Study 1 was to examine the general performance of PPMC in evaluating the fit of GR model to performance assessment data. Thus, a variety of threats to model fit were considered and several different types of discrepancy measures were employed. Among the measures examined in this study, most were the polytomous extensions of measures used in the previous research for dichotomous IRT models,

and a few were newly proposed. These discrepancy measures were designed to assess model-fit at the test level, item level, or item-pair (or pair-wise) level.

The "test-level" measure involved the "*test score distribution*" (see Section 2.4.2.1). Plots comparing observed and posterior predictive distributions as well as the chi-square statistic (χ_T^2) were used to provide evidence about the fit of unidimensional GR models at the test level. Section 2.4.2.1 provides details about this measure.

The "item-level" discrepancy measures used with PPMC in previous research for dichotomous models include "*Bayesian χ^2 Statistic", "item score distribution", "Orlando and Thissen's item-fit Statistic"*, and "*item-total score correlation"* (see Section 2.4.2.2). In the current study, polytomous extensions of the "*item score distribution"* and "*item-total score correlation"* discrepancy measures were used. In addition, one traditional fit statistic "*Yen's Q_1"* and one alternative item-fit index "*Stone's fit statistic"* were employed. These two classical item-fit statistics have never been used in a Bayesian framework and were proposed in this study in order to examine their performances with PPMC. Although the "*Bayesian χ^2 Statistic"* is an intuitive and useful statistic for evaluating overall fit for many statistical models, several researchers (Li et al., 2006; Sinharay et al., 2006) found that it was not useful for IRT model checking. As a result, this measure was not considered in the current study.

Six "pair-wise" discrepancy measures used with PPMC for dichotomous IRT models were reviewed in Section 2.4.2.3. They include "*Yen's Q_3 statistic*", "*Chen and Thissen's chi-square LD index", "Odds Ration (OR)", "Mantel-Haenszel (MH) statistic", "absolute item covariance residual"*, and "*Hoijtink's conditional item covariance index*". Levy (2006) used four of them ("*Yen's Q_3 statistic*", "*OR*", "*Chen and Thissen's chi-square LD index", "absolute item covariance residual"*) with PPMC to detect the local dependence among responses to

dichotomous items, and found that "*Yen's Q_3 statistic*" was most effective and "*Chen and Thissen's chi-square LD index"* was less effective than other pair-wise measures. Based on these findings, the least effective pair-wise measure was not used in the current study, and polytomous versions of the other three measures were used. The *MH* statistic is computationally equivalent to the conditional *OR*, and both *OR* and *MH* statistics were found to successfully detect misfit whenever there is a violation of the local independence assumptions though the *MH* statistic detected model misfit more often than the *OR* (Sinharay et al., 2006). Due to their similar performance, only the *OR* statistic was employed as a discrepancy measure in the current study. *"Hoijtink's conditional item covariance index"* was not employed in this study. Though there was no previous study comparing this measure with other pair-wise measures, it was expected that its performance would be similar to "*Yen's Q_3 statistic*" since the basic rationale underlying these two measures are similar. "*Yen's Q_3"* statistic represents the correlation between responses to two items after accounting for the latent ability (i.e., conditional on the ability), and *"Hoijtink's index"* measures the covariance between the responses to two items conditional on examinees' scores based on the remaining items or rest scores. Both measures reflect conditional relationship between responses to one item pair. The main difference of *"Hoijtink's index"* from "*Yen's Q_3"* is that it uses "observed rest scores" to estimate examinees' latent ability.

In summary, one test-level measure ("*test score distribution*"), four item-level measures ("*item score distribution"*, *"item-total score correlation"*, *"Yen's Q_1"*, and *"Stone's fit statistic"*), and three pair-wise measures ("*Yen's Q_3 statistic*", "*OR*", and "*absolute item covariance residual"*) were used in the current study. Among them, two item-fit measures can be used for polytomous items (see Section 2.2.3) and thus no extension was required. The polytomous extension of the chi-square statistic (χ_T^2) measuring the difference between

observed and predicted "test score distributions" was discussed in Section 2.4.2.1. The two pair-wise measures ("*Yen's* Q_3" and "*absolute item covariance residual"*) can be extended to accommodate polytomous item responses simply by calculating expected responses using the polytomous IRT models instead of dichotomous items. After the expected responses are obtained, the computation of these two measures is the same as for dichotomous items (see Section 2.4.2.3). The following describes polytomous item extensions for the remaining three measures.

(1) *Item Score Distribution*

In Section 2.4.2.2, a goodness-of-fit statistic ($\chi_j^2\_Fit$) used to summarize the discrepancy between observed and posterior predictive item score distributions for dichotomous items was discussed. For polytomous items, this statistic can be defined as:

$$\chi_j^2\_Fit = \sum_{k=0}^{M_j} \frac{\left[O_{jk} - E_{jk}\right]^2}{E_{jk}}, \tag{3.4}$$

where M_j is the highest score on item j, and O_{jk} (E_{jk}) is the observed (predicted) number of examinees scoring in response category k on item j. E_{jk} can be calculated by summing the probabilities of responding to category k on item j across all N examinees:

$$E_{jk} = \sum_{i=1}^{N} p_{ijk}(\theta_i) \qquad \text{k=0, …, } M_j. \tag{3.5}$$

(2) *Item-Total Score Correlation*

The item-test score correlation is the correlation between examinees' total test scores and their item scores on a particular item. For dichotomous items, the item-total score correlation is commonly estimated using point-biserial or biserial correlations. Sinharay et al. (2006) have shown that the biserial correlation between item and test scores was a powerful discrepancy

measure for detecting misfit of the Rasch model when data was generated from a 2PL or 3PL model. It was therefore hypothesized that the item-total score correlation would also be effective for detecting "local dependence" among the responses to polytomous items. The underlying rationale is that local dependence might affect item discrimination and the item-test correlation is related to item discrimination. For example, Yen (1993) demonstrated that positive LD would produce higher item discriminations for LD items.

The correlation between total test scores and scores on polytomous items should be estimated using a "polyserial" correlation in theory. However, in practice, when there are a number of response categories, the Pearson product-moment correlation is often used to estimate item-total score correlation. Five response categories have been found to be a minimum in order to use Pearson correlations (Dollan, 1994). In this study, the number of response categories was 5 and Pearson correlations were used to estimate the association between items and total test scores.

(3) *Global Odds Ratios*

For dichotomous items, the contingency table for one pair of items is 2x2 and there is only one odds ratio (OR) value for one item pair (see Equation 2.24). However, the computation of an OR with a polytomous item pair involves a R x C (R>2 and C>2) contingency table from which multiple ORs can be computed. There are three basic types of odds ratios in a R x C contingency table: *local odds ratios*, *local-global odds ratios*, and *global odds ratio* (Agresti, 2002).

Local odds ratios are defined using cells in adjacent rows and adjacent columns (Agresti, 2002). Suppose two polytomous items j and j^* have the maximum score M_j and M_{j*}, respectively. That is, the total number of response categories is (M_j+1) for item j and (M_{j*}+1) for item j^*. The

corresponding contingency table is (M_j+1) x (M_{j^*}+1). Let k and k* denote the response scores on items j and j^* respectively, the (M_j * M_j) non-redundant local odds ratios can be defined as:

$$OR_{kk^*} = \frac{n_{kk^*} n_{(k+1)(k^*+1)}}{n_{(k+1)k^*} n_{k(k^*+1)}} \qquad (k = 0, 1. . . (M_j\text{-}1), \quad k^* = 0, 1\ldots (M_{j^*}\text{-}1)), \tag{3.6}$$

where n_{kk^*} is the observed number of examinees having response k on item j and response k^* on item j^*. For two items with 5 response categories (0, 1, 2, 3 and 4), there are 16 non-redundant local odds ratios.

It can be seen that the number of local odds ratios will increase dramatically as the number of response categories for each item increases. Therefore, it is not convenient to use local odds ratios when the items have a large number of response categories. One alternative way to measure the association in R x C contingency tables is to dichotomize one of the items according to a cut point and compute *local-global odds ratios*. For example, if the responses on the column item are dichotomized, the R x C contingency table will reduce to a R x 2 table and the number of non-redundant odds ratios is only *(R-1)* rather than *(R-1)*(C-1)*. For two items j and j^* with the maximum item score M_j and M_{j^*}, respectively, the local-global odds ratio is defined as:

$$OR_{kk^*}^{(LG)} = \frac{n_{k(\leq k^*)} n_{k(>k^*)}}{n_{k(>k^*)} n_{k(\leq k^*)}} \qquad (k = 0, 1. . . (M_j\text{-}1), \quad k^* = 0, 1\ldots (M_{j^*}\text{-}1)), \tag{3.7}$$

where k^* is the cut point on the response scale for item j^*, and $n_{k(\leq k^*)}$ represents the number of examinees scoring k on item j and scoring k^* and lower on item j^*. The local-global odds ratios are local with respect to the row item and global with respect to the column item. Following the same logic, global-local odds ratios can be defined as global with respect to the row item and local with respect to the column item. For two items with 5 response categories (0, 1, 2, 3 and 4),

there are four non-redundant global-local or local-global odds ratios, much smaller than the number of local odds ratios.

A single OR is often preferred in order to simplify or summarize the association in R x C contingency tables as for 2 x 2 contingency tables. In this situation, a *global odds ratio* can be computed. For two items j and j^* with the maximum item score M_j and M_{j*}, respectively, a R x C contingency table can be reduced to a 2 x 2 contingency table by dichotomizing the response categories of each item. The global odds ratio is defined as the cross-ratio of this pooled 2 x 2 table:

$$OR_{kk^*}^{(G)} = \frac{n_{(\leq k)(\leq k^*)} n_{(>k)(>k^*)}}{n_{(\leq k)(>k^*)} n_{(>k)(\leq k^*)}} \qquad (k = 0, 1. . . (M_j\text{-}1), \quad k^* = 0, 1\ldots (M_{j^*}\text{-}1)), \tag{3.8}$$

where k and k^* are the cut points on the response scales for item j and item j^* respectively, and $n_{(\leq k)(\leq k^*)}$ denotes the number of examinees scoring k and lower on item j and scoring k^* and lower on item j^*. For different cut points, the global odds ratios may be different.

In this study, only a *global odds ratio* was employed as one possible discrepancy measure due to its simplicity. The dichotomization was based on score rubrics typically used with performance assessments. For items with 5 response categories (0-4), Categories 3 and 4 were treated as "correct" responses, and 0, 1, and 2 were treated as "incorrect" responses. Thus, the cut point was set to 2.

Previous research (Sinharay et al, 2005, 2006; Li et al, 2006; Levy, 2006) has found the *OR* measure to be a useful discrepancy measure for checking several aspects of model fit for dichotomous IRT models. It was assumed that the *global OR* measure would be useful for polytomous models. However, it might be not as effective as *OR* for dichotomous items due to the dichotomization of the response categories.

3.1.5 Conduct PPMC

As reviewed in Chapter 2, conducting PPMC involves simulating replicated data under a presumed model and comparing the discrepancy measures for observed data against the distribution of discrepancy measures across the replicated data sets using graphical displays or PPP-values to evaluate model fit.

PPP-values provide a quantitative measure of the degree to which observed data would be expected under the model. PPP-values near 0.5 indicate that the realized (i.e. observed) discrepancies fall in the middle of the distribution of discrepancy measures based on the posterior predictive response data (i.e., replicated data). Such values provide evidence for model fit. In contrast, extreme PPP-values near 0 or 1 suggest that the observed discrepancies are inconsistent with the posterior predictive discrepancies and hence are indicative of model misfit. More specifically, PPP-values near 0 indicate that the predictive discrepancy values under the model are *smaller* than the realized values most of the time, indicating that the model *under-predicts* this discrepancy measure. Using the same logic, PPP-values near 1 indicate that the predictive discrepancy values are *larger* than the realized values, indicating that the model *over-predicts* the measure. In the current study, extreme PPP-values were defined as those below 0.05 or above 0.95, corresponding to a two-tailed test with α=0.10 in a hypothesis testing framework.

In addition to PPP-values, different types of graphical plots were also used in the current study to provide graphical evidence about model fit. As discussed in Chapter 2, it is more appropriate to use the PPMC approach as a diagnostic tool for model fit rather than a hypothesis test because the PPP-values are not necessarily uniformly distributed under the null conditions. Thus, a preferable way to interpret the difference between observed and predicted discrepancy measures in PPMC is also to employ graphical plots.

Within each condition for Study 1, the generated data served as "observed data", and the posterior predictive (i.e. replicated) data sets under the unidimensional GR model were simulated within WinBUGS in the process of estimating the model parameters. The values of the proposed discrepancy measures were calculated both for the observed data as well as each of the predicted data and then compared using graphical plots and PPP-values. Among all the 8 discrepancy measures investigated in this study, four measures (*"item score distribution", "Yen's Q_3", "absolute item covariance residual", "global OR"*) and their corresponding PPP-values were computed within WinBUGS. However, the remaining four discrepancy measures (*"test score distribution", "item-total score correlation", "Yen's Q_1",* and *"Stone's fit statistic"*) were calculated by inputting the replicated response data and parameter estimates for all iterations (CODA output) from WinBUGS into SAS. If we label the first set of 4 measures as PPMC1 measures, and the remaining 4 measures as PPMC2 measures, the general steps to implement PPMC in Study 1 are as follows:

1) Generate a unidimensional GR data in SAS;
2) Run WinBUGS from SAS through a batch file to estimate the generated data using a unidimensional GR model, simulate replicated response data, and compute the PPP-values of the four PPMC1 measures. In addition, save the replicated response data and parameter estimates for all iterations (CODA files) into text files for the next implementation of PPMC based on the four PPMC2 measures. Also save the CODA files for the realized and predictive discrepancies in order to compare them using graphical plots;
3) Read these CODA text files from (2) into SAS datasets;
4) Compute the realized and predictive values of the PPMC2 discrepancy measures based on observed data (i.e., generated data) and the CODA datasets from (3) in SAS, and then

obtain their PPP-values. As for the PPMC1 measures, save the realized and predictive discrepancies in order to draw graphical plots.

The preliminary study conducted in Section 3.1.3 used two chains of 4000 iterations. The results showed that each chain converged very quickly and the item parameters were well recovered. Based on those results, only one chain of length of 4000 was run for conducting PPMC due to the intensive computation in WinBUGS. The first 3500 iterations in each chain were discarded as part of the burn-in phase, and posterior estimation of model parameters and PPMC were conducted based on the 500 remaining iterations. Item recovery using the posterior sample of 500 was evaluated using the Root Mean Square Difference (RMSD) statistic. This statistic compared the true (or generating) and estimated parameters across 20 replications, as follows:

$$RMSD = \sqrt{\frac{\sum_{n=1}^{20}(true - estimate)^2}{20}}. \qquad (3.9)$$

The results indicated that a posterior sample size of 500 was adequate for accurate recovery of item parameters for GR model (see Chapter 4). In addition, this sample size was consistent with previous studies (Fu et al., 2005; Levy, 2006; Li et al, 2006).

To investigate Type-I error rates and empirical power for each discrepancy measure proposed, the PPMC analysis was replicated 20 times (one for each generated data) within each condition. The proportion of the 20 replications with extreme PPP-values (< 0.05 or > 0.95) for each discrepancy measures provides estimates of Type-I error rates of this measure under the null condition (Condition 1) or estimates of empirical power rates of this measure under other misfit conditions (Conditions 2-5). It should be noted that for each replication, different types of discrepancy measures resulted in different numbers of PPP-values. For any replication, the test-

level chi-square measure was evaluated once leading to one PPP-value; each item-level discrepancy measure was evaluated 15 times (once for each item) leading to 15 PPP-values; and each pair-wise discrepancy measure was evaluated 105 times (one for each unique pairing of items) leading to 105 PPP-values. In order to summarize results, PPP-values for item-level and pair-wise level measures were pooled based on data structure. Type-I error rates and empirical power rates were based on these pooled PPP-values. The details are discussed in the results chapter.

Appendix C provides the WinBUGS code used for the implementation of PPMC based on the four PPMC1 measures including estimating unidimensional GR models, calculating these four discrepancy measures and their PPP-values, as well as simulating replicated response data. In addition, the SAS code used to create a batch file for running PPMC in WinBUGS from SAS is given in Appendix D. The SAS code for conducting PPMC using the four PPMC2 measures is available from the author upon request.

3.2 SIMULATION STUDY 2

As reviewed in Chapter 2, unidimensional polytomous IRT models are commonly used in the analysis of performance assessment data. However, the underlying assumptions such as unidimensionality and local item independence are most likely to be violated for performance assessment data. In that situation, more complex polytomous models might be needed to account for the violation of assumptions. For example, a multidimensional GR model (De Ayala, 1994) may be more appropriate for analyzing multidimensional performance data, or a modified GR model for testlets (Wang, Bradlow & Wainer, 2002) may be more appropriate for performance assessments that involve a subset of items with a common stimulus. In order to choose the preferred model for a particular performance assessment data, model comparison tools may be employed. A number of model comparison techniques in a Bayesian framework have been reviewed previously. The purpose of Study 2 was to investigate the relative performance of three Bayesian model selection methods (DIC, CPO, and PPMC) in choosing the preferred model for analyzing performance assessment data.

3.2.1 Design of Simulation Study 2

In order to explore the relative performance of these three Bayesian model comparison methods, four conditions were considered (Table 3.12). In each condition, typical performance assessment data were generated based on an appropriate IRT model (Mg) and then calibrated using several different data-analysis (Ma) models. Three Bayesian model comparison indices were then computed for each Ma and the preferred model was selected based on each of indices. Indices

were then examined to determine the extent to which "Mg" was selected as the "preferred" model.

Table 3.12 Design and Conditions in Simulation Study 2

Data-Generating Model (Mg)	Data-Analysis Model (Ma)	Condition Number
2P GR	(1) 2P GR (2) 1P GR (3) RS	1
2-dim Simple-Structure GR	(1) 2P GR (2) 2-dim Simple-Structure GR	2
2-dim Complex-Structure GR	(1) 2P GR (2) 2-dim Complex-Structure GR	3
Testlet GR	(1) 2P GR (2) Testlet GR	4

In Condition 1, the responses were generated under the two-parameter Samejima's (1969) GR model, but estimated using two restricted GR models (one-parameter (1P) GR model and RS model) in addition to the true model (two-parameter (2P) GR model). Both one-parameter (1P) GR and RS models require fewer parameters to estimate than two-parameter (2P) GR models. A 1P GR model is similar to Samejima's 2P GR model except that all slope parameters are fixed to a single value. As a result, only one slope parameter needs to be estimated. The RS model developed by Muraki (1990) is a restricted case of the 2P GR model for analyzing responses to the items with a rating-scale type response format. Lane and Stone (2006) pointed out that this RS model may be appropriate for performance assessments where a general rubric is used as the basis for developing specific item rubrics since the response scales and the differences between score levels may be the same across the set of items. In the RS model, the threshold parameters of the 2P GR model are partitioned into two terms: a location parameter for each item, and one set of category threshold parameters for all items. The number of parameters in the RS model is therefore reduced greatly as compared with the 2P GR model. The purpose of Condition 1 was

designed to determine if the model comparison criteria could discriminate between these three models and select the 2P GR model as the preferred model.

Similar to some of the conditions examined in Study 1, the data generated in Conditions 2, 3 and 4 (see Table 3.12) reflect typical performance assessment applications in which the assumptions underlying unidimensional GR model are violated. Specifically, in Condition 2, 2-dimensional (2-dim) simple-structure GR responses were generated based on a multidimensional GR (MGR) model (De Ayala, 1994) to reflect the violation of the unidimensionality assumption. In Condition 3, 2-dimensional (2-dim) complex-structure GR data were simulated under a MGR model to represent responses to a performance assessment which mainly measures a dominant ability (e.g., math), but a subset of items also measure a nuisance or construct-irrelevant dimension (e.g., reading). This nuisance factor would also result in local dependence among the subset of items. In Condition 4, responses to a test with a testlet were generated under a modified GR model for testlet (Wang, Bradlow & Wainer, 2002). The responses to items within a testlet (e.g., a shared stimulus or passage) would be locally dependent. In each of these three conditions, the generated data was calibrated using both the 2P GR model and the more complex data-generating model in order to determine whether the model comparison tools were useful in selecting the complex models as the preferred model when the underlying assumptions of the GR model did not hold.

3.2.2 Generate Item Response Data

As for Study 1, 20 datasets were generated for each condition using the Mg with each dataset containing responses for 2000 simulated examinees to 15 polytomous items with 5 response categories. For Condition 1, the configurations of item parameters for unidimensional 2P GR

models and the procedure for data generation were the same as for Condition 1 in Study 1. For other conditions, the data generation procedures were the same as the corresponding conditions in Study 1 except that no other factors were manipulated in Conditions 2, 3, and 4 in Study 2.

Recall, for Condition 2 in Study 1, the correlation between two dimensions was fixed at one of two levels: 0.3 and 0.6, and different simple-structure 2-dimensional responses were generated based on these two correlations. In Study 2, responses were generated only based on the correlation of 0.6 since this value represents typical correlations among different dimensions for many large-scale operational tests which cover a wide range of content domains. The item parameters were exactly same as for Study 1 (see Table 3.2).

For Condition 3, the ratio of a_2 to a_1 for the first 5 items was set to two levels (0.5 and 1.0) in Study 1, reflecting mild and large dependence between the dominant and nuisance dimensions. For the same condition in Study 2, only mild dependence (a_2/a_1 = 0.5) case was considered since it may be more realistic in practical applications. Other model parameters were the same as for Study 1 (see Table 3.2).

For the testlet condition (Condition 4), in Study 1, the testlet effect variance $\sigma^2_{d(i)}$ was specified at three levels: 0.5, 1.0, and 2.0 to reflect mild, large, and extreme dependence among the testlet items, respectively. In this study, only mild dependence was used. The item parameters can be found in Table 3.2.

3.2.3 Estimate Different Data-Analysis Models in WinBUGS

In each condition, each of the 20 generated datasets was calibrated using the different data-analysis models in WinBUGS 1.4. Since the model comparison indices are calculated based on posterior estimation of model parameters, how well the different models involved in Study 2 can

be estimated in WinBUGS provides important evidence about the validity of the model comparison results. Therefore, a preliminary study was first conducted to determine how long the chain should run to achieve convergence, how many iterations are needed after convergence to estimate the parameters, and the extent of model parameter recovery in WinBUGS for different models. The calibration of the two-parameter (2P) GR model in WinBUGS was already validated in Study 1. Thus, the estimation of the other models in WinBUGS was evaluated for Study 2.

One-Parameter (1P) GR Model

In order to validate the estimation of one-parameter GR model, data were generated based on the same model which contained responses for 2000 simulated examinees to 15 polytomous items with 5 response categories. The slope parameters for all 15 items were fixed to 1.7 and the threshold parameters were set to the same values used for the two-parameter GR models.

A one-parameter GR model was fit to the generated data in WinBUGS. The code is similar to that for estimating GR model (Appendix B) except that one slope is estimated rather than multiple slope parameters. Two chains of length of 6000 were run which took approximately 7 hours to complete. The first 1000 iterations were treated as burn-in and discarded, and the remaining chains were thinned by taking every other iteration to obtain a combined posterior distribution based on a sample of 5000. All the sampling histories, brg diagrams, and autocorrelation plots suggested that each Markov chain converged to a stationary distribution very quickly. The values of MC errors indicated the sample size of 5000 was sufficient for precise posterior inference. Figure 3.6 illustrates the corresponding convergence

diagnostic graphs for the first threshold parameter of Item 1 and the common slope parameter for all the 15 items. Similar results were observed for the other item parameters.

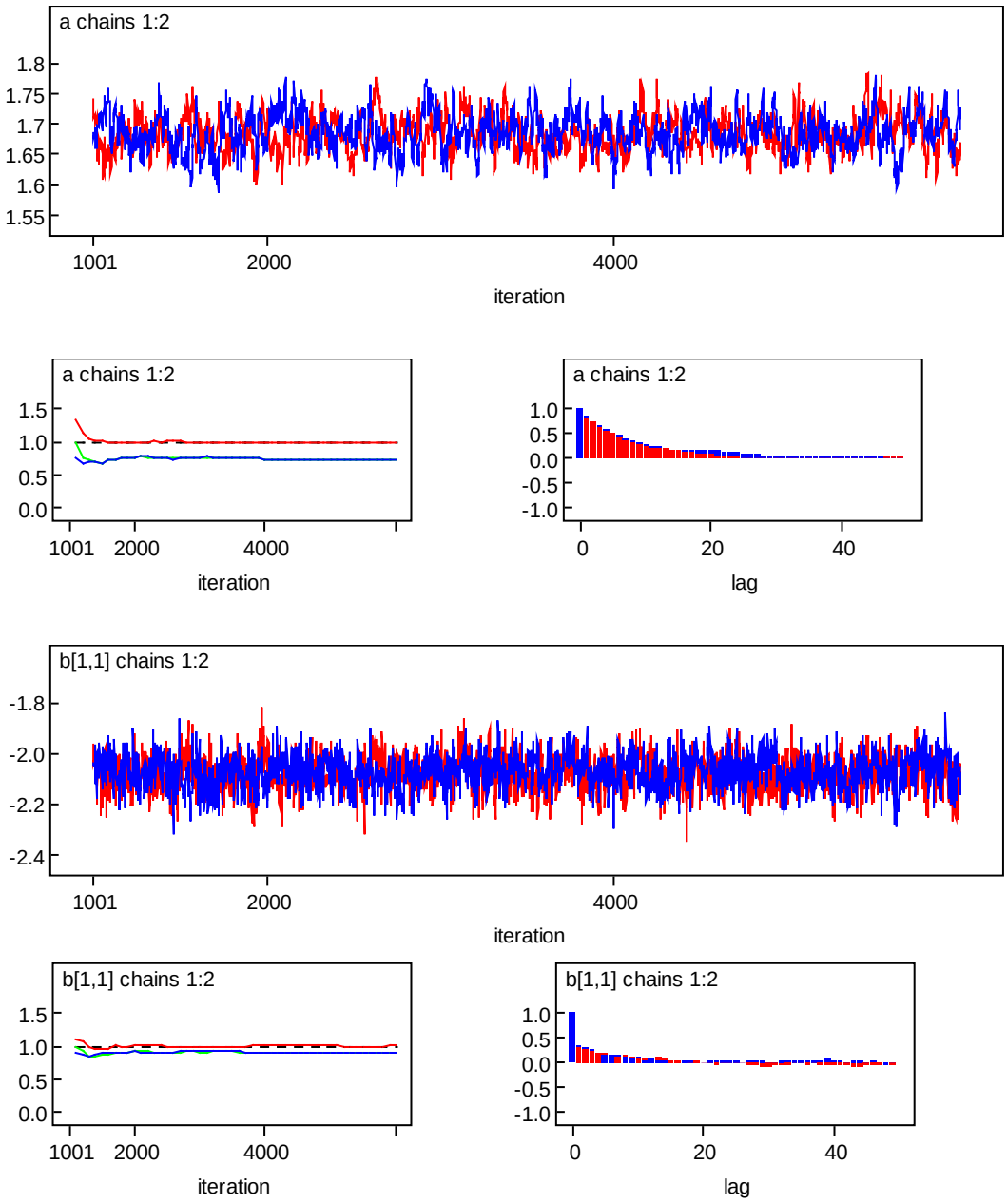


Figure 3.6 Example Convergence Diagnostic Plots for Item Parameters under 1P GR Model

Table 3.13 provides the comparison between the generating parameters and the estimates in WinBUGS for the one-parameter GR model. As can be seen, parameters were well recovered. The bias in the slope estimate was -0.01, and the average absolute bias for the thresholds

parameters across all items was 0.044. The results indicated that the one-parameter GR model could be estimated precisely in WinBUGS, and the WinBUGS code used for the estimation was valid.

Table 3.13 Item Parameter Recovery for 1P GR Model in WinBUGS

	True					Estimates				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1		-2.0	-1.0	0.0	1.0		-2.08	-1.02	0.03	1.04
2		-1.5	-0.5	0.5	1.5		-1.49	-0.48	0.52	1.59
3		-1.0	0.0	1.0	2.0		-0.97	-0.01	1.04	2.07
4		-3.0	-1.5	-0.5	1.0		-3.20	-1.49	-0.52	0.99
5		-1.0	0.5	1.5	3.0		-0.96	0.56	1.55	3.31
6		-2.0	-1.0	0.0	1.0		-1.99	-0.93	0.00	1.01
7		-1.5	-0.5	0.5	1.5		-1.53	-0.54	0.53	1.55
8	1.7	-1.0	0.0	1.0	2.0	1.69	-1.00	0.06	1.06	2.15
9		-3.0	-1.5	-0.5	1.0		-2.99	-1.54	-0.51	1.02
10		-1.0	0.5	1.5	3.0		-0.98	0.50	1.57	2.91
11		-2.0	-1.0	0.0	1.0		-1.98	-0.98	-0.02	1.02
12		-1.5	-0.5	0.5	1.5		-1.46	-0.44	0.53	1.61
13		-1.0	0.0	1.0	2.0		-1.02	0.03	1.01	1.96
14		-3.0	-1.5	-0.5	1.0		-2.97	-1.52	-0.55	0.97
15		-1.0	0.5	1.5	3.0		-1.01	0.49	1.50	2.84

RS Model

The operating characteristic curves for Muraki (1990)'s RS model can be expressed as:

$$P_{ix}^{*}(\theta) = \frac{\exp\left[Da_i\left(\theta - (b_i - c_x)\right]\right.}{1 + \exp\left[Da_i\left(\theta - (b_i - c_x)\right)\right]} \,. \tag{3.10}$$

As can be seen, the threshold parameters (b_{ix}) of the 2P GR model are partitioned into two terms in the RS model: a location parameter (b_i) for each item, and one set of category threshold parameters (c_x) that applies to all items. The RS model is a restricted version of the 2P GR model because the RS model assumes the category boundaries are equally distant from each other

across all items, whereas they are free to vary across items in the 2P GR model. As a result, the RS model requires fewer parameters to estimate than the 2P GR model. Thus, if a set of items has a common set of response options or is scored based on a general rubric, the RS model may provide an advantage over the 2P GR model.

In order to validate estimation of the RS model in WinBUGS, responses for 2000 simulated examinees to 15 polytomous items with 5 response categories were generated under the RS model. The configuration of item parameters for 15 items under the RS model involved a combination of 3 levels for the slope parameters (1.0, 1.7, and 2.4) and 5 levels for the location parameters: -1.0, -0.5, 0, 0.5, and 1.0. Note that the five levels of the location parameters were the same as the average difficulties of the 2P GR model (see Table 3.2). The category threshold parameters were set to 1.5, 0.5, -0.5, and -1.5.

The generated data were estimated in WinBUGS. As for the 2P GR model, the prior distributions of the slope parameters were defined as lognormal distributions with means 0 and variances 1. Normal priors were assigned to the location parameters with means equal to 1 and variances equal to 4. Prior distributions for the category threshold parameters were defined as the same normal distributions as for the location parameters with two constraints: they were ordered and the sum of them was 0. Finally, following standard conventions, ability parameters were assigned standard normal priors.

One chain of length of 10000 was run in WinBUGS. The first 1000 iterations were discarded (burn-in iterations) and the remaining chain was thinned by taking every other iteration to obtain a posterior sample of 4500. All the sampling histories and autocorrelation plots suggested that the Markov chain converged to a stationary distribution very quickly. That also indicated that a shorter chain may adequate for estimating this model. The values of MC errors

indicated the sample size of 4500 was sufficient for precise posterior inference. Figure 3.7 provides the history plots and autocorrelation plots for the slope and location parameters for Item 1. Similar results were observed for the other item parameters.

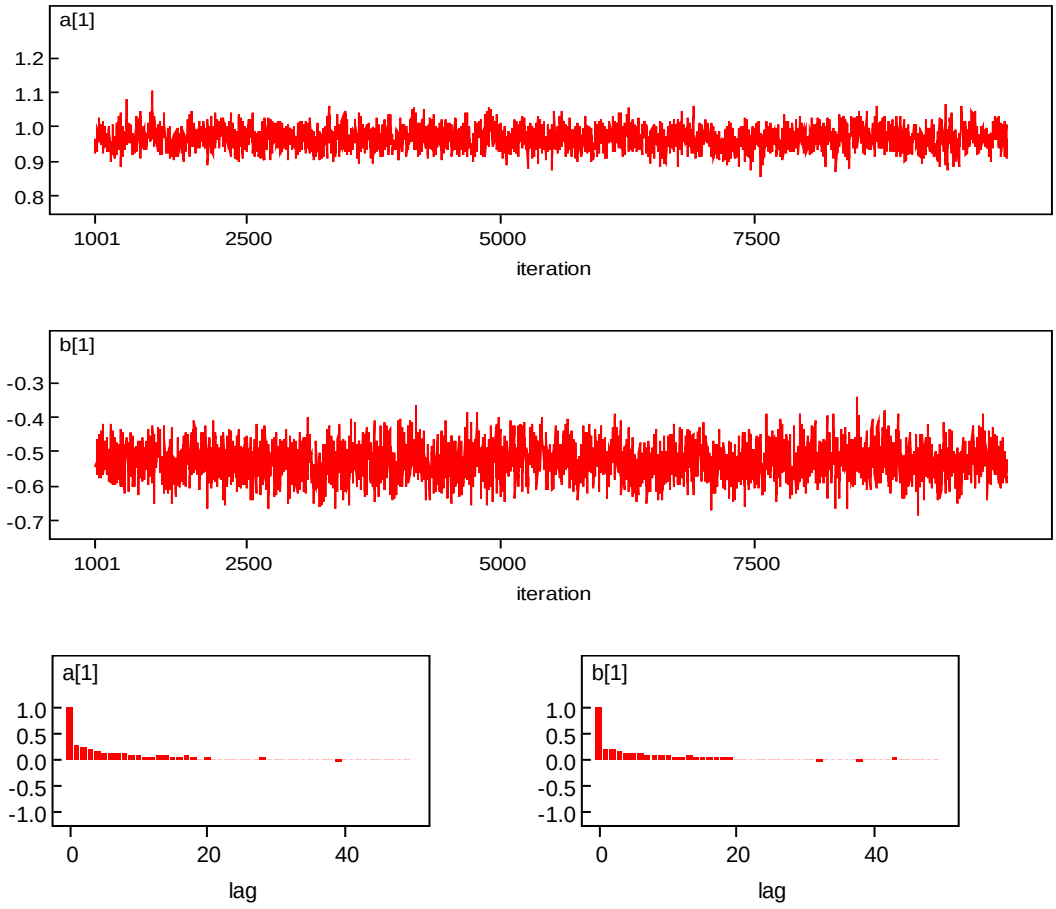


Figure 3.7 Example Convergence Diagnostic Plots for Item Parameters under RS Model

Table 3.14 provides the comparison between the generating parameters and the corresponding estimates in WinBUGS. As can be seen, the parameters were recovered well. The average absolute bias in the slope estimate across all items was 0.058, and it was 0.034 for the location parameters across all items and 0.028 for the category threshold parameters across all categories. Thus, the results indicated that the WinBUGS code for estimating RS model was valid.

Table 3.14 Item Parameter Recovery for RS Model in WinBUGS

	True			Estimates	
Item	a	b		a	b
1	1.0	-0.5		0.97	-0.53
2	1.0	0		0.97	-0.03
3	1.0	0.5		0.93	0.59
4	1.0	-1.0		0.97	-1.06
5	1.0	1.0		0.96	1.03
6	1.7	-0.5		1.60	-0.53
7	1.7	0		1.69	0.00
8	1.7	0.5		1.66	0.54
9	1.7	-1.0		1.65	-1.07
10	1.7	1.0		1.68	1.01
11	2.4	-0.5		2.29	-0.49
12	2.4	0		2.28	0.02
13	2.4	0.5		2.33	0.51
14	2.4	-1.0		2.44	-1.03
15	2.4	1.0		2.29	1.05
	C1	C2	C3	C4	
True	1.5	0.5	-0.5	-1.5	
Estimates	1.55	0.51	-0.51	-1.54	

2-dim Simple-Structure GR Model

In order to validate the estimation of two-dimensional (2-dim) simple-structure GR models in WinBUGS, 2-dim simple-structure item response were generated for 2000 simulated examinees to 15 polytomous items with 5 response categories. The configuration of item parameters was the same as in Study 1 (see Table 3.2). Ability parameters for the two dimensions were randomly selected from a bivariate normal (0, 1) with the specified correlation of 0.6.

The generated data were estimated under a 2-dimensional simple-structure GR model in WinBUGS using the code given in Appendix E. The prior distributions for the item parameters were the same as for the GR model. As for unidimensional IRT models, multidimensional models have scale or metric indeterminacy problem. To solve this problem, the abilities on two

dimensions were assigned multivariate normal priors, with means of 0 and variances of 1. However, the covariance (or correlation) between the two dimensions was not fixed. Based on previous research, this inter-dimensional correlation was assigned a normal prior with mean equal to the true correlation 0.6, and the variance of 0.25 for the current study. This approach to addressing the metric indeterminacy problem was similar to Yao and Boughton (2007)'s approach except that they fixed the correlation to the true value. They found that as long as the fixed correlations were smaller than the true correlations, the resulting estimated item parameters were close to their true item parameters. However, in real applications, true correlations are unknown, so Yao and Boughton (2007) suggested that the correlations can be approximated by the correlations between the subtotal scores for each dimension or the correlations between the estimated unidimensional scores for each dimension. In this study, the correlation was estimated rather than fixed.

Two chains of length of 10000 were run and took approximately 6 hours to complete. WinBUGS uses Metropolis MCMC algorithm for estimating this complicated model. By default, this sampling method utilizes the first 4000 iterations to determine suitable proposal distribution variances in order to obtain an acceptance rate of between 20% and 40%. As the result, the first 5000 iterations in each chain were discarded (burn-in iterations), and the remaining chains were thinned by taking every other iteration to get a combined posterior sample of 5000. All the sampling histories, bgr diagrams, and autocorrelation plots suggested the Markov chains converged to stationary posterior distributions. The values of MC errors indicated the sample size of 5000 was sufficient for precise posterior inference. Figure 3.8 includes the brg diagrams and autocorrelation graphs for the slope and first threshold parameters for Item 1 and for the correlation between two dimensions. Similar results were observed for the other item parameters.

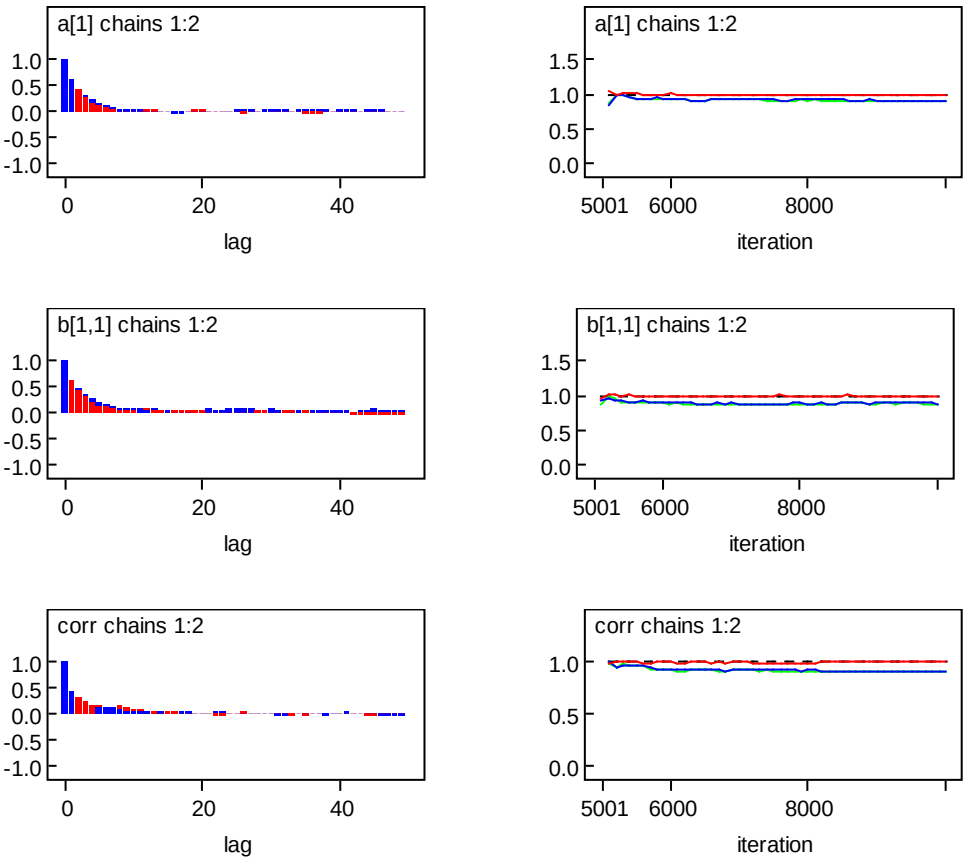


Figure 3.8 Convergence Diagnostic Plots for Parameters under 2-dim Simple-Structure GR Model

Point estimates of the model parameters and standard errors were computed based on the posterior sample of 5000 iterations. Table 3.15 presents the comparison between the generating parameters and the corresponding estimates in WinBUGS for the 2-dim simple-structure GR model. As can be seen, the parameters were recovered well. The average absolute bias in the slope estimate across all items was 0.074, and 0.058 for the threshold parameters. In addition, the correlation parameter between two dimensions was recovered well. The bias was 0.01. The results indicated that the WinBUGS code for estimating 2-dim simple-structure GR model was valid.

Table 3.15 Item Parameter Recovery for 2-dim Simple-Structure GR Model in WinBUGS

	True						Estimates					
Item	a1	a2	b1	b2	b3	b4	a1	a2	b1	b2	b3	b4
1	1.0	0	-2.0	-1.0	0.0	1.0	1.12	-	-1.74	-0.85	0.07	0.95
2	1.7	0	-1.5	-0.5	0.5	1.5	1.69	-	-1.54	-0.48	0.53	1.61
3	2.4	0	-1.0	0.0	1.0	2.0	2.28	-	-0.97	0.05	1.07	2.09
4	1.0	0	-3.0	-1.5	-0.5	1.0	0.93	-	-3.17	-1.53	-0.51	1.04
5	1.7	0	-1.0	0.5	1.5	3.0	1.85	-	-0.87	0.50	1.49	2.81
6	2.4	0	-2.0	-1.0	0.0	1.0	2.17	-	-1.96	-0.99	-0.01	1.00
7	1.0	0	-1.5	-0.5	0.5	1.5	1.03	-	-1.44	-0.44	0.61	1.53
8	1.7	0	-1.0	0.0	1.0	2.0	1.72	-	-1.00	0.03	1.00	1.98
9	0	2.4	-3.0	-1.5	-0.5	1.0	-	2.34	-3.06	-1.53	-0.48	1.09
10	0	1.0	-1.0	0.5	1.5	3.0	-	0.97	-1.05	0.48	1.54	2.91
11	0	1.7	-2.0	-1.0	0.0	1.0	-	1.54	-2.16	-1.07	0.01	1.09
12	0	2.4	-1.5	-0.5	0.5	1.5	-	2.38	-1.46	-0.45	0.51	1.42
13	0	1.0	-1.0	0.0	1.0	2.0	-	0.99	-0.95	-0.01	0.98	1.90
14	0	1.7	-3.0	-1.5	-0.5	1.0	-	1.66	-3.07	-1.56	-0.51	0.95
15	0	2.4	-1.0	0.5	1.5	3.0	-	2.44	-0.97	0.52	1.55	3.15
Correlation:			true = 0.60						estimate = 0.59			

2-dim Complex-Structure GR Model

Responses for 2000 simulated examinees to 15 polytomous items with 5 response categories were generated based on a 2-dimensional complex-structure GR model. The configuration of item parameters was the same as for the first case of Condition 3 in Study 1. Ability parameters for two dimensions were randomly selected from a bivariate normal (0, 1) with the specified correlation of 0.3.

The generated data were estimated using a 2-dimensional complex-structure GR model in WinBUGS. When estimating complex-structure MIRT models using MCMC, it is important to solve both metric indeterminacy and rotational indeterminacy problems. As for 2-dim simple-structure model, the metric indeterminacy problem was addressed by assigning the abilities on

two dimensions means of 0 and variances of 1. The rotational indeterminacy problem only exists for complex-structure models when one item measures more than one dimension. Analogous to factor analysis, the dimensions' orientation are not unique. They can be rotated in the dimension space without changing the model fit. To solve the rotational indeterminacy, the two ability axes were constrained to be orthogonal, and for the last 10 items which only measure the dominant dimension, the slope parameters were fixed at 0 on the nuisance dimension. Thus, the prior of abilities followed multivariate normal, with means of 0 and a variance-covariance matrix equal to the identity matrix. Note that the approaches to addressing the indeterminacy problems were similar to those used by Bolt and Lall (2003). The prior distributions for other item parameters were the same as for the GR model.

Response data in Condition 3 were generated based on two correlated dimensions (correlation = 0.3), however, estimation of the complex-structure GR model always imposed an orthogonal factor solution. Therefore, the slope estimates from WinBUGS could not be compared directly with the generating slope parameters. Instead, the corresponding generating slope parameters with respect to an orthogonal solution (correlation = 0) should be derived in order to compare the estimates from WinBUGS with true values. It should be noted that though the orthogonal solution affects the direct evaluation of item recovery, it does not affect the probability of responses to each response categories. Thus the solution for rotational indeterminacy should not affect the results for the model-fit and model-comparison.

In order to check the item recovery of 2-dim complex-structure GR model in WinBUGS, a new dataset was generated assuming two uncorrelated or orthogonal dimensions. This dataset was then estimated in WinBUGS using the same code. Two chains of length of 10000 were run and took about 10 hours to complete. The first 4000 iterations were used to determine suitable

proposal distribution variances for Metropolis sampling, and the next 1000 iterations were discarded for the burn-in phase. The remaining chains were thinned by taking every other iteration to obtain a combined posterior sample of 5000. All the sampling histories, bgr diagrams, and autocorrelation plots suggested the Markov chains converged to a stationary distribution. Figure 3.9 provides the convergence diagnostic graphs for the two slope parameters as well as the first threshold parameter for Item 1. Similar results were observed for the other item parameters. The moderate autocorrelations existing for the second slope and the threshold parameters indicate that the chain may require additional thinning.

Point estimates of the model parameters and standard errors were computed from the mean and standard deviations of posterior distributions for parameters. Table 3.16 compares the generating parameters and the estimates in WinBUGS for the 2-dim complex-structure GR model. The average absolute bias in the slope estimates across all items was 0.067 for dimension 1 and 0.032 for dimension 2. The average absolute bias in the threshold parameter estimates across all items was 0.049. The results indicate close recovery of the item parameters in WinBUGS and the code for estimating 2-dimensional complex-structure GR models was considered valid.

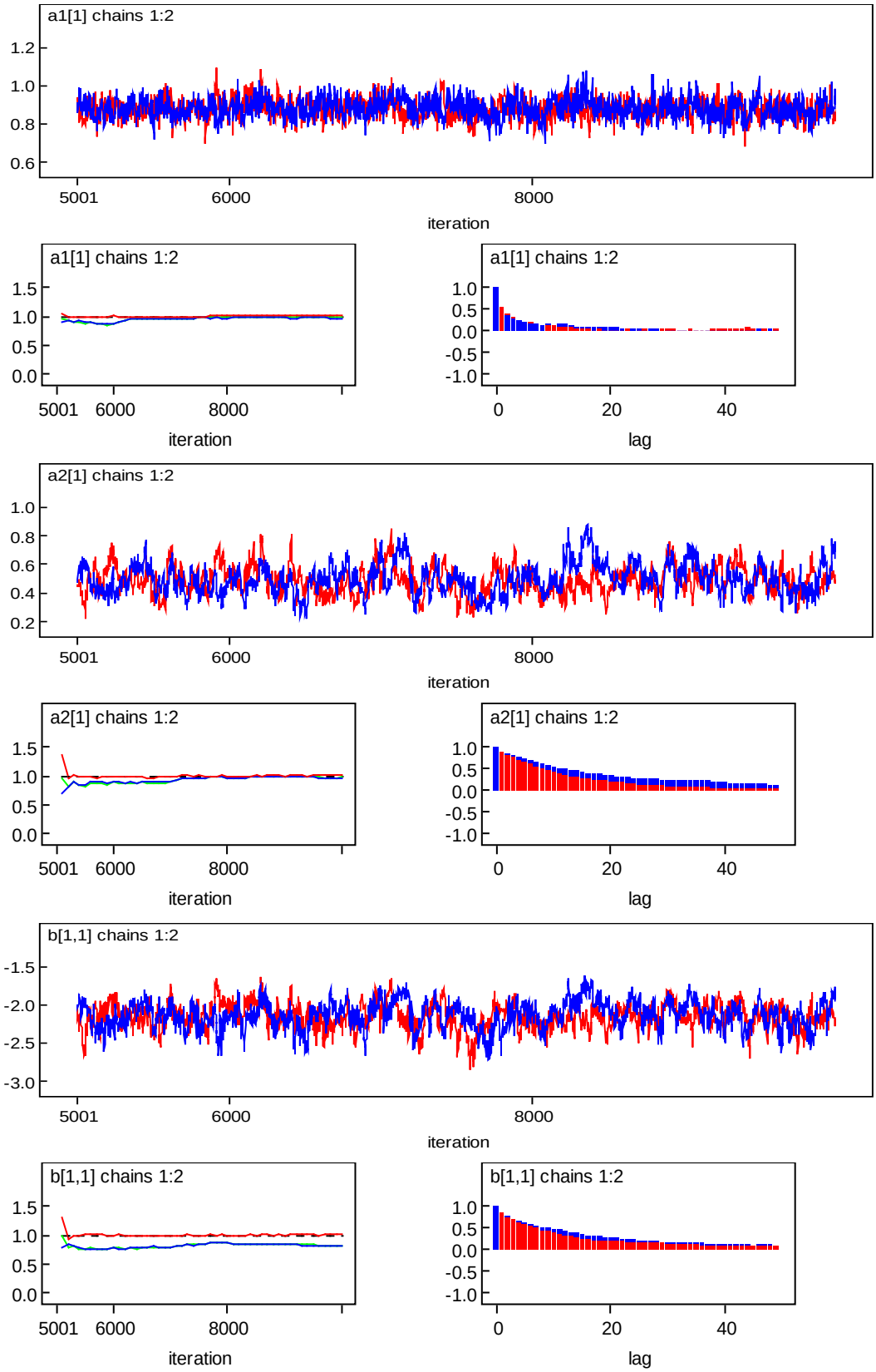


Figure 3.9 Convergence Diagnostic Plots for Parameters under 2-dim Complex-Structure GR Model

Table 3.16 Item Parameter Recovery for 2-dim Complex-Structure GR Model in WinBUGS

	True						Estimates					
Item	a1	a2	b1	b2	b3	b4	a1	a2	b1	b2	b3	b4
1	1.0	0.5	-2.0	-1.0	0.0	1.0	0.88	0.49	-2.15	-1.07	-0.01	1.07
2	1.0	0.5	-1.5	-0.5	0.5	1.5	1.06	0.49	-1.51	-0.46	0.52	1.44
3	1.0	0.5	-1.0	0.0	1.0	2.0	1.02	0.49	-0.98	0.02	0.94	2.00
4	1.0	0.5	-3.0	-1.5	-0.5	1.0	1.03	0.42	-3.13	-1.56	-0.55	1.00
5	1.0	0.5	-1.0	0.5	1.5	3.0	1.03	0.55	-0.97	0.45	1.42	2.94
6	1.7	0	-2.0	-1.0	0.0	1.0	1.67	-	-2.09	-1.06	-0.07	0.97
7	1.7	0	-1.5	-0.5	0.5	1.5	1.72	-	-1.51	-0.52	0.47	1.45
8	1.7	0	-1.0	0.0	1.0	2.0	1.74	-	-0.97	-0.01	0.96	1.98
9	1.7	0	-3.0	-1.5	-0.5	1.0	1.80	-	-2.77	-1.40	-0.45	0.98
10	1.7	0	-1.0	0.5	1.5	3.0	1.59	-	-1.00	0.48	1.54	3.20
11	2.4	0	-2.0	-1.0	0.0	1.0	2.45	-	-2.06	-1.02	-0.01	1.04
12	2.4	0	-1.5	-0.5	0.5	1.5	2.54	-	-1.48	-0.53	0.42	1.39
13	2.4	0	-1.0	0.0	1.0	2.0	2.52	-	-0.96	0.00	0.98	1.91
14	2.4	0	-3.0	-1.5	-0.5	1.0	2.31	-	-2.92	-1.54	-0.55	0.95
15	2.4	0	-1.0	0.5	1.5	3.0	2.36	-	-1.02	0.48	1.48	2.99

GR Model for Testlets

In order to validate the estimation of testlet GR models in WinBUGS, responses to a test with one testlet was generated. As in Study 1, this test included 15 5-category items and Items 6, 7 and 8 were specified as a testlet. The variance of testlet effect $\sigma^2_{d(i)}$ was fixed to 1.0, and the item parameters were the same as in Study 1 (see Table 3.2). A modified GR model for testlet was fit to this generated data in WinBUGS. The prior distributions for the item parameters and the examinees' abilities were the same as for GR models. The testlet effect was assigned a normal prior with mean of 0 and random variance of $\sigma^2_{d(i)}$. The hyper-parameter $\sigma^2_{d(i)}$ was given an inverse chi-square distribution with a degree of freedom 0.5 indicating a lack of information

about this parameter. This approach to specifying hyper prior of $\sigma^2_{d(i)}$ was the same as that used by Bradlow et al. (1999) and Li et al. (2006).

Two chains of length of 6000 were run and took approximately 6 hours to complete. The first 1000 iterations were discarded for the burn-in phase. Following the burn-in iterations, the remaining chains were thinned by taking every other iteration to get a combined posterior distribution sample of 5000. All the sampling histories, bgr diagrams, and autocorrelation plots suggested the Markov chains converged to a stationary distribution. Figure 3.10 displays the convergence diagnostic graphs for the slope and first threshold parameters for Item 6 (the first item in the testlet) and for the variance of the testlet effect. Similar results were observed for the other item parameters.

Point estimates of the model parameters and standard errors were computed from the mean and standard deviations for posterior distributions for parameters. Table 3.17 the generating model parameters and their corresponding estimates in WinBUGS. The average absolute bias was 0.055 for the slope estimate and 0.048 for the threshold parameters. In addition, the bias in the estimation of the testlet effect variance was -0.06. Thus, close recovery of parameters was indicated and the WinBUGS code used to estimate the testlet GR model was considered valid.

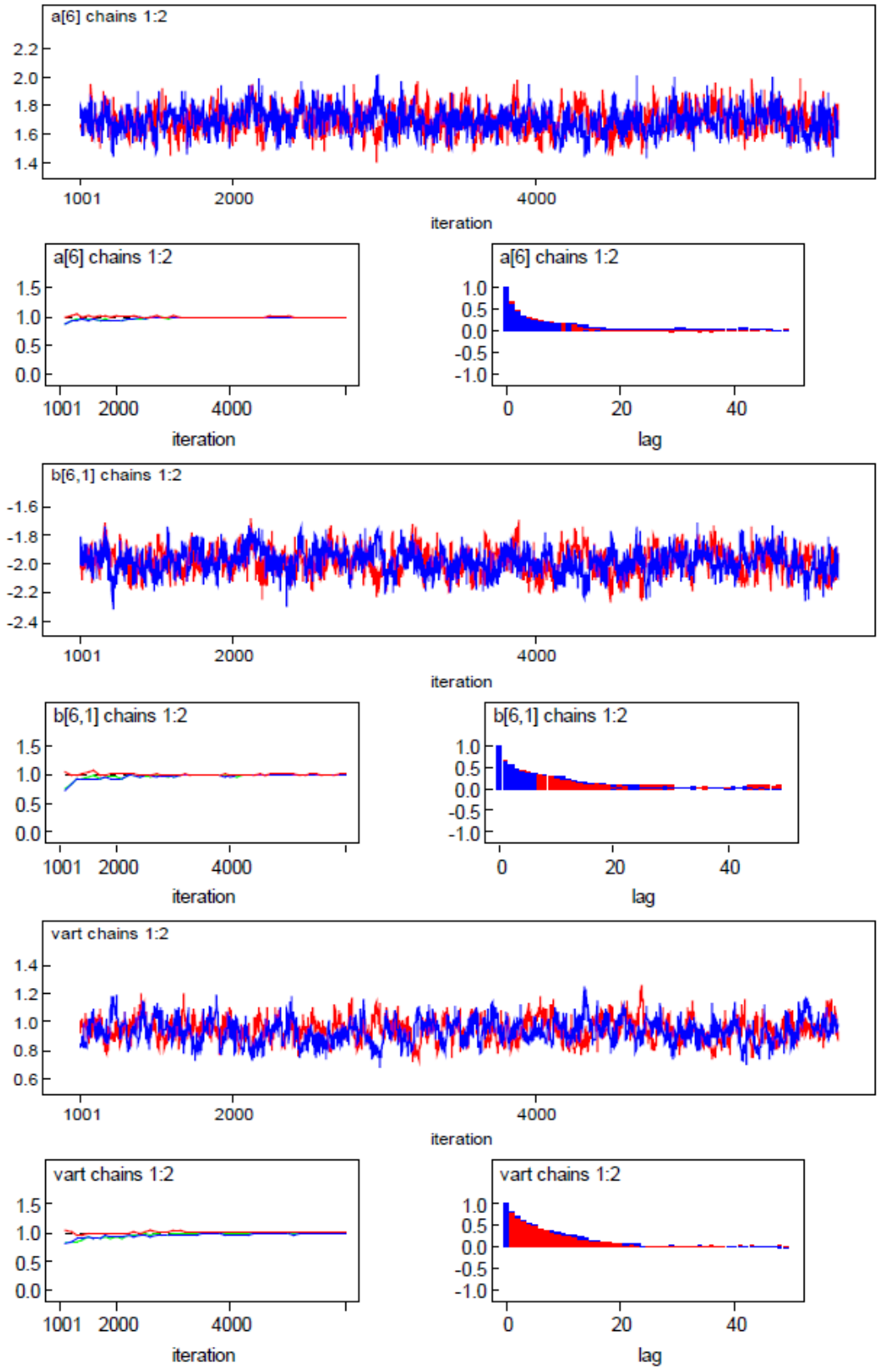


Figure 3.10 Convergence Diagnostic Plots for Parameters under Testlet GR Model

Table 3.17 Item Parameter Recovery for Testlet GR Model in WinBUGS

	True					Estimates				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	1.0	-2.0	-1.0	0.0	1.0	0.95	-2.01	-1.01	-0.00	1.11
2	1.0	-1.5	-0.5	0.5	1.5	1.00	-1.44	-0.47	0.50	1.55
3	1.0	-1.0	0.0	1.0	2.0	0.96	-0.93	0.07	1.09	2.16
4	1.0	-3.0	-1.5	-0.5	1.0	1.14	-2.70	-1.36	-0.49	0.96
5	1.0	-1.0	0.5	1.5	3.0	0.98	-1.02	0.52	1.48	3.11
6	1.7	-2.0	-1.0	0.0	1.0	1.70	-1.98	-0.94	0.04	1.10
7	1.7	-1.5	-0.5	0.5	1.5	1.67	-1.48	-0.46	0.52	1.51
8	1.7	-1.0	0.0	1.0	2.0	1.79	-0.91	0.02	1.07	2.04
9	1.7	-3.0	-1.5	-0.5	1.0	1.77	-3.08	-1.50	-0.48	1.02
10	1.7	-1.0	0.5	1.5	3.0	1.69	-1.08	0.45	1.55	3.06
11	2.4	-2.0	-1.0	0.0	1.0	2.44	-1.98	-1.0	-0.02	1.04
12	2.4	-1.5	-0.5	0.5	1.5	2.54	-1.39	-0.48	0.48	1.51
13	2.4	-1.0	0.0	1.0	2.0	2.35	-0.98	0.02	1.03	2.13
14	2.4	-3.0	-1.5	-0.5	1.0	2.27	-3.08	-1.51	-0.50	1.06
15	2.4	-1.0	0.5	1.5	3.0	2.41	-0.96	0.49	1.49	3.02
Variance of testlet effect:				true = 1.0				estimate = 0.94		

3.2.4 Conduct Model Comparison

For each of 20 generated data in each condition, different models were estimated in WinBUGS, and three Bayesian model comparison indices (DIC, CPO, and PPMC) were obtained for each model during the estimation of the models. These values for the different models were then compared in order to determine which model was preferred.

The estimates of the DIC index for different models were requested within WinBUGS. In the batch file (see Appendix D), a line "`dic.set()`" was used to set the DIC index, and another line "`dic.stats()`" was used to request the value of DIC. The smaller the value of DIC, the better the model. It should be noted that the DIC index can only be used to choose a preferred

model for the overall test. Based on DIC, we can not know which model is preferred for a specific item.

The computation of the CPO index was implemented by first computing the CPO at the level of an individual item response. A command line " inprob[i, j] <- pow(p[i, j, y[i,j]], -1)" was added to the WinBUGS code (see Appendix C) to compute the inverse likelihood of the observed item response based on the posterior model parameter values at a specific iteration. The mean value of this node " inprob[i, j]" across the posterior sample is given in the statistics output for WinBUGS and represents the estimate of CPO value for the response of student *i* to Item *j*. After the CPO_{ij} estimates were known, the CPO value for each item was computed in SAS by reading in the CPO_{ij} estimates and taking the log of the product of the CPO_{ij} across all examinees (see Equation 2.36). In addition, a *CPO* index for the overall test was summarized by taking the log of the product of the item-level CPO_j across all the items. In the current study, two levels of *CPO* index were used: the test-level *CPO* was used to compare the models for the overall test, and the item-level CPO_j were used to choose a preferred model for each item. The larger the value of the test-level *CPO*, the better the model fit for the overall test. The larger the value of the item-level CPO_j, the better the model fit for a specific item *j*.

The different models in each condition were also compared using PPMC. The details about conducting PPMC were introduced in Section 3.1.5. Recall, 8 different levels of discrepancy measure were used with PPMC in Study 1. Different from Study 1, however, the discrepancy measures employed in Study 2 only included the effective measures identified from Study 1. From the results presented in Chapter 4 for Study 1, two discrepancy measures "Yen's Q_3" and "global OR" were found to be most effective among all the 8 measures for detecting the violations of unidimensionality and local independence. Therefore, for Conditions 2-4 in Study

2, only these two measures were used with PPMC for model comparison purpose. However, for Condition 1 in which the GR, 1-parameter GR, and RS models were compared, all 8 discrepancy measures were employed with PPMC since the use of discrepancy measures with these models was not investigated and therefore unknown.

In order to compare different models using PPMC, the frequency of extreme PPP-values was computed for each model. For item-level discrepancy measures, there were 15 PPP-values for 15 items for each replication. How many items from the 15 items had extreme PPP-values (< 0.05 or > 0.95) was treated as the criterion for comparing different models. For pair-wise measures, there were 105 PPP-values for the 105 item pairs for each replication. How many item pairs out of these 105 pairs had extreme PPP-values (< 0.05 or > 0.95) was treated as the criterion to compare different models. When the true model was estimated, it was expected that no or few extreme PPP-values would be observed. In contrast, when the alternative model was estimated, more extreme PPP-values would be expected. In addition to PPP-values, graphical plots based on different models were also compared.

The relative performance of these three indices was compared with respect to the number of times each index selected the correct model across 20 replications. An effective index should be able to identify the generating model as the preferred model a large proportion of the time.

The preliminary study conducted in Section 3.2.3 used two chains of different length to estimate different models. One exception was the RS model for which one long chain (10000 iterations) was run. The results indicated that each chain converged very quickly and item parameters were well recovered. Due to the intensive computation in WinBUGS, only one chain was run to estimate the different models and compute the model comparison indices. The length

of the chain for each model depended on the model as well as the results from the preliminary study.

Condition 1**:** GR vs. one-par GR vs. RS models

For each of these three models, one chain of 5000 iterations was run, and the first 4000 was discarded as the burn-in phase and the remaining 1000 iterations were thinned by taking every other iteration to obtain a posterior sample of size 500. The computation of three model comparison indices was based on these 500 iterations.

Condition 2: unidimensional GR model vs. 2-dim simple-structure GR model

For 2-dim simple-structure GR model, one chain of 8000 iterations was run, and the first 5000 was discarded as the burn-in phase and the remaining 3000 iterations were thinned by taking every third iteration to get a posterior sample of size 1000. For the unidimensional GR model in this condition, one chain of 5000 iterations was run, and the first 3000 was discarded as the burn-in phase, and the remaining 2000 iterations were thinned by taking every other iteration to get a posterior sample of size 1000. The computation of three model comparison indices was based on these 1000 iterations.

Condition 3: unidimensional GR model vs. 2-dim complex-structure GR model

The length of the chain, thinning, and the size of posterior sample for the 2-dim complex-structure GR model was the same as for the 2-dim simple-structure GR model in Condition 2. Note that more thinning was conducted than the previous preliminary study in order to further reduce the autocorrelation among parameters. For the unidimensional GR model, one chain of 5000 iterations was run, and the first 3000 was discarded as the burn-in phase, and the remaining 2000 iterations were thinned by taking every other iteration to get a posterior sample of size 1000.

Condition 4: unidimensional GR model vs. testlet GR model

For both models, one chain of 5000 iterations was run, and the first 3000 was discarded as the burn-in phase and the remaining 2000 iterations were thinned by taking every other iteration to obtain a posterior sample of size 1000.

3.3 REAL DATA APPLICATION

This section examines the use of the proposed Bayesian approaches to model-checking and model-comparison for a real mathematics performance assessment - the QUASAR Cognitive Assessment Instrument (QCAI). QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) was a national project that sought to demonstrate that it is feasible to implement instructional programs in the middle-school grades that promote the acquisition of thinking and reasoning skills in mathematics (Silver, 1991). The QCAI was a performance assessment developed for the QUASAR project in order to evaluate the impact of innovative instructional programs on middle school students' mathematical thinking and reasoning in four sub-domains: reasoning, problem solving, communication, and understanding of the features that characterize mathematical concepts and their interrelations (Lane, 1993). The QCAI includes four test forms (A, B, C, and D), each containing 9 different open-ended tasks scored at 5 levels (0-4). These four forms were randomly distributed within each sixth- and seventh-grade class in the schools participating in the QUASAR project (Lane, Stone, Ankenmann & Liu, 1995). This test was administrated in both the fall and the spring during 1990, 1991, and 1992.

Several researchers have examined the extent to which the QCAI response data met the assumptions and properties underlying the GR model. Lane et al. (1995) conducted a

comprehensive study to evaluate the dimensionality, speededness and item parameter invariance for each of four QCAI forms across three administration occasions (Spring 1991, Fall 1991, and Spring 1992). They examined the dimensionality through the use of the confirmatory factor analysis and eigenvalue plots. Factor analysis results indicated that each of the four forms of the QCAI were essentially unidimensional. However, it was found that tasks with lower factor loadings in a one-factor model solution reflected tasks requiring some type of explanations, and the tasks with relatively high loadings generally involved problems requiring students to only display their mathematics solution strategies. Lane et al. (1995) further explored the use of two-factor models. A two-factor model was estimated in which one factor included all tasks except those requiring a nonprocedural explanation and a second factor included only the tasks requiring a nonprocedural explanation. In addition, a two-factor model was estimated in which one factor included only the tasks requiring the display of solution strategies and an explanation and a second factor included tasks requiring only the solution strategies. From the results, there was no substantial statistical evidence to support the two-factor models, thus providing additional evidence supporting one dominant dimension underlying the item responses to the QCAI.

Speededness was investigated for tasks by statistically comparing hierarchical GR models using two groups of students with different administration time lengths. For two of the eight tasks examined, only the slope parameter estimates differed, and for another two tasks, both the slope and threshold parameter estimates differed. The stability of QCAI item parameter estimates over time was investigated using restricted IRT models within a multiple-group analysis in MULTILOG. The results indicated that the parameter estimates were stable for the first year, but not stable for the second year.

It is interesting to note that in their study, in order to select a more appropriate GR model for scaling the QCAI data, they compared two hierarchical models, a two-parameter (2P) GR and a one-parameter (1P) GR that restricted the slope parameters to be equal across items. These models were compared using the log-likelihood statistics for the two models. A significant difference between the statistics indicated that the 2P GR model fit the data better than the 1P model.

Goodness of fit with respect to the QCAI items was investigated by Stone, Ankenmann, Lane, & Liu (1993) and later reexamined by Stone (2000). Due to imprecise point ability estimates caused by the small number of tasks on each QCAI form, the researchers utilized Stone's item-fit statistic G^{2*} to assess the fit of each QCAI task to the GR model. The difference between these two studies involved different Monte Carlo resampling approaches for hypothesis testing of the fit statistic.

Stone et al. (1993) used a Monte Carlo resampling method which required estimation of the GR model for each simulated dataset, thus accounting for uncertainty in both item and ability parameters in generating the simulated null distribution of the G^{2*} statistic. Fit was evaluated for each of the items on four forms (A-D) across four administration occasions (Fall 1990, Spring 1991, Fall 1991, and Spring 1992) by comparing the G^{2*} statistic with simulated null distributions. A few flawed items were excluded from the analyses for earlier administration. The total number of tasks on the four forms was 30 for the first two administrations, and 33 for the last two administrations (three flawed tasks were revised and included). The results indicated that 12 tasks fit the data across all four administrations, only 1 task did not fit the data across the four administrations, 2 tasks did not fit the data across three of the four administrations, 7 tasks

did not fit the data across two of the four administrations, and 9 tasks did not fit the data for one of the four administrations.

The resampling method used by Stone et al. (1993) was computationally intensive due to the requirement that item parameters be estimated for each Monte Carlo sample. To reduce the computational complexity, Stone (2000) proposed an alternative resampling method that used the item parameter estimates based on the real data for all Monte Carlo samples. Thus, the step involving re-estimation of the GR model for each sample was eliminated. Stone (2000) also proposed a procedure for estimating a scaling factor that could be used to rescale the fit statistic to approximate the null distribution for hypothesis testing. For this method, only uncertainty in ability estimation was considered in generating the sampling null distribution of the G^{2*} statistic. Uncertainty in item parameter estimation was considered by adjusting the derived df by the number of estimated item parameters. In order to compare this alternative resampling method with the previous method, the fit of 62 QCAI items from two of the four administrations used in Stone et al. (1993) were reanalyzed using this alternative resampling and the results were compared with those from the previous study. Although general agreement in terms of the fit of these QCAI items from the two studies was high, there was some disagreement between two studies. The disagreement existed primarily for items found to be significantly "misfitting" in Stone et al. (1993) but not significantly "misfitting" using the alternative resampling method.

In the current study, the PPMC method was used to re-examine the fit of the QCAI to the two-parameter GR model in terms of unidimensionality, local independence, and item-fit. All 8 discrepancy measures used in Simulation Study 1 were used with PPMC for this real application, and the results were compared with those from the previous studies. In addition, the 1P GR and 2P GR models were re-compared using the proposed Bayesian model-comparison tools to see if

the 2P GR model fit the QCAI data better as found in Lane et al. (1995). Moreover, a 2-dimensional complex-structure GR model was estimated in order to see if a complex multidimensional model was preferred over the simple unidimensional GR model. In this multidimensional model, the first dimension included all items, and the second dimension included only the items requiring an explanation. It should be noted that only Yen's Q_3 statistic and the global OR measure were used with PPMC for the 2-dimensional complex-structure model since these two measures were found to be the most effective measures based on the simulation studies.

For this real data application, three QCAI forms with 8 items each were reanalyzed: Form A administrated in Spring 1991 (AS91), Form A given in Spring 1992 (AS92), and Form B given in Spring 1992 (BS92). The sample sizes were 399, 459, and 446 for the AS91, AS92, and BS92 forms, respectively.

Table 3.18 compares the decisions regarding item fit for the items on these three forms from Stone et al. (1993) and Stone (2000). All decisions regarding item fit were made at the α = 0.05 level of significance. The misfitting items were indicated by asterisks. As seen in this table, in Stone et al. (1993), there were 4 misfitting items for the AS91 test form, 2 misfitting items for the AS92 form, and 5 misfitting items for the BS92 form. However, two of these items were not identified as misfitting by Stone (2000). The fit of these items was re-examined using the PPMC method, and the results were compared with the results in this table.

Table 3.18 Misfitting Items Identified in Stone et al. (1993) and Stone (2000)

AS91			AS92			BS92		
Item	Stone et al, 1993	Stone, 2000	Item	Stone et al, 1993	Stone, 2000	Item	Stone et al, 1993	Stone, 2000
1	*	*	1			1	*	*
2			2	*		2	*	*
3	*		3	*	*	3	*	*
4			4			4		
5	*	*	5			5		
6			6			6	*	*
7			7			7	*	*
8	*	*	8			8		

When a 2-dimensional complex-structure GR model was used to analyze the AS91 or AS92 datasets, four explanation items (Items 1, 5, 7, and 8) loaded on the two dimensions, and all other items only loaded on the first dimension. For the BS92 dataset, three explanation items (Items 1, 5, and 8) loaded on both dimensions, and all other items only loaded on the first dimension.

With regard to the implementation of MCMC and PPMC in WinBUGS, a chain of 15000 iterations was run to estimate, test and compare the fit of the two-parameter GR model, one-parameter GR model, and the 2-dimensional complex-structure GR model. The first 10000 iterations were discarded for the burn-in phase and the remaining 5000 iterations were thinned by selecting every 5^{th} iteration to obtain posterior distributions based on 1000 iterations. The implementation of PPMC and the computation of model-comparison indices were based on this posterior sample.

4.0 RESULTS

This chapter presents the results from two simulation studies and one real application study in three separate sections. Simulation Study 1 aimed to explore the performance of the PPMC method in detecting aspects of lack of fit for unidimensional GR models using the proposed discrepancy measures. The Type-I error rates or empirical power rates for these discrepancy measures with PPMC are presented in the first section. The second section includes the results from Simulation Study 2 in which the relative effectiveness of three Bayesian model-comparison methods (DIC, CPO, and PPMC) were compared. The third section presents results evaluating the fit of the unidimensional GR model item responses from the QCAI performance assessment using the PPMC method and model-comparison indices.

4.1 RESULTS FROM SIMULATION STUDY 1

In order to investigate the performance of the PPMC method in evaluating different assumptions underlying the unidimensional GR model, five conditions were considered in Study 1 (see Table 3.1). Condition 1 represents the null condition in which both the generating model (Mg) and analysis model (Ma) were the unidimensional GR model, and thus Type-I error rates for PPMC were investigated. In Conditions 2 to 5, different types of misfit were simulated based on different GR models, and empirical power rates for PPMC in detecting different misfit were

examined. The results from Study 1 are organized in the order of the conditions. For each condition, both PPP-values and graphical plots for each of the 8 discrepancy measures used with PPMC are summarized and reported.

4.1.1 Item Parameter Recovery

Table 4.1 RMSD for Item Parameter Recovery in WinBUGS for GR Model

Item	a	b1	b2	b3	b4
1	0.05	0.09	0.06	0.05	0.08
2	0.06	0.08	0.05	0.06	0.12
3	0.06	0.08	0.06	0.08	0.14
4	0.04	0.14	0.10	0.07	0.05
5	0.05	0.07	0.06	0.11	0.21
6	0.08	0.10	0.05	0.04	0.06
7	0.07	0.06	0.04	0.05	0.10
8	0.09	0.05	0.04	0.05	0.12
9	0.06	0.11	0.06	0.05	0.05
10	0.09	0.05	0.05	0.07	0.19
11	0.09	0.08	0.04	0.03	0.05
12	0.12	0.07	0.04	0.03	0.06
13	0.10	0.04	0.03	0.03	0.08
14	0.09	0.13	0.06	0.04	0.04
15	0.10	0.04	0.04	0.07	0.14
	$\overline{RMSD}(a) = 0.08$		$\overline{RMSD}(b) = 0.07$		

PPMC is based on the posterior estimation of model parameters, and the quality of model parameter recovery using MCMC estimation is an important factor in determining whether PPMC could be implemented successfully. As a result, parameter recovery was examined first. Table 4.1 represents the RMSD for each item parameter across the 20 replications. The average RMSD across all items was 0.08 for slope parameter and 0.07 for threshold parameters. These

results indicate one chain of 4000 and a posterior sample of 500 was adequate for the accuracy of estimation of the unidimensional GR model using MCMC within WinBUGS.

4.1.2 Condition 1 (Ma = Mg = unidimensional GR)

When the estimated model was the same as the generating model, the proportion of 20 replications with extreme PPP-values (< 0.05 or > 0.95) for each discrepancy measure provides evidence with regard to Type-I error rates, or how often misfit was wrongly detected by the PPMC method. If the PPP-values were used in the same way as classical p-values, the nominal Type I error rate would be 0.10. Table 4.2 reports the overall median PPP-value and average Type-I error rate for each measure. The values were pooled across all possible items (for item-level measures) or item pairs (for pair-wise measures) and also across the 20 replications. The underlying rationale is that the generated data in Condition 1 were unidimensional GR data and all items (item-pairs) had the same dimensional structure. Thus, they were exchangeable in terms of dimensionality.

Specifically, for the "test-level" measure, the median PPP-value was the median of PPP-values across the 20 replications, and the Type-I error rate was reflected by the proportion of 20 replications with extreme PPP-values. For each item-level discrepancy measure, each of the 15 items had a median PPP-value and a Type-I error rate was computed across the 20 replications. The overall median PPP-value was the median value of these median PPP-values of the15 items, and the overall Type-I error rate was the average of the Type-I error rates over 15 items. Following the same logic, for each pair-wise measure, each of 105 item pairs had a median PPP-value and a Type-I error rate across the 20 replications. The overall median PPP-value was the

median of the median PPP-values for 105 item pairs, and the overall Type-I error rate was the average of the Type-I error rates over these 105 item pairs.

Table 4.2 Median PPP-values and Average Proportions of Replications with Extreme PPP-values (< 0.05 or >0.95) when Ma=Mg=unidimensional GR

	Discrepancy Measure	Median PPP	Average Proportion
Test-Level	Test Score Distribution	0.51	0.10
Item-Level	Item Score Distribution	0.50	0.00
	Item-Total Score Correlation	0.48	0.00
	Yen's Q1	0.52	0.00
	Stone's Fit Statistic	0.50	0.01
Pair-wise	Global OR	0.49	0.05
	Yen's Q3	0.49	0.06
	Item Covariance Residual	0.52	0.01

As shown in Table 4.2, the median PPP-values for all the measures were around 0.50 which was expected under the null condition. Thus, the realized (observed) values of the discrepancy measure were not consistently larger or smaller than the posterior predictive values (i.e., no systematic difference between realized and predictive values), indicating no departure of model-data fit. All the proportions (except the test-level measure) were below 0.10, suggesting that the use of PPP-values in hypothesis testing would lead to highly conservative tests (i.e., they tend not to show misfit of a correct model too often). The two pair-wise measures (global OR and Yen's Q_3) appeared to have empirical type-I error rates most close to the nominal rate, though still quite lower.

The conservativeness of the discrepancy measures investigated in this study was further explored by examining the distribution of PPP-values. As reviewed in Chapter 2, the departure of the distribution of PPP-values from a uniform distribution under the null condition would result in a conservative test when PPP-values are used in a hypothesis testing framework. The closer to

uniform the distribution, the closer to the nominal level the Type-I error rate would be (Levy, 2006).

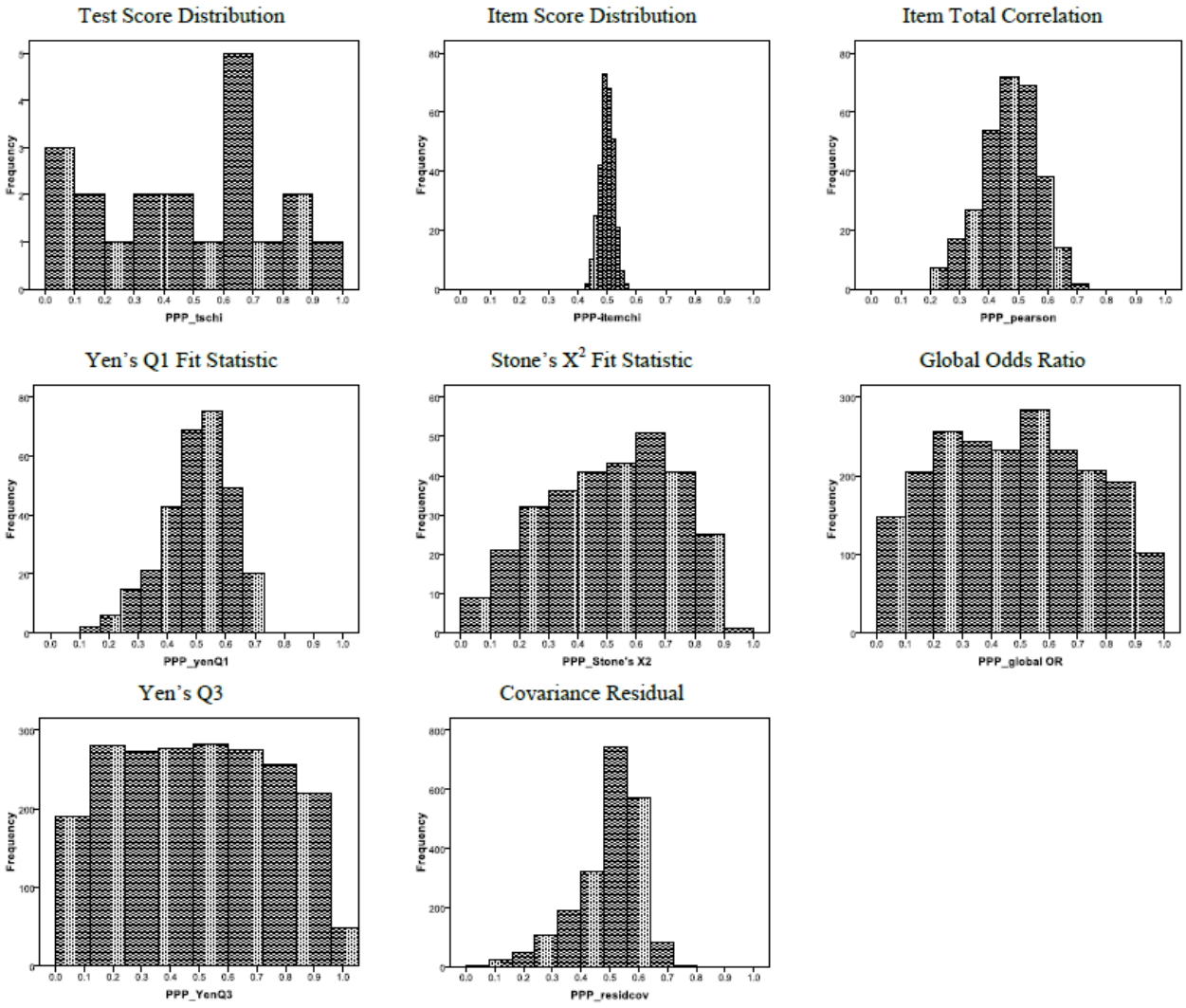


Figure 4.1 Distributions of PPP-values for Each Discrepancy Measures under the Null Condition

Figure 4.1 presents the distributions of the PPP-values for each discrepancy measure across all possible items or item pairs and across 20 replications as well. The distributions were drawn without distinguishing different items or item pairs because of the exchangeability assumption. As observed in this figure, all distributions of PPP-values were centered at around 0.5. However, the shape and the variability of the distributions differed for the different measures. The distribution of the chi-square used to measure the difference between observed

and expected *item score distributions* was least variable around 0.5. The *item-total score correlation, Yen's Q_1 item-fit statistic,* and *item covariance residual* exhibited slightly more variation. *Stone's item-fit statistic* showed more variability. Two pair-wise measures (*global OR, Yen's Q_3*) exhibited most variability. It is interesting to note the chi-square measure χ_T^2 used to measure the difference between observed and model-predicted *test score distributions* also exhibited more variability. Although it is not entirely clear why the test-level and item-level measures differed in variability, this test-level measure was observed to be more variable than the item-level measures.

From Figure 4.1, the distributions of PPP-values for the two pair-wise measures – *global OR* and *Yen's Q_3* and the *test score distribution* were more close to uniform distributions as compared to the other discrepancy measures. As a result, they exhibit empirical Type I error rates closer to the nominal rate of 0.10 than others as shown in Table 4.2. These findings are consistent with previous research (Levy, 2006; Meng, 1994; Robins et al., 2000; Rubin, 1996). They showed that different from classical p-values, PPP-values are not uniformly distributed under null conditions, even asymptotically. Though the distribution may be centered at 0.5 it is less dispersed than a uniform distribution. Thus, the PPP-values under the correct model tend to be closer to 0.5 more often than would be expected under a uniform distribution. However, Levy (2006) also showed that some effective measures approximated uniform distributions and approximated nominal level Type-I errors.

As discussed previously, graphical plots are also often used to provide diagnostic evidence about misfit. In general, when the discrepancy measure only depends on the data, the position of the observed value in the distribution of posterior predictive values is examined. When the measure depends on both the data and model parameters, pairs of realized vs.

predictive discrepancies are plotted in a scatter plot. For the null condition of this study, several different plots are shown for three levels of discrepancy measures. These plots served as reference or baseline plots for the misfitting conditions (Conditions 2-5).

Test-Level Measure

Figure 4.2 shows two diagnostic plots based on the "*observed test score distribution*" for one dataset (replication) generated from the unidimensional GR model and estimated using the same model. The first plot is the observed score distribution versus the 90% posterior predictive (PP) distributions (between 5% and 95%) of total test scores. As seen from this plot, the observed score distribution is within the PP interval. The second plot shows the realized and predictive values of the measure χ_T^2, summarizing the discrepancy between the observed and predictive test score distributions. In this plot, the x-axis represents the predictive χ_T^2 values and the y-axis represents the realized χ_T^2 values. As can be observed, the realized χ_T^2 values were not consistently larger or smaller than the predictive values. Both plots provide graphical evidence about model-fit. In addition, the corresponding PPP-value was 0.63 for this data set, also indicating a good fit between model and data.

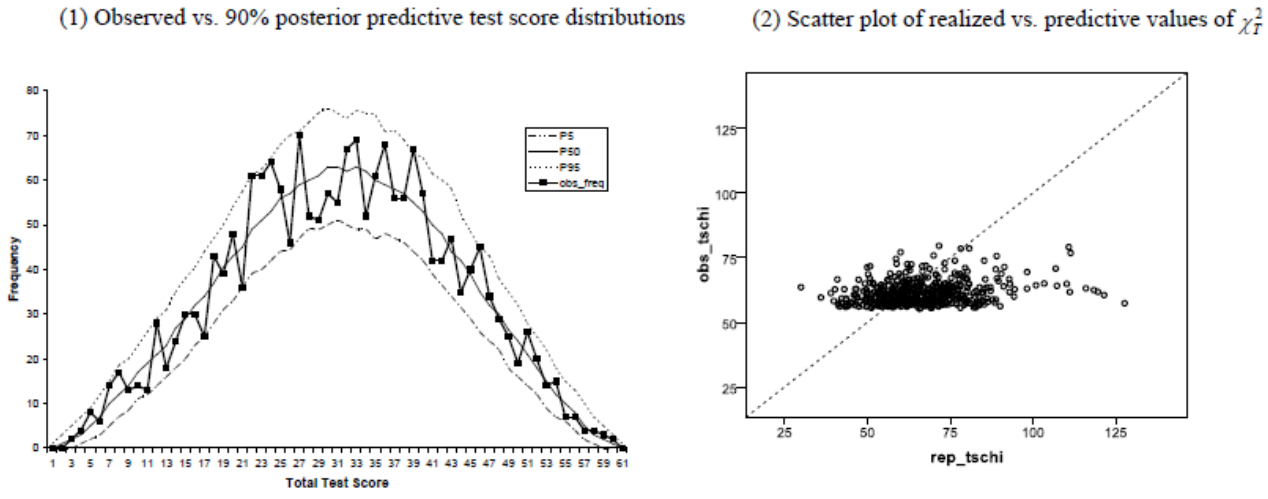


Figure 4.2 Diagnostic Plots based on Test Score Distribution when Ma=Mg=unidimensional GR

Item-Level Measures

Table 4.3 presents the median PPP-values and Type-I error rate of the item-level discrepancy measures for each item. As can been from this table, the median PPP-values were around 0.50, and the Type-I error rates were 0.00, indicating high conservativeness for the four item-level measures. It should be noted for each of the two item-fit measures (Yen's and Stone's), both Pearson's chi-square and likelihood ratio statistics were examined. Since both statistics had very similar results, only the results for Pearson's chi-square statistic are reported.

Table 4.3 Median PPP-values and Proportions of Replications with Extreme PPP-values for Item-Level Meausres when Ma=Mg=unidimensional GR

Item	Item Score Distribution		Item-Total Correlation		Yen's Q_1		Stone' Item-Fit	
	Median PPP	Type-I	Median PPP	Type-I	Median PPP	Type-I	Median PPP	Type-I
1	0.51	0.00	0.46	0.00	0.52	0.00	0.49	0.00
2	0.51	0.00	0.45	0.00	0.54	0.00	0.54	0.00
3	0.50	0.00	0.48	0.00	0.52	0.00	0.56	0.05
4	0.50	0.00	0.48	0.00	0.51	0.00	0.49	0.00
5	0.50	0.00	0.49	0.00	0.47	0.00	0.47	0.00
6	0.48	0.00	0.49	0.00	0.48	0.00	0.57	0.00
7	0.50	0.00	0.49	0.00	0.50	0.00	0.50	0.00
8	0.50	0.00	0.45	0.00	0.48	0.00	0.48	0.00
9	0.50	0.00	0.48	0.00	0.56	0.00	0.55	0.00
10	0.50	0.00	0.50	0.00	0.52	0.00	0.50	0.00
11	0.50	0.00	0.46	0.00	0.51	0.00	0.59	0.00
12	0.51	0.00	0.51	0.00	0.53	0.00	0.52	0.05
13	0.50	0.00	0.46	0.00	0.54	0.00	0.58	0.00
14	0.50	0.00	0.52	0.00	0.53	0.00	0.46	0.00
15	0.50	0.00	0.50	0.00	0.50	0.00	0.43	0.05

Figure 4.3 illustrates the observed item-total score correlations, corresponding 90% posterior predictive intervals and the median posterior correlations for each of 15 items based on one replication. A clear pattern in this plot is that the items fell into three groups in terms of the value of item-total correlation. This was expected since the first five items had the same true slope value of 1, Items 6-10 had the same true slope of 1.7, and the last five items had slope of 2.4. Item-total score correlations reflect the item discriminations and are related to the slope parameters. The observed correlation (solid dot) for each item approximated the median

posterior correlation, indicative of a good fit of the unidimensional GR model to the data for this discrepancy measure.

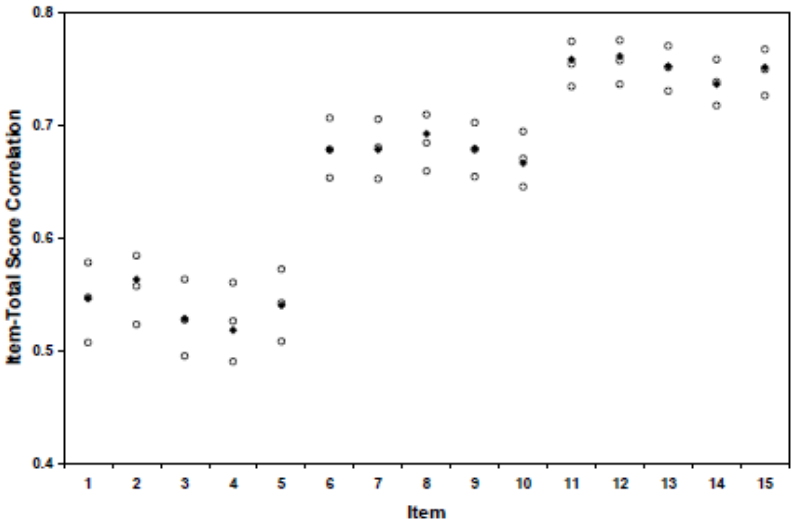


Figure 4.3 Oberved vs. 90% Posterior Preditive Interval of Item-Total Correlation for Each Item when Ma=Mg=undimensional GR

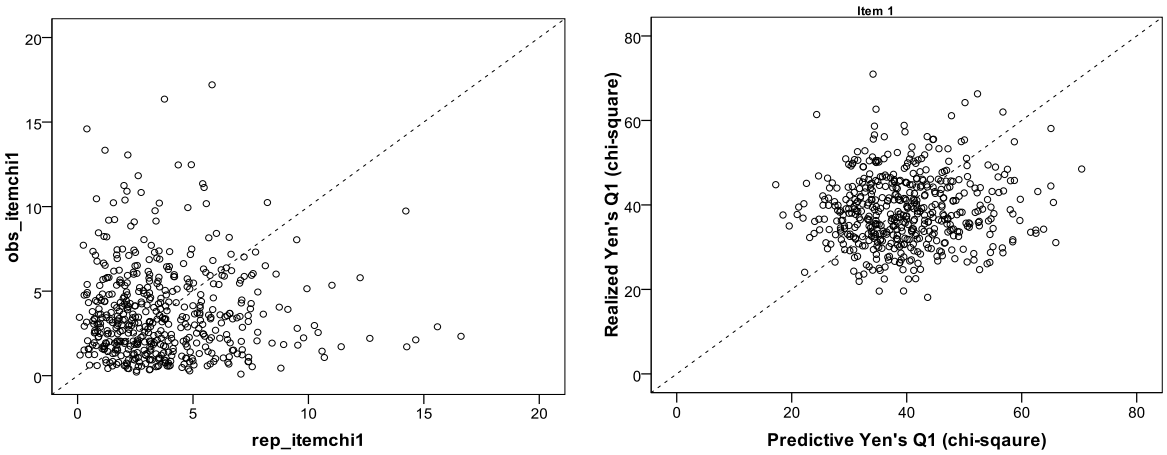


Figure 4.4 Realized vs. Posterior Predictive Values of Item-Level Chi-Square Measure and Yen's Q_1 for Item 1 when Ma=Mg=unidimensional GR

Unlike the item-total score correlation measure which is dependent only on the data, the other three item-level measures depend on both the data and model parameters. Figure 4.4 shows the scatter plots of realized vs. posterior predictive values for the "*item-level chi-square*

measure" (measuring the discrepancies between observed and predictive item score distributions) and "*Yen's Q_1 item-fit statistic*". The PPP-values for these two measures were 0.51 and 0.54, respectively. As can be seen, there was no systematic difference between the realized and posterior predictive values. The scatter plot for "*Stone's item-fit measure*" was similar to these two plots and is not provided here. It should be noted that the plots discussed above were drawn from one dataset (i.e., one replication). Similar plots were observed for the other 19 datasets.

Pair-Wise Measures

For each pair-wise measure, there are 105 PPP-values for each replication. In order to summarize the results across the 20 replications more efficiently, pie plots similar to those used by Sinharay and his colleague (2006) were employed. Figure 4.5 displays the median PPP-values (Left) and Type-I error rates (Right) for each item pair across the 20 replications for the three pair-wise measures. In the left plot, there is one pie for each item pair, and the proportion of a circle that is filled is equal to the magnitude of corresponding median PPP-value. The right plot provides information related to how the discrepancy measure detected misfit for each item pair. The filled proportion of a pie represents the proportion of 20 replications with extreme PPP-values (i.e., Type-I error rate) for that item pair. There is a clear pattern in this figure: under the null condition, the median PPP-values were all around 0.5 (left plot), and the proportion of extreme PPP-values were small (right plot). In addition, a larger number of pie plots for the "*item covariance residual*" measure were not filled, indicating that this measure was more conservative than the other two measures. The same phenomenon was found previously when comparing the overall Type-I error rates for these three pair-wise measures in Table 4.2.

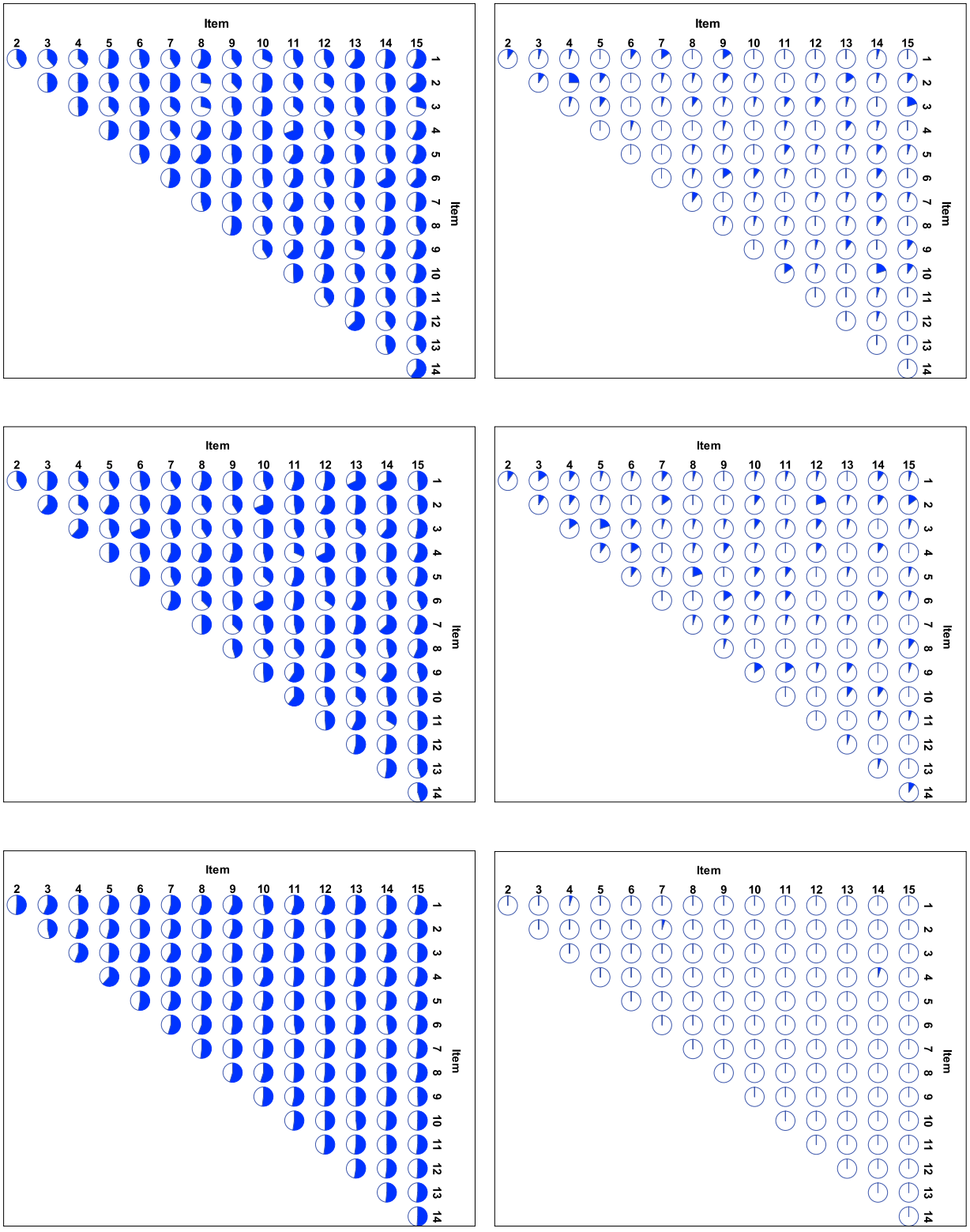


Figure 4.5 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) when Ma=Mg= unidimensional GR

It is also useful to examine the pattern for a single dataset rather than a summary across 20 datasets. Figure 4.6 shows the PPP-values of *Yen's* Q_3 and *Item Covariance Residual* for each item pair based on one of the 20 replications. The *global OR* displayed a similar pattern as *Yen's* Q_3 and is not shown here. As observed in these two plots, most of the PPP-values were not extreme, providing evidence that the GR model fit the data. It is interesting to note that the PPP-values of *Yen's* Q_3 were more variable than those of *Item Covariance Residual*. This was expected based on the difference between their PPP-values distributions. As observed in Figure 4.1, the distributions of *global OR* and *Yen's* Q_3 measures were more variable and closer to uniform distributions than the *Item Covariance Residual*. Similar plots were found for the other 19 datasets.

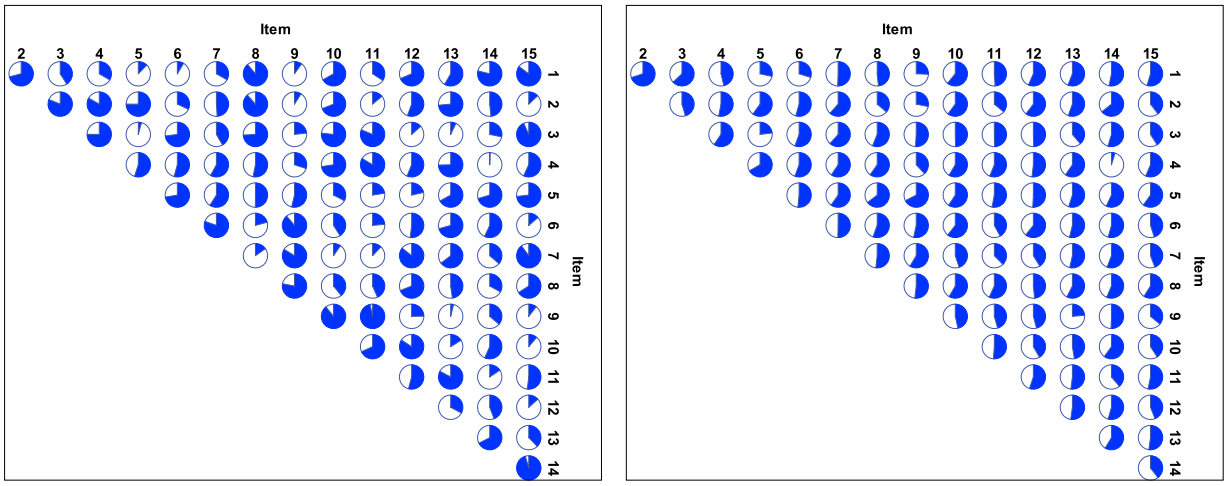


Figure 4.6 Display of PPP-values (based on a single dataset) for Yen's Q_3 (Left), and Item Covariance Residual (Right) when Ma=Mg= unidimensional GR

Figure 4.7 plots the observed global ORs involving the first item, 90% PP interval, and PP medians for one replication under the null condition. No observed global ORs (solid triangle) fall outside the PP interval, suggesting the model fits the data. Similar findings were obtained for other replications and other items. Figure 4.8 provides the scatter plots of the realized vs. posterior predictive values for *Yen's* Q_3 and *Item Covariance Residual* measures for one item

pair based on a single data. As can be seen, there were no systematic differences between the realized and posterior predictive values. Similar plots were obtained for the other 19 datasets and for other item pairs and are not presented here.

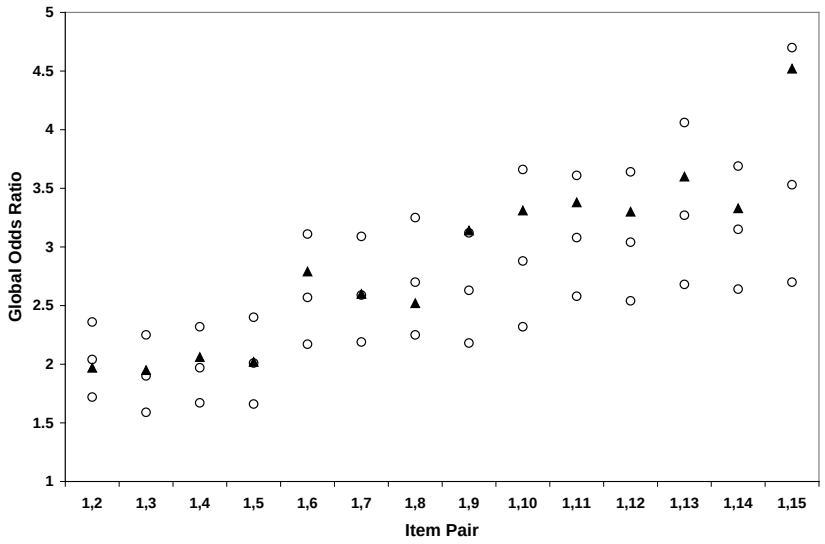


Figure 4.7 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) when Ma=Mg= unidimensional GR

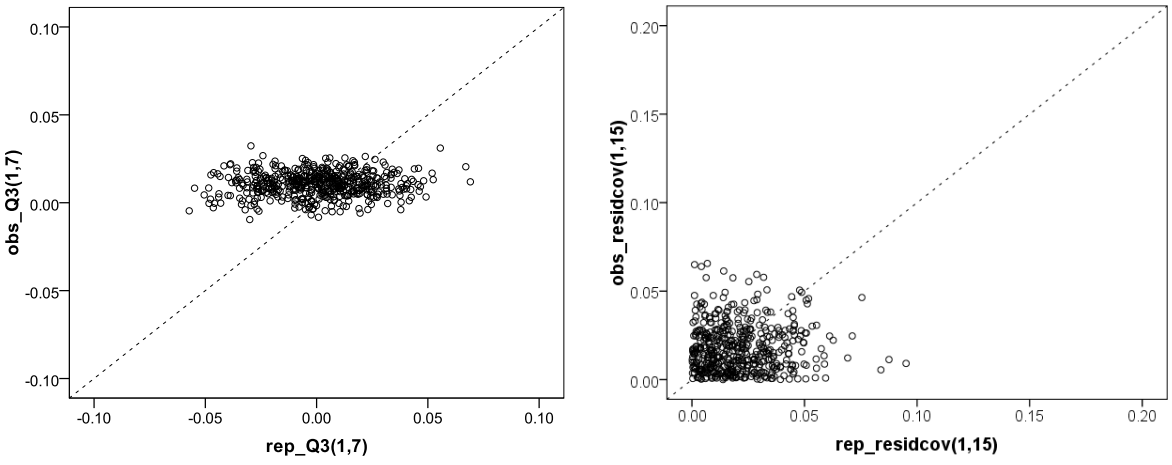


Figure 4.8 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 and Item Covariance Residual (for a single data) when Ma=Mg= unidimensional GR

4.1.3 Condition 2 (Mg = 2-dim simple-structure GR , Ma = 1-dim GR)

In this condition, the generated data reflected two dimensions (the first 8 items in Dim1 and the last 7 items in Dim2), but the estimated model was a unidimensional model. The ability of the PPMC method in detecting the violation of unidimensionality was explored by using all 8 proposed measures. Two cases were considered in this condition, one with low inter-dimensional correlation (ρ=0.3), and another with a more typical moderate inter-dimensional correlation (ρ=0.6).

Table 4.4 Overall Median PPP-values and Average Proportions of Replications with Extreme PPP-values for all Measures – Condition 2

	Measure	**Type**	**Case 1** (ρ=0.3)		**Case 2** (ρ=0.6)	
			Median PPP	Power	Median PPP	Power
Test-Level	Test score dist	-	0.06	0.25	0.41	0.10
Item-Level	Item score dist	Dim1	0.50	0.00	0.50	0.00
		Dim2	0.50	0.00	0.50	0.00
	Item-test corr	Dim1	1.00	0.91	0.28	0.09
		Dim2	0.00	0.99	0.66	0.03
	Yen's Q_1	Dim1	0.50	0.00	0.52	0.00
		Dim2	0.50	0.00	0.53	0.00
	Stone's fit stat	Dim1	0.50	0.00	0.51	0.01
		Dim2	0.50	0.01	0.53	0.00
Pair-Wise	Global OR	(Dim1, Dim1)	0.07	0.45	0.01	0.74
		(Dim1, Dim2)	0.99	0.79	0.98	0.68
		(Dim2, Dim2)	0.00	0.87	0.00	0.80
	Yen's Q_3	(Dim1, Dim1)	0.01	0.74	0.00	0.96
		(Dim1, Dim2)	1.00	0.98	1.00	0.97
		(Dim2, Dim2)	0.00	0.95	0.00	0.97
	Item cov resid	(Dim1, Dim1)	0.08	0.44	0.00	0.87
		(Dim1, Dim2)	0.00	0.93	0.01	0.83
		(Dim2, Dim2)	0.00	0.86	0.00	0.92

Table 4.4 presents the pooled median PPP-values and the average proportion of extreme PPP-values across the 20 replications (i.e., empirical power) for each discrepancy measure and for the two correlation cases. Under the assumption that the items in the same dimension were interchangeable, there were two types of items – items in Dim1 and items in Dim2 for each item-level measure. Therefore, the median PPP-values and the proportions were pooled across items

in each dimension. For the pair-wise measures, there were three types of item pairs: item pairs from the first dimension (Dim1, Dim1), item pairs from the second dimension (Dim2, Dim2), and item pairs from different dimensions (Dim1, Dim2). The PPP-values were pooled from the same type of item pairs across the 20 replications.

As observed from this table, the three pair-wise measures were sufficiently powerful in detecting the misfit of the unidimensional GR model to the two-dimensional data for both cases. Median PPP-values were extreme and the empirical power rates were high. Yen's Q_3 index performed best in terms of empirical power, and the item covariance residual measure performed better than the global OR. It is worthy to note that the global OR and Yen's Q_3 measures are both directional measures, and their PPP-values reflect the relationship between realized and posterior predictive discrepancies. For example, for item pairs from the same dimension, the median PPP-values for these two measures were close to 0. This indicated that the observed association between these item pairs was systematically higher than predicted under the unidimensional GR model. Thus the unidimensional model underestimated item relationships. For two items from the different dimensions, the median PPP-values were close to 1, indicating that the observed association was consistently lower than expected under the GR model, and the model overestimated their relationship. The absolute item covariance residual does not have this feature.

As the inter-dimensional correlation increased from 0.3 to 0.6, these three pair-wise measures were consistently powerful in detecting the misfit. The results in Table 4.5 also illustrate that the test-level and item-level measures did not appear as useful as the pair-wise measures in detecting multidimensionality among the data where ρ=0.6. The median PPP-values were not extreme and the proportions of extreme PPP-values (i.e., empirical power) were very

small. However, for ρ=0.3, the test-level measure and the item-total score correlation measure exhibited increased power. Specifically, when the correlation decreased from 0.6 to 0.3, the median PPP-value for the test-level chi-square measure decreased from 0.41 to 0.06, and the corresponding power rate increased from 0.10 to 0.25. For the item-total correlation, the overall median PPP value became extreme, increasing from 0.28 to 1.00 for the items in Dim1, and decreasing from 0.66 to 0.00 for the items in Dim2. The average power rate increased from 0.09 to 0.91 for Dim1 items, and from 0.03 to 0.99 for Dim2 items. The median PPP value of 1.00 indicated the observed item-total correlations were consistently lower than the predictive values for the items in Dim1, suggesting the 1-dim GR model over-estimated this measure. On the other hand, the median PPP value of 0.00 indicated the observed item-total correlations were consistently higher than the predictive values for the items in Dim2, suggesting the 1-dim GR model under-estimated this measure. Since the performance of the item-total score correlation changed dramatically when the inter-correlation decreased from 0.6 to 0.3, further study is needed in order to explore the impact of higher correlations among dimensions.

As for Condition 1, graphical plots were provided to show the graphical evidence for the misfit of the 1-dim GR model to the 2-dim data. It should be noted that only the plots related to the effective measures are presented since the plots for the ineffective measures were similar to the corresponding plots under the null condition (Condition 1).

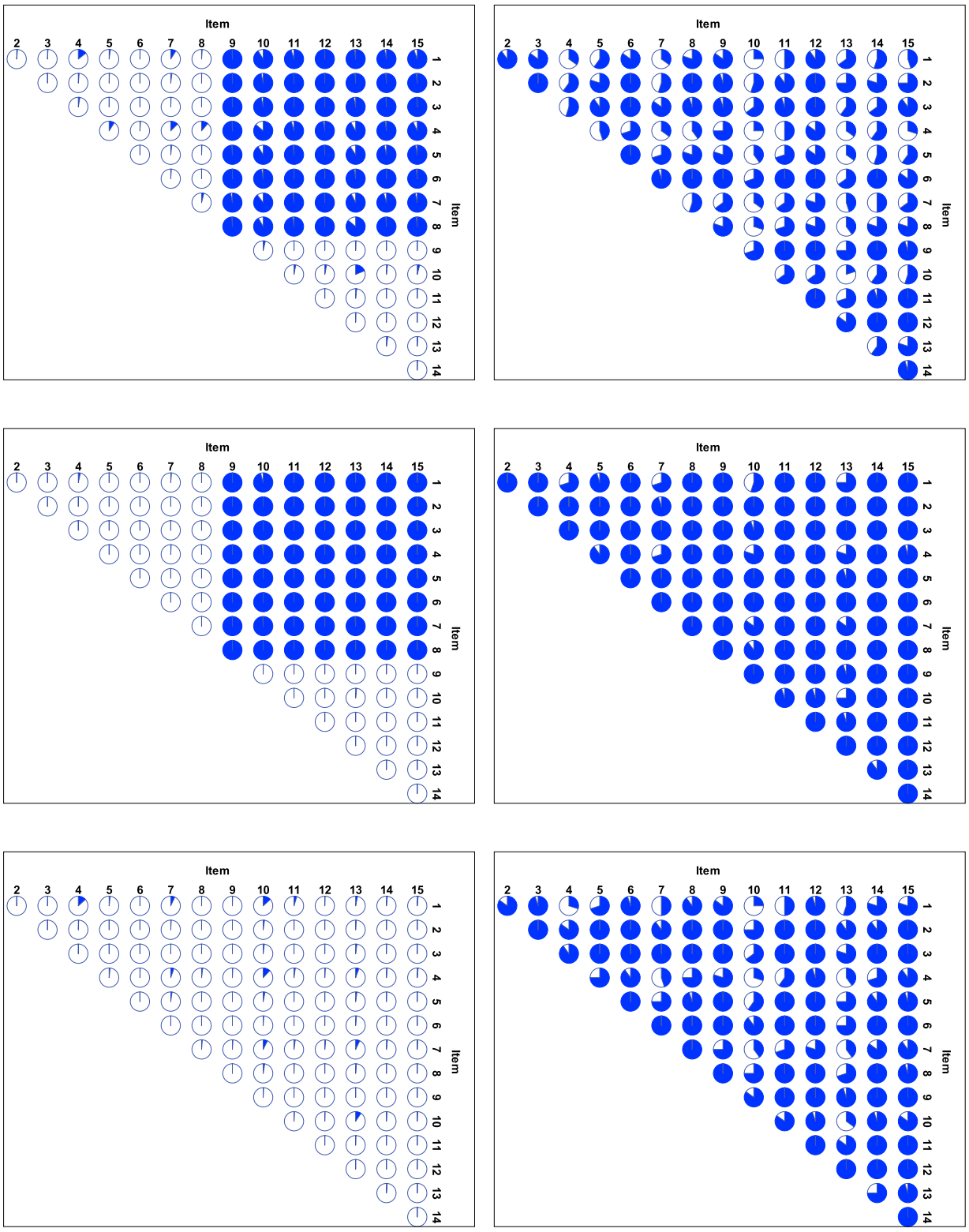


Figure 4.9 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 2 (ρ=0.6)

Figure 4.9 displays the median PPP-values (Left) and empirical power (Right) of the three pair-wise measures for each item pair across the 20 replications for Case 2. The large number of the extreme PPP-values in this figure clearly indicates that the unidimensional GR model did not fit the data. Moreover, the pattern in the plots for the two directional measures (global ORs and Yen's Q_3) differed clearly from the pattern under the null condition: all the 15 items fell into two clusters - Items 1-8 formed one cluster, and Items 9-15 formed another cluster. This pattern matched the factor structure of the generated data. The pie plots for Case 1 were similar to the plots for Case 2 and are not shown here.

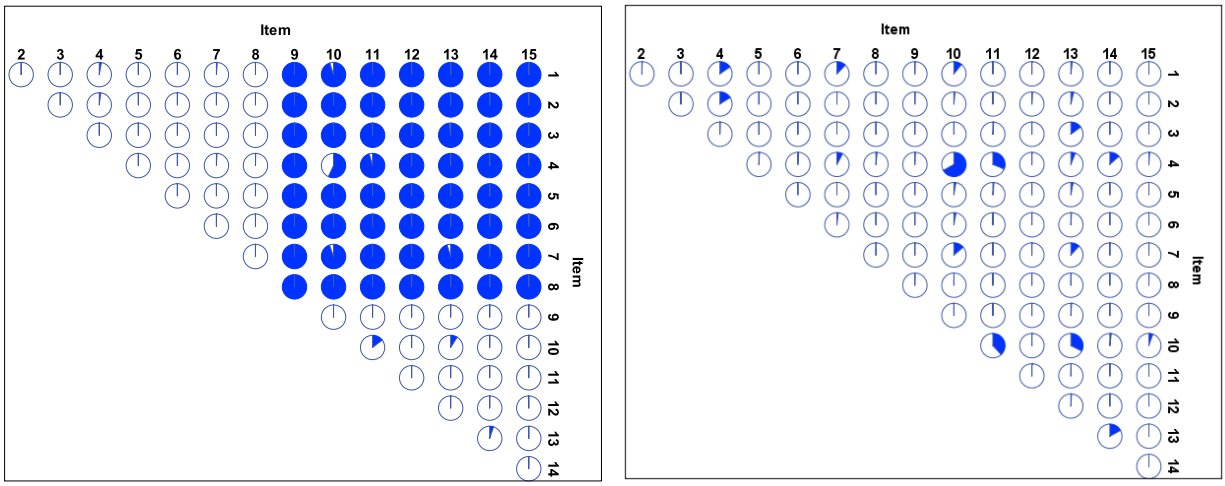


Figure 4.10 Display of PPP-values (based on a single dataset) for Yen's Q_3 (Left), and Item Covariance Residual (Right) - Condition 2 (ρ=0.6)

Figure 4.10 shows the PPP-values for *Yen's Q_3* and *Item Covariance Residual* for each item pair based on one replication when the correlation was 0.6. Results for the *global OR* displayed a similar pattern as *Yen's Q_3* and thus are not shown here. As observed in these two plots, the pattern for a single dataset was similar to the pattern based on the 20 replications (see Figure 4.9): most of the PPP-values were extreme, providing evidence of misfit of the unidimensional GR model to the data.

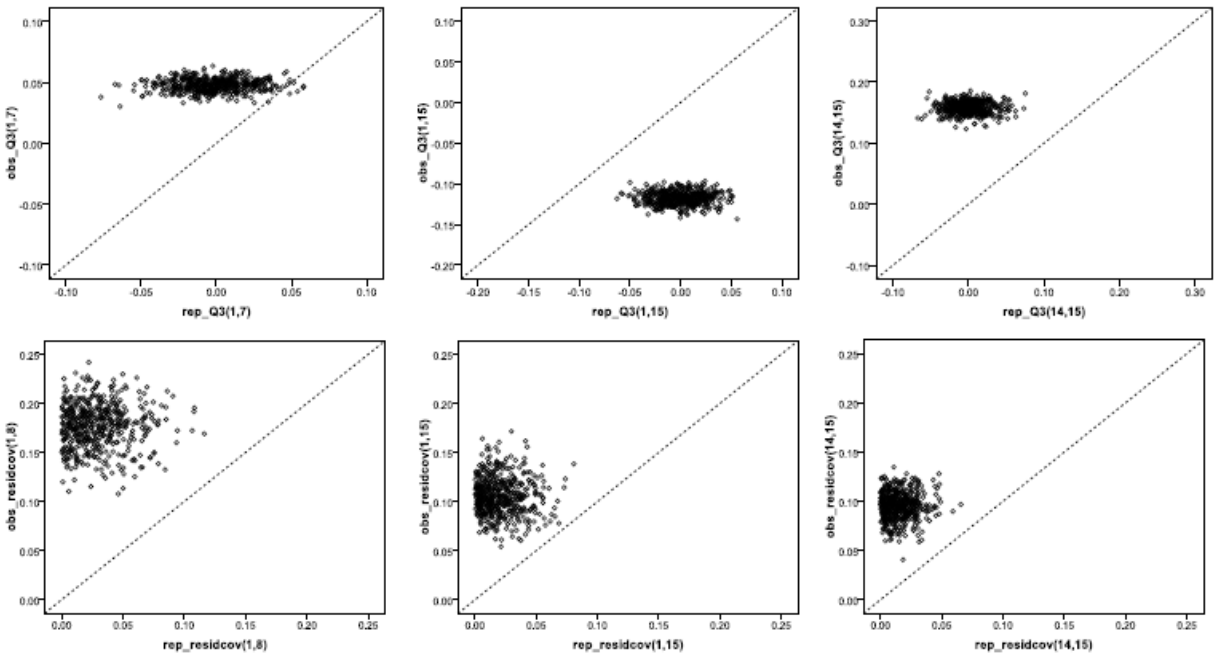


Figure 4.11 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (top), and Item Covariance Residual (bottom) (for a single data) – Condition 2 / Case 2 (ρ=0.6)

Figure 4.11 displays the comparison of realized and PP values of Yen's Q_3 and the item covariance residual measure for different types of item pairs based on a single replication when ρ=0.6. As can been seen from the top plots, for items in the same dimension (Items 1, 7 or Items 14, 15), the realized values of Q_3 were mostly larger than the predictive values since the scatter plot is above the diagonal line. In contrast, for items from the different dimensions (Items 1, 15), the realized values of Q_3 were lower than the predictive values. Unlike Yen's Q_3, the item covariance residual measure has no direction. As observed from the bottom plots, the realized values of residuals were all systematically larger than the predictive residuals under the unidimensional GR model. These results provided evidence of model misfit.

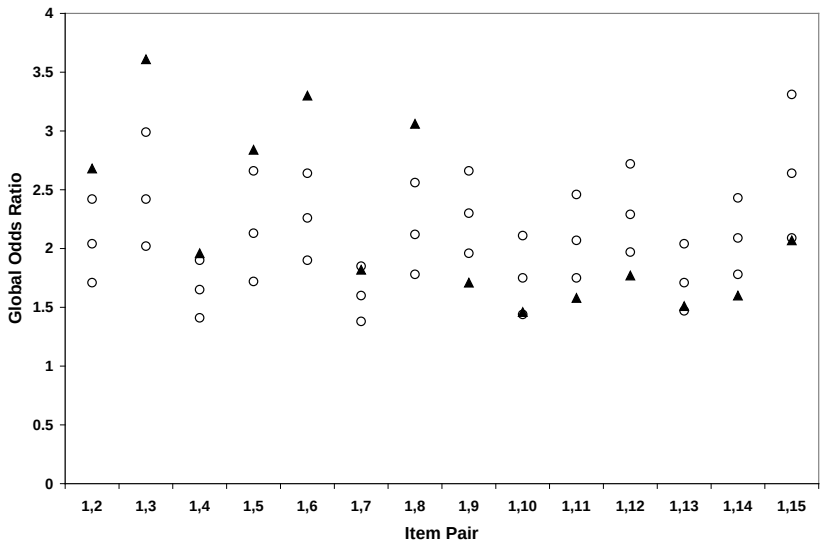


Figure 4.12 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) – Condition 2 / Case 2 (ρ=0.6)

The global OR measure for the first item (90% PP interval, and PP medians) are shown in Figure 4.12. As seen from this figure, the observed global ORs (solid triangle) fall outside or above the PP interval for Item1 paired with Items 2-8 (all in Dim1 items). Whereas, the observed ORs fall outside or blow the PP interval for Item1 paired with the items in Dim2 (Items 9-15). The pattern in this figure indicated that the observed global OR were mostly larger than the predictive values for the item pairs from the same dimension, but smaller for the item pairs from the different dimensions.

The above plots for the three pair-wise measures illustrate results for some item pairs and for one replication. Similar results were found for other item pairs and for the other 19 replications. Overall, the results above indicated that the PPMC method using three pair-wise measures detected a lack of fit of the unidimensional GR model to the two-dimensional test data. In addition, the directional measures, global OR and Yen's Q_3, provided plots which indicated how the items may be grouped dimensionally.

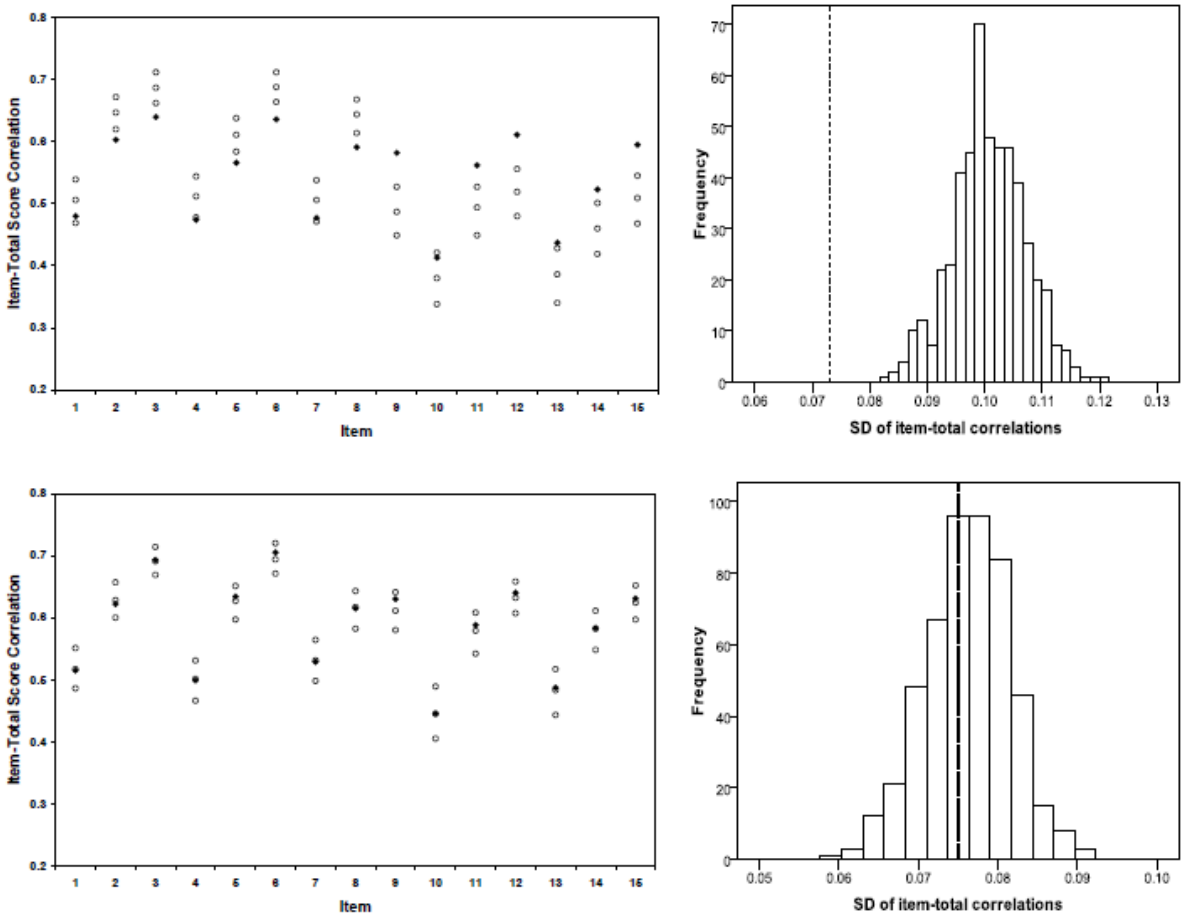


Figure 4.13 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation (Left) and Histogram of Predicted SDs (for a single replication) for Case 1 (top) and Case 2 (bottom) – Condition 2

Recall that the item-total score correlation measure was found to be powerful when the inter-dimensional correlation was 0.3 (Case 1), but exhibited lower power when the correlation increased to 0.6 (Case 2). This finding is clearly illustrated in Figure 4.13 which includes two types of plots for each case. The left plot presents the observed item-total correlation and 90% PP interval for each item based on a single replication. The right plot shows the position of the standard deviation (SD) of the observed item-total correlations for all items in the distribution of the SDs of the predictive item-total correlations. As can been seen, when the correlation was 0.3, the observed correlation fell outside or at the lower end of the PP intervals for the items in Dim1, and fell outside or at the upper end of the intervals for the items in Dim2. The observed SD was

located to the far left in the histogram of the predictive SDs, indicating that the observed item-total correlations were less variable than the predictive correlations. However, when the correlation increased to 0.6, there was not much difference between observed and predictive values. As can been from the bottom plots, the observed correlations approximated the medians of the predictive correlations, and the observed SD is in the middle of the histogram.

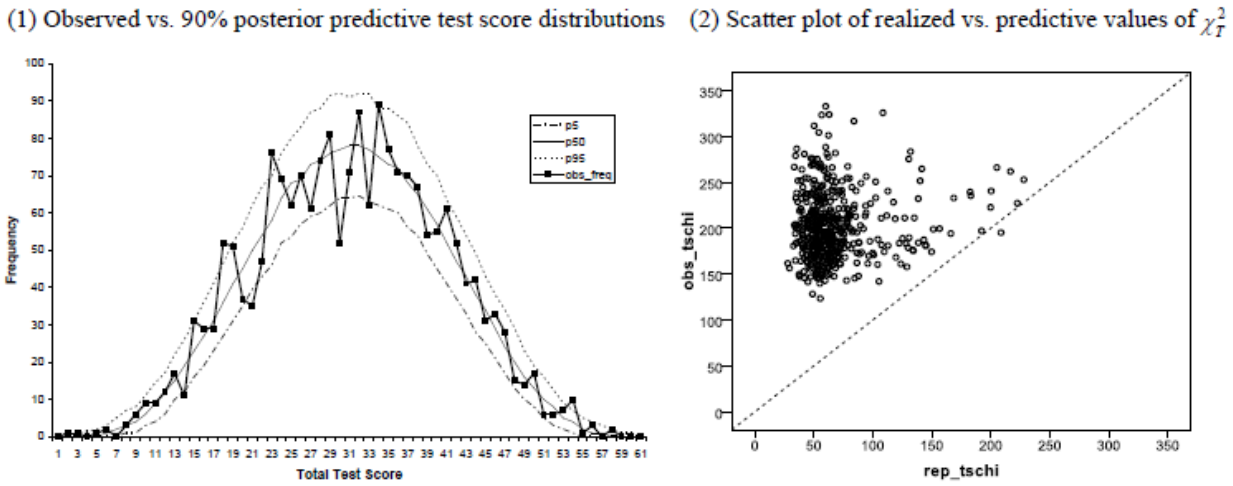


Figure 4.14 Diagnostic Plots based on Test Score Distribution (for a single data) – Condition2 /Case 1

As discussed previously, the test-level measure demonstrated adequate power in detecting the misfit of the GR model to this two-dimensional data when the correlation was 0.3. This finding is illustrated in Figure 4.14 which includes two diagnostic plots based on the total test score distribution for one replication (the PPP-value for this replication was 0.03). The left one displays moderate power since the observed frequencies lie outside the 90% PP intervals for several but not a majority of total test score values. The right plot demonstrates more power since most of the realized χ_T^2 values were larger than predicted values. Compared with Figure 4.2 which includes the same plots under the null condition, Figure 4.14 indicates that the unidimensional GR model can not adequately explain the observed test score distribution given this 2-dim empirical simple-structure data.

4.1.4 Condition 3 (Mg = 2-dim complex-structure GR , Ma = 1-dim GR)

In this condition, the generated data were two-dimensional with complex-structure (Items 1-5 measured a dominant dimension as well as a nuisance dimension, and Items 6-15 only measured the dominant dimension), and a unidimensional model was estimated. The ability of the PPMC method to detect a violation of local independence was explored by using all the 8 proposed measures. Two cases were considered in this condition according to the ratio of a_2 (the slope of the nuisance dimension) to a_1 (the slope of the dominant dimension) for the first 5 items. One ratio was set to 0.5 and another ratio was 1.0, reflecting mild and large dependence between two dimensions, respectively.

Table 4.5 Overall Median PPP-values and Average Proportion of 20 Replications with Extreme PPP-values for all Measures – Condition 3

	Measure	**Type**	**Case 1** (mild dependence)		**Case 2** (large dependence)	
			Median PPP	Power	Median PPP	Power
Test-Level	Test score dist	-	0.29	0.10	0.25	0.10
Item-Level	Item score dist	2dim	0.50	0.00	0.50	0.00
		1dim	0.50	0.00	0.50	0.00
	Item-test corr	2dim	0.33	0.00	0.17	0.00
		1dim	0.57	0.00	0.65	0.00
	Yen's Q_1	2dim	0.53	0.00	0.55	0.00
		1dim	0.51	0.00	0.52	0.00
	Stone's fit stat	2dim	0.56	0.01	0.51	0.03
		1dim	0.52	0.02	0.49	0.01
Pair-Wise	Global OR	(2dim, 2dim)	0.18	0.20	0.00	0.94
		(2dim, 1dim)	0.60	0.06	0.76	0.13
		(1dim, 1dim)	0.44	0.05	0.42	0.06
	Yen's Q_3	(2dim, 2dim)	0.02	0.66	0.00	1.00
		(2dim, 1dim)	0.65	0.09	0.94	0.45
		(1dim, 1dim)	0.44	0.06	0.28	0.10
	Item cov resid	(2dim, 2dim)	0.15	0.18	0.00	1.00
		(2dim, 1dim)	0.53	0.00	0.38	0.01
		(1dim, 1dim)	0.52	0.00	0.50	0.00

Table 4.5 presents the pooled median PPP-values and the average proportions of extreme PPP-values across the 20 replications (i.e., empirical power) for each discrepancy measure and for the two cases. Based on the dimension structure, Items 1-5 were treated as interchangeable,

and Items 6-15 were assumed interchangeable. Thus, the items were classified into two types: "2dim" in the table represents the items measuring two dimensions (Items 1-5); "1dim" reflects the items measuring the dominant dimension (Items 6-15). For each item-level measure, the median PPP-value and empirical power rate were pooled across items of the same type. In addition, there were three types of item pairs: item pairs measuring two dimensions (2dim, 2dim), item pairs measuring the dominant dimension (1dim, 1dim), and pairs reflecting the "2dim" and "1dim" items (2dim, 1dim). The results for the pair-wise measures were pooled from the same type of item pairs and from the 20 replications as well.

As can be seen from Table 4.5, the test-level and item-level measures were not effective in detecting the local dependence among the first 5 items since the power rates were quite small. However, the three pair-wise measures performed effectively. The global OR and item covariance residual measures exhibited low power (0.20 and 0.18, respectively), and Yen's Q_3 showed moderate power (0.66) in detecting the mild local dependence (Case 1) among the first 5 items ("2dim" items). The median PPP-value of Yen's Q_3 for all the pairs among Items 1-5 (2dim, 2dim) was 0.02. This approximately 0 value indicated that most of the realized Q_3 values were consistently larger than the predictive values under the unidimensional GR model, further indicating that the GR model underestimated the association among the first 5 items. In other words, the first 5 items had more dependence than expected under the unidimensional model. Though the global OR and item covariance residual measures did not exhibit adequate power, their median PPP-values for the (2dim, 2dim) pairs were far from 0.50 (0.18 and 0.15, respectively), providing some evidence for model misfit.

As the strength of dependence on the nuisance dimension increased (Case 2), the performance of the pair-wise measures with PPMC improved as would be expected. For the

large dependence case in Table 4.5, both Yen's Q_3 index and the item covariance residual measure had full power (1.00) in detecting the large local dependence among the first five items. Their median PPP-values were 0.00, implying that all the realized values were larger than the predictive values. In addition, the global OR measure exhibited sufficient power (0.94) for this case, and the median PPP-value was also close to 0. Overall, all the three pair-wise measures were effective in detecting the large dependence among the first five items, but for the mild dependence, only Yen's Q_3 appeared to display adequate power.

It is worthy to note that as the degree of dependence increased, Yen's Q_3 measure also had the potential to detect the associations between the modeled dependent and independent items (2dim, 1dim). For Case 2, Yen's Q_3 showed moderate power (0.45) for the (2dim, 1dim) pairs, and the corresponding median PPP-value was 0.94 for Yen's Q_3 index. This high value indicated that most of the realized Q_3 values for the (2dim, 1dim) pairs were consistently smaller than the predictive values under the unidimensional GR model.

Unlike the pair-wise measures, the performances for the test-level and item-level measures did not improve significantly with increased dependence (Case1 vs. Case 2). However, it is interesting to note that though the item-total score correlation was not as effective as the pair-wise measures in detecting the local dependence among the first five items, the decrease in the median PPP-values from 0.33 to 0.17 from Case 1 to Case 2 suggested a potential to detect lack of fit with increased dependence. The low value 0.17 indicated that the observed item-test score correlations for the first five items were larger than the predicted correlations under a unidimensional GR model. How much dependence among items is required for this measure to become effective needs further study.

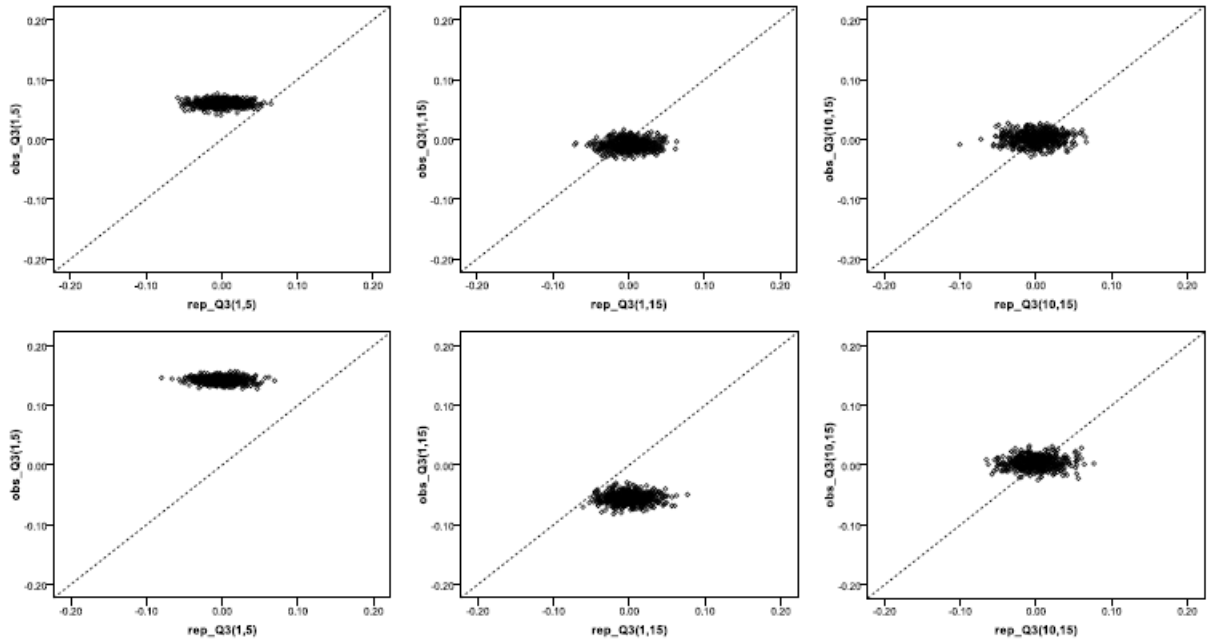


Figure 4.15 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (for a single data) for Case 1 (top) and Case 2 (bottom) – Condition 3

The findings from Table 4.5 are illustrated in Figures 4.15-4.19. Figure 4.15 presents the scatter plots of the realized and predictive Yen's Q_3 values based on one replication for Case 1 (top) and Case 2 (bottom). In each case, there are three example scatter plots for three types of item pairs, respectively. For the (1dim, 1dim) type of pairs (e.g., (Item10, Item15)), about half of the points were above the diagonal line and another half of points were below the line for both cases, indicating there was no systematic difference between the realized and predictive values for the item pair only measuring one dominant dimension. But for the (2dim, 2dim) type of item pairs (e.g., (Item1, Item5)), the scatter plots were consistently above the diagonal line for the mild dependence case, and even further above the diagonal line for the large dependence case. Both of these plots indicated that the realized Q_3 values were consistently larger than the predictive values, and provided graphical evidence for model misfit. In addition, with the degree of dependence increasing, the plot for the (2dim, 1dim) type of item pairs (e.g., (Item1, Item15)) falls below the diagonal line. This indicated that the realized Q_3 values were consistently smaller

than the predictive values, providing more evidence about the misfit of the unidimensional GR model to this simulated locally dependent data.

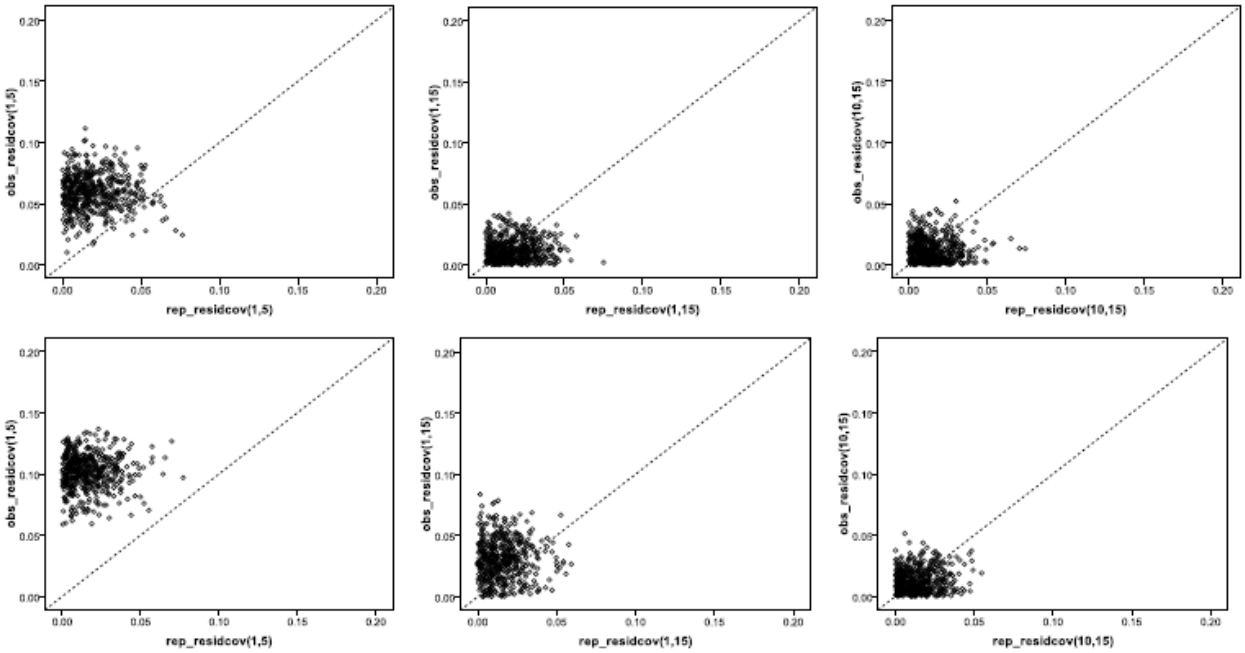


Figure 4.16 Scatter plots of Realized vs. Posterior Predictive Values of Item Covariance Residual (for a single data) for Case 1 (top) and Case 2 (bottom) – Condition 3

Figure 4.16 includes similar scatter plots for the item covariance residual measure based on the same replications used for Yen's Q_3. As can be seen, for the (2dim, 2dim) type of item pairs (e.g., (Item1, Item5)), most points were above the diagonal line for the mild dependence case, and the entire plot was above the line when the dependence was large (Case 2). This result indicates the realized item covariance residuals were systematically larger than the predictive values under the unidimensional GR model, thus providing evidence of model misfit.

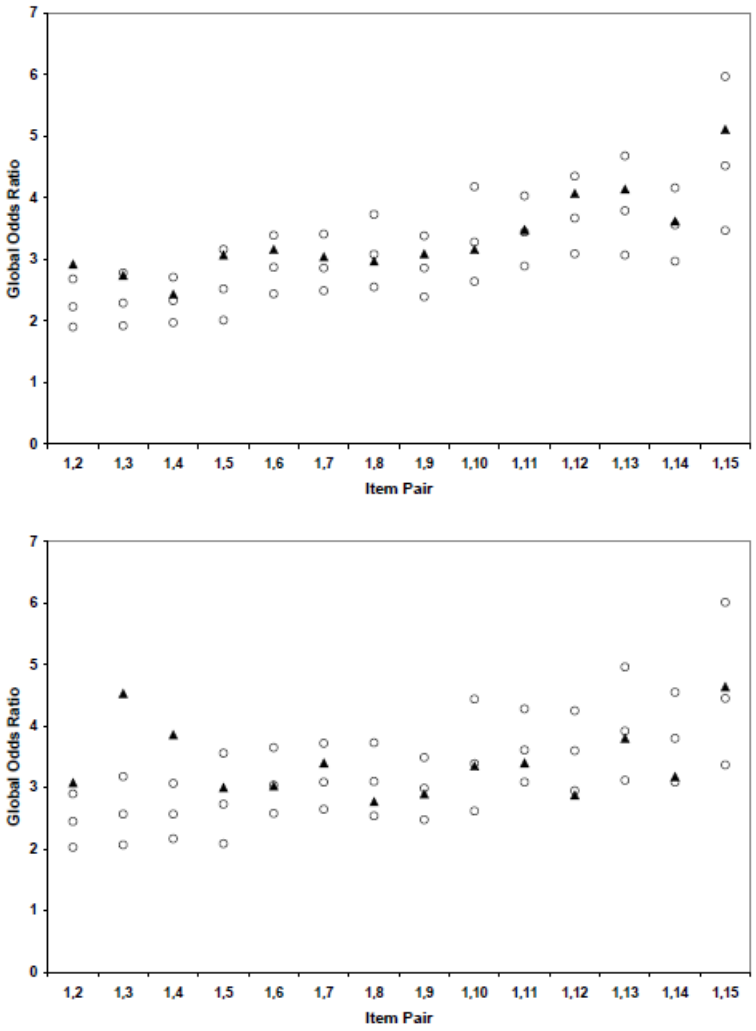


Figure 4.17 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 1 with Other Items (for a single replication) for Case 1 (top) and Case 2 (bottom) – Condition 3

Figure 4.17 displays the observed global ORs for Item 1, the 90% PP interval, and PP medians for the two dependence conditions. As seen for Case 1 from this figure, most of the observed global ORs (solid triangles) fall outside or at the upper end of the PP interval for Item1 paired with other items measuring two dimensions (Items 2-5), and tend to be far above the interval when the dependence increased (Case 2). In contrast, almost all the observed ORs lay within the PP interval for Item1 paired with items measuring only one dimension (Items 6-15). It should be noted that although Figures 4.15 – 4.17 for each case were drawn from one dataset, the same phenomena were observed for the other 19 datasets.

As for the previous conditions, pie plots were used to examine any pattern in the PPP-values. Figures 4.18 and 4.19 display the median PPP-values (Left) and empirical power (Right) of the three item-pair measures for each item pair across the 20 replications for Case 1 and Case 2, respectively. The pattern in the PPP-values can be easily observed from Case 2, the large dependence case (Figure 4.19). For the directional measures (global OR, and Yen's Q_3), the median PPP-values were around 0.50 for the (1dim, 1dim) pairs, close to 0 for the (2dim, 2dim) pairs, and close to 1 for the (2dim, 1dim) pairs. This pattern is more evident for the most effective measure - Yen's Q_3. For the non-directional measure – item covariance residual, the median PPP-values were close to 0 for the (2dim, 2dim) pairs, but around 0.50 for the (1dim, 1dim) and (2dim, 1dim) pairs. In addition, the empirical power rates of these three measures were all close to 1 for the (2dim, 2dim) pairs, but Yen's Q_3 measure also had moderate power for the (2dim, 1dim) pairs.

For the mild dependence case, Case 1 (Figure 4.18), the pattern is not as evident as for Case 2. However, it is still clear that the first 5 items were different from the remaining items. Their extreme PPP-values indicated that the unidimensional GR model did not fit these 5 items. The patterns found in these two figures were different from the patterns under the null condition, thus providing evidence of model misfit.

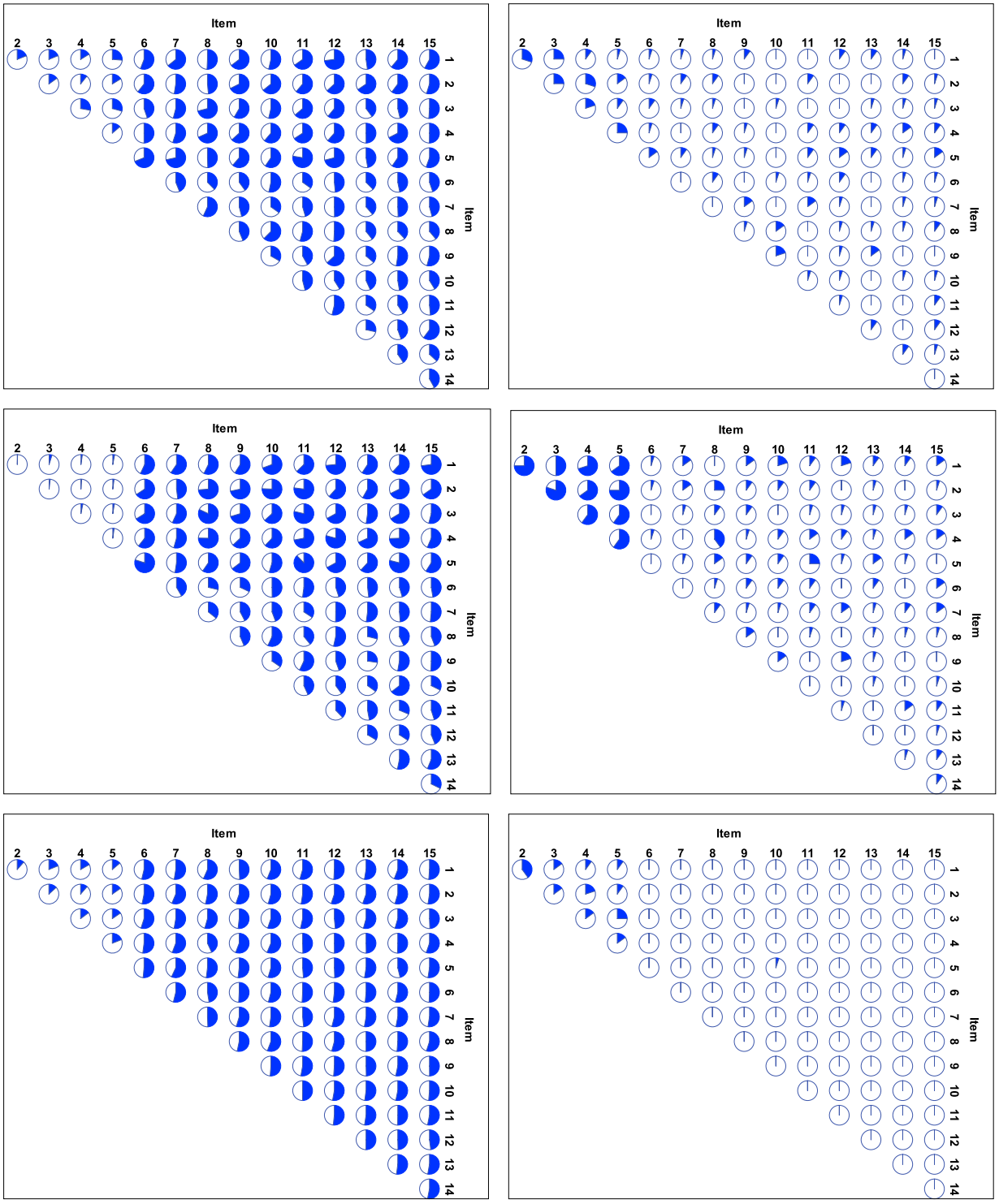


Figure 4.18 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 3/ Case 1

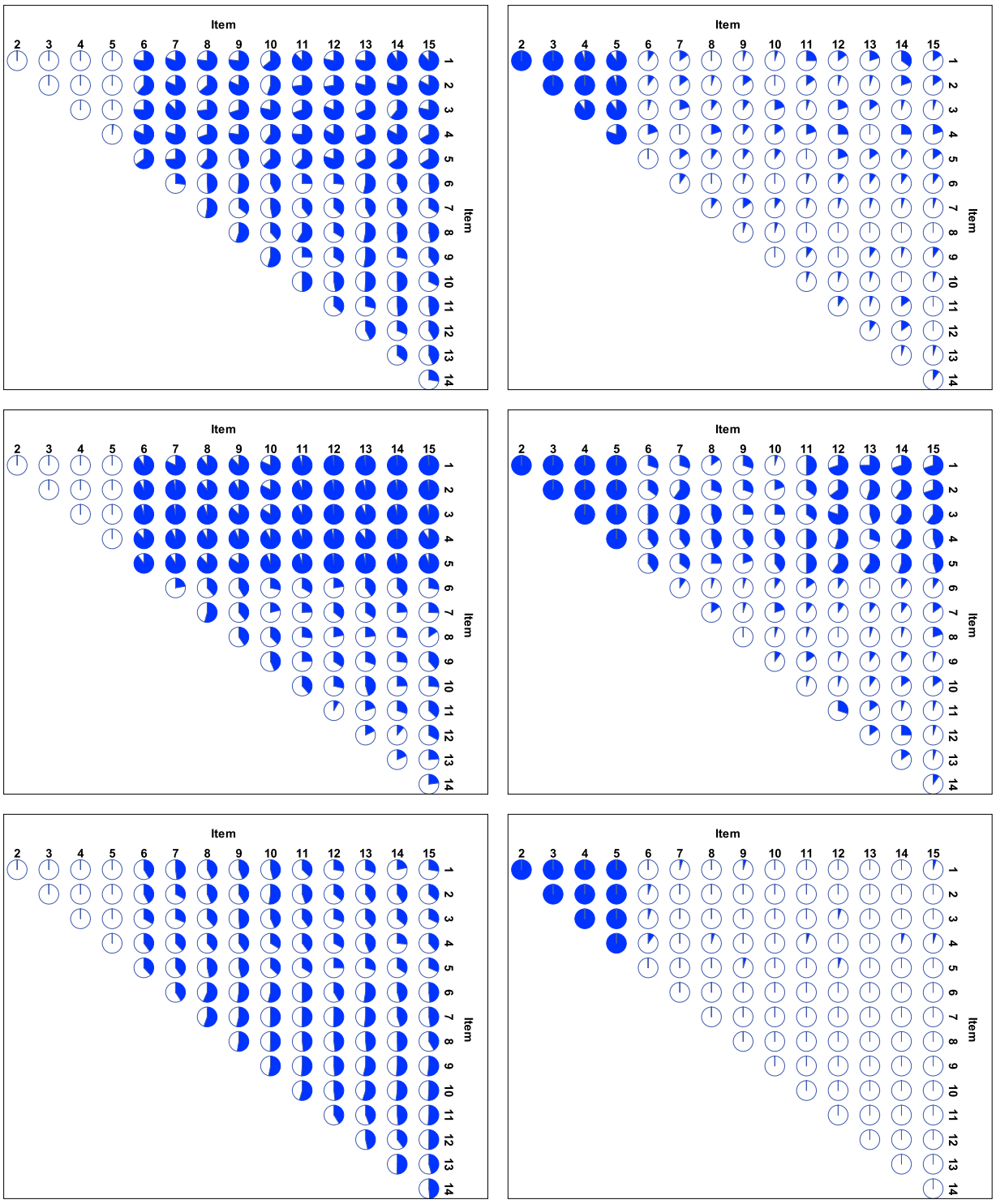


Figure 4.19 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 3/ Case 2

4.1.5 Condition 4 (Mg = testlet GR , Ma = 1-dim GR)

In this condition, the effectiveness of different discrepancy measures with PPMC in detecting local dependence among responses to testlet items was investigated. Recall that for this condition, Items 6, 7 and 8 were designed to be in a testlet and three levels of dependence among them were considered: mild ($\sigma^2_{d(i)} = 0.5$), large ($\sigma^2_{d(i)} = 1.0$), and extremely large ($\sigma^2_{d(i)} = 2.0$). The other items were simulated to be locally independent.

Table 4.6 Overall Median PPP-values and Average Proportion of 20 Replications with Extreme PPP-values for all Measures – Condition 4

	Measure	**Type**	**Case 1** (mild)		**Case 2** (large)		**Case 3** (extreme large)	
			Median PPP	Power	Median PPP	Power	Median PPP	Power
Test-Level	Test score dist	-	0.64	0.05	0.49	0.05	0.46	0.15
Item-Level	Item score dist	testlet	0.50	0.00	0.50	0.00	0.49	0.00
		indep	0.50	0.00	0.50	0.00	0.50	0.00
	Item-test corr	testlet	0.14	0.00	0.05	0.52	0.00	1.00
		indep	0.54	0.00	0.63	0.00	0.75	0.02
	Yen's Q_1	testlet	0.45	0.00	0.47	0.00	0.48	0.00
		indep	0.50	0.00	0.50	0.00	0.51	0.00
	Stone's fit stat	testlet	0.50	0.02	0.40	0.02	0.50	0.02
		indep	0.51	0.02	0.49	0.01	0.49	0.01
Pair-Wise	Global OR	(testlet, testlet)	0.00	1.00	0.00	1.00	0.00	1.00
		(testlet, indep)	0.70	0.10	0.80	0.20	0.86	0.23
		(indep, indep)	0.40	0.05	0.43	0.05	0.40	0.06
	Yen's Q_3	(testlet, testlet)	0.00	1.00	0.00	1.00	0.00	1.00
		(testlet, indep)	0.91	0.40	0.98	0.58	0.99	0.72
		(indep, indep)	0.39	0.08	0.36	0.09	0.36	0.09
	Item cov resid	(testlet, testlet)	0.00	1.00	0.00	1.00	0.00	1.00
		(testlet, indep)	0.45	0.01	0.28	0.05	0.18	0.14
		(indep, indep)	0.52	0.00	0.50	0.00	0.52	0.00

Table 4.6 presents the overall median PPP-values and average proportions of extreme PPP-values for the three cases. In this condition, there are two types of items – those labeled "testlet" represents the testlet items (Items 6-8); and those labeled "independent" are the other items. There are also three types of item pairs – testlet item pairs (testlet, testlet), independent item pairs (indep, indep), and pairs reflecting one testlet item and one independent item (testlet,

testlet). For each item-level measure, the median PPP-values and empirical power rates in Table 4.6 were pooled from the same type of items and from the 20 replications. For each pair-wise measure, the median PPP-values and empirical power rates were pooled from the same type of item pair and also aggregated over the 20 replications.

As found in Table 4.6, the three pair-wise measures had full power (1.00) in detecting the misfit of unidimensional GR model to the modeled dependence among the testlet items, even for the mild dependence case. The median PPP-values of these three measures were 0 for the (testlet, testlet) pairs across the three cases, indicating that the realized associations among the testlet items were consistently larger than the predicted under the GR model. In addition, the ability of the two directional measures (global OR and Yen's Q_3) in detecting the misfit of the GR model to the relationships between the testlet items and the independent items increased as the degree of modeled dependence among the testlet items increased. Specifically, Yen's Q_3 measure showed low (0.40), moderate (0.58), and large (0.72) power for the (testlet, indep) pairs for the mild, large, and extremely large dependence cases, respectively. The global OR measure also exhibited low power (0.20 and 0.23) for the (testlet, indep) pairs for Case 2 and Case 3, but very low power for the mild dependence condition. In contrast, the item-covariance residual exhibited very low power for the (testlet, indep) pairs, even for the extremely large dependence condition. The median PPP-values of Yen's Q_3 measures were close to 1 for the (testlet, independent) pairs, implying that the realized associations between the testlet item and independent items were mostly lower than the predicted under the GR model. However, the pooled median PPP-values for the (indep, indep) item pairs for all the three pair-wise measures were close to 0.50, indicating the realized associations between the independent items were consistent with predicted values under the GR model.

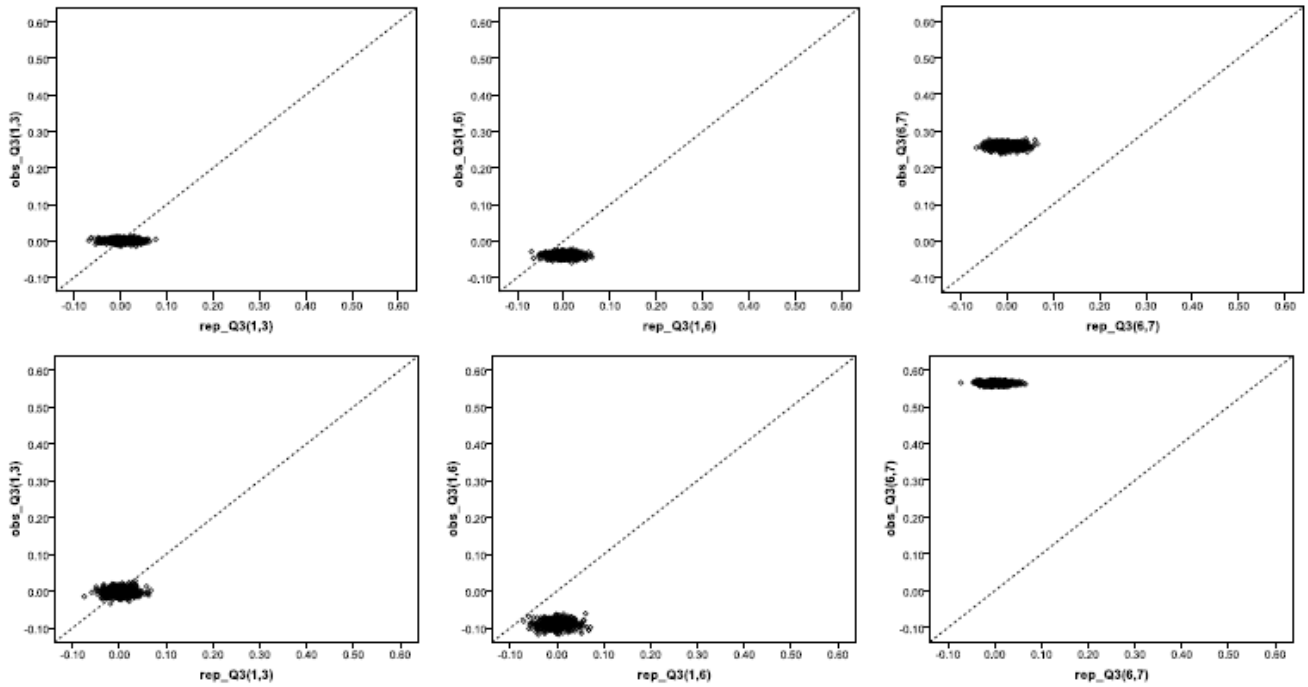


Figure 4.20 Scatter Plots of Realized vs. Posterior Predictive Values of Yen's Q_3 (for a single data) for Case 1 (top) and Case 3 (bottom) – Condition 4

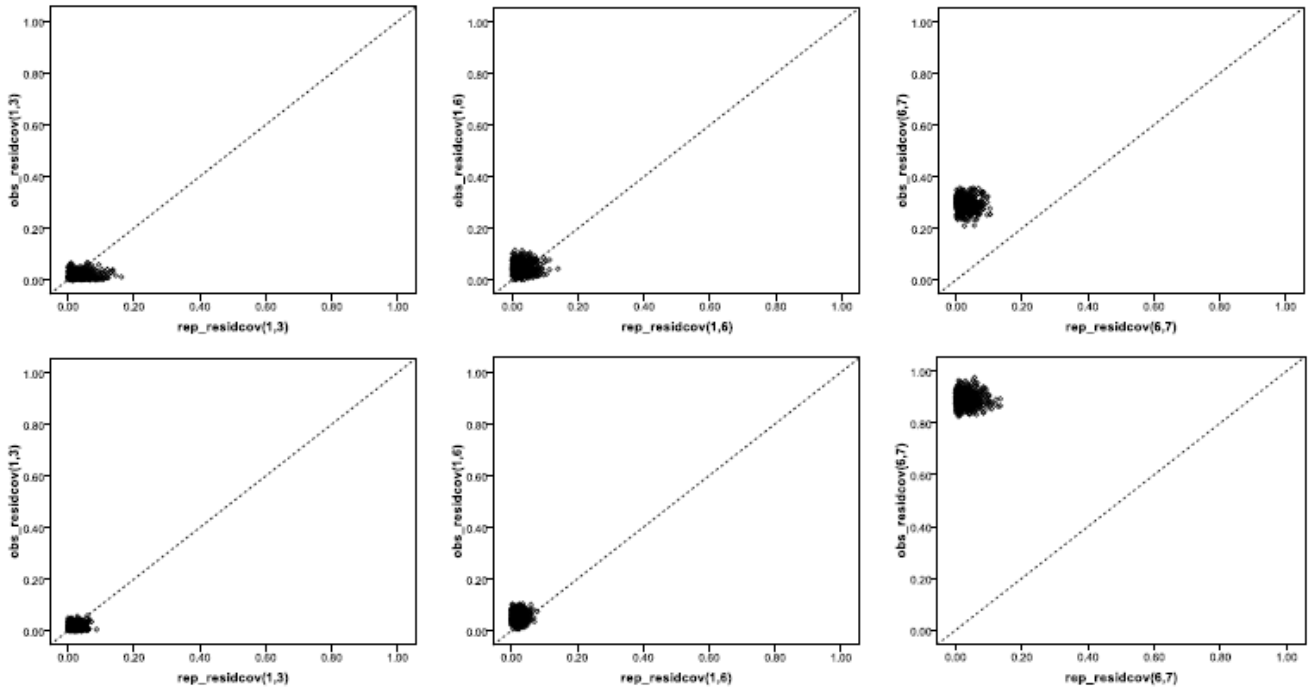


Figure 4.21 Scatter Plots of Realized vs. Posterior Predictive Values of Item Covariance Residual (for a single data) for Case 1 (top) and Case 3 (bottom) – Condition 4

The findings about the pair-wise measures from Table 4.6 were also revealed from Figures 4.20 – 4.22 which were based on a single replication for each of two cases – mild dependence (Case 1) and extremely large (Case 3). Note that similar figures were observed for the other 19 replications.

Figure 4.20 shows the realized and posterior predictive Yen's Q_3 values for three different types of item pairs. The (Item1, Item3) pair reflects an (indep, indep) type of pair, the (Item1, Item6) reflects a (testlet, indep) type of pair, and the (Item6, Item7) represents a (testlet, testlet) pair. As can be seen, the realized Q_3 values for the (Item6, Item7) pair were consistently and sufficiently larger than the predictive values, that is, the entire scatter plot was far above the diagonal line. In contrast, the realized Q_3 values for the (Item1, Item6) pair were systematically smaller than the predictive values since most part of the scatter plot was below the diagonal line. Moreover, the discrepancies between the observed and predictive values tended to increase as the dependence among the testlet items increased. However, for the (Item1, Item3) pair, there was no systematic difference between the realized and predictive Q_3 values for both cases, and both predictive and realized values were around 0. In summary, these plots provide evidence about the directional misfit of the unidimensional GR model. The model under-estimated the relationship between the testlet items, but over-estimated the relationship between the testlet and independent items.

Figure 4.21 includes the scatter plots of the realized and posterior predictive item covariance residuals for three different types of item pairs. As can be observed, the predictive item covariance residuals under the unidimensional GR model were close to 0 for each item pair. For the independent item pair (Item1, Item3), the realized and predictive residuals were in the same range. However, for the testlet item pairs, the realized values were consistently larger than

the predictive value of 0 for both cases. They ranged from 0.2 to 0.4 for the mild dependence case, and from 0.8 to 1.0 for the extremely large dependence case. These large realized residuals indicated misfit of the GR model. As discussed previously, unlike Yen's Q_3 measure, the item covariance residual measure demonstrated very power in detecting the misfit of the model for the testlet and independent item pairs. This was also illustrated in the two plots for (Item1, Item6) in which there was no clear difference between the realized and predictive residuals though the range of realized residuals tended to a bit larger than the predictive range for Case 3.

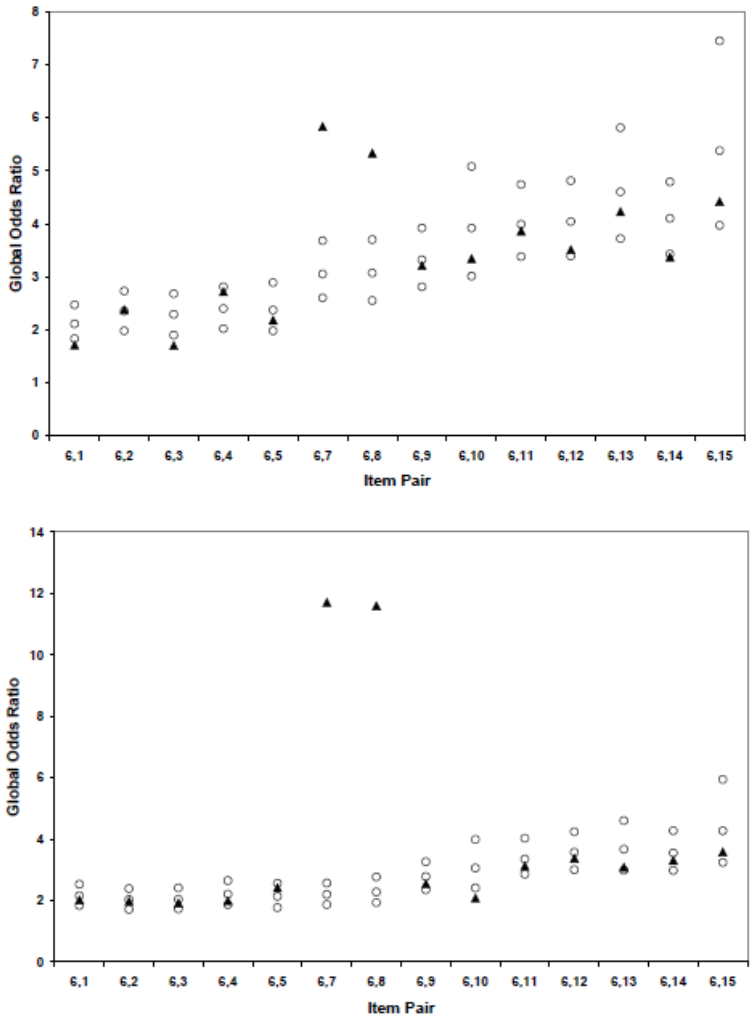


Figure 4.22 Observed vs. 90% Posterior Predictive Interval of Global OR for Item 6 with Other Items (for a single replication) for Case 1 (top) and Case 3 (bottom) – Condition 4

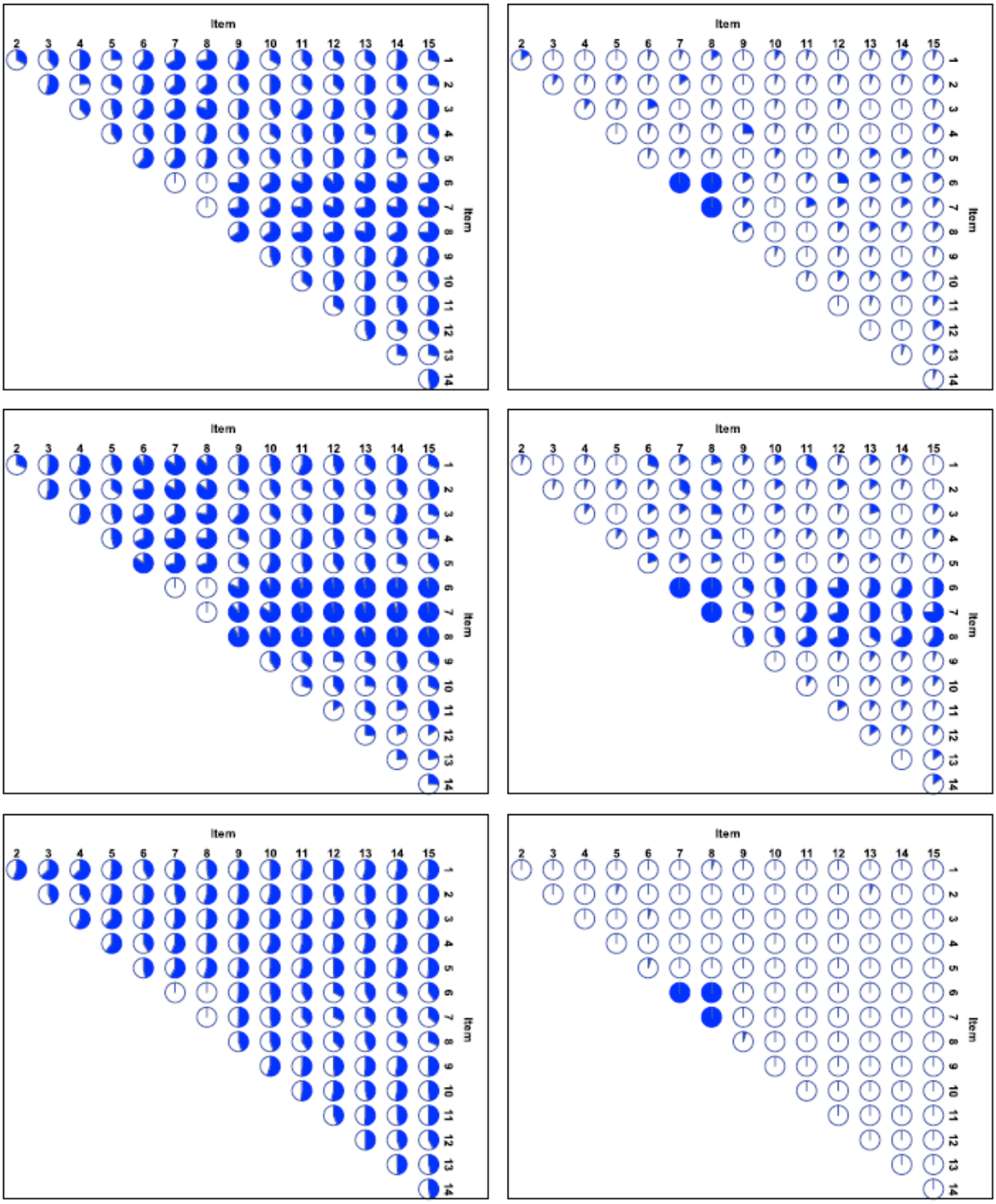


Figure 4.23 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 4/Case 1

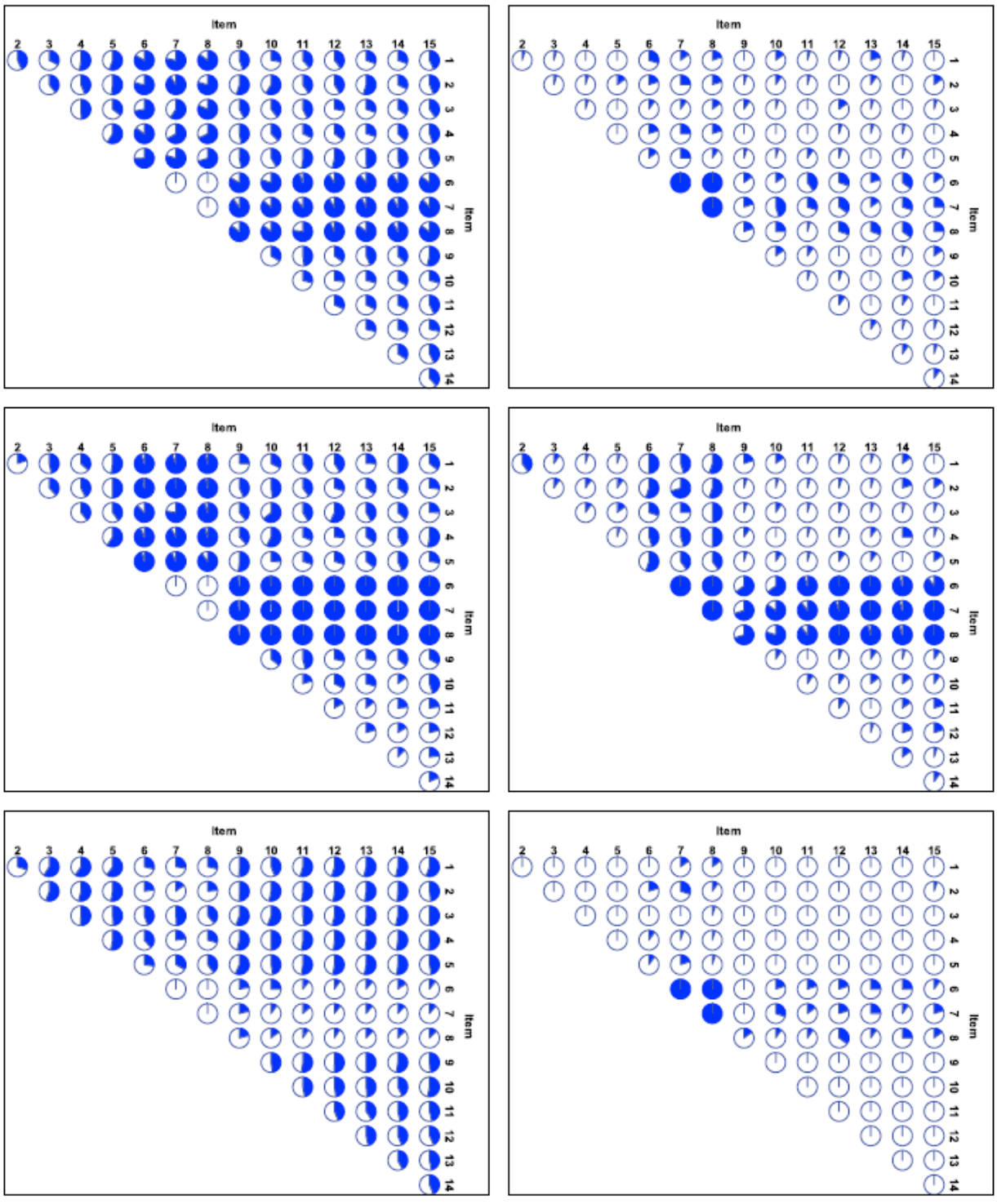


Figure 4.24 Display of Median PPP-values (Left) and Proportion of 20 Replications with Extreme PPP-values (Right) for Global OR (Row1), Yen's Q_3 (Row2), and Item Covariance Residual (Row3) – Condition 4/Case 3

Figure 4.22 displays the observed global OR value versus 90% PP interval for the global OR measure for Item 6 paired with the other items for two cases. As seen from this figure, the observed global ORs were far above the PP intervals for the two testlet item pairs ((Item6, Item7) and (Item6, Item8)), implying that the unidimensional GR model could not adequately to capture the dependencies among the responses to the testlet items.

The pattern of the PPP-values was also explored from pie plots for the pair-wise measures. Figures 4.23 and 4.24 display the median PPP-values (Left) and empirical power (Right) for each item pair across the 20 replications for three measures for the mild and extremely large dependence cases, respectively. From Figure 4.24, the median PPP-values of two directional measures (global OR and Yen's Q_3) were around 0.50 for the independent item pairs, close to 0 for the testlet item pairs, and close to 1 for the item pairs between the testlet and independent items. For the item covariance residual measure, the median PPP-values were also around 0.50 for the (indep, indep) pairs, and close to 0 for the (testlet, indep) or (testlet, testlet) pairs. Items appear to fall into two clusters: Items 6-8 in one and the remaining items in another. This pattern was clearly different from the corresponding plots under the null condition (Figure 4.5), providing strong evidence about the misfit of the GR model to the data with the large testlet effect.

Although the pattern for the mild dependence case (Figure 4.23) was not as evident as for the extremely large dependence case, the extreme PPP-values for the three testlet items also provide evidence about lack of model fit. In addition to the median PPP-values, the pie plots reflecting empirical power rates illustrate that all the three pair-wise measures had full power in detecting the local dependence among the testlet items, and Yen's Q_3 measure also exhibited moderate power in detecting a lack of fit in the unidimensional GR model to the (testlet, indep)

item pairs. Since all three pair-wise measures exhibited full power in detecting local dependence among the testlet item pairs, it may be useful to determine when these three measures will lose their full power. This could be evaluated by manipulating more levels of testlet effect less than $\sigma^2_{d(i)} = 0.5$.

As was seen from Table 4.6, the power of the item-total score correlation measure in detecting the misfit of the GR model to the testlet items increased as the degree of testlet dependence increased. The pooled median PPP-values were 0.14, 0.05, and 0.00 for the mild, large, and extremely large dependence cases, respectively. The corresponding power increased from no power (0.00) to moderate power (0.52) and to full power (1.00) for the three cases, respectively. The median PPP-value tended to be 0 for testlet items, indicating that the observed correlations for these items were higher than the predictive correlations. In contrast, for the independent items, the median PPP-values for the three cases were not extreme, indicating adequate fit of the GR model to these items. This phenomenon can also be demonstrated from Figure 4.25 which presents the observed correlation and 90% PP interval for each item based on a single replication. For the independent items, the observed correlations approximated the medians of the predictive correlations across the three cases. But for the testlet items, the observed correlations were at the upper end of the intervals for the mild dependence case, and fell outside the interval with the large dependence case.

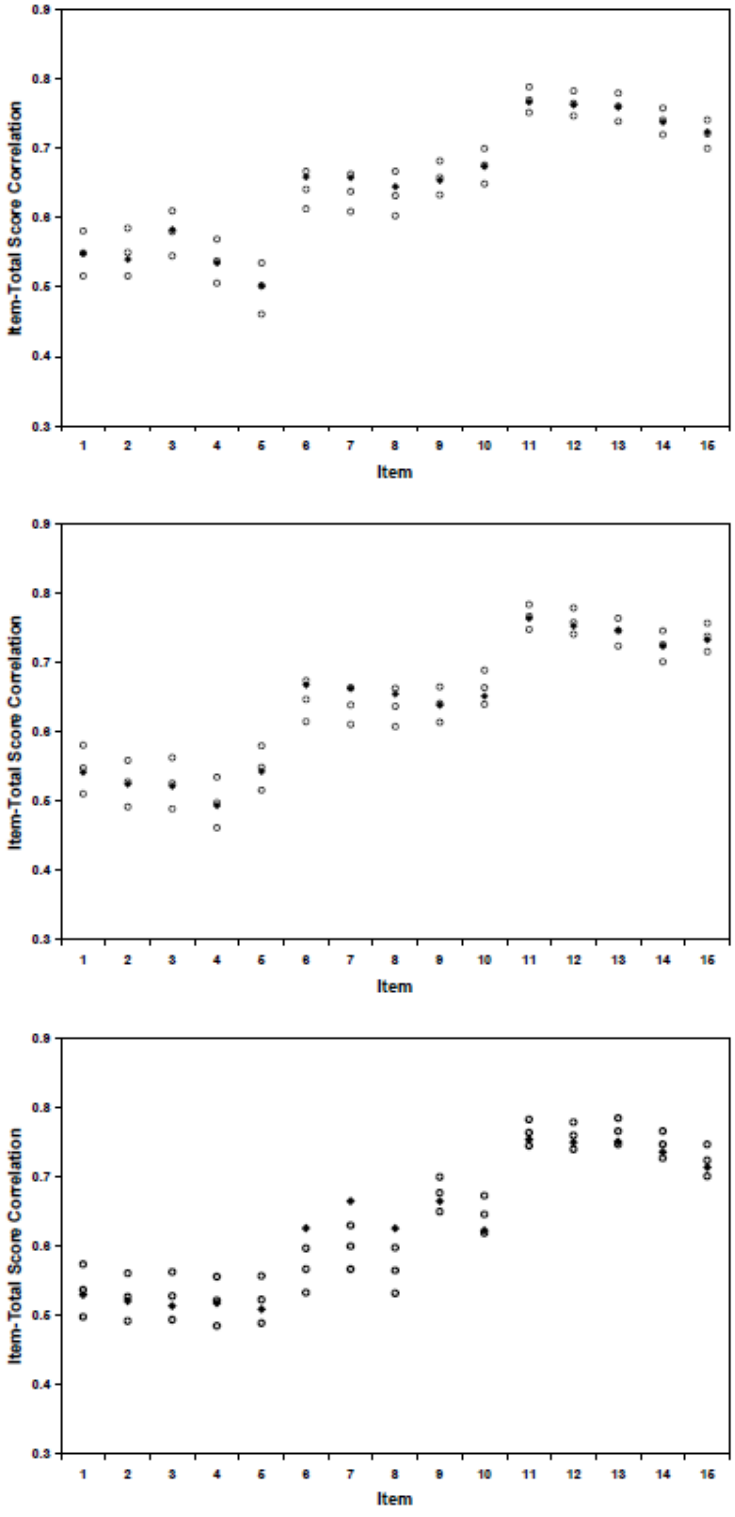


Figure 4.25 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation for Case 1 (top), Case 2 (middle), and Case 3 (bottom) based on a single replication – Condition 4

4.1.6 Condition 5 (Mg = items with improper BCCs , Ma = 1-dim GR)

This condition was intended to explore the performance of PPMC in assessing misfit due to an incorrect form of the logistic BCC functions. As discussed in Chapter 3, Items 7 and 8 were simulated to follow BCCs functions that differed from the logistic functions under the unidimensional GR model. Specifically, The BCCs of Item 7 followed cubic functions, and the BCCs of Item 8 were two-step Guttman functions. The remaining 13 items ("Other Items") were simulated based on logistic BCC functions under the unidimensional GR model.

Table 4.7 Overall Median PPP-values and Average Proportion of Replications with Extreme PPP-values for all Measures – Condition 5

	Measure	**Type**	Median PPP	Power
Test-Level	Test score dist	-	0.61	0.20
Item-Level	Item score dist	Item 7 (cubic)	0.44	0.00
		Item 8 (step)	0.49	0.00
		Other Items (logistic)	0.50	0.00
	Item-test correlation	Item 7 (cubic)	0.63	0.00
		Item 8 (step)	0.46	0.00
		Other Items (logistic)	0.44	0.00
	Yen's Q_1	Item 7 (cubic)	0.29	0.00
		Item 8 (step)	0.04	0.65
		Other Items (logistic)	0.49	0.00
	Stone's fit statistic	Item 7 (cubic)	0.01	0.90
		Item 8 (step)	0.00	1.00
		Other Items (logistic)	0.49	0.04
Pair-Wise	Global OR	(Item7, Item8)	0.52	0.00
		(misfit, fit)	0.59	0.09
		(fit, fit)	0.47	0.07
	Yen's Q_3	(Item7, Item8)	0.67	0.10
		(misfit, fit)	0.51	0.05
		(fit, fit)	0.49	0.07
	Item covariance residual	(Item7, Item8)	0.44	0.00
		(misfit, fit)	0.52	0.00
		(fit, fit)	0.53	0.00

Table 4.7 presents the overall median PPP-values and average proportions of extreme PPP-values across the 20 replications for this condition. As can be seen from this table, for each

item-level measure, the median PPP-values and power for the simulated GR items ("Other Items") were pooled across the 13 items and across the 20 replications. For each pair-wise measure, three values were computed for the overall median PPP-value and the average empirical power, respectively. One was for the pair of two misfitting items, (Item 7, Item 8), another for the pairs between one misfitting item and one fitting item, and the third one for the fitting item pairs.

The results in Table 4.7 show that only two classical item-fit statistics detected misfit between the observed BCCs and the predictive BCCs under the GR model. For the simulated GR items, the median PPP-values were 0.49 for both fit measures, and the average proportions of extreme PPP-values for Yen's Q_1 and Stone's X^2 were 0.00 and 0.04, respectively. The average proportions for the fitting items reflect the Type-I error rates in a hypothesis testing framework. Though both item fit measures were conservative in the PPMC context, Stone's measure had a larger Type-I error rate than Yen's measure. Regarding the power in detecting the misfitting items, Stone's measure exhibited sufficient power in detecting the two modeled misfitting items – 0.90 for Item 7, and 1.00 for Item 8. Yen's Q_1 measure was found to have less power (0.65) for detecting the misfitting item with two-step Guttman BCC functions (Item 8), but did not exhibit any power for the misfitting item with cubic BCC functions (Item 7). Since only two types of BCC functions were considered and several factors were fixed in this study, the comparison of the performance of these two item-fit statistics in a Bayesian framework requires further investigation.

Figure 4.26 displays the scatter plots of realized and posterior predictive values for the two item-fit measures for one replication. Note that the other 19 replications had similar plots. For the fittting item (Item 1), the observed values were not systematically different from the

predictive values for both measures, indicating close correspondence between the observed and model-predicted BCCs. For the misfitting Item 7, the scatter plot for Stone's fit statistic was mostly above the diagonal line. This indicated that most of the observed values were larger than the predictive values, further suggesting item misfit. In contrast, the plot of Yen's measure did not provide evidence of model misfit for this item. For the misfitting Item 8, the scatter plots for both measures provide clear evidence of model misfit for this item.

Except for the two item-fit statistics, the other measures appeared to be ineffectiveness in detecting the departure of the observed BCCs from the predicted BCCs under the unidimensional GR model. Though the three pair-wise measures showed sufficient power for the violation of unidimensionality and local independence, they were not useful for this condition. Figure 4.27 displays the pie plots for the pair-wise measures. As can been seen, the pattern in the pie plots was very similar to that under the null condition, providing no evidence for model misfit.

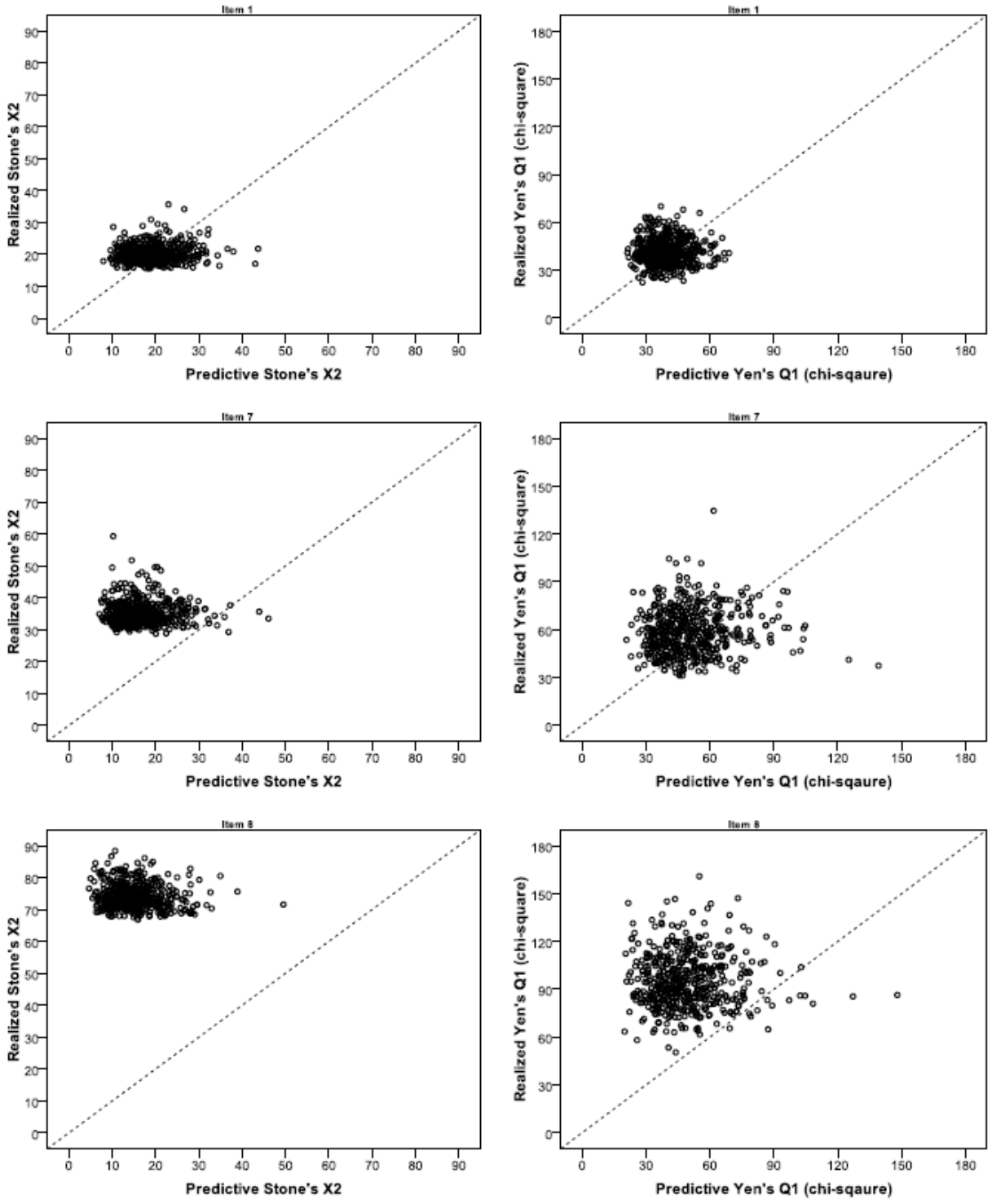


Figure 4.26 Scatter plots of Realized vs. Posterior Predictive Values of Yen's Q_1 and Stone's X^2 Item-Fit Statistics (for a single data) – Condition 5

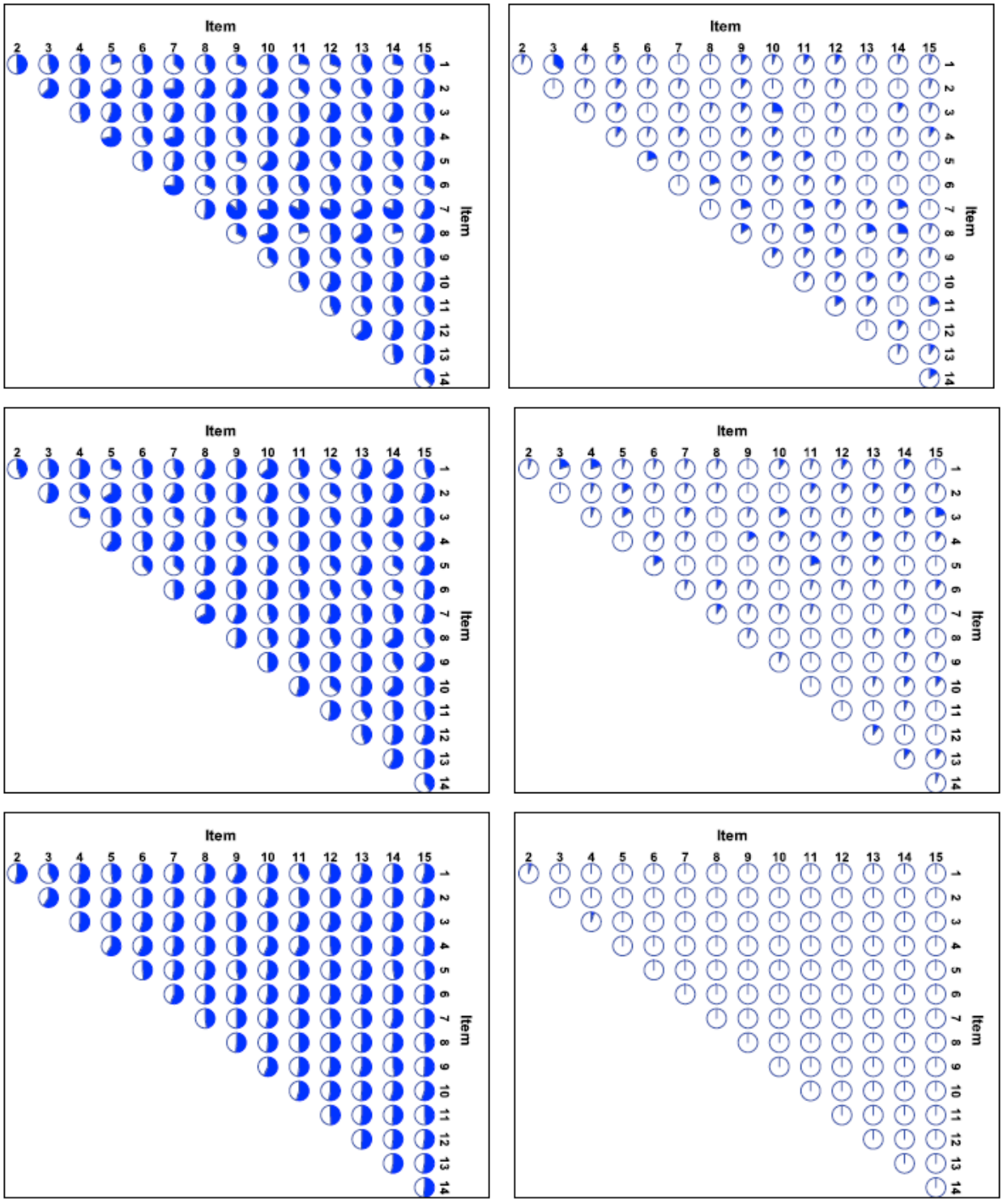


Figure 4.27 Display of Median PPP-values (left) and Proportion of 20 Replications with Extreme PPP-values (right) for Global OR (row1), Yen's Q_3 (row2), and Item Covariance Residual (row3) – Condition 5

4.2 RESULTS FROM SIMULATION STUDY 2

Study 2 aimed to explore the relative performance of three Bayesian model comparison methods (DIC, CPO, and PPMC) under four different model comparison conditions (see Table 3.12). The different models that were considered included: the two-parameter (2P) graded response (GR) model, the one-parameter (1P) GR model, the rating scale (RS) model, the testlet graded model, and multidimensional graded model. In each condition, typical performance assessment data were generated based on an appropriate IRT model (Mg) and then calibrated using several different data-analysis (Ma) models. Three Bayesian model comparison indices were then computed for each Ma and a preferred model was selected based on each of indices. The relative performance of these three indices was compared with respect to the number of times each index selected the generating or correct model across 20 replications.

4.2.1 Condition 1 (2P GR vs. 1P GR vs. RS Models)

In Condition 1, the data were generated based on 2P GR models, but calibrated using 2P GR, 1P GR, and RS models. These models differ in terms of the number of parameters to be estimated. The purpose of this condition was to determine how effectively the model comparison criteria could discriminate between these three models and select the 2P GR as the preferred model.

Item parameter recovery for the 2P GR model was examined first. Table 4.8 gives the RMSD for each item parameters across the 20 replications. The average RMSD across all items was 0.07 for both slope and threshold parameters. These results indicate one chain of 5000 and a posterior sample of 500 were adequate for estimating the 2P GR model using MCMC within WinBUGS. They were also adequate for the other two models because of the fewer parameters.

Table 4.8 RMSD for Item Parameter Recovery in WinBUGS for 2P GR Model

Item	a	b1	b2	b3	b4
1	0.07	0.13	0.09	0.06	0.09
2	0.05	0.09	0.06	0.05	0.09
3	0.06	0.10	0.06	0.08	0.12
4	0.04	0.14	0.08	0.06	0.07
5	0.05	0.07	0.05	0.08	0.16
6	0.08	0.09	0.06	0.05	0.06
7	0.06	0.06	0.03	0.03	0.06
8	0.07	0.06	0.05	0.06	0.06
9	0.08	0.15	0.06	0.05	0.05
10	0.08	0.05	0.05	0.07	0.15
11	0.07	0.09	0.04	0.03	0.05
12	0.08	0.04	0.03	0.04	0.06
13	0.09	0.04	0.03	0.05	0.06
14	0.10	0.19	0.06	0.04	0.04
15	0.08	0.05	0.04	0.04	0.14
	$\overline{RMSD}(a) = 0.07$		$\overline{RMSD}(b) = 0.07$		

Table 4.9 presents summary descriptive statistics (i.e., min, max, and mean) of the DIC and test-level CPO values across the 20 replications for each model, as well as the rank of the three models based on the mean value. The frequencies of choosing each model across the 20 replications are also reported in this table. As can been seen, the mean DIC values were 73960, 75498, and 74484 for the 2P GR, 1P GR, and RS models, respectively, indicating the 2P GR model (Rank 1) fit the data better than the RS model (Rank 2) which in turn was better than the 1P GR model (Rank 3). In addition, the mean test-level CPO values were -16076, -16405, and -16190 for the 2P GR, 1P GR, and RS models, respectively. Unlike the DIC index, larger CPO values reflect the preferred model. Thus, the CPO and DIC indices reached the same conclusion about the comparison of these three models. In addition, these two indices appeared to perform equally well regarding the frequency of choosing the 2P GR model as the preferred model for the overall test. As seen in this table, both indices chose the generating or true model as the preferred

model for each of the 20 replications. The distributions of the index values for the three models are shown in Figure 4.28 using the box plots. The distribution of DIC values for the 1P GR model was above the distribution for the 2P GR model, whereas the distribution of CPO values for the 1P GR model was below the distribution of values for the 2P GR model. Both suggested that the 2P GR model fit the data consistently better across the 20 replications.

Table 4.9 Model Selection for Overall Test using DIC and Test-Level CPO – Condition 1

	DIC				
Model	Min	Max	Mean	Rank	Frequency of Selecting Model
2P GR*	73322	74566	73960	1	20 (100%)
1P GR	74854	76098	75498	3	0 (0%)
RS	73863	75113	74484	2	0 (0%)
	CPO				
2P GR*	-16208	-15937	-16076	1	20 (100%)
1P GR	-16536	-16265	-16405	3	0 (0%)
RS	-16327	-16054	-16190	2	0 (0%)

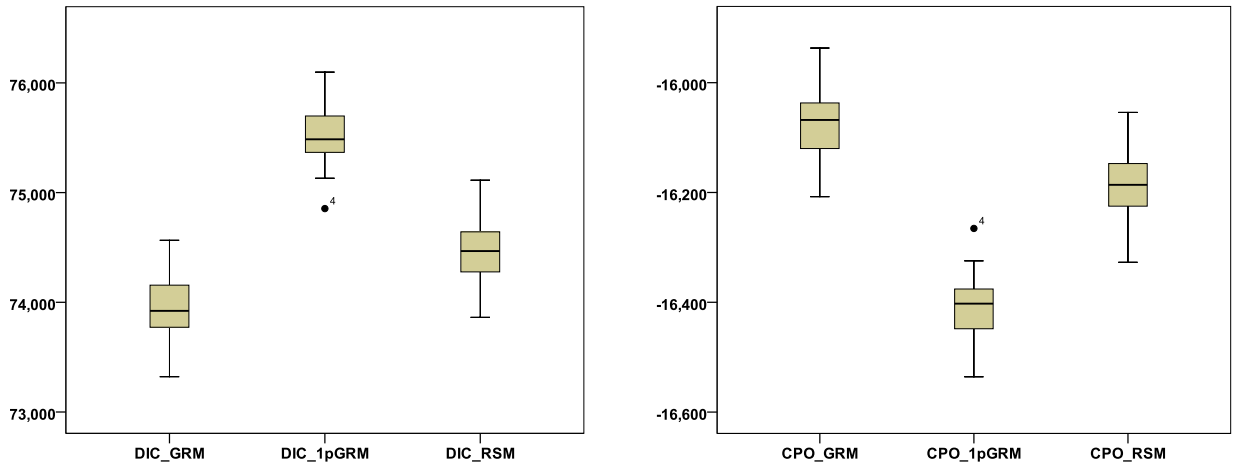


Figure 4.28 Box-plots of DIC and Test-Level CPO across 20 Replications – Condition 1

The previous comparison using the DIC and test-level CPO indices focused on the fit of models at the test level. It was used to answer the question "which model best fit the responses to the test?" As is well known, much of the power of IRT is that it models examinee responses at

the item level. Therefore, a model which fits the overall test is not necessarily appropriate for each item. As a result, comparing the models for each item provides additional information about item model-fit.

Table 4.10 Model Selection for Each Item using Item-Level CPO Index – Condition 1

Item		2P GR	1P GR	RS
1	Mean	-1266	-1295	-1269
	Frequency	19	0	1
2	Mean	-1289	-1315	-1292
	Frequency	19	0	1
3	Mean	-1267	-1295	-1269
	Frequency	20	0	0
4	Mean	-1201	-1228	-1205
	Frequency	19	0	1
5	Mean	-1202	-1230	-1207
	Frequency	20	0	0
6	Mean	-1110	-1115	-1116
	Frequency	20	0	0
7	Mean	-1140	-1145	-1146
	Frequency	20	0	0
8	Mean	-1114	-1119	-1120
	Frequency	20	0	0
9	Mean	-982	-986	-993
	Frequency	20	0	0
10	Mean	-980	-985	-990
	Frequency	20	0	0
11	Mean	-956	-990	-965
	Frequency	20	0	0
12	Mean	-984	-1021	-994
	Frequency	20	0	0
13	Mean	-960	-994	-969
	Frequency	20	0	0
14	Mean	-807	-839	-823
	Frequency	20	0	0
15	Mean	-817	-848	-831
	Frequency	20	0	0

Whereas, the DIC index can only be used to compare the models for the overall test, the CPO index can be used to compare the models at the test- and item-levels. Table 4.10 includes the mean CPO index values (across the 20 replications) for each of 15 items based on the three

different models, as well as the frequency the true model (i.e., 2P GR) was chosen as the preferred model for each item. As can been seen, the mean CPO value for the 2P GR model was larger than the value for the RS model for each item, which was in turn larger than the value for the 1P GR model. This indicated the 2P GR model fit the responses to each item better than the other two models. In addition, for 12 out of 15 items, the item-level CPO index indicated that the generating model was the preferred model for each of the 20 replications. For Items 1, 2, and 4, the RS model was chosen as the preferred model for one replication, but the generating model was chosen as the preferred models for the other 19 replications. The results indicated that the true model was selected to be the preferred model for the overall test and also for each item.

Both the DIC and CPO indices are effective model-comparison tools and are generally used to compare the relative fit of different models. It should be noted that these two indices involve a relative comparison. When one or both of the models to be compared are appropriate, the DIC and CPO indices can be used to obtain the preferred model. However, when either models to be compared are not appropriate or do not fit a data, a preferred model can not be chosen based on either the DIC or CPO indices. For example, in Condition 1, if only the 1P GR and RS models were compared using the DIC or CPO indices, the RS model would be preferred over the 1P GR model. However, the RS model is not really appropriate since the true model was the 2P GR model. In this sense, the general model-comparison tools (i.e., DIC and CPO) only consider the relative fit of different models rather than the absolute fit of each model. Compared with these two indices, the PPMC method can be used to evaluate the fit of different models and compare them at the same time.

For Condition 1, four item-level discrepancy measures (i.e., *item score distribution*, *Yen's Q_1 index*, *Stone's item-fit statistic*, and *item-total score correlation*) and three pair-wise measures

(i.e., *Yen's Q_3 index*, *global odds ratio*, and *item covariance residual*) were used with PPMC to compare the three different models – the 2P GR, the 1P GR, and the RS models. Note that the test-level distribution measure in Study 1 was not used here since it was found not effective in most of the conditions in Study 1.

Table 4.11 Number of Items with Extreme PPP-values across 20 Replications (Item-level Measures)

Model	Min	Max	Mean
	Item Score Distribution		
2P GR*	0	0	0
1P GR	0	0	0
RS	12	15	14
	Yen's Q_1		
2P GR*	0	0	0
1P GR	10	10	10
RS	8	9	8
	Stone's Item-Fit Stat		
2P GR*	0	1	0
1P GR	10	11	10
RS	13	14	13
	Item-Test Correlation		
2P GR*	0	0	0
1P GR	11	15	12
RS	11	15	13

Tables 4.11 presents the minimum, maximum, and mean numbers of the 15 items with extreme PPP-values across 20 replications for each item-level measure. For example, for Yen's Q_1 measure, when the analysis model was the true model (2P GR), there were no extreme PPP-values for each replication. However, there were an average 10 out of 15 items with extreme PPP-values when the analysis model was the 1P GR model, and an average 8 items with extreme PPP-values when the analysis model was the RS model.

As can be observed from this table, either the 1P or 2P GR model appeared to fit the data when using the item score distribution discrepancy measure. However, based on the two item-fit measures and the item-test score correlation, more items were identified as misfitting when the

analysis model was the 1P GR or RS model. Therefore, the 2P GR model was clearly the preferred model for the generated data in Condition 1.

For each pair-wise measure, there were 105 PPP-values corresponding to the 105 item pairs in each replication. Tables 4.12 provides the minimum, maximum, and mean number of item pairs with extreme PPP-values across the 20 replications. A large number of extreme PPP-values would indicate model misfit. As can be seen, the 2P GR model had the least number of extreme PPP-values for all three measures, providing evidence that the 2P GR model was preferred. In contrast, the larger number of extreme PPP-values for the other two models indicated misfit of these models to the data.

Table 4.12 Number of Item-pairs with Extreme PPP-values across 20 Replications (Pair-wise Measures)

Model	Min	Max	Mean
	Yen's Q_3		
2P GR*	3	12	7
1P GR	40	55	48
RS	8	17	13
	Global OR		
2P GR*	3	11	6
1P GR	67	87	79
RS	21	40	31
	Item Covariance Residual		
2P GR*	0	1	0
1P GR	78	94	86
RS	38	51	48

All the measures considered in Condition 1, except the item score distribution measure, appeared to be effective in discriminating between these three models. Based on these measures, the PPMC method chose the generating model as the preferred model for each of the 20 replications. Thus, this method had the same performance as the other two indices regarding the frequency of choosing the true model. In addition, when comparing the models and evaluating the fit of each model, the PPMC method can provide more information about the potential misfit of a model.

Table 4.13 Median PPP-values for Each Item-level Measure across 20 Replications

	Item-Level Discrepancy Measures											
	Item Score Dist			Yen's Q_1			Stone's Item-Fit			Item-Test Correlation		
Item	2P GR	1P GR	RS	2P GR	1P GR	RS	2P GR	1P GR	RS	2P GR	1P GR	RS
1	0.50	0.23	0.00	0.49	0.00	0.37	0.46	0.00	0.25	0.45	1.00	0.03
2	0.50	0.23	0.00	0.51	0.00	0.15	0.54	0.00	0.02	0.44	1.00	0.02
3	0.51	0.21	0.00	0.51	0.00	0.33	0.54	0.00	0.17	0.50	1.00	0.05
4	0.50	0.26	0.00	0.53	0.00	0.02	0.51	0.00	0.00	0.54	1.00	0.99
5	0.50	0.25	0.00	0.50	0.00	0.02	0.47	0.00	0.00	0.50	1.00	1.00
6	0.50	0.50	0.00	0.54	0.55	0.06	0.52	0.37	0.00	0.50	0.06	0.01
7	0.51	0.50	0.03	0.55	0.57	0.03	0.61	0.48	0.00	0.47	0.03	0.00
8	0.51	0.49	0.01	0.50	0.53	0.11	0.43	0.34	0.01	0.44	0.09	0.00
9	0.50	0.50	0.21	0.50	0.61	0.01	0.49	0.50	0.00	0.45	0.10	1.00
10	0.50	0.49	0.00	0.49	0.51	0.01	0.52	0.34	0.00	0.47	0.06	0.99
11	0.51	0.38	0.00	0.52	0.00	0.03	0.49	0.00	0.00	0.47	0.00	0.00
12	0.51	0.37	0.00	0.54	0.00	0.11	0.60	0.00	0.00	0.50	0.00	0.00
13	0.50	0.37	0.00	0.51	0.00	0.07	0.52	0.00	0.00	0.45	0.00	0.00
14	0.50	0.41	0.00	0.53	0.00	0.00	0.49	0.00	0.00	0.45	0.00	1.00
15	0.50	0.42	0.00	0.52	0.01	0.00	0.51	0.00	0.00	0.51	0.00	1.00

Table 4.13 includes the median PPP-values for each item-level discrepancy measure across the 20 replications when each of the models was used to estimate the data. As can be seen, when the 2P GR model fit to the data, the median PPP-values of the item-level measures for each item were close to 0.50, indicating good fit of the model. When the 1P GR model was fit to the data, the median PPP-values for the two item-fit measures were extreme (close to 0.00) for Items 1-5, and Items 11-15, but around 0.50 for Items 6-10. The pattern in these PPP-values indicated that the 1P GR model could not fit the responses to Items 1-5 and 11-15, but fit the responses to Items 6-10. By examining the slopes of the 2P GR model and the common slope of the 1P GR model, Items 6-10 had a true slope parameter of 1.7 and the estimated common slope for the 1P

GR was about 1.6. However, the true slopes were 1.0 for Items 1-5 and 2.4 for Items 11-15, which were much different from the common slope estimate 1.6. In addition, from the pattern in the median PPP-values for the item-test correlation, the potential misfit of the 1P GR model can be observed. As shown in Figure 4.29, when the 2P GR model was estimated (top plot), the observed item-test score correlations were well within the 90% posterior predictive intervals. In contrast, when the 1P GR model was estimated (middle plot), the posterior predictive intervals were consistent across all the 15 items, but the observed correlations fell into three clusters: 1) For Items 1-5, the observed correlations were systematically lower than the predictive values; 2) For Items 11-15, the observed values were consistently higher than the predictive values; 3) For Items 6-10, the observed values were within the posterior predictive intervals.

As shown in Table 4.13, all four item-level measures had extreme PPP-values for each item when the RS model was estimated, reflecting misfit of the RS model. For the item-test correlation measure, the PPP-values for Items 4-5, 9-10, and 14-15 were close to 1.00, indicating the observed correlations were systematically larger than the predictive values under the RS model. However, the PPP-values for the remaining items were close to 0.00, indicating that the observed correlations were systematically smaller than the predictive values. These phenomena can be also observed in the bottom plot in Figure 4.29.

Figure 4.30 displays the pie plots for the three pair-wise measures. As can been seen, all the median PPP-values were around 0.50, providing evidence of model fit for the 2P GR model. The existence of the large number of extreme values in the middle and bottom plots indicated model misfit for the 1P GR and RS models.

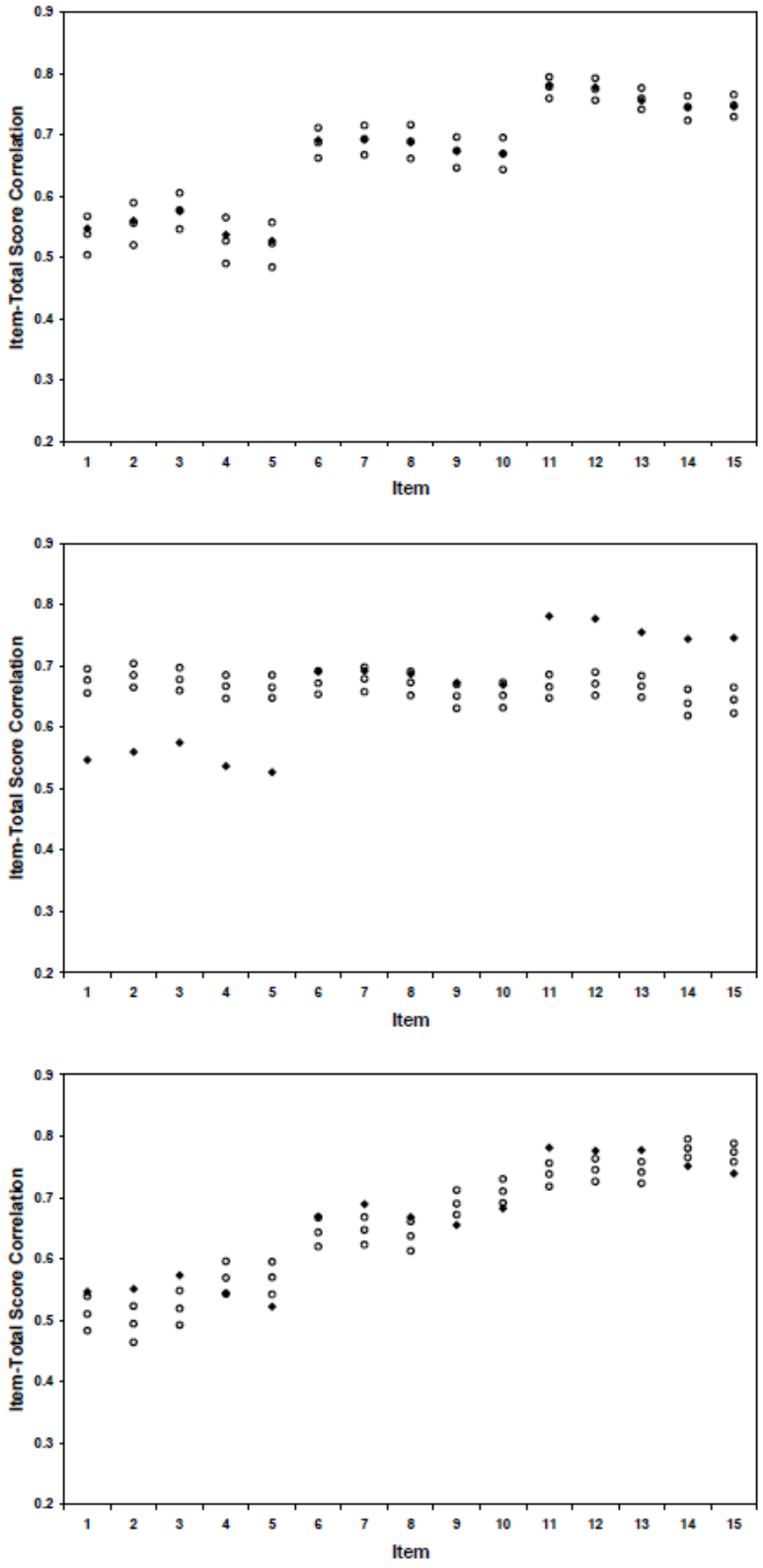


Figure 4.29 Observed vs. 90% Posterior Predictive Interval of Item-Total Score Correlation for 2P GR (top), 1P GR (middle), and RS (bottom) Model

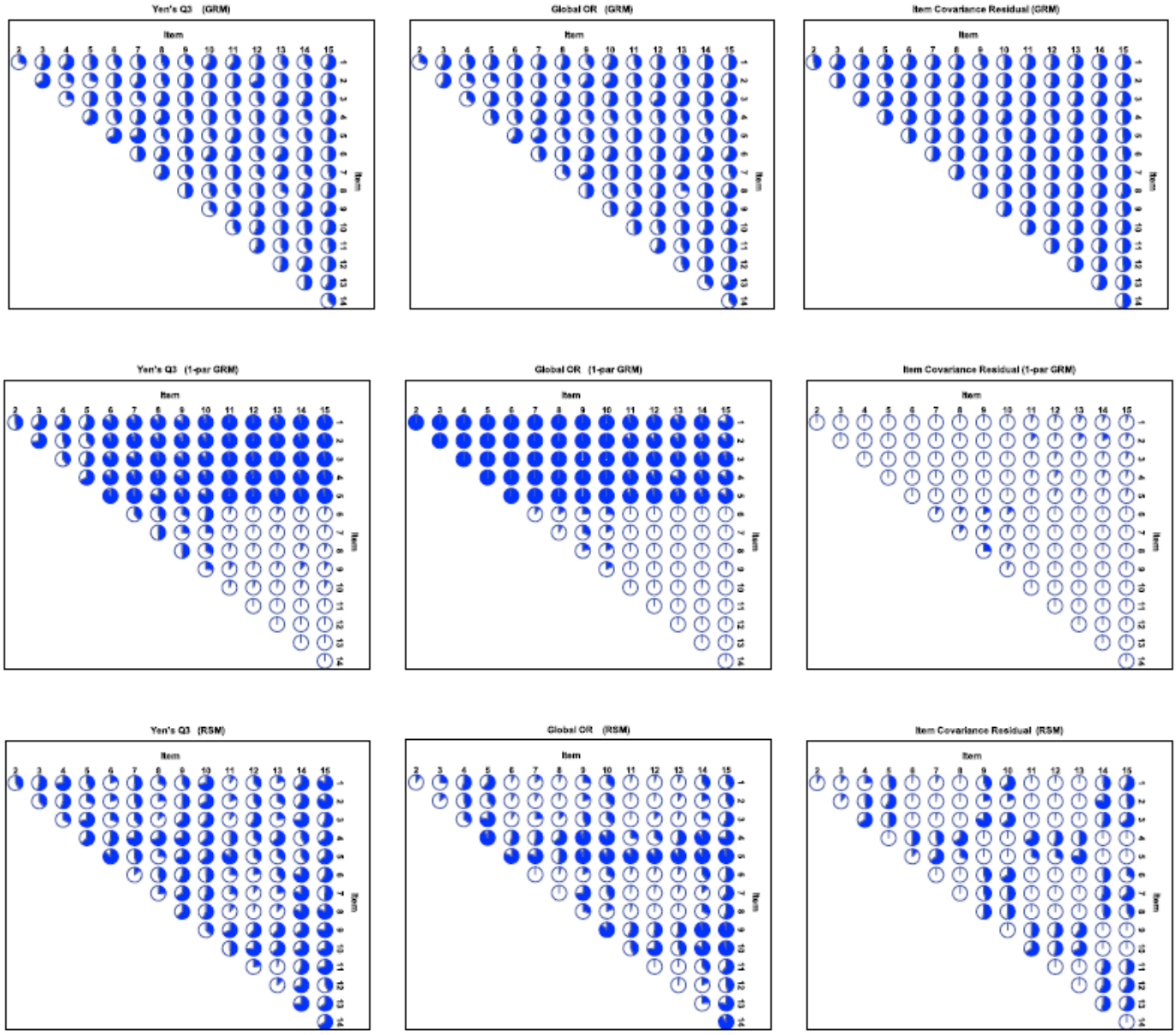


Figure 4.30 Display of Median PPP-values for Pair-wise Measures when fitting 2P GR (top), 1P GR (middle), and RS(bottom) models to the Data

4.2.2 Condition 2 (1-dim GR vs. 2-dim simple-structure GR model)

In Condition 2, the data were generated based on 2-dim simple-structure GR models, but calibrated using both the common 1-dim GR model and the true 2-dim simple-structure GR model. The three model comparison criteria were compared in terms of their abilities to choose the true model as the preferred model.

Table 4.14 RMSD for Item Parameter Recovery in WinBUGS for 2-dim Simple-Structure Model

Item	a1	a2	b1	b2	b3	b4
1	0.06	-	0.14	0.08	0.05	0.07
2	0.08	-	0.08	0.05	0.04	0.07
3	0.09	-	0.04	0.04	0.03	0.08
4	0.06	-	0.17	0.11	0.06	0.08
5	0.06	-	0.05	0.04	0.05	0.11
6	0.07	-	0.06	0.04	0.04	0.04
7	0.06	-	0.09	0.06	0.06	0.11
8	0.05	-	0.04	0.03	0.05	0.06
9	-	0.11	0.16	0.07	0.04	0.03
10	-	0.05	0.07	0.04	0.09	0.20
11	-	0.09	0.09	0.05	0.04	0.07
12	-	0.10	0.05	0.03	0.04	0.06
13	-	0.07	0.10	0.04	0.07	0.14
14	-	0.07	0.15	0.07	0.04	0.05
15	-	0.09	0.05	0.04	0.06	0.15
$RMSD(corr) = 0.016$						

Item parameter recovery for the 2-dim simple-structure GR model was examined first. Table 4.14 gives the RMSD value for each item parameter across the 20 replications. The average RMSD was 0.07 and 0.08 for the first and second slope, respectively, and the average RMSD across all the threshold values was 0.07. The RMSD for the inter-dimensional correlation

was 0.016. These results indicate one chain of 8000 and a posterior sample of 1000 were adequate for estimating the 2-dim simple-structure GR model using MCMC within WinBUGS.

Table 4.15 Model Selection for Overall Test using Different Indices – Condition 2

Model	Min	Max	Mean	Frequency of Choosing True
	DIC			
2-dim GR*	74887	75955	75434	20 (100%)
1-dim GR	78532	79363	78854	
	CPO			
2-dim GR*	-16541	-16312	-16430	20 (100%)
1-dim GR	-17255	-17074	-17143	
	PPMC (global OR)			
2-dim GR*	2	11	7	20 (100%)
1-dim GR	69	85	75	
	PPMC (Yen's Q_3)			
2-dim GR*	0	10	4	20 (100%)
1-dim GR	98	105	102	

Table 4.15 presents the minimum, maximum, and mean values of each index for the two models, and the frequency of choosing the true model (i.e., 2-dim GR) across the 20 replications. As can been seen, the mean DIC values were 75434 and 78854, and the mean CPO values were -16430 and -17143, for the 2-dim and 1-dim GR model respectively. The lower DIC and the higher CPO value for the 2-dim GR model indicated that the 2-dim model fit the data better than the common 1-dim GR model. Recall, for this condition, only two pair-wise discrepancy measures (global OR and Yen's Q_3 index) were used with PPMC. For PPMC, the index was the total number of item pairs having extreme PPP-values. As shown in the table, when the true model was used to analyze the data, on average, only 7 (or 4) out of 105 item pairs with extreme PPP-values for the global OR measure (or Yen's Q_3 index) were observed. However, when the 1-dim GR model was estimated, there were a large number of pairs with extreme PPP-values –

75 and 102 pairs for the global OR and Yen's Q_3 index, respectively. Thus, the PPMC results also indicated that the 2-dim model was preferred over the 1-dim GR model.

As can also be seen in the table, the three indices appeared to perform equally well regarding the frequency of choosing the 2-dim GR model as the preferred model for the overall test. All of the indices selected the true model as the preferred model for each of the 20 replications. It is also worthy to note that there was no overlap between the ranges of each of these three indices for the two models. For example, the range of DIC across the 20 replications was (-16541, -16312) for the 2-dim GR model, and (-17255, -17074) for the 1-dim model. The non-overlapping ranges can also be seen in Figure 4.31.

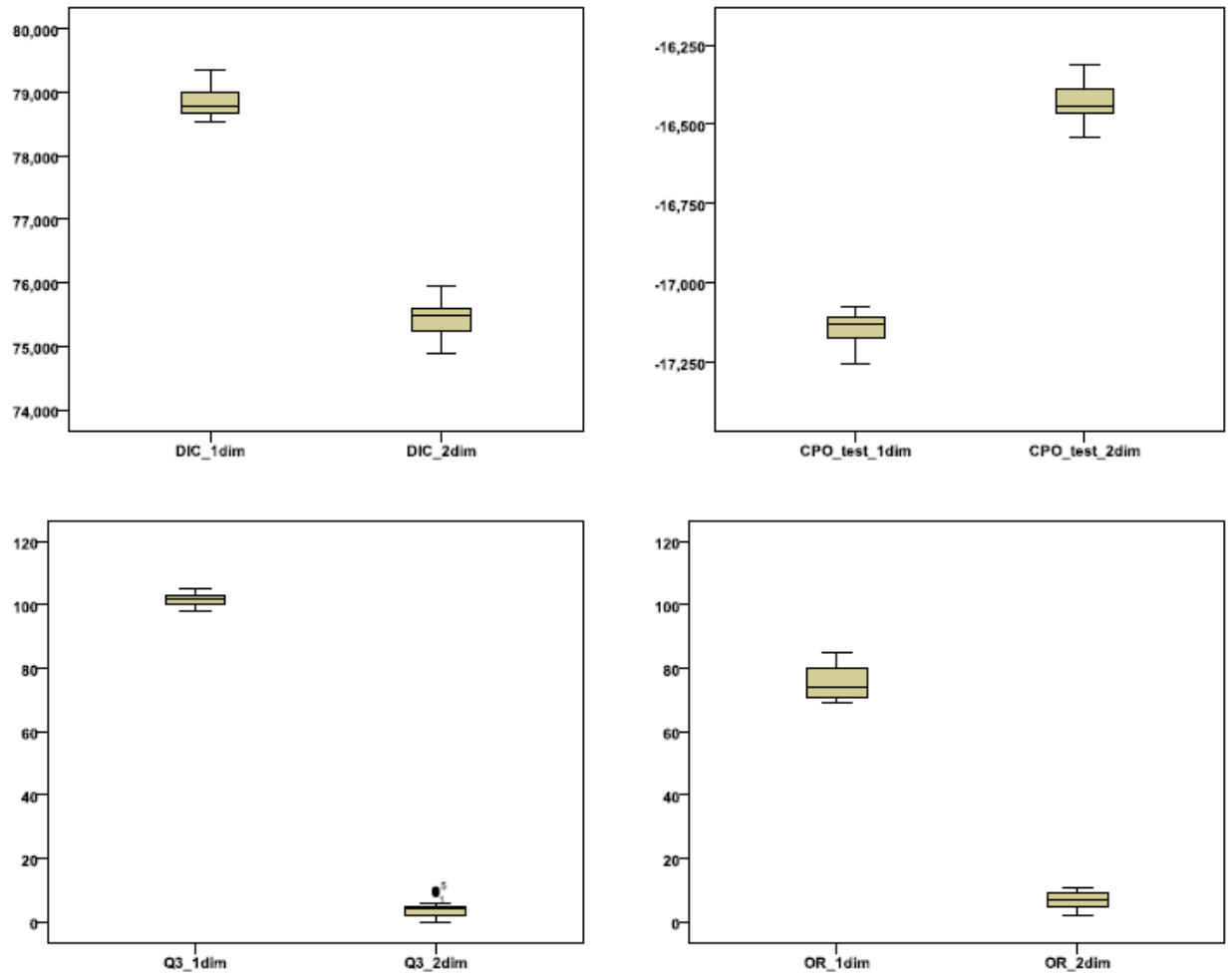


Figure 4.31 Box-plots of Model Comparison Indices across 20 Replications – Condition 2

The distributions of DIC and PPMC values for the 2-dim model were far below the distribution of values for the 1-dim model, suggesting that the 2-dim model fit the data consistently better across the 20 replications. The box-plot for CPO values for the 2-dim model was far above that for the 1-dim model, also indicating that the 2-dim model was preferred.

Table 4.16 includes the minimum, maximum, and mean CPO index values (across the 20 replications) for each of the 15 items based on the two models, as well as the frequency the true model (i.e., 2-dim GR) was chosen to be the preferred model for each item. As can been seen, the mean CPO value for the 2-dim GR model was larger than the value for the 1-dim model for each item, indicating that the 2-dim GR model fit the responses to each item better. Moreover, for all items, the item-level CPO index chose the true model as the preferred model over the 20 replications.

Figure 4.32 displays the median PPP-values for two pair-wise discrepancy measures when estimating the two different models. When a 1-dim GR model was estimated, all the PPP-values were extreme and the items fell into two clusters – Items 1- 8 in one, and Items 9-15 in another. This pattern indicated that a 2-dimensional model should be considered. In contrast, when a 2-dim model was estimated, all the PPP-values were around 0.5, suggesting the fit of the 2-dim model.

Table 4.16 Model Selection for Each Item using Item-level CPO Index – Condition 2

Item	Model	Min	Max	Mean	Frequency of Choosing True
1	2-dim GR*	-1299	-1259	-1274	20 (100%)
	1-dim GR	-1312	-1279	-1291	
2	2-dim GR*	-1182	-1136	-1162	20 (100%)
	1-dim GR	-1227	-1186	-1209	
3	2-dim GR*	-1025	-985	-998	20 (100%)
	1-dim GR	-1093	-1054	-1077	
4	2-dim GR*	-1236	-1192	-1214	20 (100%)
	1-dim GR	-1254	-1206	-1231	
5	2-dim GR*	-1019	-972	-1001	20 (100%)
	1-dim GR	-1069	-1020	-1044	
6	2-dim GR*	-1023	-983	-999	20 (100%)
	1-dim GR	-1111	-1060	-1075	
7	2-dim GR*	-1321	-1288	-1301	20 (100%)
	1-dim GR	-1332	-1306	-1319	
8	2-dim GR*	-1151	-1119	-1133	20 (100%)
	1-dim GR	-1203	-1155	-1180	
9	2-dim GR*	-866	-816	-848	20 (100%)
	1-dim GR	-954	-883	-923	
10	2-dim GR*	-1228	-1193	-1209	20 (100%)
	1-dim GR	-1243	-1213	-1227	
11	2-dim GR*	-1159	-1114	-1133	20 (100%)
	1-dim GR	-1212	-1158	-1184	
12	2-dim GR*	-1049	-1016	-1032	20 (100%)
	1-dim GR	-1130	-1091	-1114	
13	2-dim GR*	-1301	-1243	-1277	20 (100%)
	1-dim GR	-1315	-1267	-1297	
14	2-dim GR*	-1023	-973	-1002	20 (100%)
	1-dim GR	-1070	-1014	-1049	
15	2-dim GR*	-871	-818	-845	20 (100%)
	1-dim GR	-955	-897	-923	

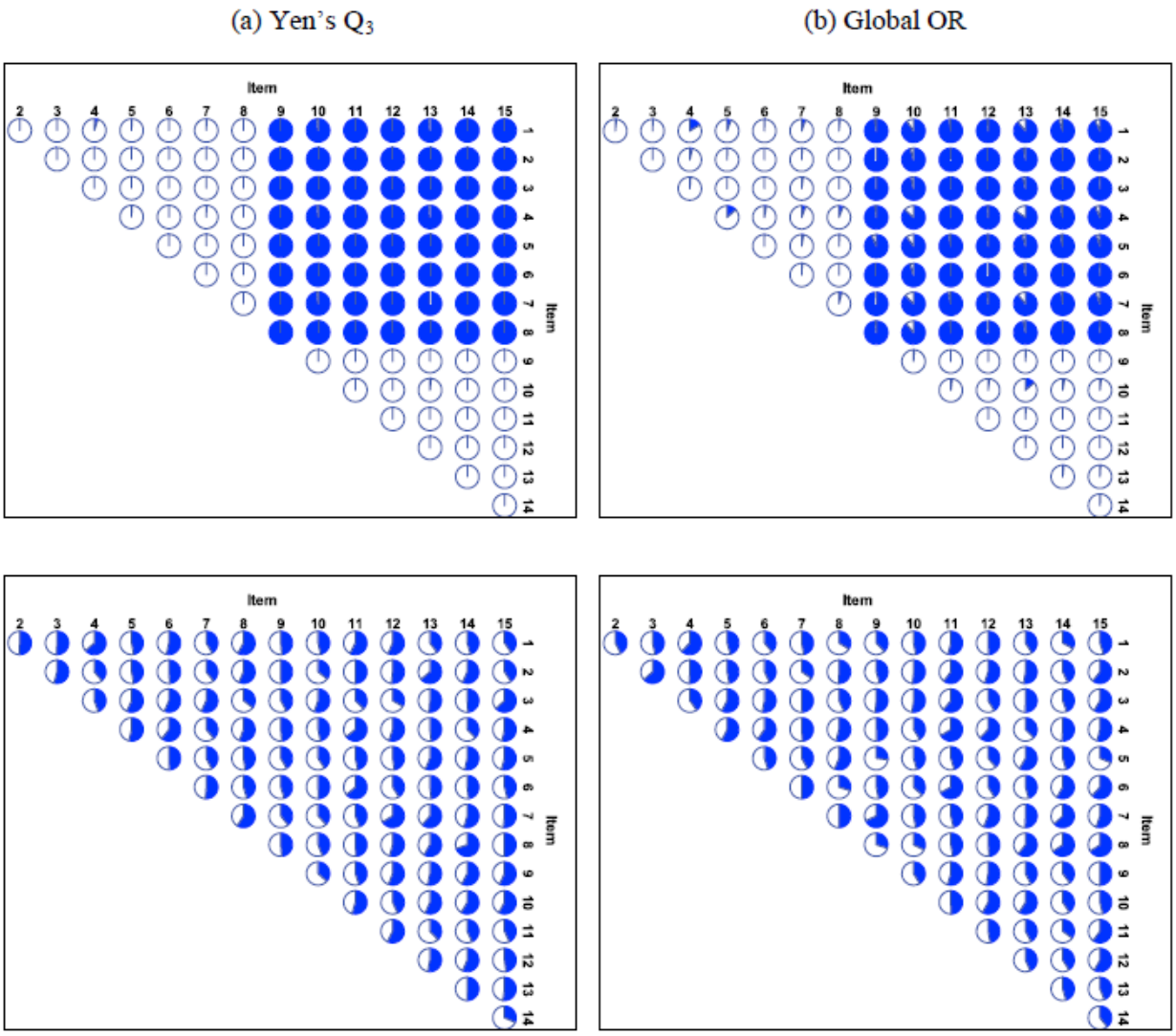


Figure 4.32 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when Fitting 1-dim GR model (top) and 2-dim simple-structure GR model (bottom) to the Data

4.2.3 Condition 3 (1-dim GR vs. 2-dim complex-structure GR model)

In this condition, the data were generated based on 2-dim complex-structure GR models, but calibrated using both the common 1-dim GR model and the generating 2-dim complex-structure GR model. The three model comparison criteria were compared in terms of their abilities to select 2-dim model as the preferred model.

Table 4.17 RMSD for Item Parameter Recovery in WinBUGS for 2-dim Complex-Structure Model

Item	a1	a2	b1	b2	b3	b4
1	0.18	0.08	0.16	0.09	0.04	0.10
2	0.15	0.10	0.13	0.04	0.06	0.15
3	0.14	0.12	0.08	0.04	0.09	0.15
4	0.15	0.10	0.30	0.15	0.06	0.10
5	0.16	0.10	0.09	0.04	0.13	0.26
6	0.07	-	0.07	0.05	0.04	0.04
7	0.06	-	0.08	0.04	0.05	0.07
8	0.06	-	0.06	0.03	0.05	0.07
9	0.06	-	0.17	0.05	0.03	0.05
10	0.07	-	0.04	0.04	0.06	0.11
11	0.07	-	0.04	0.03	0.03	0.05
12	0.08	-	0.05	0.03	0.03	0.05
13	0.08	-	0.05	0.03	0.04	0.06
14	0.10	-	0.12	0.05	0.03	0.03
15	0.09	-	0.05	0.03	0.05	0.12

Item parameter recovery for the 2-dim complex-structure GR model was examined first. Table 4.17 gives the RMSD value for each item parameter across the 20 replications. The average RMSD across all the threshold values was 0.074. For the slope parameter a_1, the average RMSD was 0.075 across the items (6-15) measuring only the dominant dimension, and 0.157 across the items (1-5) measuring the dominant AND the nuisance dimension. The average RMSD for the slope parameter a_2 was 0.099. The relatively larger values of RMSD for the two

slopes for the first five items were due to fixing the correlation to be 0 when estimating the model in WinBUGS (the true correlation was 0.30). However, this rotation of the two dimensions would not affect the computation of the model-comparison indices.

Table 4.18 Model Selection for Overall Test using Different Indices – Condition 3

Model	Min	Max	Mean	Frequency of Choosing True
	DIC			
2-dim GR*	71391	72365	71905	20 (100%)
1-dim GR	71563	72580	72093	
	CPO			
2-dim GR*	-15746	-15534	-15645	20 (100%)
1-dim GR	-15776	-15556	-15670	
	PPMC (global OR)			
2-dim GR*	1	8	4	18 (90%)
1-dim GR	2	14	8	
	PPMC (Yen's Q_3)			
2-dim GR*	1	8	4	20 (100%)
1-dim GR	9	21	15	

Table 4.18 presents the minimum, maximum, and mean values for each index for the two models, as well as the frequency of choosing the true model (i.e., 2-dim complex-structure GR) across the 20 replications. As can been seen, the mean DIC values were 71905 and 72093, and the mean CPO values were -15645 and -15670 for the 2-dim complex-structure and 1-dim GR model, respectively. The lower DIC value and the higher CPO value for the 2-dim GR model indicated that this complex model was preferred over the simple unidimensional GR model. For the PPMC application, when the true model was estimated, 4 out of 105 item pairs with extreme PPP-values for both pair-wise measures were observed. However, when the 1-dim GR model was estimated, more item pairs had extreme PPP-values – 8 and 15 pairs for the global OR and Yen's Q_3 index respectively. The distributions of these indices are shown in Figure 4.33.

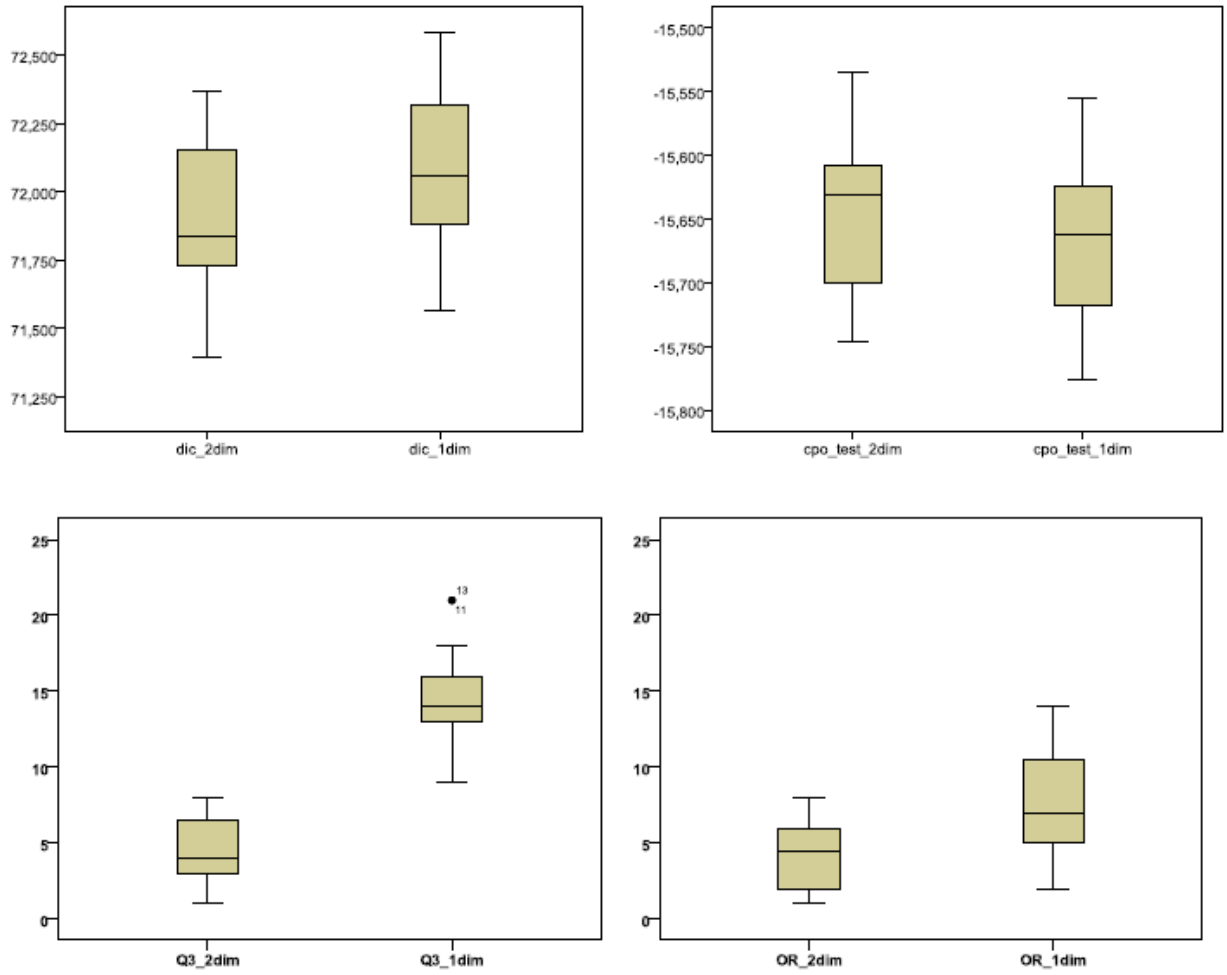


Figure 4.33 Box-plots of Model Comparison Indices across 20 Replications – Condition 3

As shown in Table 4.18, the DIC, CPO and PPMC using Yen's Q_3 measures appeared to perform equally well regarding the frequency of choosing the 2-dim GR model as the preferred model for the overall test. However, when the global OR measure was used with PPMC, for 2 replications, the 1-dim GR model was wrongly chosen as the preferred model. The PPMC results indicated that the choice of discrepancy measures would affect the performance of the PPMC application in comparing different models. If the measure was not effective, the PPMC method would lose power and would not be effective as the typical model-comparison indices (DIC and CPO). For this condition, Yen's Q_3 measure appeared to be more effective than the global OR measure.

Table 4.19 Model Selection of Each Item using Item-level CPO Index – Condition 3

Item	Model	Min	Max	Mean	Frequency of Choosing True
1	2-dim GR*	-1226	-1183	-1201	19 (95%)
	1-dim GR	-1230	-1188	-1205	
2	2-dim GR*	-1260	-1222	-1238	19 (95%)
	1-dim GR	-1264	-1226	-1242	
3	2-dim GR*	-1229	-1191	-1209	19 (95%)
	1-dim GR	-1234	-1191	-1212	
4	2-dim GR*	-1092	-1056	-1076	20 (100%)
	1-dim GR	-1097	-1059	-1081	
5	2-dim GR*	-1092	-1042	-1071	20 (100%)
	1-dim GR	-1095	-1050	-1076	
6	2-dim GR*	-1137	-1088	-1114	14 (70%)
	1-dim GR	-1137	-1089	-1115	
7	2-dim GR*	-1160	-1118	-1140	10 (50%)
	1-dim GR	-1160	-1117	-1139	
8	2-dim GR*	-1131	-1086	-1113	13 (65%)
	1-dim GR	-1132	-1087	-1114	
9	2-dim GR*	-1007	-964	-983	17 (85%)
	1-dim GR	-1007	-964	-984	
10	2-dim GR*	-1004	-943	-979	15 (75%)
	1-dim GR	-1004	-944	-978	
11	2-dim GR*	-986	-938	-961	14 (70%)
	1-dim GR	-986	-930	-959	
12	2-dim GR*	-1011	-965	-988	14 (70%)
	1-dim GR	-1013	-965	-989	
13	2-dim GR*	-984	-938	-958	14 (70%)
	1-dim GR	-985	-938	-959	
14	2-dim GR*	-834	-780	-807	18 (90%)
	1-dim GR	-834	-781	-808	
15	2-dim GR*	-831	-784	-806	12 (60%)
	1-dim GR	-832	-785	-808	

As discussed above, the 2-dim complex-structure GR model fit better for the overall test. Table 4.19 includes the minimum, maximum, and mean CPO index values, as well as the

frequency the true model was chosen as the preferred model for each item. As can been seen, for Items 1-5, which measured both the dominant and nuisance dimensions, the item-level CPO selected the 2-dim model as the preferred model 95% to 100% of the time. However, for the other items (Items 6-15), which only measured the dominant dimension, the 2-dim model was chosen as the preferred model with a lower percentage (50% to 90%). This would be expected since the 1-dim GR model should be appropriate for those items simulated to measure one dimension. In addition, for Items 1-5, the mean CPO value for the 2-dim GR model was larger than the value for the 1-dim model, and the difference between the two mean CPO values was greater than 3 units. For Items 6-15, though most of the items had larger mean CPO values for the 2-dim GR model, the difference between two models was only about 1 unit. It is should be noted that this small difference might not provide sufficient evidence for favoring the 2-dim GR model over the 1-dim GR model.

Recall, the smaller value of DIC, the better the fit of a model. However, any difference in DIC less than 5 units for two models may not indicate sufficient evidence in favor of one model over another (Spiegelhalter et al., 2003). There are no discussed guidelines for CPO as for DIC, but the item-level CPO results for this condition may indicate that a difference of less than 3 units may not provide sufficient evidence supporting one model over another. However, the amount of difference in CPO necessary to suggest a significant difference between models needs further investigation.

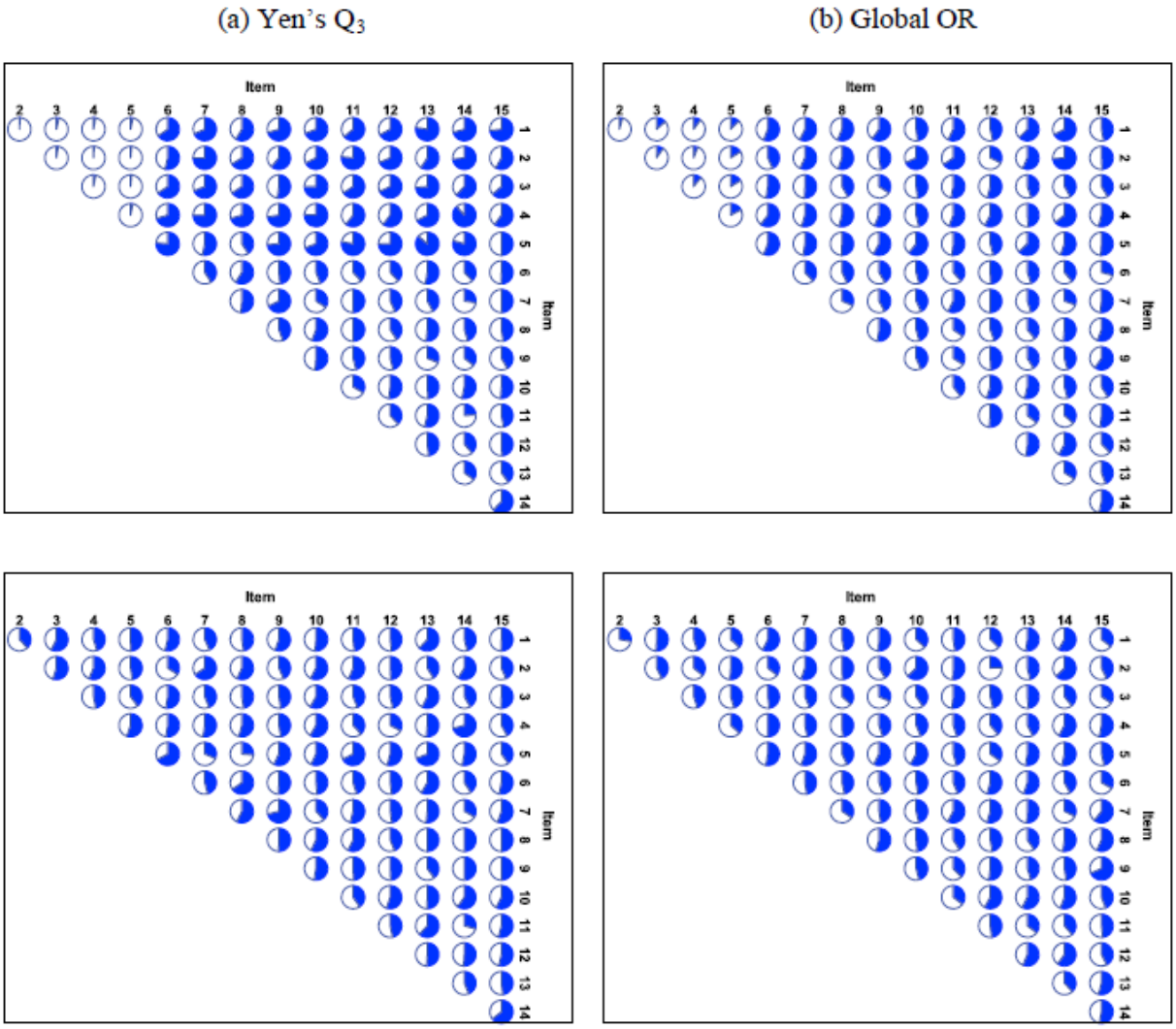


Figure 4.34 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when Fitting 1-dim GR Model (top) and 2-dim complex-structure GR Model (bottom) to the Data

Figure 4.34 displays the median PPP-values for the two pair-wise discrepancy measures when both models were estimated. As can be observed, when the 2-dim complex-structure model was estimated (bottom plots), all the PPP-values were around 0.5, providing evidence of fit for the model. In contrast, when the unidimensional GR model was estimated, all the PPP-values were extreme for the item pairs involving the first 5 items, but around 0.5 for the other item pairs. This pattern indicated that the unidimensional GR model was not appropriate for Items 1-5, but was appropriate for Items 6-15. Additionally, the close to 0 PPP-values for the item pairs

among Items 1-5 indicated that the realized correlations among these five items were consistently larger than the predicted correlations under the unidimensional GR model. This also suggested that another factor may be measured by these 5 items in addition to the dominant dimension.

In summary, all three indices showed that a 2-dim complex-structure GR model fit the overall test better than a unidimensional GR model. The item-level CPO index further showed that this complex model was needed to model the responses to the first 5 items, but a simple unidimensional GR model might be adequate for the other items. In addition, the PPMC results showed the misfit of a unidimensional GR model to the responses to the first 5 items as well as the fit of this simple model to the other items.

4.2.4 Condition 4 (1-dim GR model vs. GR model for testlet)

In this condition, Items 6, 7 and 8 were designed as a testlet, and the responses to these testlet items were generated under a modified GR model for testlets. The responses to other items were simulated to be locally independent based on the unidimensional (1-dim) GR model. For each of the 20 generated data sets, both the 1-dim GR model and the testlet GR model were estimated to the same data in WinBUGS, and three Bayesian model comparison indices were obtained for each model. The values for different models were then compared in order to determine which model was preferred.

Table 4.20 RMSD for Item Parameter Recovery in WinBUGS for Testlet GR Model

Item	a	b1	b2	b3	b4
1	0.05	0.13	0.06	0.04	0.07
2	0.04	0.11	0.07	0.04	0.07
3	0.04	0.07	0.05	0.07	0.10
4	0.05	0.18	0.08	0.05	0.06
5	0.05	0.07	0.05	0.07	0.15
6	0.08	0.09	0.06	0.04	0.04
7	0.10	0.09	0.06	0.04	0.08
8	0.10	0.05	0.04	0.07	0.10
9	0.06	0.12	0.05	0.04	0.05
10	0.05	0.05	0.03	0.06	0.13
11	0.11	0.09	0.04	0.02	0.05
12	0.07	0.06	0.04	0.03	0.05
13	0.09	0.05	0.03	0.05	0.06
14	0.11	0.15	0.05	0.04	0.05
15	0.08	0.03	0.03	0.06	0.15
			$RMSD(\sigma^2) = 0.037$		

Item parameter recovery for the testlet GR model was examined first. Table 4.20 gives the RMSD value for each item parameter across the 20 replications. The average RMSD was 0.073 and 0.067 for the slope and threshold parameters, respectively. The RMSD for the testlet

variance across the 20 replications was 0.037. These results indicated one chain of 5000 and a posterior sample of 1000 were adequate for the accuracy of estimation of the GR model for testlet using MCMC within WinBUGS.

Table 4.21 Model Selection for Overall Test using Different Indices – Condition 4

Model	Min	Max	Mean	Frequency of Choosing True
	DIC			
testlet GR*	73750	75177	74170	20 (100%)
1-dim GR	74476	75833	74924	
	CPO			
testlet GR*	-16361	-16056	-16145	20 (100%)
1-dim GR	-16482	-16191	-16287	
	PPMC (global OR)			
testlet GR*	2	12	6	18 (90%)
1-dim GR	5	17	10	
	PPMC (Yen's Q_3)			
testlet GR*	3	9	5	20 (100%)
1-dim GR	17	28	21	

Table 4.21 presents the minimum, maximum, and mean values for each index for the two models, as well as the frequency of choosing the true model (i.e., testlet GR) across the 20 replications. As can been seen, the mean DIC values were 74170 and 74924, and the mean CPO values were -16145 and -16287 for the testlet GR model and 1-dim GR model, respectively. The lower DIC and the higher CPO value for the testlet GR model indicated that this complex model fit the overall test better than the simple unidimensional GR model. For the PPMC application, when the testlet model was estimated, 5 (6) out of 105 item pairs with extreme PPP-values for Yen's Q_3 (global OR) were observed. However, when the unidimensional GR model was estimated, more item pairs had extreme PPP-values – 10 and 21 pairs for the global OR and Yen's Q_3 index, respectively. The distributions of these indices are shown in Figure 4.35.

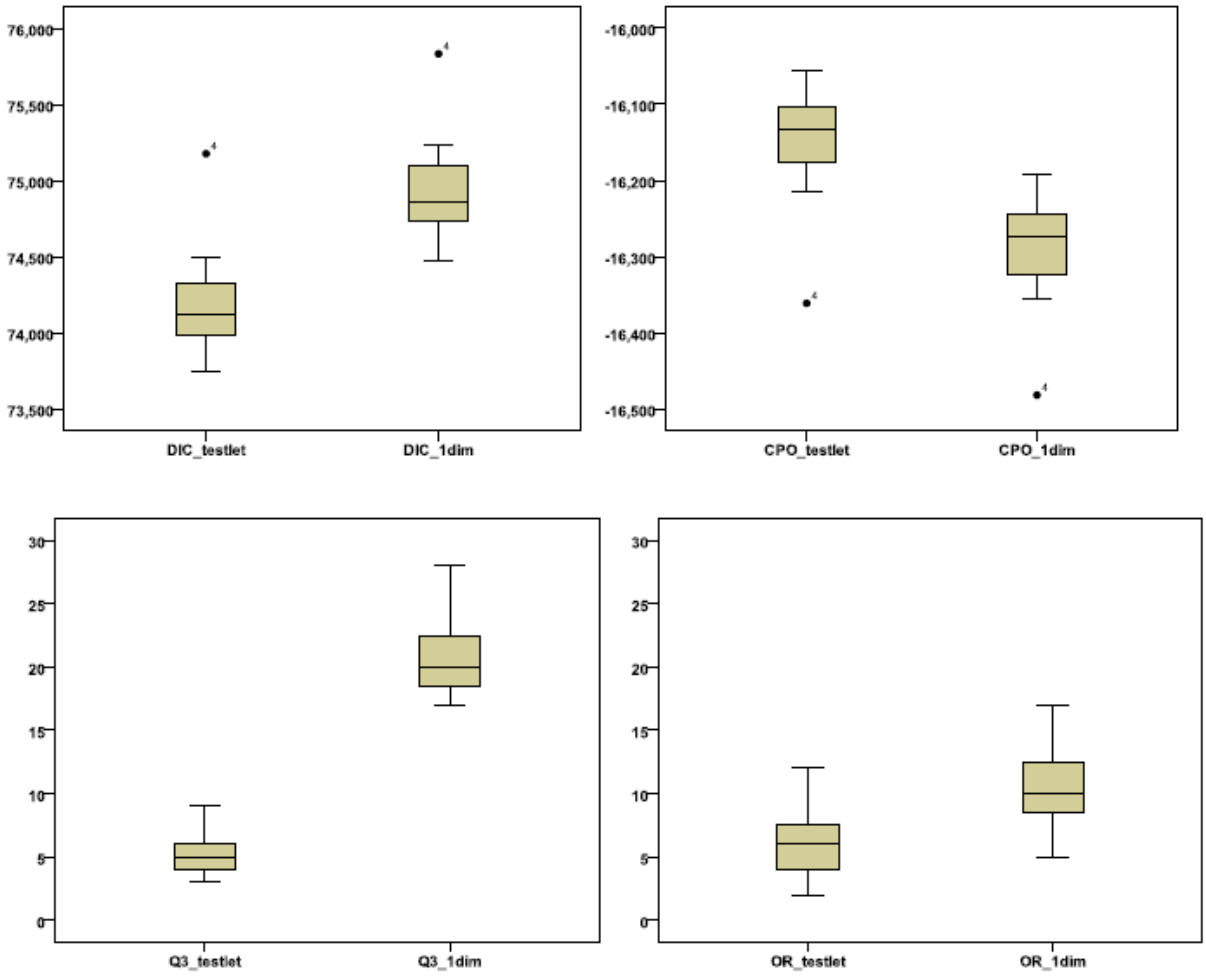


Figure 4.35 Box-plots of Model Comparison Indices across 20 Replications – Condition 4

As shown in Table 4.21, the DIC, CPO and PPMC using Yen's Q_3 measures appeared to perform equally well. All approaches resulted in selecting the testlet GR model as the preferred model 100% of the time. However, when the global OR measure was used with PPMC, the testlet GR model was chosen as the preferred model 90% of the time. As for Condition 3, Yen's Q_3 measure appeared to be slightly more effective than the global OR measure for this condition.

Table 4.22 Model Selection for Each Item using Item-level CPO Index – Condition 4

Item	Model	Min	Max	Mean	Frequency of Choosing True
1	testlet GR*	-1284	-1247	-1269	15 (75%)
	1-dim GR	-1284	-1247	-1269	
2	testlet GR*	-1307	-1275	-1293	15 (75%)
	1-dim GR	-1307	-1276	-1294	
3	testlet GR*	-1287	-1250	-1266	16 (80%)
	1-dim GR	-1288	-1250	-1267	
4	testlet GR*	-1214	-1178	-1199	14 (70%)
	1-dim GR	-1215	-1178	-1120	
5	testlet GR*	-1221	-1185	-1204	16 (80%)
	1-dim GR	-1221	-1186	-1204	
6	**testlet GR***	**-1137**	**-1094**	**-1121**	**20 (100%)**
	1-dim GR	**-1180**	**-1133**	**-1162**	
7	**testlet GR***	**-1185**	**-1130**	**-1151**	**20 (100%)**
	1-dim GR	**-1220**	**-1166**	**-1193**	
8	**testlet GR***	**-1150**	**-1099**	**-1124**	**20 (100%)**
	1-dim GR	**-1190**	**-1140**	**-1166**	
9	testlet GR*	-1004	-961	-986	18 (90%)
	1-dim GR	-1007	-962	-987	
10	testlet GR*	-1009	-964	-986	16 (80%)
	1-dim GR	-1009	-965	-987	
11	testlet GR*	-999	-939	-968	18 (90%)
	1-dim GR	-1001	-944	-970	
12	testlet GR*	-1008	-969	-991	16 (80%)
	1-dim GR	-1011	-969	-993	
13	testlet GR*	-989	-938	-962	19 (95%)
	1-dim GR	-992	-941	-965	
14	testlet GR*	-832	-764	-811	20 (100%)
	1-dim GR	-834	-765	-814	
15	testlet GR*	-845	-788	-814	20 (100%)
	1-dim GR	-851	-790	-816	

Table 4.22 includes the item-level CPO index information for each item. As can be seen, for the items in the testlet (Items 6, 7 and 8), the mean CPO values for the testlet GR model were much larger than the values for the unidimensional model. The difference was about 42 units for these three items, and the testlet GR model was chosen as the preferred model 100% of the time. For the other independent items, the mean CPO values were about the same for most of these items, and the maximum CPO difference between two models was less than 3 units. Though the testlet GR model was selected as the preferred model for these independent items 70% to 100% of the time, the difference of less than 3 units did not provide sufficient evidence in favor of a testlet GR model over a unidimensional model. As a result, it may be reasonable to apply the simple unidimensional GR model to these items.

Figure 4.36 displays the median PPP-values for the two pair-wise discrepancy measures when both models were estimated. As can be observed, when the testlet GR model was estimated (bottom plots), all the PPP-values were around 0.5, suggesting the fit of the model. In contrast, when the unidimensional GR model was estimated, all the PPP-values were extreme for the item pairs with the three testlet items (Items 6, 7, and 8), but around 0.5 for the pairs among the independent items. Additionally, the close to 0 PPP-values for the item pairs for the testlet items indicated that the realized correlations among these items were consistently larger than the predicted correlations under the unidimensional GR model. These results indicated that the unidimensional GR model was not appropriate for Items 6, 7, and 8, but was appropriate for the other items.

In summary, all three indices indicated that a testlet GR model fit the overall test better than a unidimensional GR model when item responses with a testlet were simulated. The item-level CPO index further showed that a testlet GR model fit Items 6, 7 and 8 significantly better

than a unidimensional GR model, but this testlet model might be not necessary for the other items. Moreover, the PPMC results indicated that the misfit of a unidimensional GR model to the testlet items was due to the higher than expected correlations among the testlet items. The PPMC results also indicated a good fit of the testlet GR model to all items.

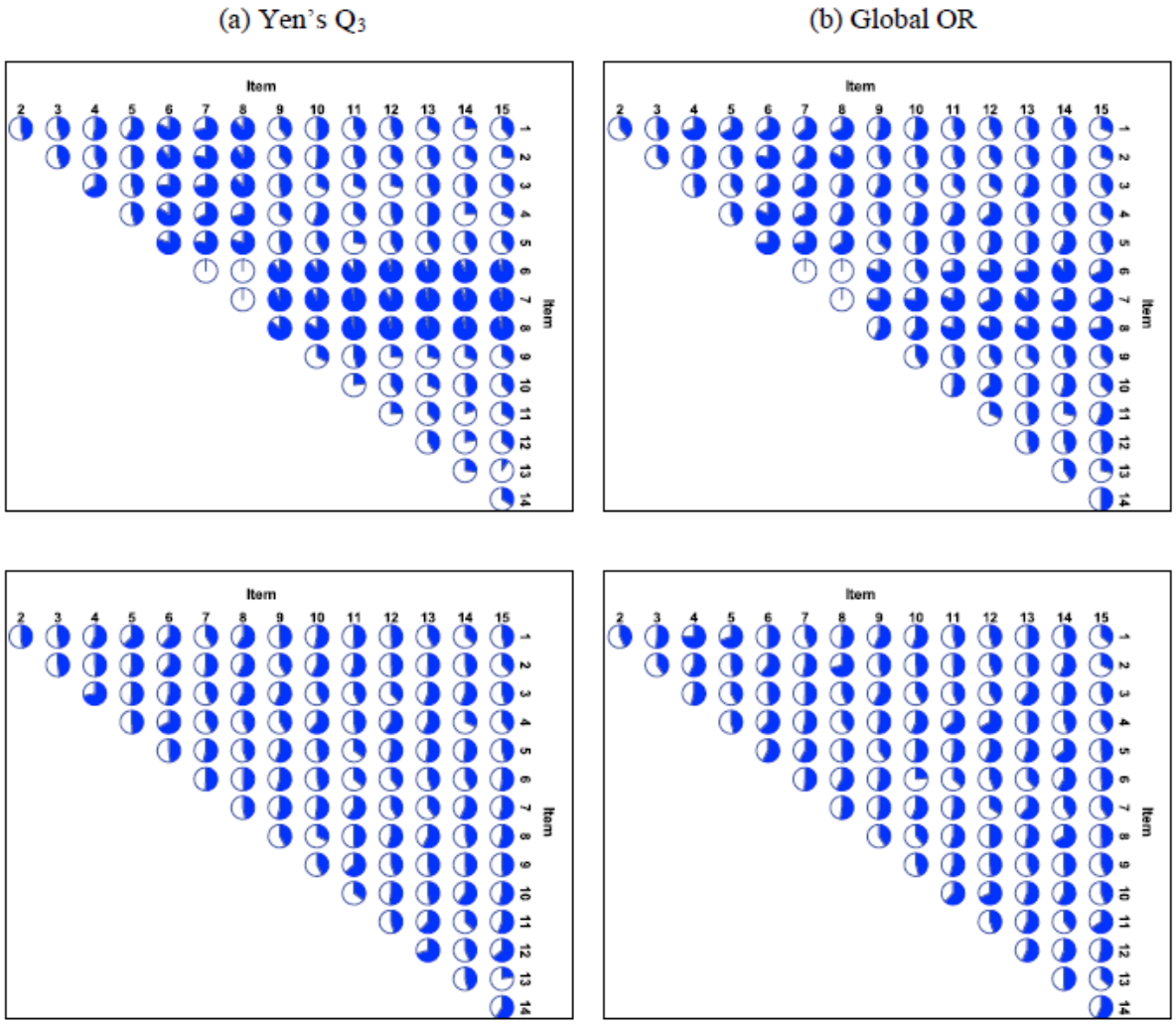


Figure 4.36 Display of Median PPP-values for Yen's Q_3 (left) and Global OR (right) when fitting 1-dim GR Model (top) and testlet GR Model (bottom) to the Data

4.3 RESULTS FROM REAL APPLICATION

This section presents the results from the application of the Bayesian model-fit and model-comparison methodology investigated in the current study to three QCAI data sets (AS91, AS92, and BS92). Each dataset was calibrated using both a 2P GR (hereafter simply referred to as GR) model and a 1P GR model in WinBUGS, and different aspects of fit of each model were evaluated by using the PPMC method. In addition, the model-comparison indices (DIC, CPO, and PPMC) were computed for both models and a preferred model was chosen for each dataset. It should be noted that all 8 discrepancy measures were used with the PPMC application in order to assess different aspects of fit.

4.3.1 QCAI Data 1 – AS91

As for the previous simulation studies, the estimation of item parameter for GR models using MCMC in WinBUGS was evaluated first. Since there were no true values for real data, the item parameters were also estimated using MULTILOG. Comparing the results from both programs provided information about the consistency of item parameter estimates.

Table 4.23 provides the item parameter estimates for the GR model based on the AS91 data. As can be seen, the estimates from the two programs were very similar. The average absolute difference between WinBUGS and MULTILOG estimates across all the items was 0.051 for the slope parameters, and 0.052 for all the threshold parameters. It should be noted that the estimates in MULTILOG were slightly different from the values in Hansen (2004). Though Hansen (2004) estimated the same model based on the same data in MULTILOG, she used all the available responses including the missing responses. The estimates in Table 4.23 were based

on the data excluding the missing responses. The reason is that WinBUGS can not handle missing values. The same issue existed for the other two datasets.

Table 4.23 Item Parameter Estimates using WinBUGS and Multilog – AS91

	WinBUGS					Multilog				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	0.87	1.30	2.09	2.43	2.97	0.87	1.31	2.08	2.40	2.91
2	1.57	-1.04	-0.08	0.73	1.53	1.63	-0.98	-0.06	0.71	1.48
3	1.27	0.04	1.25	1.50	1.80	1.32	0.06	1.23	1.46	1.73
4	1.22	0.18	0.93	1.39	2.17	1.30	0.19	0.90	1.31	2.04
5	1.00	-1.22	-0.10	1.93	3.47	1.03	-1.15	-0.08	1.89	3.36
6	1.20	-0.70	0.19	0.86	3.56	1.23	-0.66	0.20	0.84	3.46
7	1.16	-1.82	0.66	1.44	2.37	1.22	-1.72	0.65	1.38	2.25
8	1.56	0.93	1.57	1.85	2.21	1.66	0.90	1.51	1.76	2.09

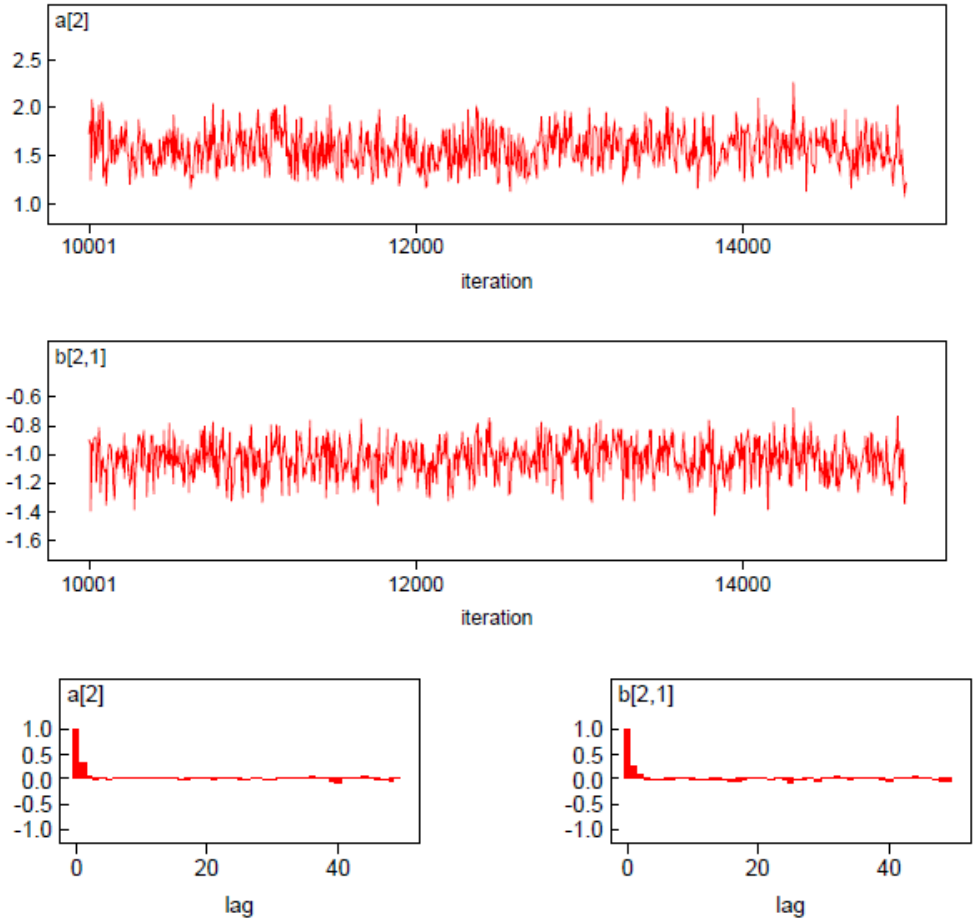


Figure 4.37 Example History and Autocorrelation Plots – AS91

For each real dataset, a long chain of 15000 iterations was run in WinBUGS and the first 10000 iterations were discarded as the burn-in phase. The remaining 5000 iterations were thinned every fifth iteration to obtain a total posterior sample of 1000. Figure 4.37 displays the sample history and autocorrelation plots for the slope and one threshold parameters for Item 2. Other parameters had similar plots. These plots demonstrated that the convergence in the chain.

Model-Fit

The PPP-values of the chi-square statistic summarizing the discrepancy between the observed and predictive test score distributions were 0.56 and 0.64 for the GR and 1P GR models, respectively. Both values were not extreme, indicating these two models fit the data in terms of the total test score distributions.

Table 4.24 PPP-values for Item-level Measures based on GR and 1P GR Models – AS91

	Item-Level Discrepancy Measures							
	Item Score Dist		Yen's Q_1		Stone's Item-Fit		Item-Test Corr	
Item	GR	1P GR	GR	1P GR	GR	1P GR	GR	1P GR
1	0.50	0.47	0.25	0.09	0.08	0.01	0.47	0.95
2	0.53	0.51	0.51	0.57	0.57	0.43	0.49	0.04
3	0.50	0.48	0.40	0.43	0.11	0.15	0.38	0.29
4	0.52	0.51	0.49	0.48	0.48	0.52	0.46	0.51
5	0.47	0.48	0.40	0.29	0.11	0.05	0.31	0.77
6	0.53	0.52	0.51	0.50	0.56	0.55	0.37	0.38
7	0.51	0.51	0.49	0.48	0.36	0.44	0.27	0.33
8	0.46	0.47	0.36	0.42	0.05	0.08	0.25	0.02

Table 4.24 includes the PPP-values for the item-level discrepancy measures for each of the 8 QCAI items. As can be seen, when a GR model was used to analyze the AS91 data, the PPP-values of the item score distribution, item-test score correlation, and Yen's Q_1 index had no extreme values. However, Stone's fit statistic showed extreme values for a few of items, indicating some misfitting items. Recall, in previous studies (see Table 3.18), Items 1, 3, 5, and 8

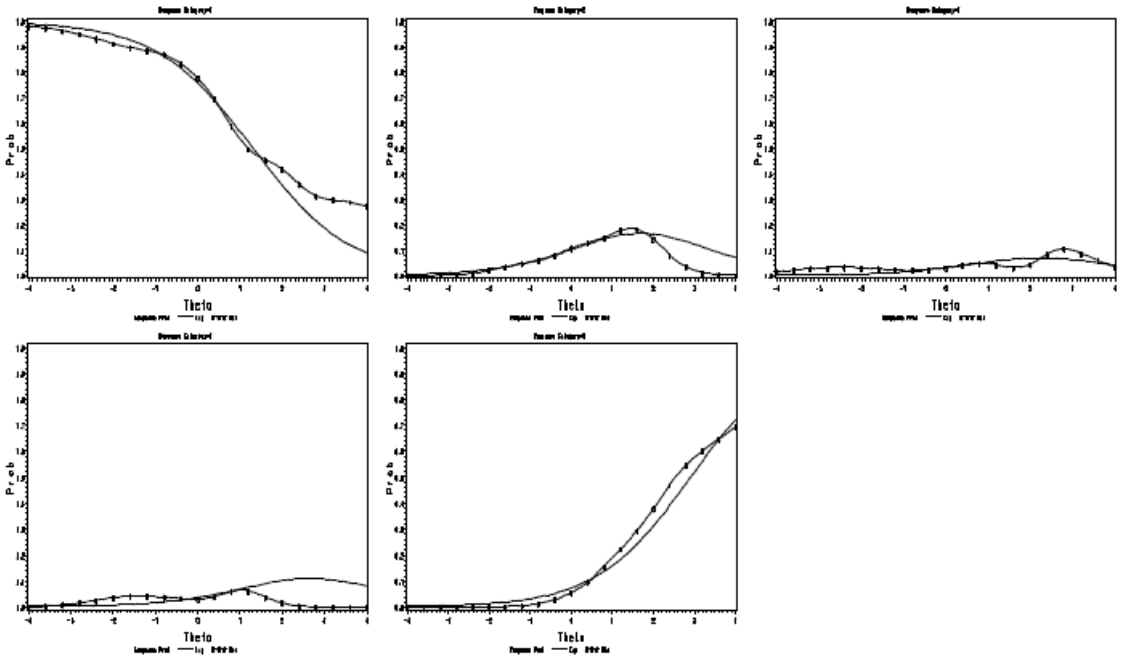
were identified as misfitting by Stone et al. (1993), and Items 1, 5, and 8 were flagged as misfitting by Stone (2000). Both studies used an $\alpha = 0.05$ significance level. If the same significance level was used for the current study, only Item 8 was flagged as misfitting. However, if a higher level of significance of $\alpha = 0.10$ was used, Items 1 and 8 would be classified as misfitting. Items 3 and 5 had PPP-values around 0.11, thus indicating the potential for item misfit.

In order to explain the different results and also to identify if there were some real issues with item misfit, the observed and expected item response category curves (ICCs) were drawn. Appendix F includes the ICCs for all the items on each of these three QCAI test forms. The ICCs for four misfitting items (Items 1, 3, 5, and 8) on the AS91 form are also shown in Figure 4.38. As can be seen, the discrepancies between observed and expected ICCs for response categories 1-4 were quite large for Item 5. There were some discrepancies between the observed and expected ICCs for response categories 1, 2, 3 and 4 for Item 3, though the ICCs for the last category (5) matched very well for the ability range of -2.5 to 2.5 in which most students fell. The fact that Items 3 and 5 were not flagged as misfitting based on the PPP-values further indicated the conservativeness of the PPMC method. As a result, a level of $\alpha = 0.10$ was employed for the real application.

It is important to determine if the misfitting identified by a statistical test has substantial practical consequences. The comparison of the expected and observed ICCs could be used for this purpose. Among these four misfitting items, Item 3 may not have significant practical consequences of misfitting since the discrepancies between expected and observed ICCs were not large in the ability range (-2.5, 2.5). The relatively large differences between ICCs for the other three items (Items 1, 5, and 8) may indicate that the item misfit may have practical

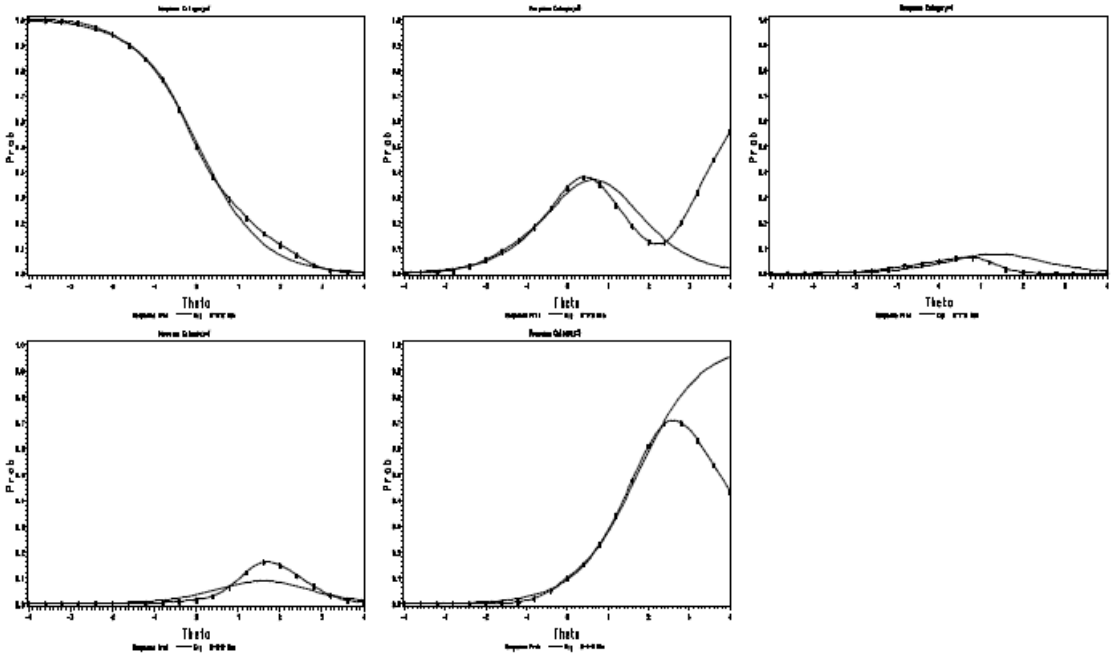
consequences. These results also imply that the method used by Stone (1993) for evaluating item-fit is relatively liberal. In contrast, the PPMC method used in the current study is relatively conservative. The method used by Stone (2000) appears to lie between these two approaches.

Item 1:



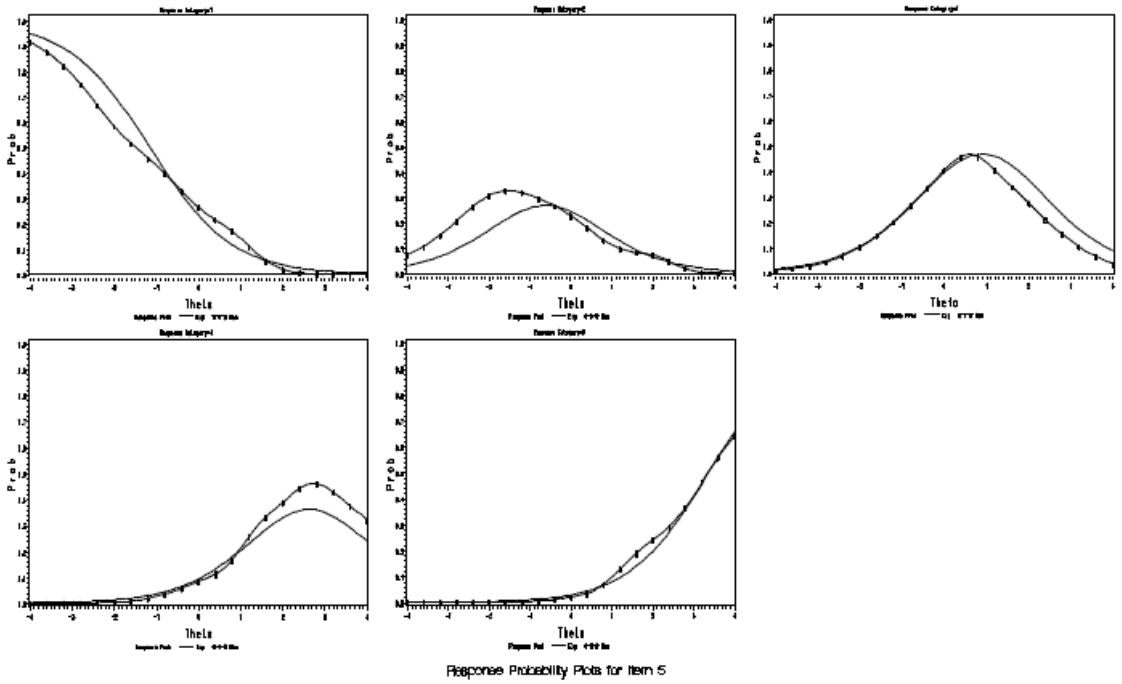
Response Probability Plots for Item 1

Item 3:



Response Probability Plots for Item 3

Item 5:



Item 8:

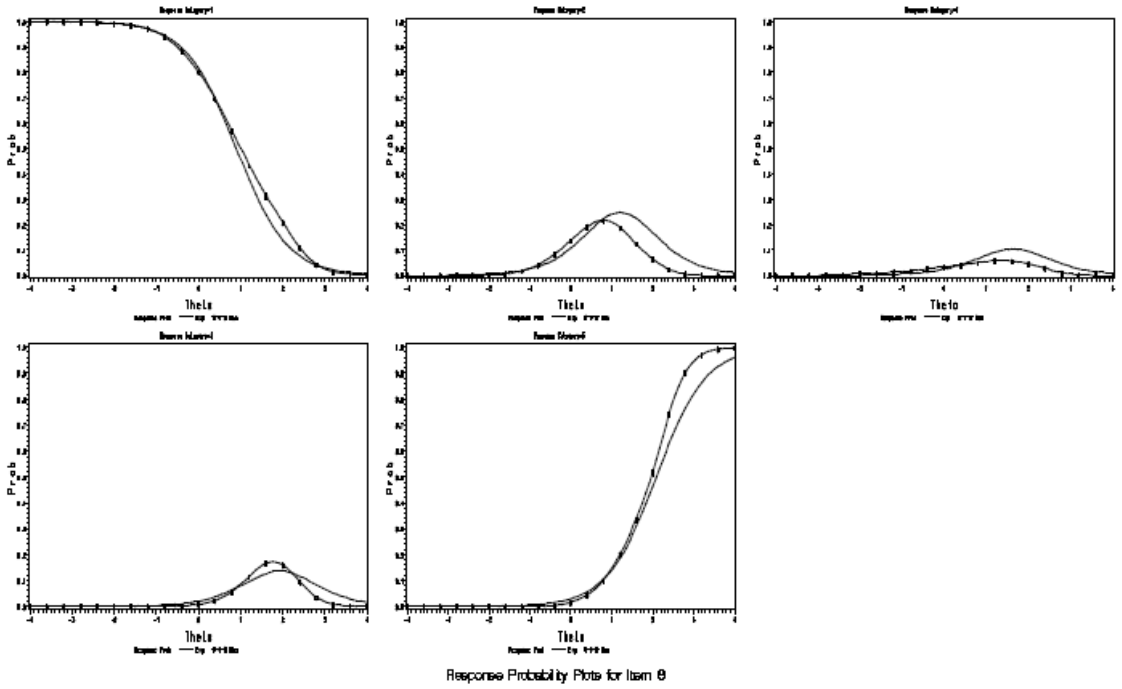


Figure 4.38 Observed vs. Expected ICCs for Misfitting Items on the AS91 Form

When a 1P GR model was fit this data, as shown in Table 4.24, the PPP-values of the item score distribution were not extreme, suggesting that this model could predict the item score distribution. A few of items were flagged based on the two item-fit measures. The most important measure for this model is the item-total score correlation since it is related to the slope parameter. When the common-slope GR model was used, Item 1 had a PPP-value of 0.95, indicating the observed item-test correlation for this item was systematically lower than the predicted correlations. In contrast, Items 2 and 8 had PPP-values of 0.04 and 0.02 respectively, indicating the observed correlation was significantly higher than the predictive correlations. This result was expected given the slope estimates in Table 4.23. The slopes of Items 3-7 were closer to the common slope estimate (1.23). In contrast, Item 1 had a slope of 0.87, and Items 2 and 8 had slopes of almost 1.60.

Figure 4.39 displays the PPP-values for three pair-wise discrepancy measures for all the 28 item pairs using pie plots. As can be seen, there were no clear patterns in these plots. Most of the PPP-values for Yen's Q_3 and item covariance residual were not extreme for both models, indicating that there was no clear evidence of violation in the unidimensionality and local independence assumptions. Though the global OR measure showed more extreme values than the other two measures, the results might not be convincing since several items on the AS91 form were very difficult. The dichotomization of the responses based on the rubric (0, 1, 2 treated as 0, and 3, 4 treated as 1) resulted in some zero frequency cells in the contingency tables. Therefore, it was necessary to use the other two measures to evaluate model-fit for these real datasets.

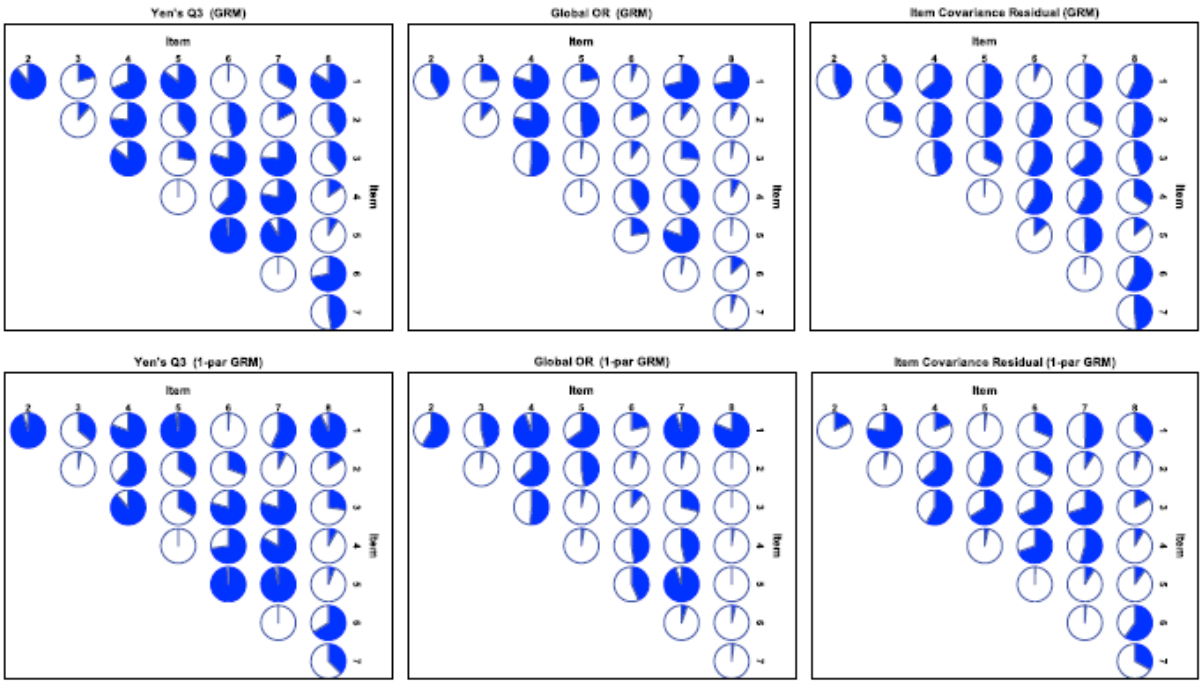


Figure 4.39 Display of PPP-values for Pair-wise Measures when fitting GR Model (top) and 1P GR Model (bottom) to the data – AS91

Overall, the results of PPMC using Yen's Q_3 and the item covariance residual showed that that the QCAI AS91 data was essentially unidimensional and items exhibited local independence. This conclusion was consistent with that from Lane et al. (1995). The GR model appeared to fit the AS91 data well regarding most aspects of the fit measured by these discrepancy measures. For example, the GR model could be used to predict the test and item score distributions, the relationships among the items, and the item-test score correlations. However, several misfitting items to the GR model were identified using PPMC with Stone's item fit measure, a finding which is consistent with previous studies. The results also showed that a 1P GR model could account for the item/test score distributions, and the correlations among the items, but this model could not explain the item-test score correlations correctly.

Model-Comparison

In order to compare the GR, 1P GR, and 2-dimensional complex-structure GR models, three model-comparison indices (DIC, CPO, and PPMC) were computed and compared in Table 4.25. It should be noted that only Yen's Q_3 statistic and the global OR measure were used with PPMC for the 2-dimensional complex-structure model. These two measures were found to be the most effective measures based on the simulation studies. As can be seen, the smallest DIC value for the 2-dim complex-structure GR model suggested this complex model was preferred over the GR model which in turn was preferred over the 1P GR model. The test-level CPO values for these three models only differed by less than 1 and thus did not provide sufficient evidence in favor of one model over the other for the overall test. The item-level CPO values (see Table 4.26) were also very close for the three models, further indicating there were no significant differences between these three models.

Table 4.25 Model Selection Indices for Overall Test – AS91

Model	DIC	CPO	PPMC							
			Test score dist	Item score dist	Yen's Q_1	Stone's fit stat	Item-test corr	Yen's Q_3	Global OR	Item cov resid
GR	7455	-1625.2	0.56	0/8	0/8	2/8	0/8	4/28	6/28	2/28
1P GR	7471	-1625.3	0.64	0/8	1/8	3/8	3/8	8/28	10/28	5/28
2-dim GR	7415	-1626.1	-	-	-	-	-	4/28	5/28	-

Table 4.26 Item-level CPO Index for Each Item – AS91

Model	Item							
	1	2	3	4	5	6	7	8
GR	-153.9	-246.9	-192.6	-208.0	-236.4	-229.0	-222.0	-136.4
1P GR	-154.2	-247.9	-192.5	-207.5	-237.2	-227.8	-221.8	-136.4
2-dim GR	-154.3	-246.4	-192.5	-208.3	-236.5	-229.2	-222.3	-136.5

Table 4.25 also summarizes the PPMC results for this dataset. For the GR and 1P GR models, all 8 discrepancy measures were used. For the test score distribution measure, the numbers in the table are the PPP-values. For the item-level measures, the numbers reflect the number of items with extreme PPP-values across the total number of 8 items. For example, for the item-test score correlation, there were no extreme PPP-values under the GR model, but there were 3 of 8 items with extreme values under the 1P GR model. For the pair-wise measures, the numbers represent the number of item pairs having extreme PPP-values.

Firstly, we can see that there were more extreme values for the 1P GR model over the GR model. However, the most useful measure for comparing these two models may be the item test score correlation since it reflects the difference in item parameters between these two models. By examining this measure, it is clear that the GR model fit this data significantly better. The DIC values also indicated that the GR model was preferred over the 1P GR model, but this index only provided an absolute measure of model fit. In contrast, the results of the PPMC application not only indicated that the GR model was preferred, but also indicated that this model fit the data reasonably well. The conclusion that the GR model fit the data significantly better than the 1P GR model is consistent with the previous finding by Lane et al. (1995).

It can also seen from Table 4.25 that when a 2-dimensional complex-structure GR model was used to analyze this AS91 dataset, 4 and 5 out of 28 item pairs had extreme PPP-values for Yen's Q_1 index and the global OR measure, respectively. When a unidimensional GR model was estimated, the numbers of extreme PPP-values were about the same as for the 2-dimensional model. This result indicates that there was not sufficient evidence to prefer the 2-dimensional model. Thus a simpler unidimensional GR model may be adequate for this dataset. This conclusion is also consistent with the previous finding by Lane et al. (1995).

Overall, the DIC values indicated that a 2-dimensional model was preferred for the AS91 dataset, but the PPMC results indicated that a unidimensional GR model was adequate. The CPO values did not provide sufficient evidence in favor of one model over the other.

4.3.2 QCAI Data 2 – AS92

Table 4.27 provides the item parameter estimates for the GR model based on the AS92 data. As can be seen, the estimates from two programs were very similar. The average absolute difference between WinBUGS and MULTILOG estimates across all the items was 0.021 for the slope parameters, and 0.032 for all the threshold parameters. In addition, Figure 4.40 shows the sample history and autocorrelation plots for the slope and one threshold parameters for item 2. Other parameters had similar plots. These plots demonstrated that the convergence of the chain was attained for this AS92 data.

Table 4.27 Item Parameter Estimates using WinBUGS and Multilog – AS92

	WinBUGS					**Multilog**				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	1.12	1.25	1.84	2.64	3.34	1.12	1.28	1.87	2.66	3.34
2	1.47	-1.03	-0.08	0.66	1.41	1.51	-0.97	-0.05	0.66	1.39
3	1.14	0.30	1.26	1.66	1.80	1.16	0.32	1.26	1.64	1.76
4	0.70	-0.18	0.85	1.16	3.97	0.69	-0.13	0.89	1.19	3.99
5	0.70	-1.72	-0.54	2.16	4.17	0.67	-1.73	-0.52	2.26	4.33
6	1.11	-0.93	0.03	0.84	3.50	1.13	-0.89	0.05	0.84	3.46
7	1.32	-1.48	0.28	1.06	1.77	1.37	-1.41	0.29	1.04	1.72
8	1.38	1.26	2.17	2.46	3.03	1.38	1.28	2.17	2.44	2.99

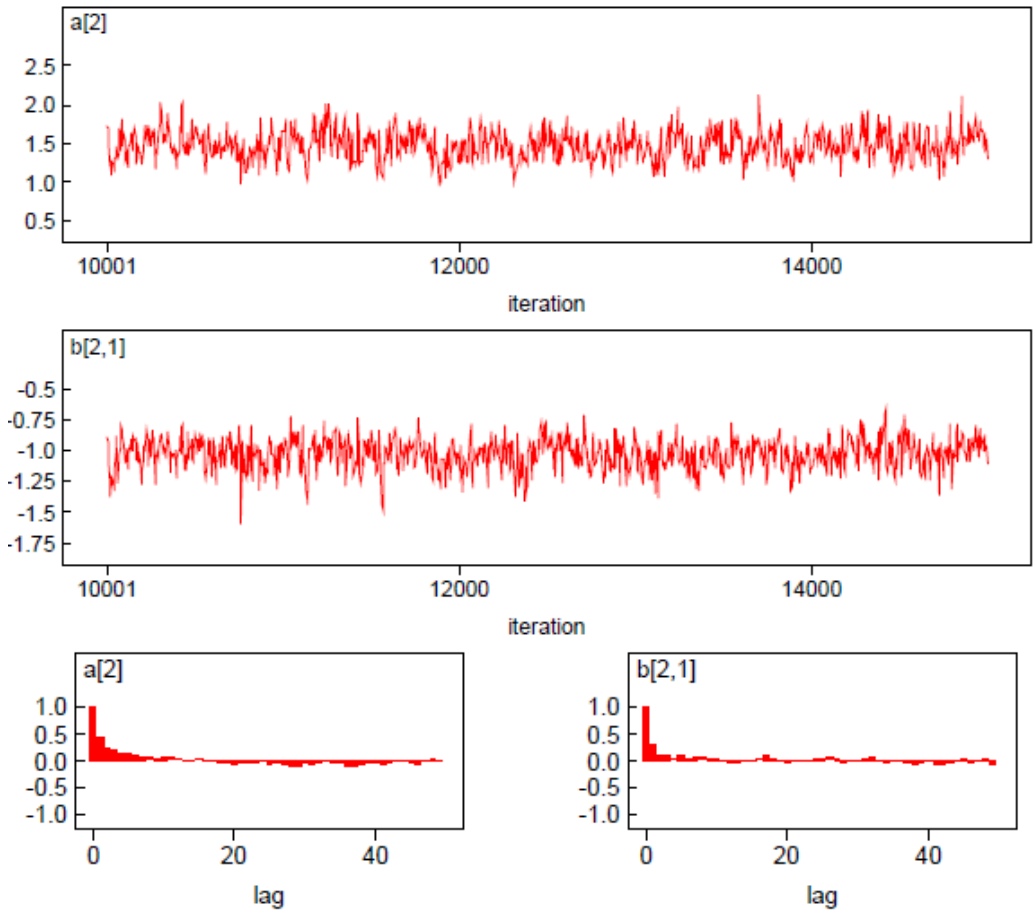


Figure 4.40 Example History and Autocorrelation Plots – AS92

Model-Fit

The PPP-values of the chi-square statistic summarizing the discrepancy between the observed and predictive test score distributions were 0.24 and 0.18 for the GR and 1P GR models, respectively. Both values were not extreme, indicating these two models fit the data regarding the total test score distribution.

Table 4.28 includes the PPP-values for the four item-level discrepancy measures for each item. When a GR model was used to analyze this data, all the PPP values were not extreme, suggesting a good fit of the GR model in the aspects measured by these four measures. However, when a 1P GR model was estimated, two items (Items 2 and 5) were identified as misfitting by

Stone's item-fit measure. Most important, among the eight items, six items had extreme PPP values for the item-total score correlation measure, providing evidence that the 1P GR model could not predict the item-test score correlations in the data.

Table 4.28 PPP-values for Item-level Measures based on GR and 1P GR Models – AS92

	Item-Level Discrepancy Measures							
	Item Score Dist		Yen's Q_1		Stone's Item-Fit		Item-Test Corr	
Item	GR	1P GR	GR	1P GR	GR	1P GR	GR	1P GR
1	0.51	0.49	0.54	0.54	0.57	0.62	0.40	0.29
2	0.50	0.47	0.40	0.35	0.19	0.06	0.17	0.00
3	0.48	0.51	0.42	0.45	0.18	0.28	0.17	0.04
4	0.51	0.46	0.53	0.32	0.54	0.14	0.46	1.00
5	0.48	0.43	0.46	0.26	0.33	0.06	0.48	1.00
6	0.47	0.50	0.51	0.55	0.36	0.42	0.53	0.44
7	0.48	0.47	0.50	0.48	0.41	0.38	0.31	0.02
8	0.52	0.47	0.42	0.54	0.46	0.42	0.24	0.01

It should be noted that Items 2 and 3 were identified as misfitting for the GR model by Stone et al. (1993), and Item 3 was also flagged by Stone (2000). Even if the significance level $\alpha = 0.10$ was used, no item would be flagged as misfitting using the PPMC method with Stone's item fit measure. The PPP values for Items 2 and 3 were around 0.19 and 0.18, respectively. Though their values were lower than the other values, they were not extreme enough to indicate item misfitting.

The ICCs for Items 2 and 3 are shown in Figure 4.41, and the ICCs for other items are in Appendix F. It can be seen the observed and predicted ICCs matched reasonably well for Item 2. Thus, the item misfit identified by Stone (1993) may not indicate a practical consequence. For Item 3, the observed ICCs for three response categories 1, 3 and 4 were very close to the corresponding predicted ICCs. However, there were some discrepancies between the observed and predicted ICCs for the other categories (2, and 5). These results further indicate the

conservativeness of the PPMC method in evaluating item-fit, and the liberalness of the method used by Stone (1993). As before, the method used by Stone (2000) falls in between these two approaches and yields results that are more reasonable for practical purposes.

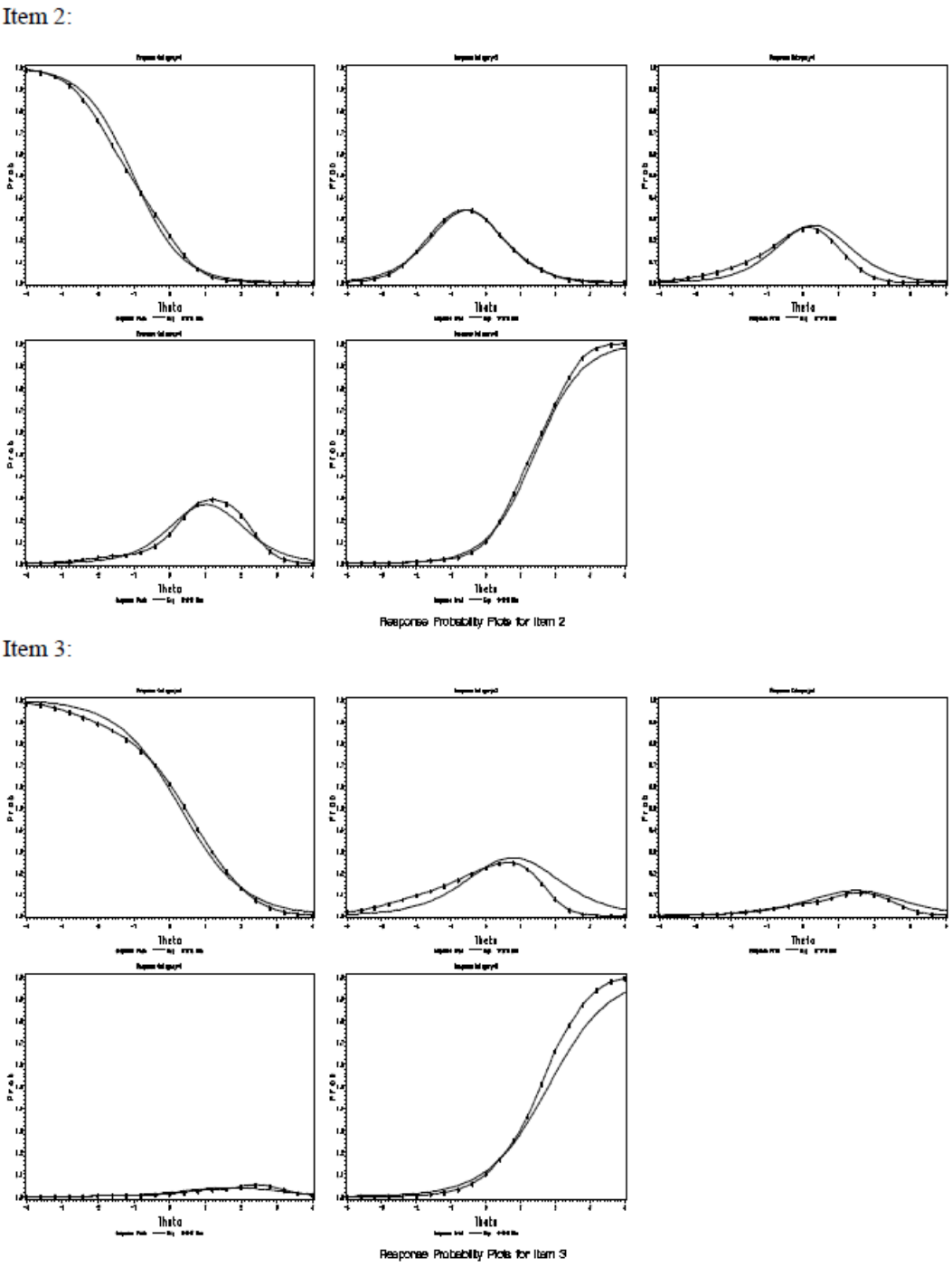


Figure 4.41 Observed vs. Expected ICCs for Misfitting Items on the AS92 Form

Figure 4.42 displays the PPP-values for three pair-wise discrepancy measures for all the 28 item pairs using the pie plots. As can be seen, there were no clear patterns in these plots, and most of the PPP-values for Yen's Q_3 and item covariance residual were not extreme for both models. The results indicated that there was no clear evidence of violations in the unidimensionality and local independence assumptions for this data. This conclusion is consistent with the finding in Lane et al. (1995).

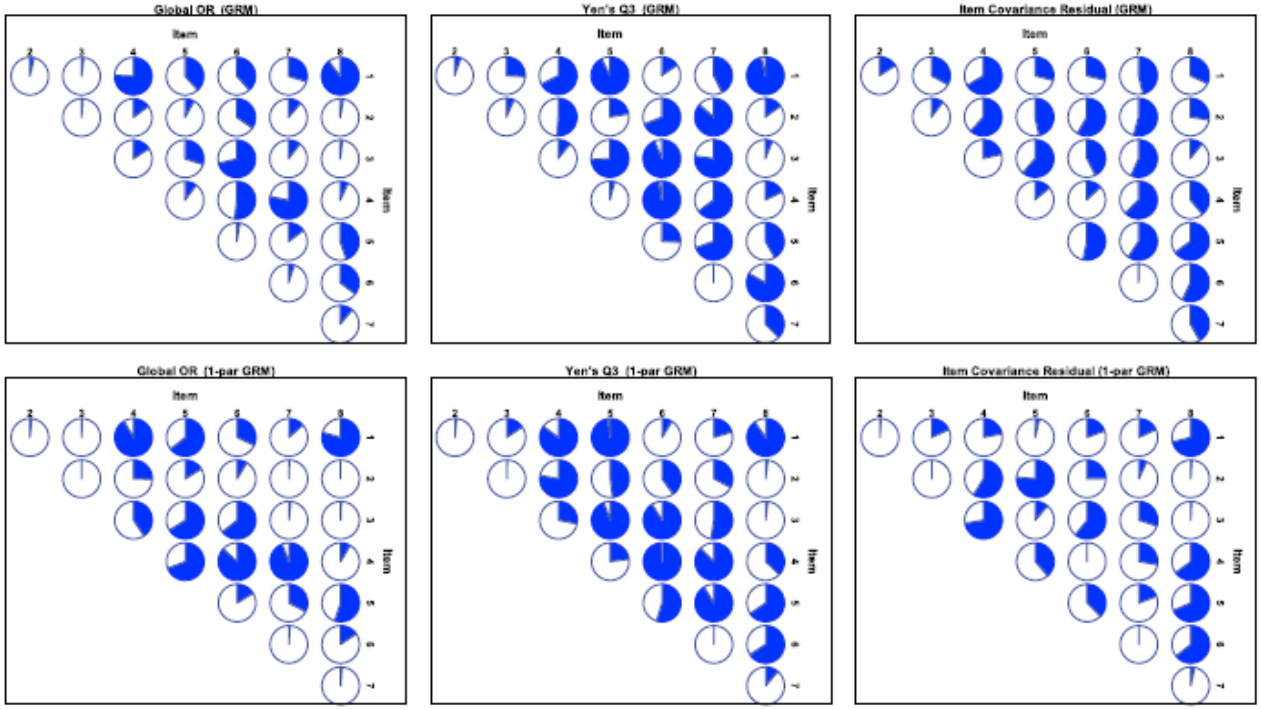


Figure 4.42 Display of PPP-values for Pair-wise Measures when fitting GR Model (top) and 1P GR Model (bottom) to the Data – AS92

Overall, the GR model appeared to fit the AS92 data well regarding different aspects of the fit such as dimensionality, item-fit, item/test score distribution, and item-test score correlations. Although a 1P GR model could explain several aspects of properties in the data, it could not explain the relationship between items and test scores.

Model-Comparison

Table 4.29 compares the values of DIC, CPO and PPMC for the GR and 1P GR models. As can be seen, the DIC values indicated that the 2-dimensional complex-structure GR model was preferred over the GR model, which in turn was preferred over the 1P model for this particular dataset. However, the CPO values for these three models suggested that the GR model was preferred over the 2-dim GR model which in turn was better than the 1P model. For the PPMC indices, in general, there were more extreme PPP values for the 1P GR model compared to the other two models. As an example, six of eight items had extreme values for the item-test score correlation. However, the same numbers of extreme PPP-values for the 2-dim GR and the unidimensional GR model indicated that the unidimensional GR model was adequate for this dataset.

Table 4.29 Model Selection Indices for Overall Test – AS92

Model	DIC	CPO	PPMC							
			Test score dist	Item score dist	Yen's Q_1	Stone's fit stat	Item-test corr	Yen's Q_3	Global OR	Item cov resid
GR	8644	-1884	0.24	0/8	0/8	0/8	0/8	4/28	6/28	1/28
1P GR	8693	-1891	0.18	0/8	0/8	0/8	6/8	7/28	9/28	8/28
2-dim GR	8598	-1888	-	-	-	-	-	4/28	6/28	-

In summary, these three model-comparison indices reached the same conclusion that the GR model was preferred over the 1P GR model for the AS92 data, a finding which is consistent with Lane et al. (1995). In addition, both the CPO and PPMC results indicated that the GR model was also preferred over the 2-dimensional GR model. The DIC index tended to choose a more complex model as the preferred model based on the results for the previous AS91 dataset and this AS92 dataset.

4.3.3 QCAI Data 3 – BS92

Table 4.30 Item Parameter Estimates using WinBUGS and Multilog – BS92

	WinBUGS					Multilog				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	0.93	-1.69	-0.12	1.40	3.21	0.97	-1.60	-0.10	1.34	3.08
2	1.50	-0.47	0.59	1.06	1.35	1.57	-0.44	0.58	1.03	1.29
3	1.64	-0.28	0.62	1.17	1.64	1.70	-0.25	0.61	1.14	1.58
4	1.76	0.70	1.33	1.46	1.81	1.82	0.69	1.30	1.42	1.74
5	1.36	-2.15	-0.29	0.49	1.66	1.40	-2.07	-0.26	0.49	1.61
6	1.06	-0.43	0.42	0.92	1.63	1.12	-0.39	0.41	0.87	1.53
7	1.67	0.36	1.04	1.20	1.54	1.73	0.37	1.02	1.16	1.49
8	0.89	-1.45	-0.64	-0.21	0.67	0.95	-1.34	-0.60	-0.20	0.62

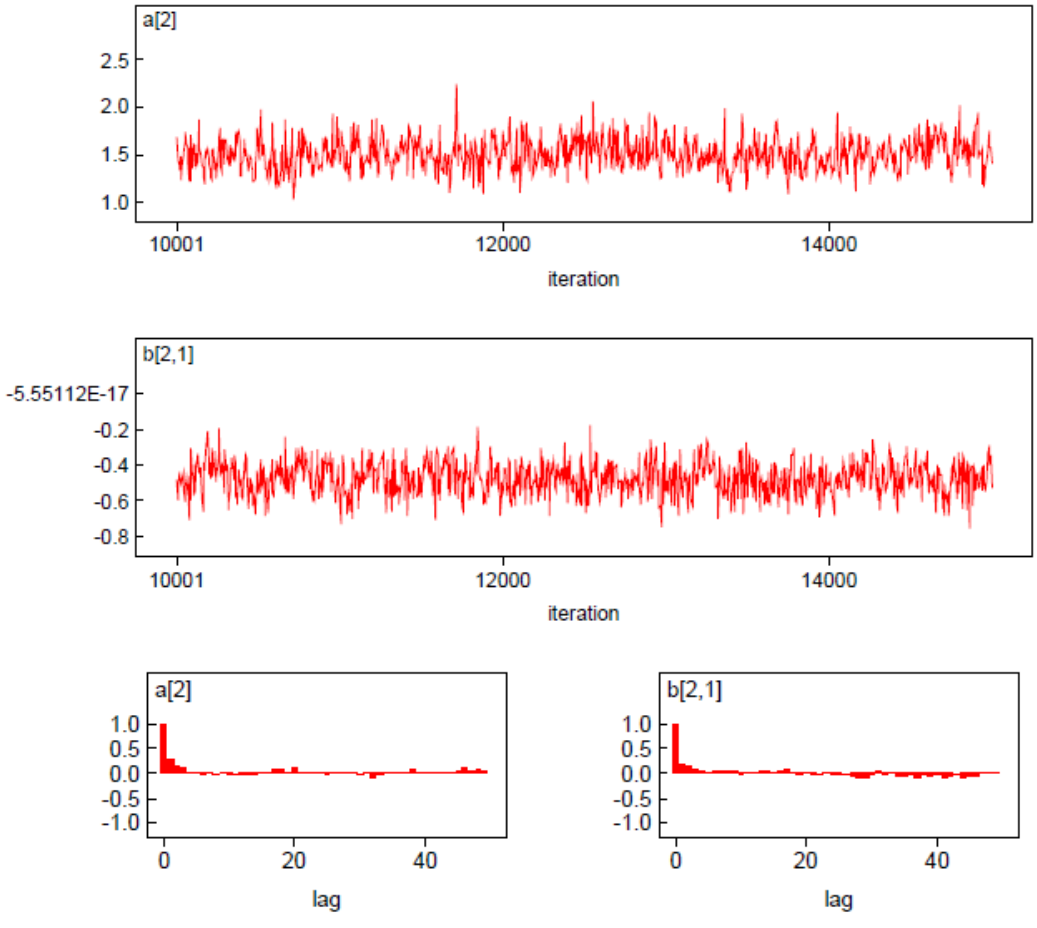


Figure 4.43 Example History and Autocorrelation Plots – BS92

Table 4.30 provides the item parameter estimates for the GR model based on the BS92 data. As can be seen, the estimates from two programs were close. The average absolute difference between WinBUGS and MULTILOG estimates across all the items was 0.056 for the slope parameters, and 0.044 for all the threshold parameters. Figure 4.41 shows the sample history and autocorrelation plots for the slope and one threshold parameters for item 2. Other parameters had similar plots. These plots indicate that convergence of the chain was attained.

Model-Fit

The PPP-values of the chi-square statistic summarizing the discrepancy between the observed and predictive test score distributions were 0.02 and 0.01 for the GR and 1P GR models, respectively. Both values were extreme, indicating these two models could not adequately predict the test score distribution for this dataset.

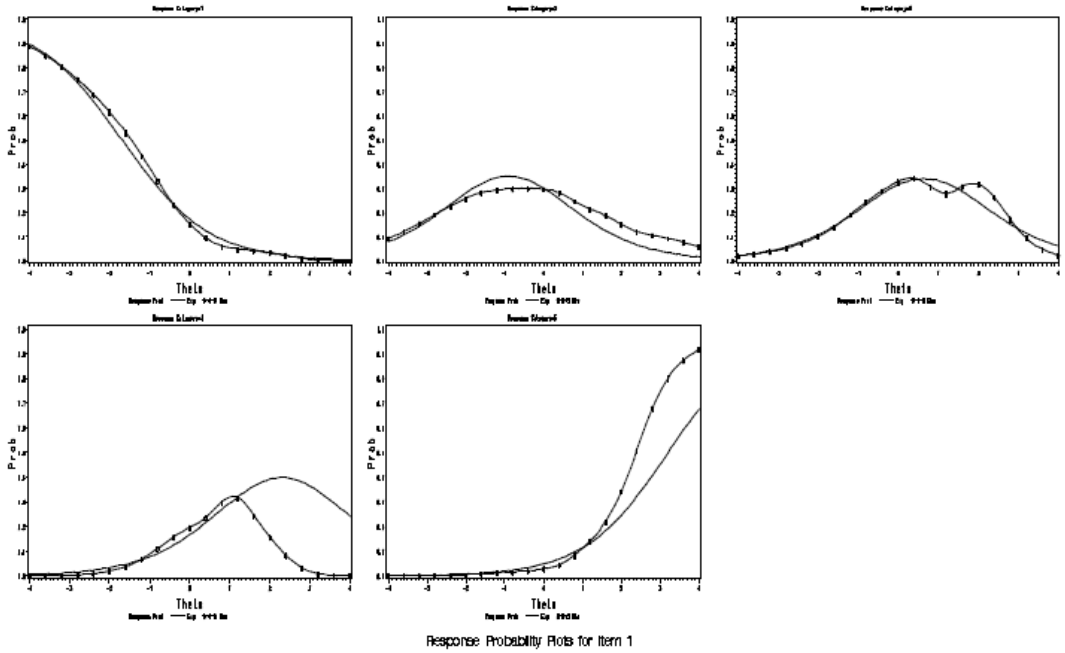
Table 4.31 PPP-values for Item-level Measures based on GR and 1P GR Models – BS92

	Item-Level Discrepancy Measures							
	Item Score Dist		Yen's Q_1		Stone's Item-Fit		Item-Test Corr	
Item	GR	1P GR	GR	1P GR	GR	1P GR	GR	1P GR
1	0.48	0.41	0.26	0.11	0.05	0.01	0.49	1.00
2	0.49	0.44	0.31	0.34	0.05	0.03	0.12	0.00
3	0.46	0.42	0.26	0.27	0.01	0.00	0.23	0.00
4	0.52	0.49	0.43	0.52	0.69	0.27	0.27	0.00
5	0.50	0.52	0.45	0.50	0.22	0.36	0.33	0.16
6	0.47	0.46	0.22	0.18	0.02	0.01	0.13	0.56
7	0.48	0.46	0.31	0.33	0.05	0.02	0.13	0.00
8	0.51	0.49	0.48	0.25	0.39	0.13	0.33	0.99

Table 4.31 presents the PPP-values for the four item-level discrepancy measures for each item. When a GR model was used to analyze this data, the PPP-values of the item score distribution, item-test score correlation, and Yen's Q_1 index were not extreme. However, five items (Items 1-3, 6-7) demonstrated extreme PPP-values for Stone's item fit statistic and would

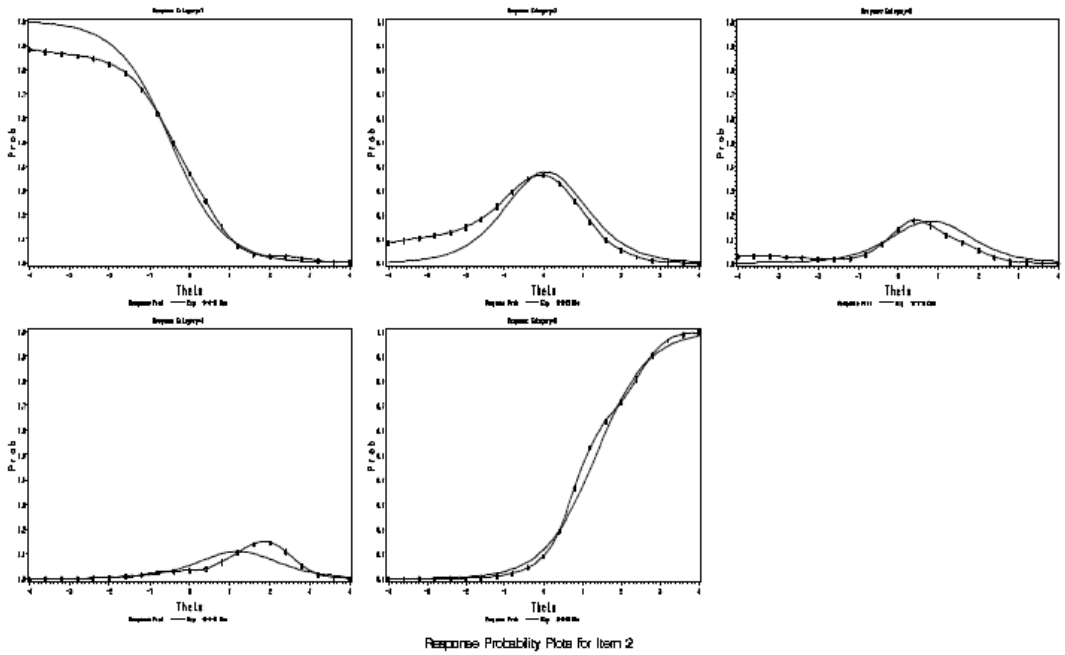
therefore be flagged as misfitting. It is worthy to note that the same five items were also identified as misfitting by Stone et al. (1993) and Stone (2000) (see Table 3.18). The number of misfitting items might explain why the GR model could not predict adequately the test score distribution for this dataset.

Item 1:



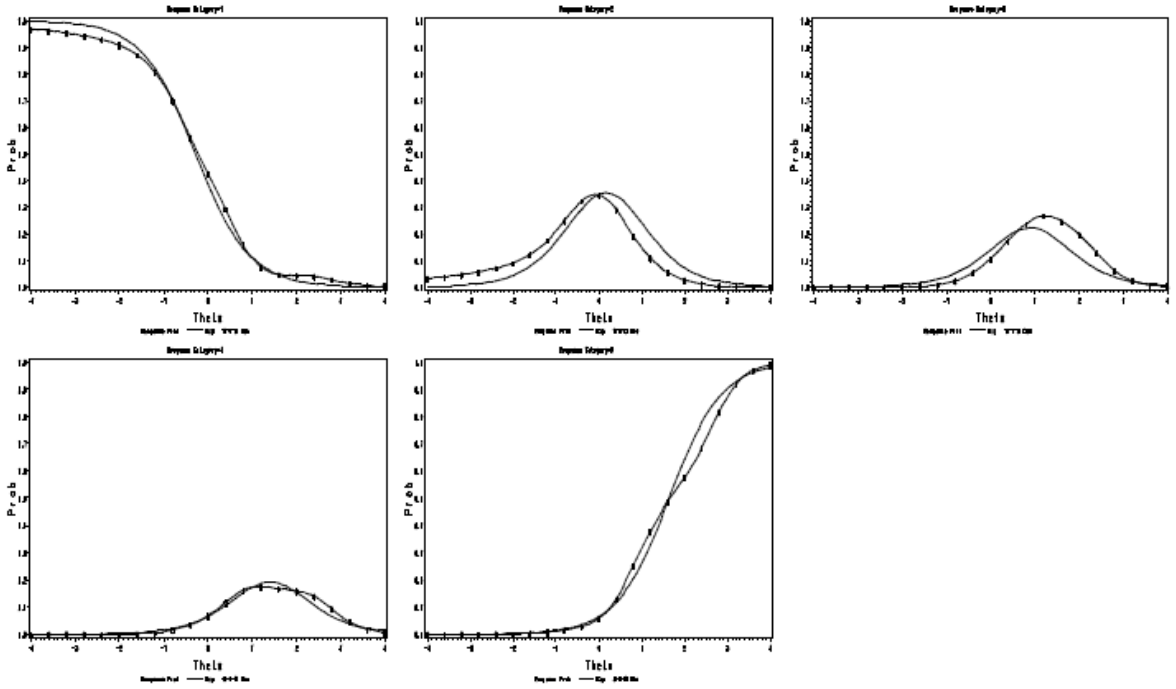
Response Probability Plots for Item 1

Item 2:



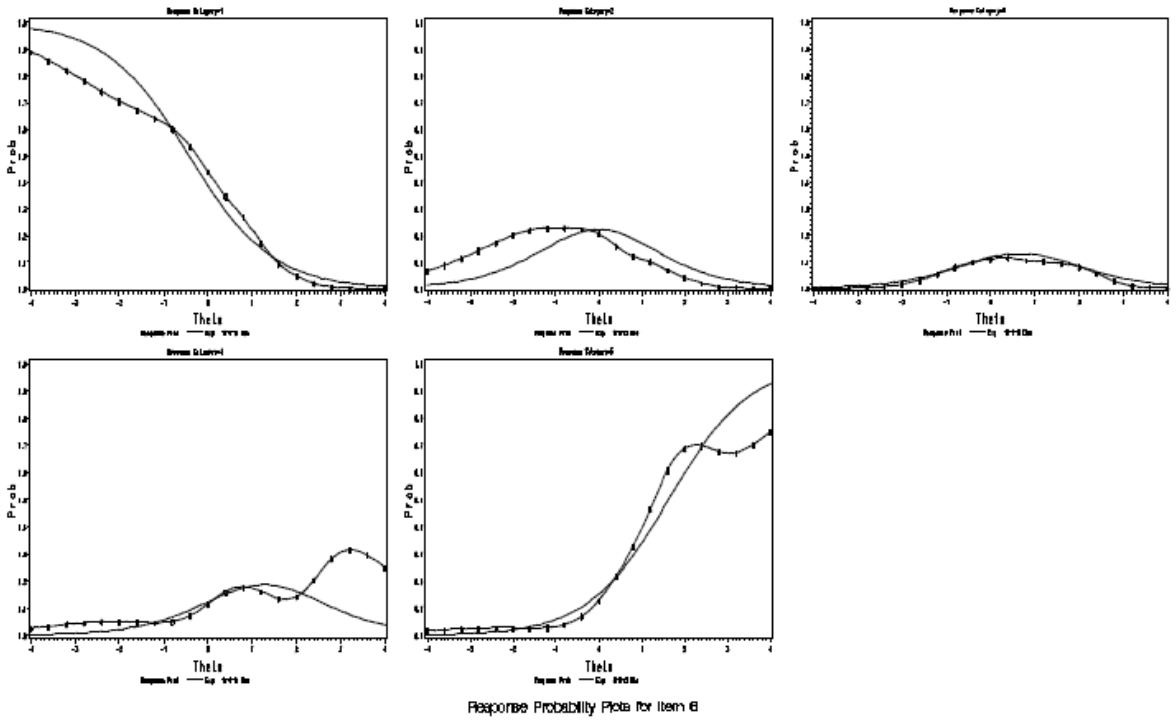
Response Probability Plots for Item 2

Item 3:



Response Probability Plots for Item 3

Item 6:



Response Probability Plots for Item 6

Item 7:

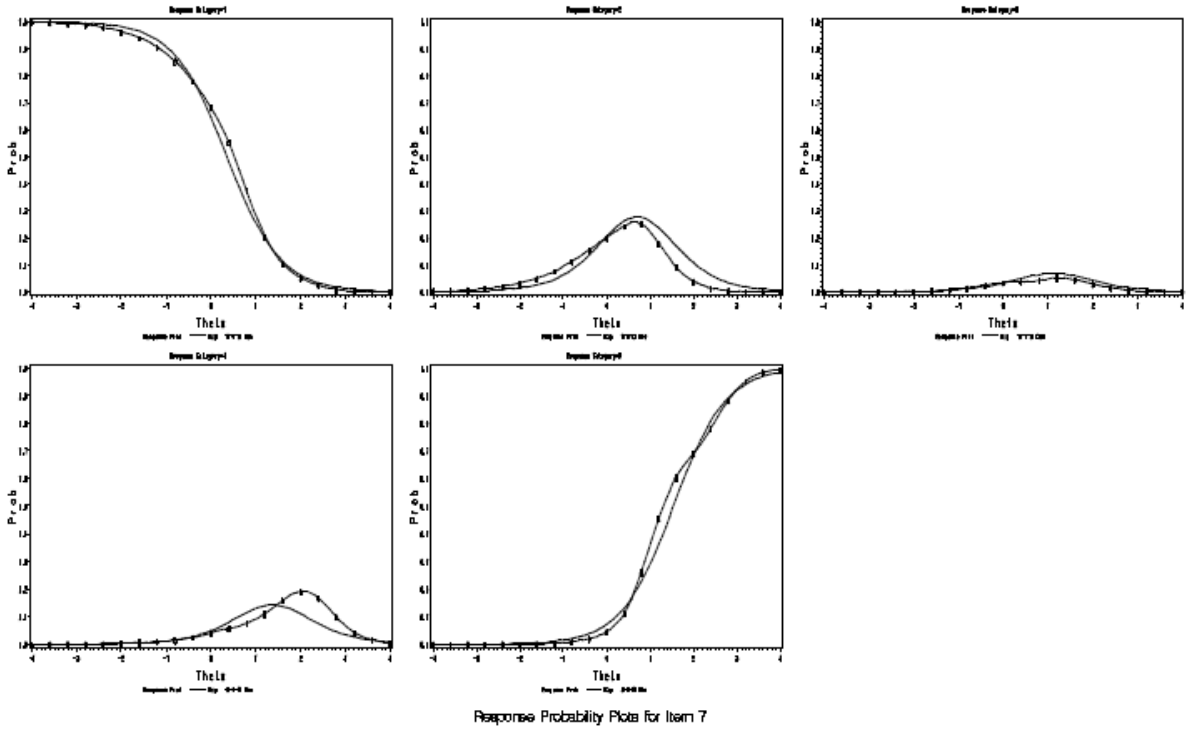


Figure 4.44 Observed vs. Expected ICCs for Misfitting Items on the BS92 Form

The ICCs for these five misfitting items are shown in Figure 4.44, and the ICCs for other items are in Appendix F. As can be seen, the discrepancies between the observed and predicted ICCs for Items 1-3, and 6 were large, and these large discrepancies may reflect significant practical consequences of item misfit. However, for Item 7, the observed ICCs matched the predicted ICCs reasonably well except that there were small discrepancies for response categories 2 and 4. These discrepancies may not indicate practical significance in item misfit.

When a 1P GR model was estimated, as shown in Table 4.31, the same five items were also flagged as misfitting by Stone's item-fit measure. In addition, six items had extreme PPP values for the item-test score correlation. The results provided sufficient evidence that the 1P GR model was not appropriate for this dataset.

It is interesting to note that Yen's Q_1 index, in contrast with Stone's fit statistic, appeared to have no power for the short tests that were studied (AS91, AS92, BS92). A number of items across the three datasets were identified as misfitting using Stone's fit statistic, but not flagged using Yen's index. Previous research has found that item fit statistics such as Yen's or Bock's indices do not perform well with short tests (Orlando & Thissen, 2000; Stone & Hansen, 2000; Stone & Zhang, 2003). Imprecision in ability estimates with short tests result in classification errors in the item fit tables which in turn affects the null chi-square distribution and hypothesis testing (Stone, 2000). However, in the PPMC framework, the sampling distributions are based on Monte Carlo resampling methods. It is not clear why results based on Yen's Q_1 index and Stone's fit statistic differed, and therefore, more research is needed to explain this finding.

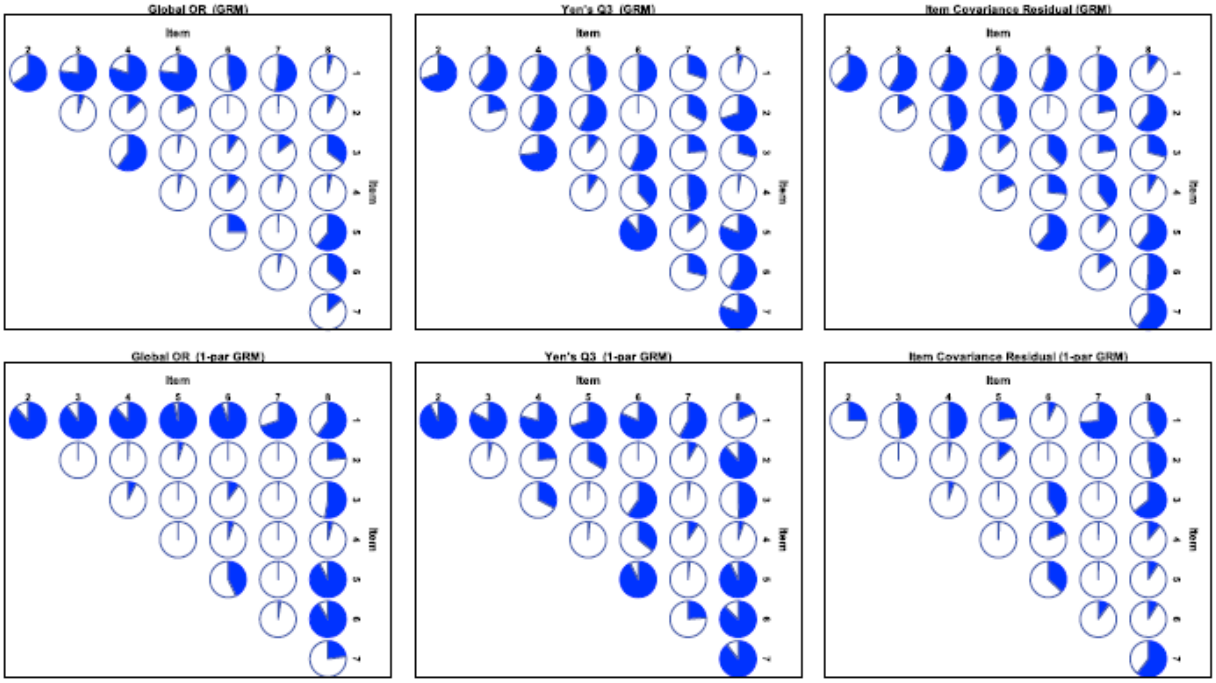


Figure 4.45 Display of PPP-values for Pair-wise Measure when fitting GR Model (top) and 1P GR Model (bottom) to the Data – BS92

Figure 4.45 displays the PPP-values for three pair-wise discrepancy measures for all the 28 item pairs using pie plots. As can be seen, most of the PPP-values for Yen's Q_3 and item

covariance residual under the GR model were not extreme, implying the assumptions of unidimensionality and local independence underlying the GR model were not violated for this dataset. This conclusion is consistent with the finding in Lane et al. (1995). However, there were more item pairs with extreme PPP values under the 1P GR model.

In summary, neither the GR model nor 1P GR model fit the BS92 dataset very well. Though the GR model fit the data in terms of item score distribution and item-test score correlations, there were five misfitting items identified by Stone's fit statistic. Moreover, this model could not explain the test score distribution observed in this dataset. The 1P GR model exhibited these same problems. In addition, it could not explain the relationship between items and test scores.

Model-Comparison

Table 4.32 Model Selection Indices for Overall Test – BS92

Model	DIC	CPO	PPMC							
			Test score dist	Item score dist	Yen's Q_1	Stone's fit stat	Item-test corr	Yen's Q_3	Global OR	Item cov resid
GR	8727	-1902	0.02	0/8	0/8	5/8	0/8	3/28	9/28	1/28
1P GR	8771	-1909	0.01	0/8	0/8	5/8	6/8	6/28	12/28	9/28
2-dim GR	8679	-1902	-	-	-	-	-	2/28	7/28	-

Table 4.32 compares the values for model-comparison indices. The smaller DIC and larger CPO values for the GR model suggested that the GR model was preferred over the 1P GR model for the BS92 dataset. For the PPMC indices, in general, there were more extreme PPP values for the 1P GR model, further indicating that the GR model was the preferred model. The PPMC results also tell us that even though the GR model was better than the one-par GR for this dataset, it did not fit the data in several aspects such as test score distribution and item-fit.

Regarding the fit of the 2-dimensional GR model and the unidimensional GR model, it can be seen from this table that the 2-dim GR model had smaller DIC value than the GR model, indicating the 2-dim GR model may be preferred. However, the CPO values for these two models were the same, providing insufficient evidence in favor of one model over the other model. Additionally, the PPMC results also did not provide enough evidence to support the more complex 2-dim GR model. Therefore, based on the CPO and PPMC results, the relatively more parsimonious model (i.e., the GR model) would be preferred. As for the other dataset, this is consistent with the finding by Lane et al. (1995). The different results between the DIC index and the other indices further indicated that the DIC index tends to select a more complex model.

5.0 DISCUSSION

The present work, through two simulations and three real data examples, evaluates the application of Bayesian model-fit and model-comparison techniques to assess fit of unidimensional GR models and compare different GR models for performance assessment applications. This section summarizes the major findings from this work and also provides the future research directions.

5.1 SUMMARY OF MAJOR FINDINGS

5.1.1 Simulation Study 1

The first study in the current work was to explore the general performance of the PPMC method in evaluating different aspects of fit of unidimensional GR models to performance assessments by using a variety of discrepancy measures. PPMC has been found to be useful in assessing the fit for dichotomous IRT models. Study 1 extended previous research to the use of PPMC for polytomous IRT models. The discrepancy measures examined involved one test-level measure (*observed test score distribution*), several item-level measures (*item score distribution, item total test correlation, Yen's Q_3, and Stone's item-fit statistics*), and three pair-wise measures (*global*

odds ratios, Yen's Q_3, and absolute item covariance residual). Specifically, this study was intended to address the following three research questions:

(1) What is the Type-I error rate for each proposed discrepancy measure used with PPMC in assessing the fit of unidimensional GR model?

(2) What is the empirical power for each proposed discrepancy measure used with PPMC in detecting the violation of the assumptions underlying the unidimensional GR model (i.e., unidimensionality, local independence, and item fit)?

(3) Among different types of discrepancy measures (test-level, item-level, and pair-wise measures) proposed in the current study, which measures are most effective in detecting model misfit?

Type-I Error Rates:

The results from Condition 1, where the generating model was the same as the analyzing model, demonstrated that the Type-I error rates of the discrepancy measures examined in this study were below the nominal level. This indicates that the use of PPP-values in hypothesis testing would lead to highly conservative inferences (i.e., they tend not to indicate misfit of a correct model too often). The two pair-wise measures (global OR and Yen's Q_3) appeared to have empirical Type-I error rates that were closest to the nominal rate, though still quite lower. This finding confirmed the conclusion from the previous PPMC research (Bayarri & Berger, 2000; Fu et al., 2005; Levy, 2006; Sinharay, 2005; Sinharay et al., 2006) about the conservativeness of the PPMC method.

Previous studies pointed out that this conservativeness in the hypothesis tests is due to the departure of the distribution of PPP-values from the uniform distribution, which is also supported by the current study. The distributions of PPP-values for the discrepancy measure examined were

generally centered at 0.5 but less dispersed than a uniform distribution. The PPP-values under the correct model tend to be closer to 0.5 more often than would be expected under a uniform distribution. However, the distributions of PPP-values for the two pair-wise measures – *global OR* and *Yen's* Q_3 and the *test score distribution* were closest to uniform distributions as compared to the other measures. The approximate uniform distributions for the *global OR* and *Yen's* Q_3 discrepancy measures were also observed by Levy (2006).

Empirical Power Rates:

Unidimensionality

The ability of each discrepancy measure with PPMC to detect violations of unidimensionality was explored in Condition 2. Two multidimensional cases (ρ=0.3 or 0.6) were examined, reflecting a high and moderate degree of multidimensionality, respectively. Overall, the PPMC method using three pair-wise measures (Yen's Q_3, global OR, and item covariance residual) detected the lack of fit of unidimensional GR model to the two-dimensional test data successfully for both cases. Among them, Yen's Q_3 index performed best in terms of the empirical power, and the item covariance residual measure in turn performed better than the global OR. The relatively low performance of the global OR measure might be due to the dichotomization of polytomous item responses. However, Levy (2006) found that Yen's Q_3 index was more powerful than the OR measure based on the dichotomous IRT model. It is worthy to note that the global OR and Yen's Q_3 measures are both directional measures, and their PPP-values reflect the relationship between realized and posterior predictive discrepancies. The patterns of PPP-values could also be used to indicate how the items may be grouped into clusters or dimensions, and therefore used to explore the dimensionality of the item responses. In

this sense, these two measures are better than the item covariance residual which is non-directional.

The test-level and item-level discrepancy measures were found to be less effective for detecting this multidimensionality than the pair-wise measures. The three item-level measures (item score distribution, Yen's Q_1, and Stone's fit statistic) did not demonstrate any power for both cases. The item-total score correlation measure exhibited no power in detecting the moderate degree of multidimensionality (ρ=0.6), but became extremely powerful in detecting the high degree of multidimensionality (ρ=0.3). The test-level measure (i.e. test score distribution) shows certain power in detecting the misfit of the GR model when the data was highly two-dimensional (ρ=0.3).

The performance of PPMC was affected by the degree of the uniqueness in the dimensions. Specifically, as the inter-dimensional correlation increased from 0.3 to 0.6 (i.e., as the degree of uniqueness decreased), the power of three pair-wise measures decreased slightly, but they still appeared consistently powerful in detecting model misfit. In other words, the performance of PPMC was stable in the range of inter-dimensional correlations from 0.3 to 0.6. Therefore, future research that manipulates more levels between 0.6 and 1.0 is needed in order to identify the level at which the PPMC method with these three pair-wise measures would lose power. On the other hand, an increase in the inter-dimensional correlation from 0.3 to 0.6 had great impact on the effectiveness of the item-total score correlation measure. It exhibited almost full power for the low correlation condition, but had no power for the high correlation condition. Further research specifying more levels in the correlation is needed in order to more fully understand PPMC applications with this measure.

Local Independence

The performance of PPMC in detecting violations in the local independence assumption was examined. In one condition, Condition 3, the local dependence was due to an added nuisance dimension, and two levels of dependence on the nuisance dimension were considered: large dependence (a2/a1=1) and mild dependence (a2/a1=0.5). The test-level and item-level measures were found to be not useful in detecting local dependence among items loading also on the nuisance dimension, while three pair-wise measures performed effectively. All three pair-wise measures exhibited sufficient power in detecting a large dependence among the items. However, as the strength of dependence on the nuisance dimension decreased, their performance decreased. Yen's Q_3 had moderate power in detecting the mild local dependence among the items loading also on the nuisance dimension, but the global OR and item covariance residual measures did not demonstrate enough power. Overall, all three pair-wise measures were sufficiently effective in detecting a large dependence among the items, but for the mild dependence condition, only Yen's Q_3 appeared to be powerful. These findings were similar to the findings from Levy (2006) in which the performance of PPMC in detecting the local dependence among the dichotomous items was examined.

In Condition 4, local dependence was modeled through a testlet effect, and the degree of testlet effect varied from mild ($\sigma^2_{d(i)} = 0.5$) though large ($\sigma^2_{d(i)} = 1.0$) to extremely large ($\sigma^2_{d(i)} = 2.0$). The results indicated that the three pair-wise measures had full power (1.00) in detecting the modeled dependence among responses to testlet items, even for the mild dependence case. In addition, as the dependence decreased, they did not seem to be a significant effect on the performance of the measures. As a result, more levels of testlet effect less than

$\sigma^2_{d(i)} = 0.5$ should be manipulated in order to explore how the effectiveness of the pair-wise measures changes and at what level of a testlet effect these measures would lose their power.

The power of the item-total score correlation measure in detecting the misfit of the GR model to the testlet items gradually increased from no power (0.00) to moderate power (0.52) to full power (1.00) as the degree of testlet dependence increased from the mild to large to extremely large. This indicates that the change of testlet effect had an influence on the performance of the item-total score correlation measure in the PPMC context. The test-level measure and other item-level measures appeared to be insensitive to this misfit.

Item-Fit

Condition 5 was designed to evaluate the ability of the PPMC method to assess the misfit of the GR model to items which did not conform to the GR model. One misfitting item had cubic BCC functions, and another misfitting item had two-step Guttman BCC functions.

Only two classical item-fit statistics (Yen's Q_1 and Stone's fit statistic) were found to be effective for detecting this type of item misfit. Stone's measure exhibited sufficient power to detect the two modeled misfitting items. Yen's Q_1 measure was found to have adequate power (0.65) for detecting the misfitting item with two-step Guttman BCC functions, but did not exhibit any power for the misfitting item with cubic BCC functions. Since only two types of BCC functions were considered and several factors were fixed in this study, the comparison of the performance of these two item-fit statistics in a Bayesian framework requires further investigation.

Summary for Study 1

For applications of Bayesian methods for assessing IRT model-fit, the choice of the discrepancy measures is important. Consistent with the findings from Levy (2006), the pair-wise

measures were found to be more powerful in detecting violations of unidimensionality and local independence assumptions than test- and item-level measures. This may be expected since the unidimensional GR model has no parameters to model the associations between responses to pairs of items, but the pair-wise measures can capture these associations. Among the three pair-wise measures, the directional measures (global OR and Yen's Q_3) may be preferred over a non-directional measure (absolute item covariance residual). In addition, Yen' Q_3 measure appeared to perform best. Though the item-total score correlation appeared to be more sensitive to large local dependence, power was low under mild local dependence cases. The test score distribution and item score distribution appeared least useful, as well as the two item-fit statistics, in detecting a violation of unidimensionality and local independence assumptions.

Regarding the item-fit assumption, only two classical item-fit statistics (Yen's Q_1 and Stone's) were found to be useful measures in detecting non-conforming to the GR model. It is worthwhile to note that there are different sources of item misfit. Condition 5 only considered item misfit due to the discrepancy from the true GR model curves. In Conditions 2-4, other sources of misfit for item were examined. Specifically, the item misfit in Condition 2 was due to multidimensionality, and the item misfit in Conditions 3-4 was due to local dependence. However, as seen from the results, these two item-fit measures did not exhibit any power in detecting item misfit due to multidimensionality or local dependence. This finding may seem surprising, but it is consistent with findings from previous research. For example, Zhang (2003) extended Orland and Thiseen (2000)'s item-fit statistics to multidimensional dichotomous IRT models, and examined their statistical properties. Though these item-fit statistics were found to exhibit adequate power for most conditions investigated in his study, they lacked power in all conditions when data were generated under 2-dim MIRT models but scaled by one-dimensional

IRT models. Another related study was conducted by Kang and Chen (2008). They generalized Orland and Thiseen's (2000) chi-square item-fit index for polytomous items, and evaluated its performance in assessing item-fit for the GR model. The results indicated that the power of this index was much lower when the misfit was due to multidimensionality or local dependence than when it was due to departure from the form of GR model boundary curves. They further found that 20,000 examinees were required to obtain acceptable power in detecting misfit items due to multidimensionality. Though the current study used a different design and conditions, the results confirmed the insensitiveness of the classical item-fit statistics to detect misfit due to multidimensionality or local dependence, even in the PPMC context.

The evaluation of fit of IRT models usually involves collecting a wide variety of evidence about different aspects of fit. Simulation Study 1 demonstrated that the PPMC method provides a framework to collect different kinds of information about model fit. Study 1 also illustrated that the extension of the use of PPMC from dichotomous IRT models to polytomous IRT models is flexible and straightforward. Many discrepancy measures for dichotomous models are also appropriate for the GR model.

Many results from this study are also consistent with previous research. As in several studies (e.g., Sinharay, 2005, 2006), a number of different types of graphical plots were used in this study in order to provide graphical evidence about model-fit. The use of graphical displays with PPMC is useful since the plots may be easier to understand and more appealing than tables of PPP-values. Another reason is that from plots, researchers may be able to discern patterns which may indicate an alternative model. For example, as shown in Condition 2, when a unidimensional GR model was estimated with 2-dim data, the pie plots displayed two clear item clusters, implying that a 2-dim model may be appropriate.

One disadvantage of the PPMC method is its conservativeness in evaluating model-fit. However, Sinharay (2006) argued that a conservative test with reasonable power is often better than a test that rejects too often. For example, as shown in the current study, Yen's Q_3 measure had close to uniform Type-I error rates (a little bit conservative), but had sufficient power in detecting multidimensionality and local dependence.

A practical consideration with PPMC applications is the intensive computation demands that are required. Nevertheless, as discussed by Sinharay (2006), once the posterior sample obtained during the estimation of a model is saved, the computation of each discrepancy measures and PPP-values based on this sample is not computationally demanding. More importantly, the stored sample values can be used in the future for different aspect of fit using different discrepancy measures.

5.1.2 Simulation Study 2

Study 2 was used to address the research question "Do the three Bayesian model-comparison indices (DIC, CPO, and PPMC) perform equally well in choosing a preferred GR model for a particular performance assessment application?" The results showed that for all the conditions examined in this study, these three indices appeared to perform equally in selecting the true model as the preferred model for *an overall test*. However, the CPO and PPMC indices were found to be more informative than the DIC index.

Specifically, DIC can only be used to choose an overall best model for an entire test, while the CPO index can be used to compare the models at either the test- or item-level. A model may be preferred at the test level but it may not necessarily be the preferred model for each item. As a result, comparing the models for each item using the item-level CPO index provides

additional information about model-fit. For example, in Conditions 3 and 4, the three indices indicated that a more complex GR model was preferred than a simple one-dimensional GR model for the overall test. But the results at the item-level using the CPO index indicated that the more complex model was only better for several items, and a simple unidimensional GR model might be adequate for the other items. One additional finding about the CPO index is that any trivial difference in CPO values between different models may not provide sufficient evidence supporting one model over another. In that situation, a more parsimonious model should be chosen.

Consistent with previous studies (Li et al., 2006; Sinharay, 2005), the PPMC approach was also found to be effective for performing model comparisons in this study. Moreover, the advantage of PPMC applications is in that they can be used to compare the relative fit of different models, but also evaluate the absolute fit of each individual model. In contrast, the DIC and CPO model-comparison tools only consider the relative fit of different models. They do not consider the absolute fit of each model. For example, two models, Model A and Model B, may be compared using the DIC and CPO indices. But it is not known whether either of these models fit the data. In addition, the graphical plots used with PPMC applications may provide some useful information regarding "what is the reason for misfit", "which items do not fit", and "which model is appropriate"?

It should also be noted that the results from this study indicate that the choice of discrepancy measures affects the performance of PPMC applications in comparing different models. If the measure is not effective, the PPMC method is less effective than the DIC and CPO indices. As shown in Conditions 3 and 4, when Yen's Q_3 measure was used with PPMC, the PPMC index performed equally well with DIC and CPO. However, when the global OR measure

was used with PPMC, its performance was less effective than the other two indices. Yen's Q_3 measure appeared to be more effective than the global OR measure for detecting violations in local dependence among items. Note that this conclusion was also obtained from Study 1.

It is also worthy to point out that the results in Condition 1 provided incremental evidence about the effectiveness of the proposed discrepancy measures beyond that found in Study 1. In Condition 1, the data was generated based on a 2P GR model, but three models were estimated: a 1P GR model, a 2P GR model and a RS model. The misfit of the 1P GR model and RS model to the simulated 2P GR item responses was examined using PPMC. The same discrepancy measures were employed as in Study 1 (except no test-level measure). This condition was not considered in Study 1. The results indicated that all 7 measures (4 item-level and 3 pair-wise) had sufficient power to detect the misfit of the RS model to the simulated 2P GR data. Six measures except "the item score distribution" were found to be very effective in detecting the misfit of the 1P GR model. It is worthy to note that the two item-fit measures exhibited adequate power to detect the item misfit due to the different unidimensional GR models.

5.1.3 Real Application

The methodology investigated in the two simulations was further applied to three datasets from the QCAI performance assessment. Overall, the results indicated that that these datasets were essentially unidimensional and exhibited local independence among items, and that a 2P GR model provided better model-fit than a 1P GR model. These findings were consistent with that from Lane et al. (1995).

The 2P GR model appeared to fit one dataset well regarding different aspects of fit such as dimensionality, item-fit, item/test score distribution, and item-test score correlations. However, for the other two datasets, though a GR model seemed appropriate in terms of most aspects of fit, several misfitting items were identified. Moreover, this model could not explain the test score distribution observed in one dataset.

Due to the conservativeness of PPMC applications, a higher level of significance of α = 0.10 was used to identified the misfitting items (Note that the previous studies used α = 0.05). Even with the higher level of significance, there were several items flagged as misfitting. These same items were also identified as misfitting in previous studies (Stone et al., 1993; Stone, 2000), but as shown in Table 3.18, the previous studies flagged more misfitting items than using PPMC with Stone's fit measure. Thus, Stone's fit statistics became more conservative in the PPMC context. In addition, the approach used by Stone et al. (1993) flagged more misfitting items than the approach used by Stone (2000). These results indicated that the method used by Stone (1993) for evaluating item-fit is relatively liberal. In contrast, the PPMC method used in the current study is relatively conservative. The method used by Stone (2000) appears to lie between these two approaches and yield results that are more reasonable for practical purposes.

Though Stone's fit measure identified several misfitting items, Yen's Q_1 measure did not flag any item as misfitting. The classical Yen's Q_1 index did not perform similarly to Stone's item-fit statistic. This may be due to the application with short tests where the imprecision in ability estimates can affect the use of more traditional measures of item fit such as Yen's Q_1 statistic. However, in the PPMC framework, the sampling distributions are based on simulations, and it is therefore still unclear why Yen's Q_1 measure did not show sufficient power. More research is needed in order to explain this finding.

In order to see if a more complex 2-dimensional model fit these QCAI datasets better than the unidimensional 2P GR and 1P GR models, three model-comparison indices were computed. The DIC index selected the 2-dimensional complex-structure model as the preferred model. However, based on the CPO and PPMC results, the unidimensional 2P GR model would be preferred. This conclusion that a unidimensional GR model was adequate for the datasets is consistent with the finding by Lane et al. (1995). The different results between the DIC index and the other indices indicated that the DIC index tends to select a more complex model. This finding is not uncommon for other information-based criteria such as the AIC (Akaike, 1974), and BIC (Schwarz, 1978).

5.2 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

This research used two Monte Carlo simulations to address the proposed research questions. Though the conditions were carefully designed and some factors were fixed at realistic values relative to typical performance assessments, the results may not generalize to other situations not considered in the current study. For example, this study is limited in terms of the length of tests (15 items), the number of response category (5-category), the polytomous model (GR), and the number of dimensions (2 dimensions).

Another limitation is that due to computing constraints of the WinBUGS program and a large number of conditions in this study, only 20 replications at each combination of experimental conditions were implemented. Though this is smaller than that other Monte Carlo simulations, it was reasonable in the context of previous research and Bayesian methods (e.g., a

number of researchers used 5 to 30 replications). However, more replications may be needed in order to obtain more reliable and accurate results.

In addition, the performance of the PPMC method and the Bayesian model-comparison indices for the GR models requires further study. For example, the effect of factors such as sample size, the number of total items, the number of dimensions, the structure of dimensions, and the inter-dimensional correlation given modeled multidimensionality could be further explored. For each condition investigated in the current work, a more comprehensive simulation study could be conducted in order to more fully explore how combination in factors affect the performance of PPMC and the effectiveness of the model-comparison indices.

Other discrepancy measures could also be proposed and evaluated. For example, the current research considered the global OR as one measure. As reviewed in Chapter 2, several previous studies also employed a conditional OR (MH) statistic as a discrepancy measure for dichotomous items. It is possible for future research to explore the use of the conditional global OR measure. The conditional OR may be more powerful than the global OR for checking the unidimensionality or local independence assumptions for polytomous items. Another useful discrepancy measure would be the Liu-Agresti estimate of the cumulative common odds ratio (Liu & Agresti, 1996) for ordinal variables. The global OR in the current study considered only one possible of dichotomization, while the cumulative common OR measure would consider all possible dichotomizations of the polytomous responses.

Furthermore, this study focused on evaluating the fit of IRT models relative to specific aspects of model fit: dimensionality, local independence, and the form of boundary curves in the GR model. Other assumptions underlying the use of IRT models with performance assessments

could be also considered in the future such as the normal ability assumption, and the non-speededness assumption.

Finally, the current study examined the general performance of some classical model-fit statistics used with PPMC. Further research is also needed in order to systematically compare the performance of these measures in the PPMC context and the classical framework. The PPMC method has several advantages when compared with the classical model-fit methods in theory, but the results from comprehensive simulation studies varying different conditions may provide useful guidelines about the use of PPMC. One possible comparison could involve various item-fit statistics. Several sources of item misfit could be modeled, and the misfit in both classical and Bayesian frameworks could be explored using traditional item-fit statistics such as Yen's Q_1 index, and some alternative item-fit indices such as Orlando and Thissen's fit statistics and Stone's statistics. In addition, the effect of smaller sample sizes could be explored since the Bayesian methods are often recommended for applications involving small sample sizes.

APPENDIX A

SAS CODE USED TO GENERATE UNIDIMENSIONAL GR DATA

```
*************************************************************************
* This sas code is used to generate the unidimensional graded responses
*************************************************************************

*  USER CONTROL VARIABLES;
%let ncat=5;
%let nthres=4;
%let nperson=2000;
%let nitem=15;

%let seed=0;

/*input the true item parameters */
data itempar;
  input a b1 b2 b3 b4;
  cards;
1.0 -2.0 -1.0 0.0 1.0
1.0 -1.5 -0.5 0.5 1.5
1.0 -1.0  0.0 1.0 2.0
1.0 -3.0 -1.5 -0.5 1.0
1.0 -1.0  0.5 1.5 3.0
1.7 -2.0 -1.0 0.0 1.0
1.7 -1.5 -0.5 0.5 1.5
1.7 -1.0  0.0 1.0 2.0
1.7 -3.0 -1.5 -0.5 1.0
1.7 -1.0  0.5 1.5 3.0
2.4 -2.0 -1.0 0.0 1.0
2.4 -1.5 -0.5 0.5 1.5
2.4 -1.0  0.0 1.0 2.0
2.4 -3.0 -1.5 -0.5 1.0
2.4 -1.0  0.5 1.5 3.0
 ;
run;

/*put all the item paramters in one row*/
data itempar;
set itempar;
```

```
array par{*} a b1-b&nthres;
do j=1 to &ncat;
      p=par{j};
      output;
end;
keep p;
run;

proc transpose out=itempar prefix=p;
var p;
run;

/*generate the graded responses (0 1 2 3 4) */
data resp;
set itempar;
array p{&nitem,&ncat} p1-p%eval(&nitem*&ncat);
array y{&nitem} y1-y&nitem;
array cumprob{&ncat} cumprob1-cumprob&ncat;

seed=&seed;

do i=1 to &nperson;
   call rannor(seed,theta);   /* Randomly generate theta value - normal(0,1)
*/
   *theta=0;    /*set all examinees at ability 0 to validate the data
generation */
   do j=1 to &nitem;
     do k=1 to &ncat;
         cumprob[k]=.;
       end;
     do resp=0 to (&ncat-1);
          do;  /*calculate the proprobility for each category*/
          if resp=(&ncat-1) then
              prob=1/(1+exp(-p[j,1]*(theta-p[j,&ncat])));
        else if resp=0 then
          prob=1-1/(1+exp(-p[j,1]*(theta-p[j,2])));
            else
          prob=1/(1+exp(-p[j,1]*(theta-p[j,resp+1])))
              -1/(1+exp(-p[j,1]*(theta-p[j,resp+2])));
        end;

        if resp=0 then cumprob[1]=prob; /*calculate the cumulative prob
                                              (the  prob  of  a  response  in
categories<=k)*/
            else cumprob[resp+1]=prob+cumprob[resp];
       end;

      call ranuni(seed,r01); /* Generate a random number between 0 and 1 */

        do k=1 to &ncat-1;
          if k=1 and r01<=cumprob[k] then
           y[j]=0;
            else if r01>cumprob[k] and r01<=cumprob[k+1] then
               y[j]=k;       /*response: 0, 1, 2, 3, 4* (5 categories)*/
       end;
   end;
   output;
```

```
   *file wrkdir(&responsefile);
   *put  (y1-y&nitem)(1.);
end;
keep y1-y&nitem;
run;

/*transform the responses (0 1 2 3 4)to (1, 2, 3, 4, 5,) format used in
Winbugs*/
data newresp;
set resp;
array y{*} y1-y&nitem;
do j=1 to &nitem;
 y[j]=y[j]+1;
end;
keep y1-y&nitem;
run;
```

APPENDIX B

WINBUGS CODE USED TO ESTIMATE UNIDIMENSIONAL GR MODELS

```
# Unidimensional Graded Response Model

model
{
# Specify unidimensional GR Model using Logistic function
  for (i in 1:nperson) {
      for(j in 1:nitem){
          for (k in 1:ncat-1) {
             logit(pstar[i, j, k]) <-a[j]*(theta[i]- b[j, k]);
          }

          p[i, j, 1] <- 1-pstar[i, j, 1]
          for(k in 2:ncat-1){
              p[i, j, k] <- pstar[i, j, k-1] - pstar[i, j, k]
          }
          p[i, j, ncat] <- pstar[i, j, ncat-1]

       y[i, j] ~ dcat(p[i, j, 1:ncat])

     }

   theta[i]~dnorm(0,1)

}

 #specify prior

for (j in 1:nitem) {
a[j] ~ dlnorm(0, 1)

b[j,1] ~ dnorm(0, 0.25)
for (k  in 1:ncat-2){
   b[j,k+1] ~ dnorm(0, .25) I(b[j, k], )
 }
}
}
```

APPENDIX C

WINBUGS CODE USED TO IMPLEMENT PPMC

```
# Unidimensional Graded Response Model
# Use PPMC method to check the model
# The discrepancy measures in this code include
#   (1) "Item Score Distribution"
#   (2) "Yen's Q3 Statistics"
#   (3) "Absolute Item Covariance Residual"
#   (4) "Global Odds Ratios"

model
{
# Specify unidimensional GR Model using Logistic function
  for (i in 1:nperson) {
      for(j in 1:nitem){
          for (k in 1:ncat-1) {
             logit(pstar[i, j, k]) <-a[j]*(theta[i]- b[j, k]);
          }

          p[i, j, 1] <- 1-pstar[i, j, 1]
          for(k in 2:ncat-1){
              p[i, j, k] <- pstar[i, j, k-1] - pstar[i, j, k]
          }
          p[i, j, ncat] <- pstar[i, j, ncat-1]

       y[i, j] ~ dcat(p[i, j, 1:ncat])

       # compute CPO for observed item responses
       inprob[i, j] <- pow(p[i, j, y[i,j] ], -1)

        # replicated response data
       yrep[i, j] ~ dcat(p[i, j, 1:ncat])
     }

   theta[i]~dnorm(0,1)

}
```

```
 #specify prior

for (j in 1:nitem) {
a[j] ~ dlnorm(0, 1)

b[j,1] ~ dnorm(0, 0.25)
for (k  in 1:ncat-2){
   b[j,k+1] ~ dnorm(0, .25) I(b[j, k], )
 }
}

# (1) calculate the chi-sqaure statistic for item score distribution
for(j in 1:nitem){
   for (k in 1:ncat) {
      for (i in 1:nperson) {
        count_obs[i,j,k] <- equals(y[i,j], k)
        count_rep[i,j,k] <- equals(yrep[i,j], k)
      }
   n[j,k] <- sum(count_obs[ ,j,k])     # observed number of examinees having responses (k-1) (i.e. in
category k) on
                                                                                                  # item j
for observed data
   n_rep[j,k] <- sum(count_rep[ ,j,k]) # observed number of examinees having responses (k-1) on item j for
                                       # replicated data

   En[j,k] <- sum(p[,j,k])               # the expected number of examinees having responses (k-1) (i.e. in
category k)
                                     # on item j

   resid[j,k] <- pow(n[j,k]-En[j,k], 2)/(En[j,k]+0.0001*equals(En[j,k],0))
   resid_rep[j,k] <- pow(n_rep[j,k]-En[j,k], 2)/(En[j,k]+0.0001*equals(En[j,k],0))

}

itemchi2[j] <- sum(resid[j, ])      # the "realized" chi-square item-fit statistic

itemchi2_rep[j] <- sum(resid_rep[j, ])      # the "predicted" chi-square item-fit statistic

PPP.itemchi2[j] <-  step(itemchi2_rep[j]-itemchi2[j])  # the posterior predictive P-values for each item
}


# (2) Yen's Q3 Statistic

for (i in 1:nperson) {
   for(j in 1:nitem){
      for (k in 1:ncat) {
         xx[i,j,k] <- (k-1)*p[i,j,k]
      }
         E[i,j] <- sum(xx[i,j, ])   # expected item response
         r.obs[i,j] <- y[i,j]-E[i,j]    # the residual for observed data
         r.rep[i,j] <- yrep[i,j]-E[i,j]    # the residual for replicated data
}}

for(j in 1:nitem){
      r.obs.mean[j] <- mean(r.obs[1:nperson, j])  # the mean of the residulas for item j for observed data
```

```
    r.obs.sd[j] <-  sd(r.obs[1:nperson, j])          # the sd of the residulas for item j

    r.rep.mean[j] <- mean(r.rep[1:nperson, j])  # the mean of the residulas for item j for replicated data
    r.rep.sd[j] <-  sd(r.rep[1:nperson, j])          # the sd of the residulas for item j
}

for(j1 in 1:(nitem-1)){
 for(j2 in (j1+1):nitem){

  Q3.obs[j1,j2]        <-        (inprod(r.obs[1:nperson,        j1],        r.obs[1:nperson,        j2])        -
nperson*r.obs.mean[j1]*r.obs.mean[j2])/((nperson-1)*r.obs.sd[j1]*r.obs.sd[j2])      #Q3 for observed data

  Q3.rep[j1,j2]        <-        (inprod(r.rep[1:nperson,        j1],        r.rep[1:nperson,        j2])        -
nperson*r.rep.mean[j1]*r.rep.mean[j2])/((nperson-1)*r.rep.sd[j1]*r.rep.sd[j2])      #Q3 for replicated data

  PPP.Q3[j1,j2] <- step(Q3.rep[j1,j2] - Q3.obs[j1,j2])     #PPP values
}}
# (3) Absolute Item Residual Covariance
for(j in 1:nitem){
            y.mean[j] <- mean(y[1:nperson, j])
            yrep.mean[j] <- mean(yrep[1:nperson, j])

            E.mean[j] <- mean(E[1:nperson, j])
}

for(j1 in 1:(nitem-1)){
 for(j2 in (j1+1):nitem){

# sample item covariance
 S2.obs[j1,j2] <- (inprod(y[1:nperson, j1], y[1:nperson, j2]) - nperson*y.mean[j1]*y.mean[j2])/(nperson-1)
 S2.rep[j1,j2]        <-        (inprod(yrep[1:nperson,        j1],        yrep[1:nperson,        j2])        -
nperson*yrep.mean[j1]*yrep.mean[j2])/(nperson-1)
# model-based item covariance
  sigma2[j1,j2] <- (inprod(E[1:nperson, j1], E[1:nperson, j2]) - nperson*E.mean[j1]*E.mean[j2])/nperson

# Absolute Residuals between sample and model-based item covariance for each item pair
  residcov.obs[j1, j2] <- abs(S2.obs[j1, j2] - sigma2[j1, j2])     # for the observed data
  residcov.rep[j1, j2] <- abs(S2.rep[j1, j2] - sigma2[j1, j2])     # for the replicated data

PPP.residcov[j1, j2] <- step( residcov.rep[j1, j2] - residcov.obs[j1, j2])

}}

# (4) Global Odds Ratio

# Firstly, dichotomize the response data (the cut scores for each item is based on rubric
for(i in 1:nperson){
  for(j in 1:nitem){
    y.di[i,j] <- step(y[i,j]-cutscore[j])      # dichotomize the observed response based on cutscore
    yrep.di[i,j] <- step(yrep[i,j]-cutscore[j])    # dichotomize the replicated response
  }
}

for(i in 1:nperson){
  for(j in 1:nitem){
    x.di[i,j]<- 1-y.di[i,j]                # the intemedium variables used for computing OR below
```

```
    xrep.di[i,j]<- 1-yrep.di[i,j]
}}

# Compute the Global Odds Ratio
for(j1 in 1:(nitem-1)){
  for(j2 in (j1+1):nitem){
    OR[j1, j2] <- inprod(y.di[1:nperson,j1], y.di[1:nperson,j2]) * inprod(x.di[1:nperson,j1], x.di[1:nperson,j2])
/ (inprod(y.di[1:nperson,j1], x.di[1:nperson,j2]) * inprod(x.di[1:nperson,j1], y.di[1:nperson,j2]))

    OR.rep[j1, j2] <- inprod(yrep.di[1:nperson,j1], yrep.di[1:nperson,j2]) * inprod(xrep.di[1:nperson,j1],
xrep.di[1:nperson,j2]) / (inprod(yrep.di[1:nperson,j1], xrep.di[1:nperson,j2]) * inprod(xrep.di[1:nperson,j1],
yrep.di[1:nperson,j2]))

     PPP.OR[j1, j2] <- step(OR.rep[j1,j2] - OR[j1,j2])
}}

}
```

APPENDIX D

SAS CODE USED TO CREATE A BATCH FILE TO RUN PPMC FROM SAS

```
************************
* Creat the Batch file  *
************************;
%include 'C:\dissertation\study1\SASBUGS_Macro\*.sas';
FILENAME GRM 'C:\dissertation\study1\GRM';
*FILENAME bugsloc 'c:\program files\winbugs14'; /*used for window XP*/
FILENAME bugsloc 'c:\winbugs14';  /*used for window vista*/

/*Scripts to run WinBUGS*/

DATA _NULL_;
FILE bugsloc(grmBatch.txt);
PUT@1 "display('log')";
PUT@1 "check('C:/dissertation/study1/GRM/GRmodel.txt')" ;
PUT@1 "data('C:/dissertation/study1/GRM/data1.txt')";
PUT@1 "data('C:/dissertation/study1/GRM/data2.txt')";
PUT@1 "compile(1)";
PUT@1 "gen.inits()";
PUT@1 "update(4000)";
PUT@1 "set(a)";
PUT@1 "set(b)";
PUT@1 "set(yrep)";
PUT@1 "set(theta)";
PUT@1 "set(itemchi2)";
PUT@1 "set(itemchi2_rep)";
PUT@1 "set(PPP.itemchi2)";
PUT@1 "set(Q3.obs)";
PUT@1 "set(Q3.rep)";
PUT@1 "set(PPP.Q3)";
PUT@1 "set(residcov.obs)";
PUT@1 "set(residcov.rep)";
PUT@1 "set(PPP.residcov)";
PUT@1 "set(OR.rep)";
PUT@1 "set(PPP.OR)";
PUT@1 "set(inprob)";
PUT@1 "dic.set()";
PUT@1 "update(1000)";
```

```
PUT@1 "thin.samples(2)";
PUT@1 "stats(a)";
PUT@1 "stats(b)";
PUT@1 "stats(PPP.itemchi2)";
PUT@1 "stats(PPP.Q3)";
PUT@1 "stats(PPP.residcov)";
PUT@1 "stats(PPP.OR)";
PUT@1 "stats(inprob)";
PUT@1 "dic.stats()";
PUT@1 "save('C:/dissertation/study1/GRM/log.txt')";
PUT@1 "coda(a,'C:/dissertation/study1/GRM/coda_a.txt')";
PUT@1 "coda(b,'C:/dissertation/study1/GRM/coda_b.txt')";
PUT@1 "coda(theta,'C:/dissertation/study1/GRM/coda_theta.txt')";
PUT@1 "coda(yrep,'C:/dissertation/study1/GRM/coda_yrep.txt')";
PUT@1         "coda(itemchi2,'C:/dissertation/study1/GRM/coda_itemchi2.txt')";
PUT@1"coda(itemchi2_rep,'C:/dissertation/study1/GRM/coda_itemchi2rep.txt')";
PUT@1 "coda(Q3.obs,'C:/dissertation/study1/GRM/coda_Q3obs.txt')";
PUT@1 "coda(Q3.rep,'C:/dissertation/study1/GRM/coda_Q3rep.txt')";
PUT@1 "coda(residcov.obs,'C:/dissertation/study1/GRM/coda_residobs.txt')";
PUT@1 "coda(residcov.rep,'C:/dissertation/study1/GRM/coda_residrep.txt')";
PUT@1 "coda(OR.rep,'C:/dissertation/study1/GRM/coda_ORrep.txt')";
PUT@1 "quit()";
RUN;

/*create a batch file*/
DATA _NULL_;
FILE GRM(run_c1GRM.bat);
*PUT '"C:\program files\WinBUGS14\WinBUGS14.exe" /PAR grmBatch.txt';
PUT '"C:\WinBUGS14\WinBUGS14.exe" /PAR grmBatch.txt';
PUT 'exit';
RUN;
```

APPENDIX E

WINBUGS CODE USED TO ESTIMATE 2-DIMENSIONAL GR MODELS

```
model
{
# Specify simple-structure 2-dim GR Model
  for (i in 1:nperson) {
      for(j in 1:nitem1){
          for (k in 1:ncat-1) {
             logit(pstar[i, j, k]) <-a[j]*(theta[i,1]- b[j, k]);
          }
       }

      for(j in (nitem1+1):nitem){
          for (k in 1:ncat-1) {
             logit(pstar[i, j, k]) <-a[j]*(theta[i,2]- b[j, k]);
          }
       }

       for(j in 1:nitem){
          p[i, j, 1] <- 1-pstar[i, j, 1]
          for(k in 2:ncat-1){
              p[i, j, k] <- pstar[i, j, k-1] - pstar[i, j, k]
          }
          p[i, j, ncat] <- pstar[i, j, ncat-1]

       y[i, j] ~ dcat(p[i, j, 1:ncat])

     }

     theta[i,1:2]~dmnorm(mu[1:2], tau[1:2, 1:2])
}

 #specify prior

for (j in 1:nitem) {
a[j] ~ dlnorm(0, 1)

b[j,1] ~ dnorm(0, 0.25)
```

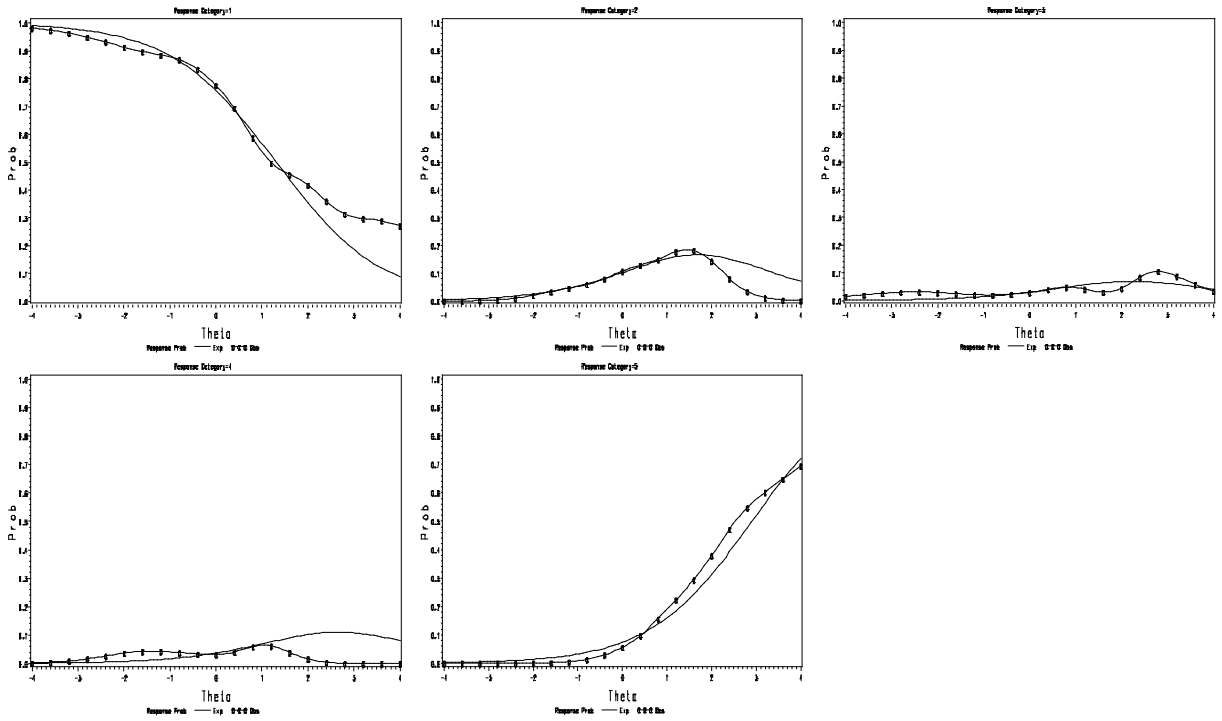
```
for (k  in 1:ncat-2){
   b[j,k+1] ~ dnorm(0, .25) I(b[j, k], )
 }
}

tau[1:2, 1:2] <- inverse(sigma[1:2, 1:2])

sigma[1,1] <- 1
sigma[2,2] <- 1
sigma[1,2] <- corr
sigma[2,1] <- corr
corr ~ dnorm(0.6,4) I(0,)
}
```

APPENDIX F

ITEM CATEGORY CURVES (ICCS) FOR THE QCAI ITEMS

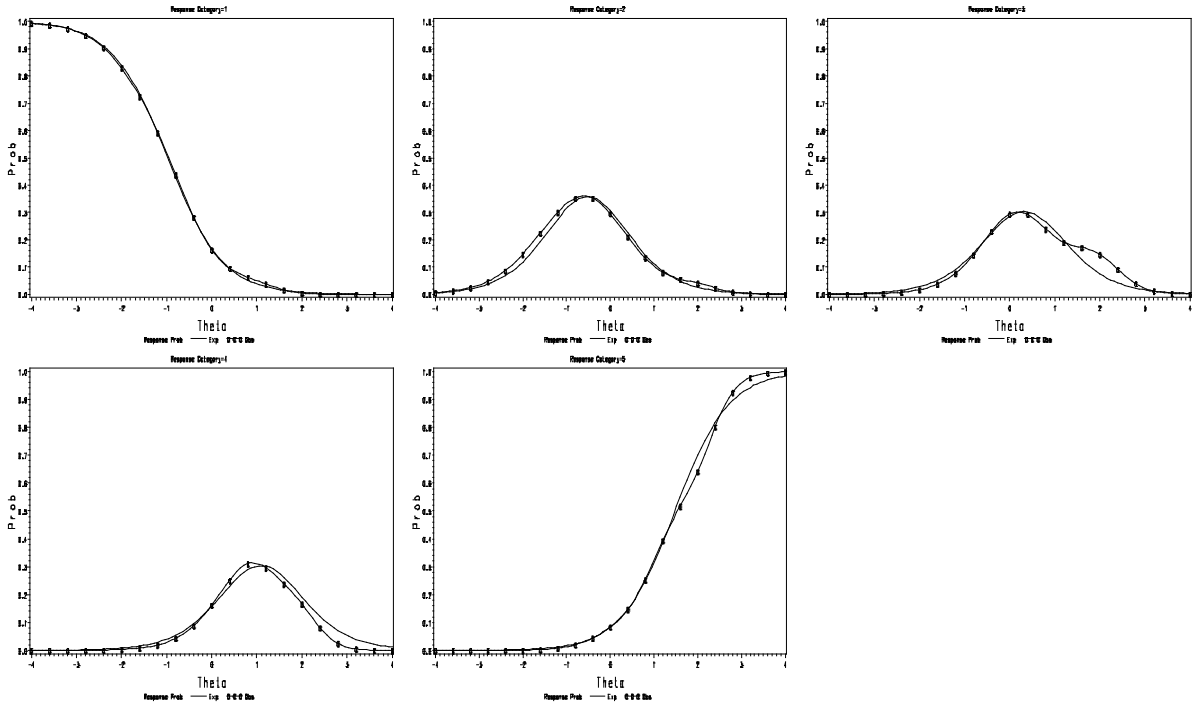
Test Form "AS91"

Item 1:



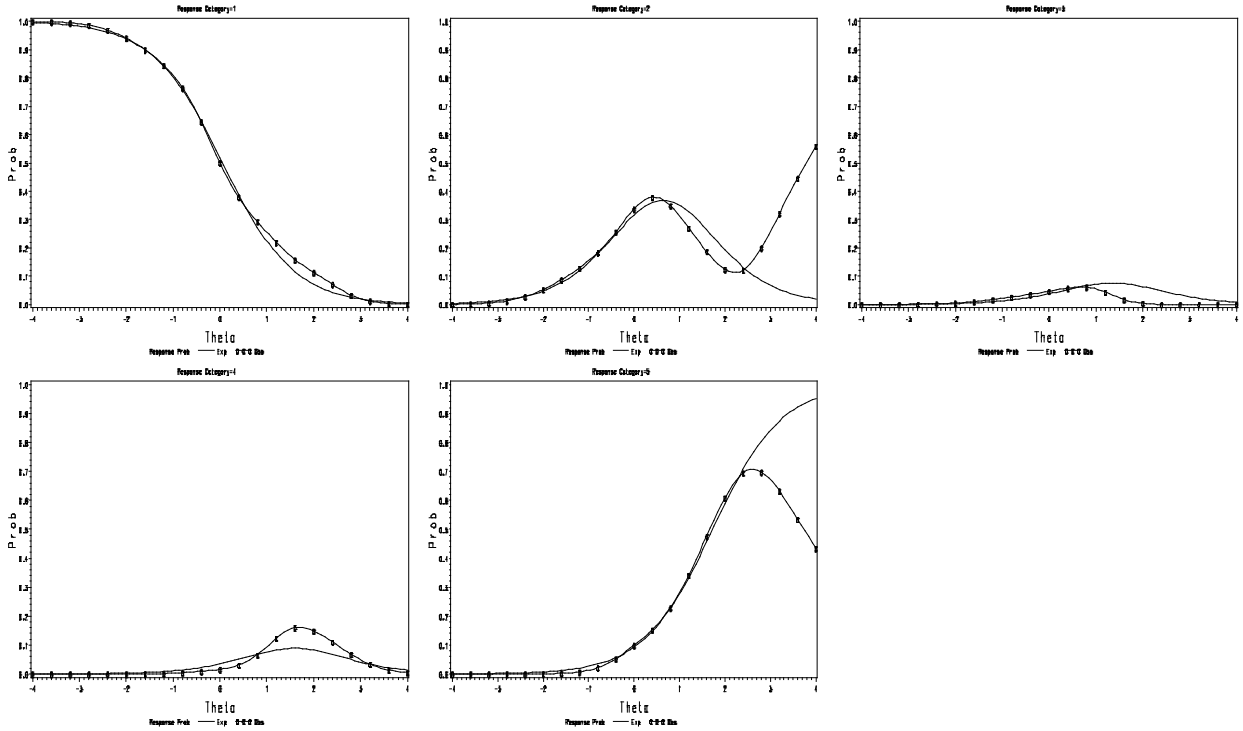
Response Probability Plots for Item 1

Item 2:



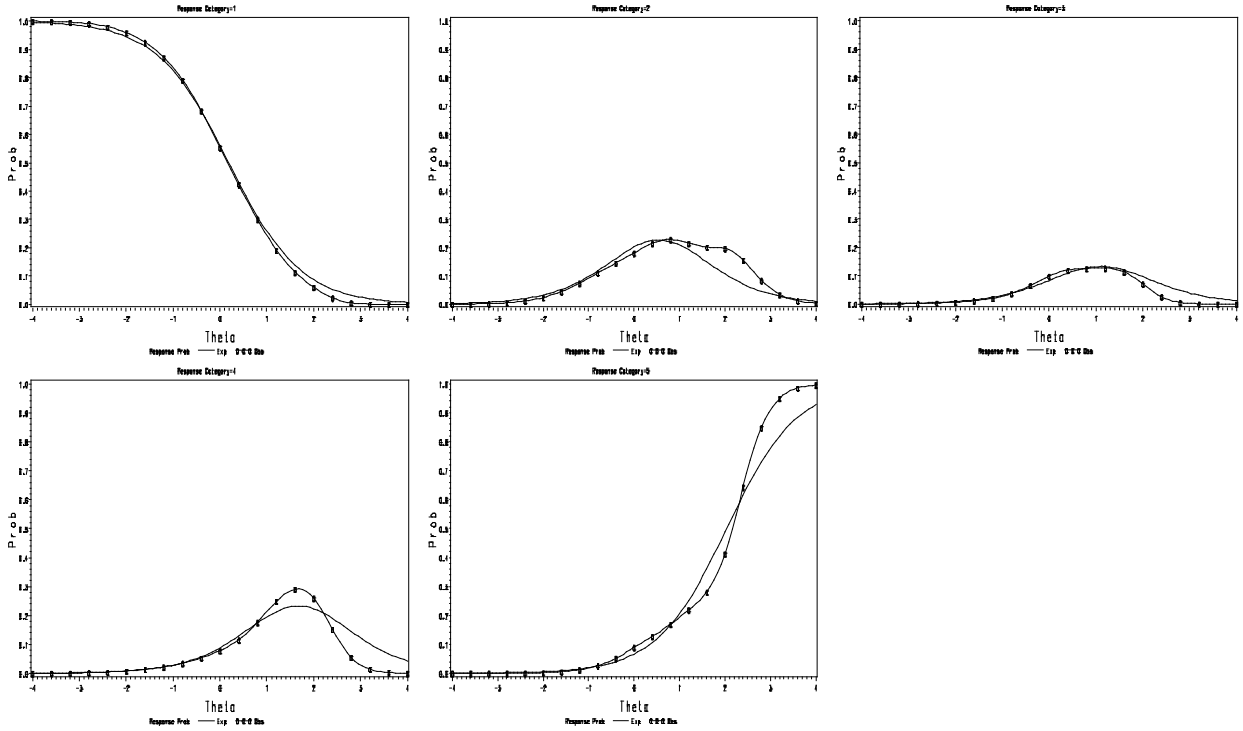
Response Probability Plots for Item 2

Item 3:



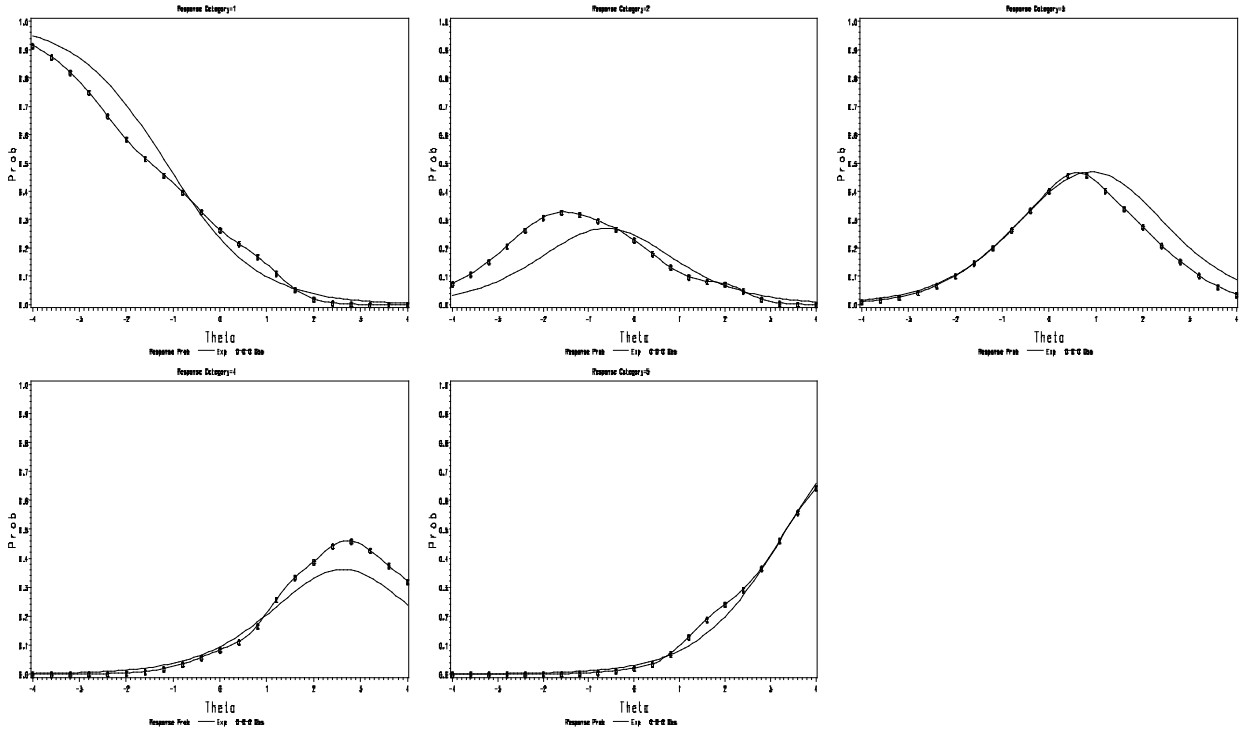
Response Probability Plots for Item 3

Item 4:



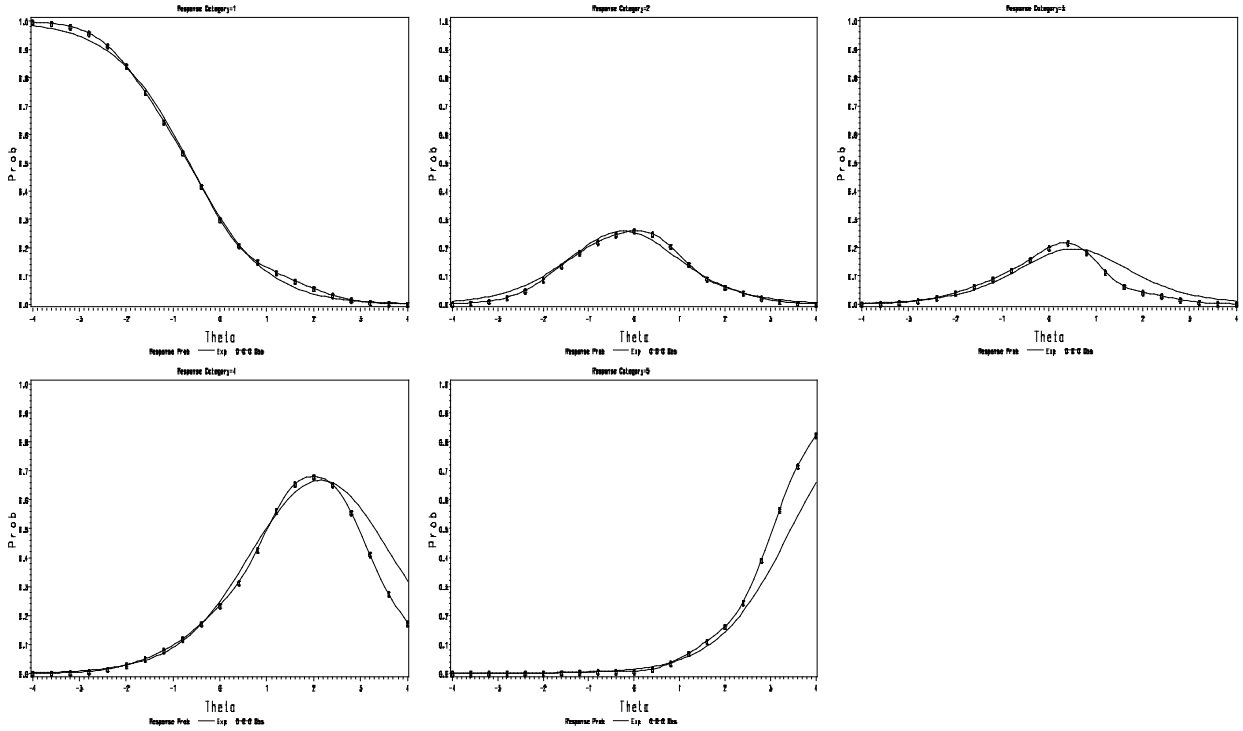
Response Probability Plots for Item 4

Item 5:



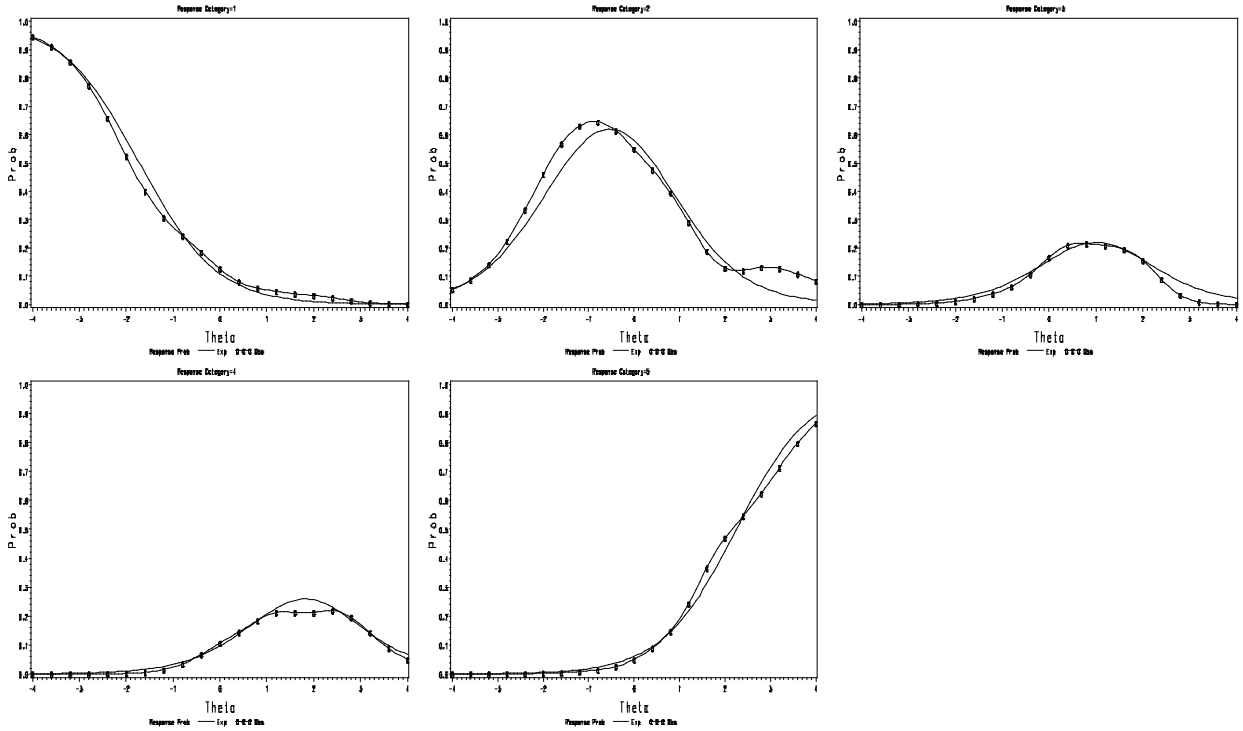
Response Probability Plots for Item 5

Item 6:



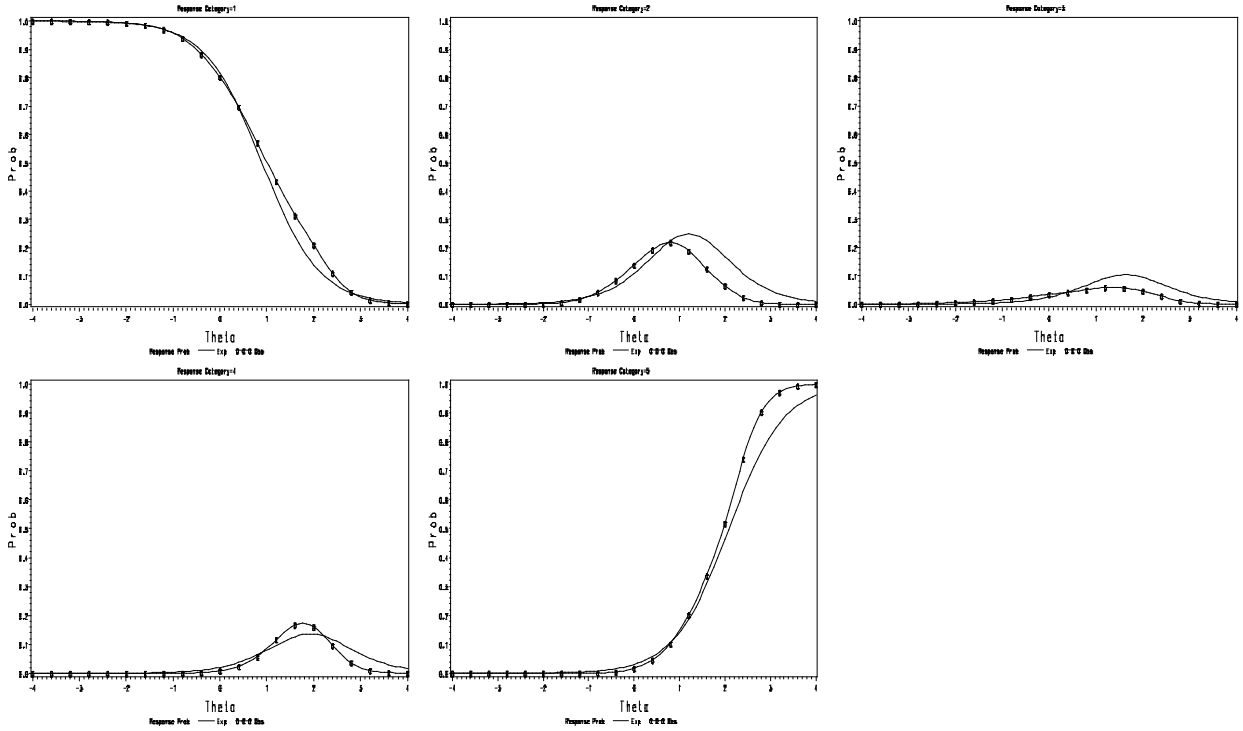
Response Probability Plots for Item 6

Item 7:



Response Probability Plots for Item 7

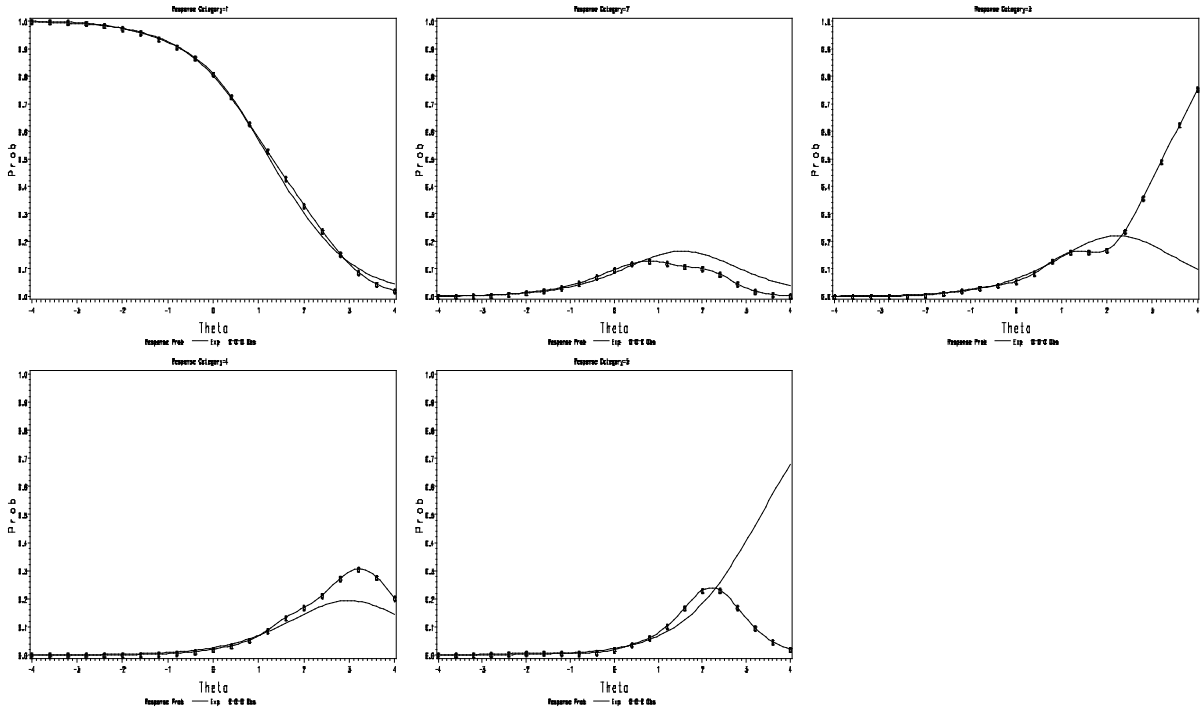
Item 8:



Response Probability Plots for Item 8

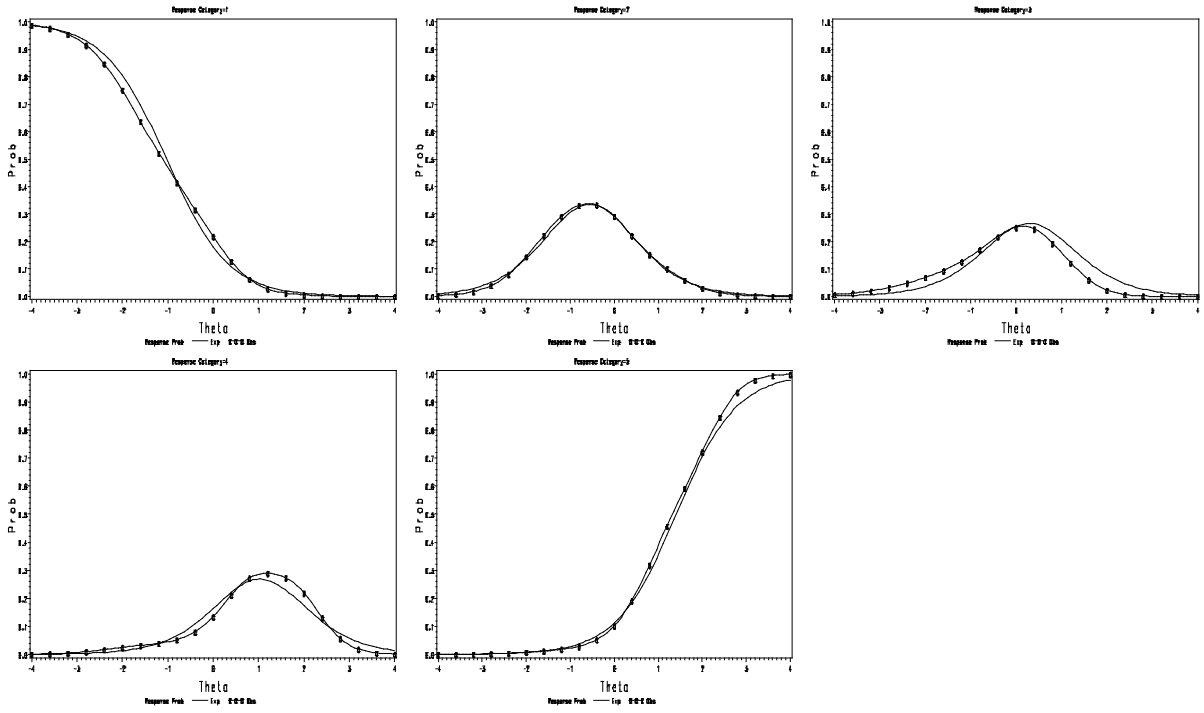
Test Form "AS92"

Item 1:



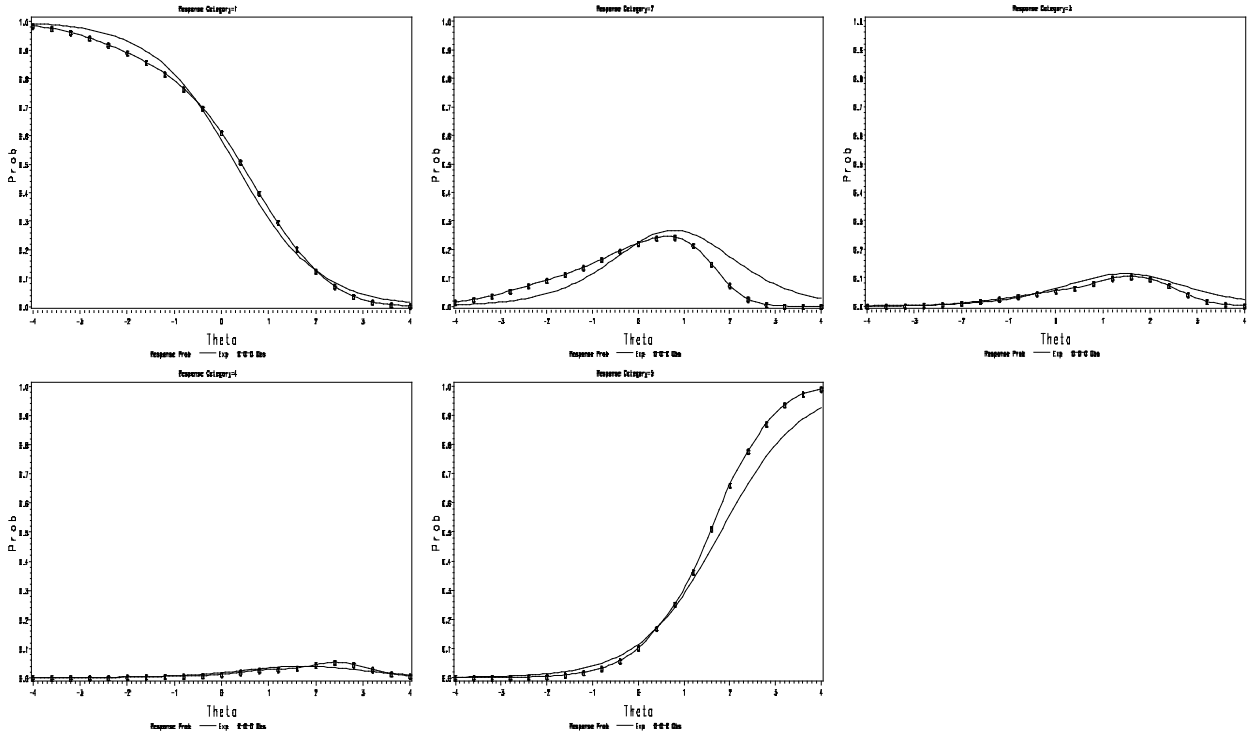
Response Probability Plots for Item 1

Item 2:



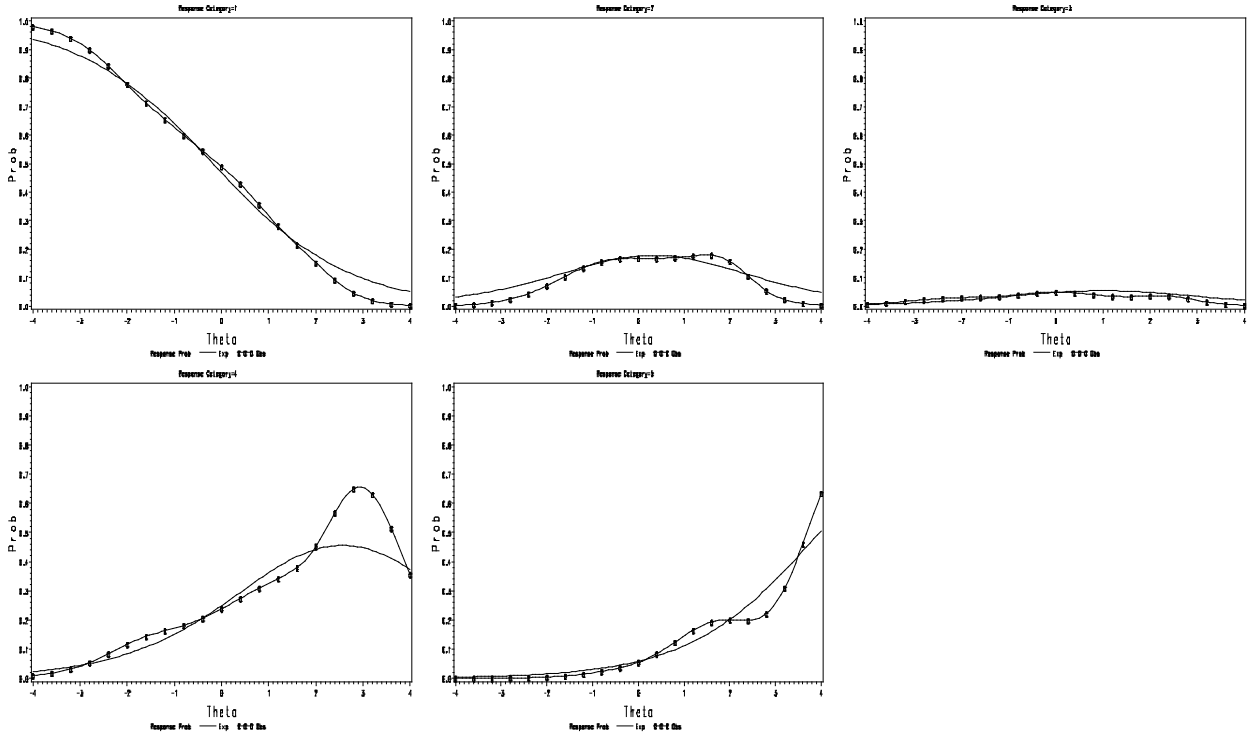
Response Probability Plots for Item 2

Item 3:



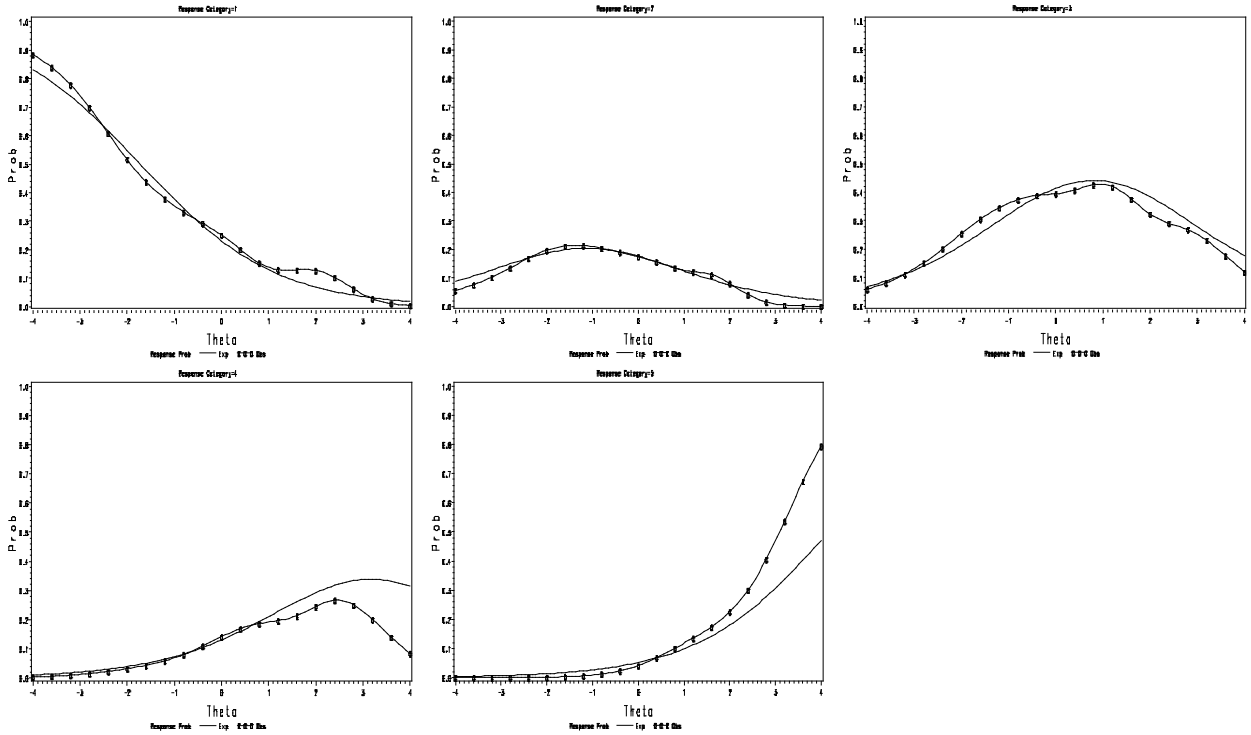
Response Probability Plots for Item 3

Item 4:



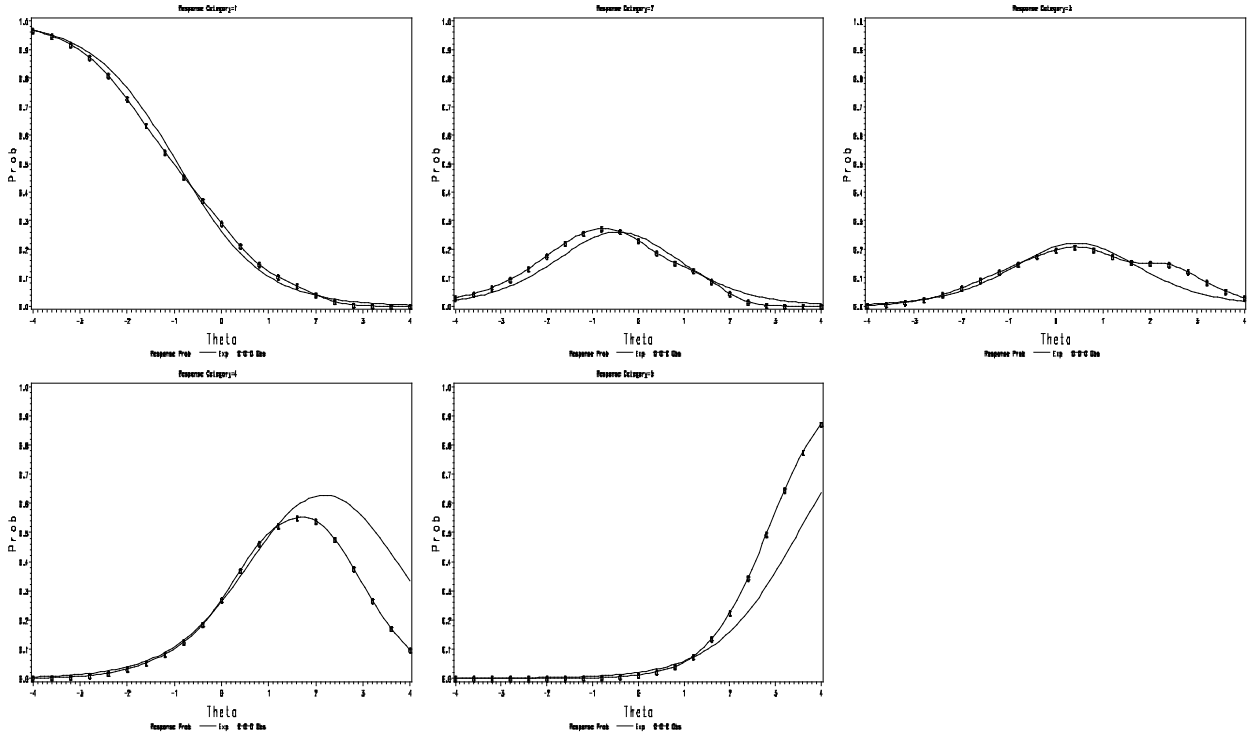
Response Probability Plots for Item 4

Item 5:



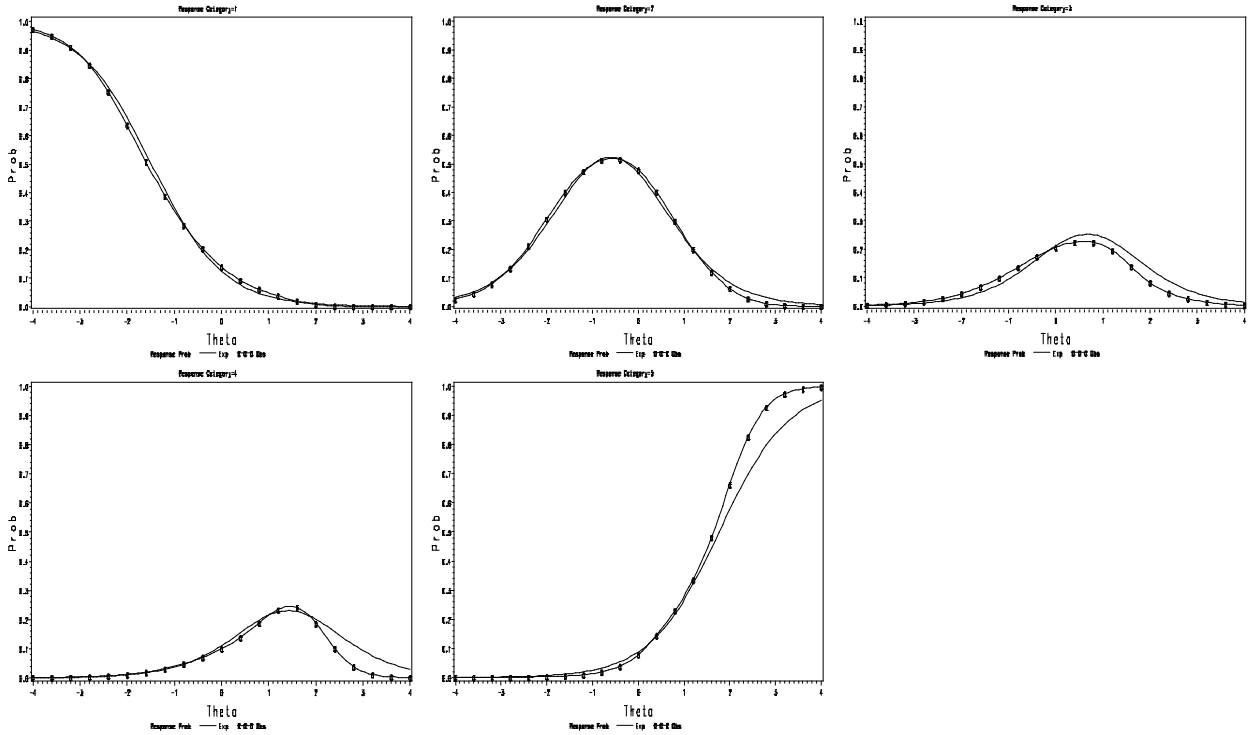
Response Probability Plots for Item 5

Item 6:



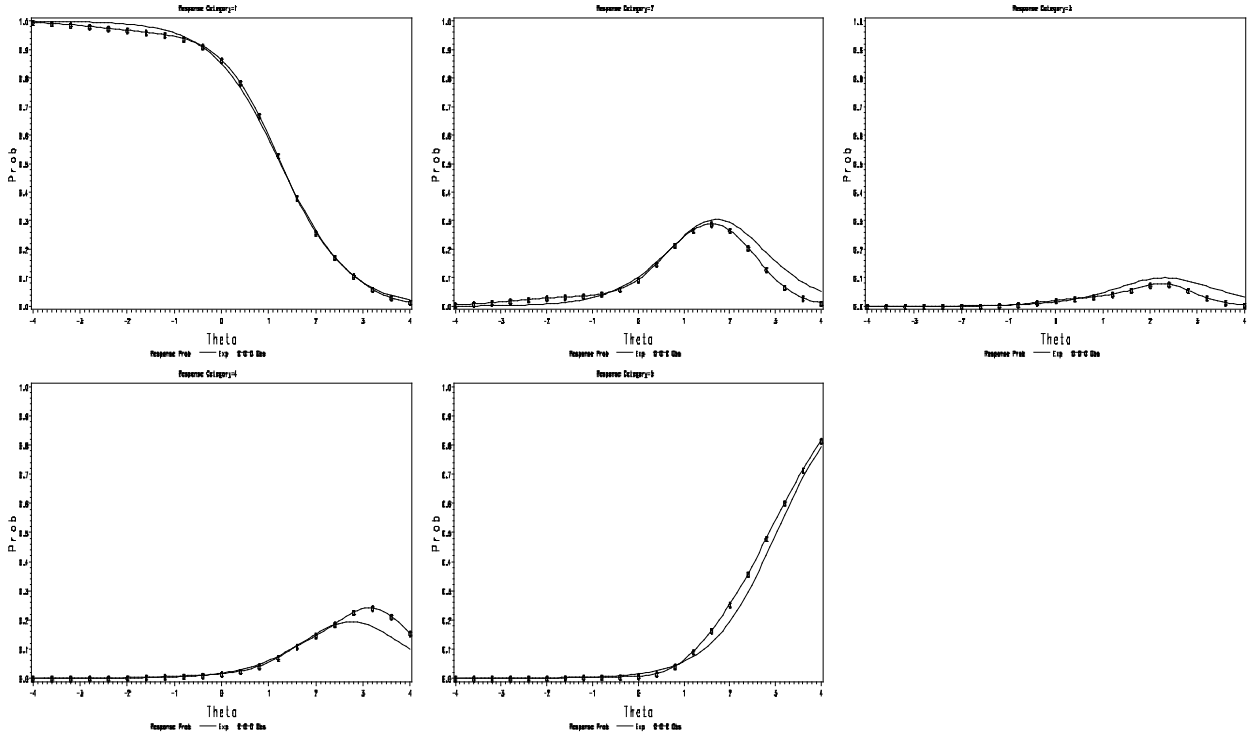
Response Probability Plots for Item 6

Item 7:



Response Probability Plots for Item 7

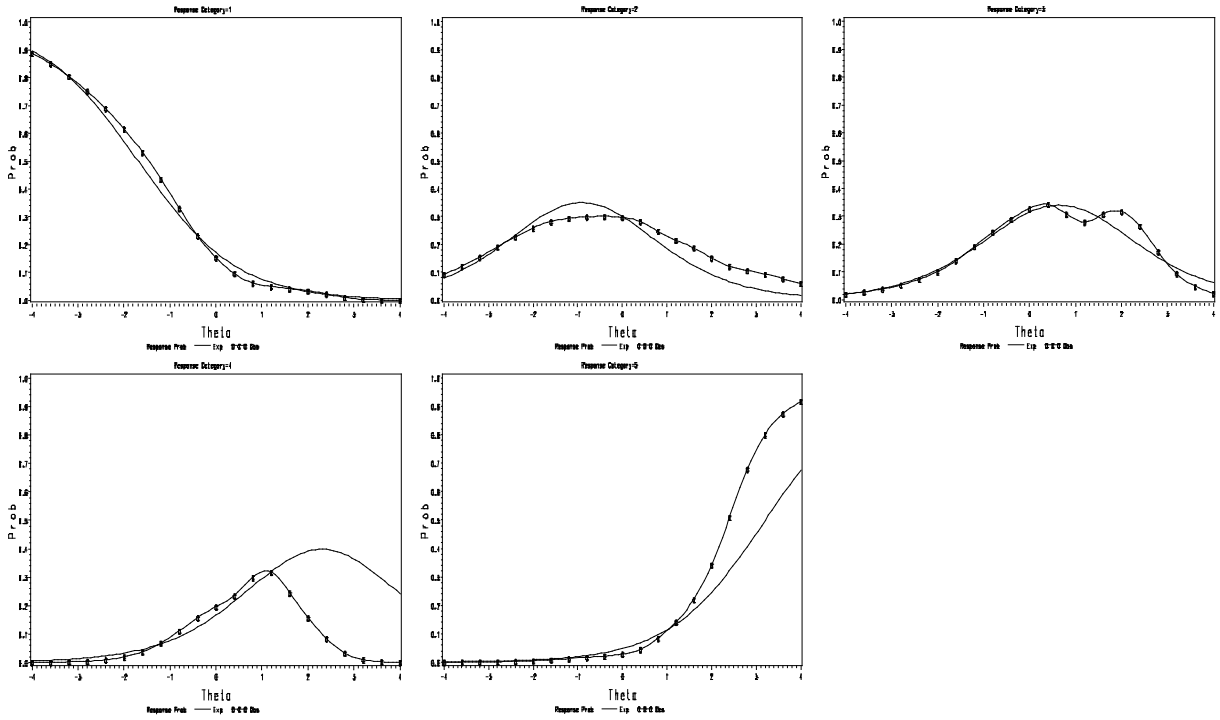
Item 8:



Response Probability Plots for Item 8

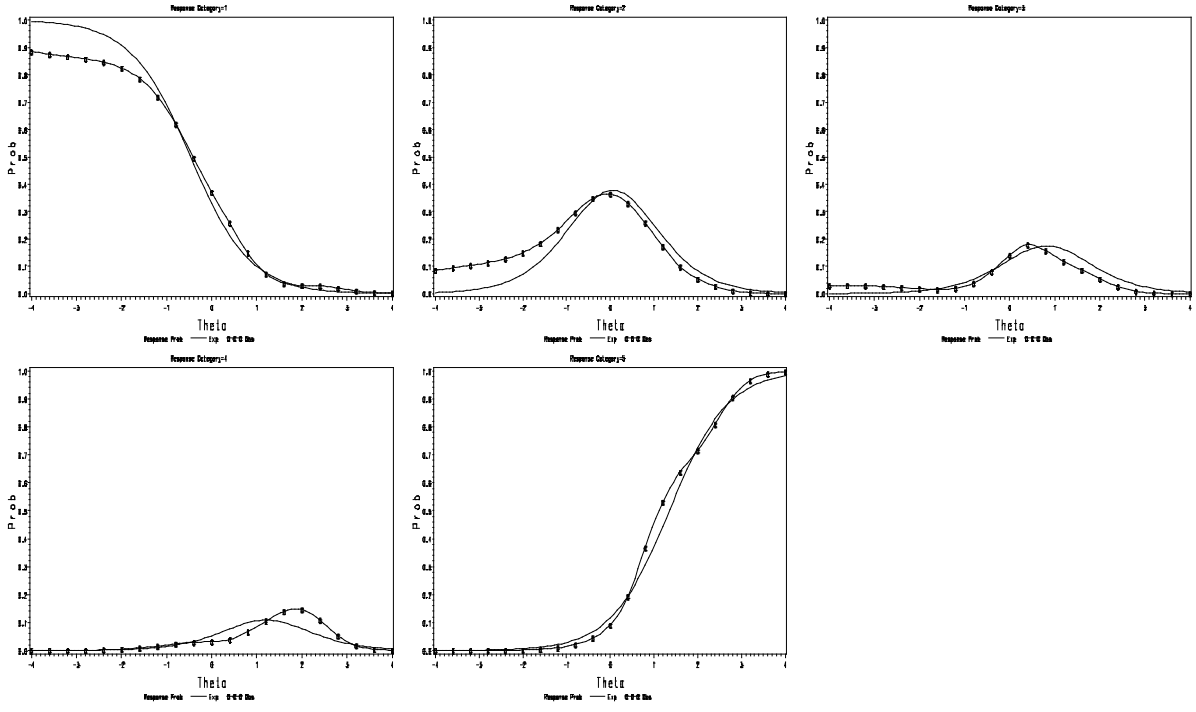
Test Form "BS92"

Item 1:



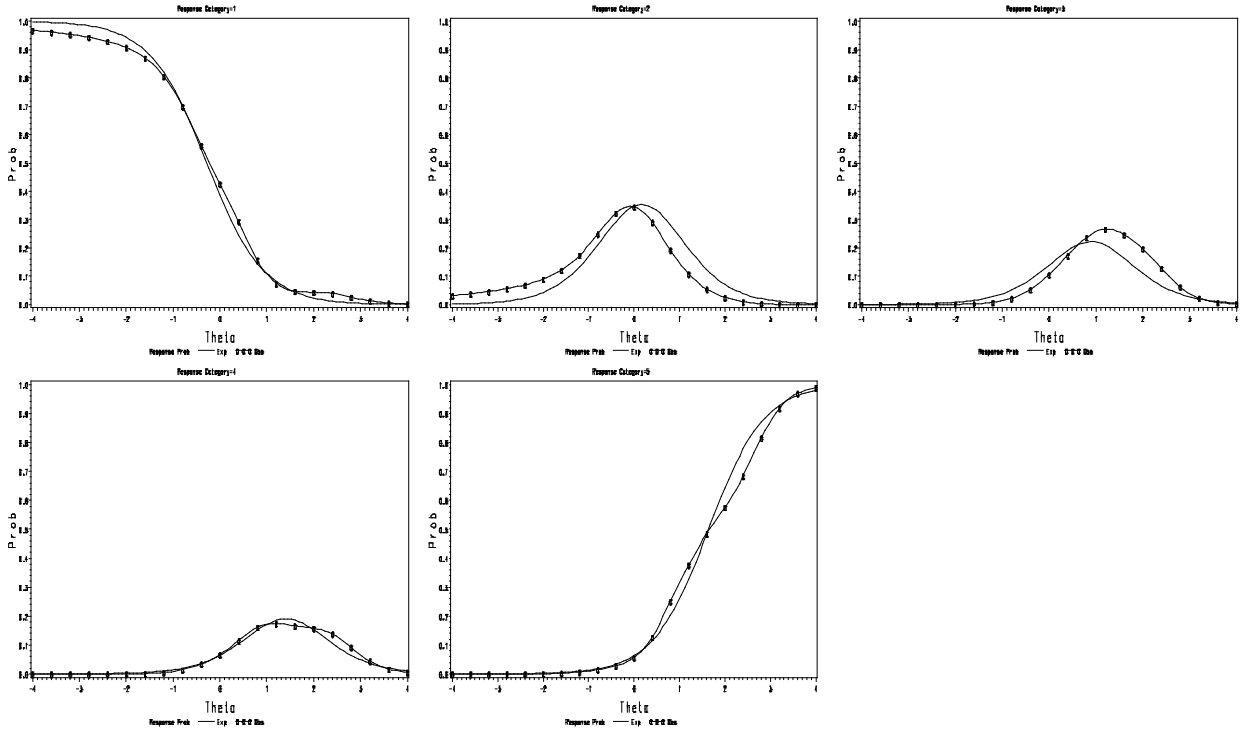
Response Probability Plots for Item 1

Item 2:



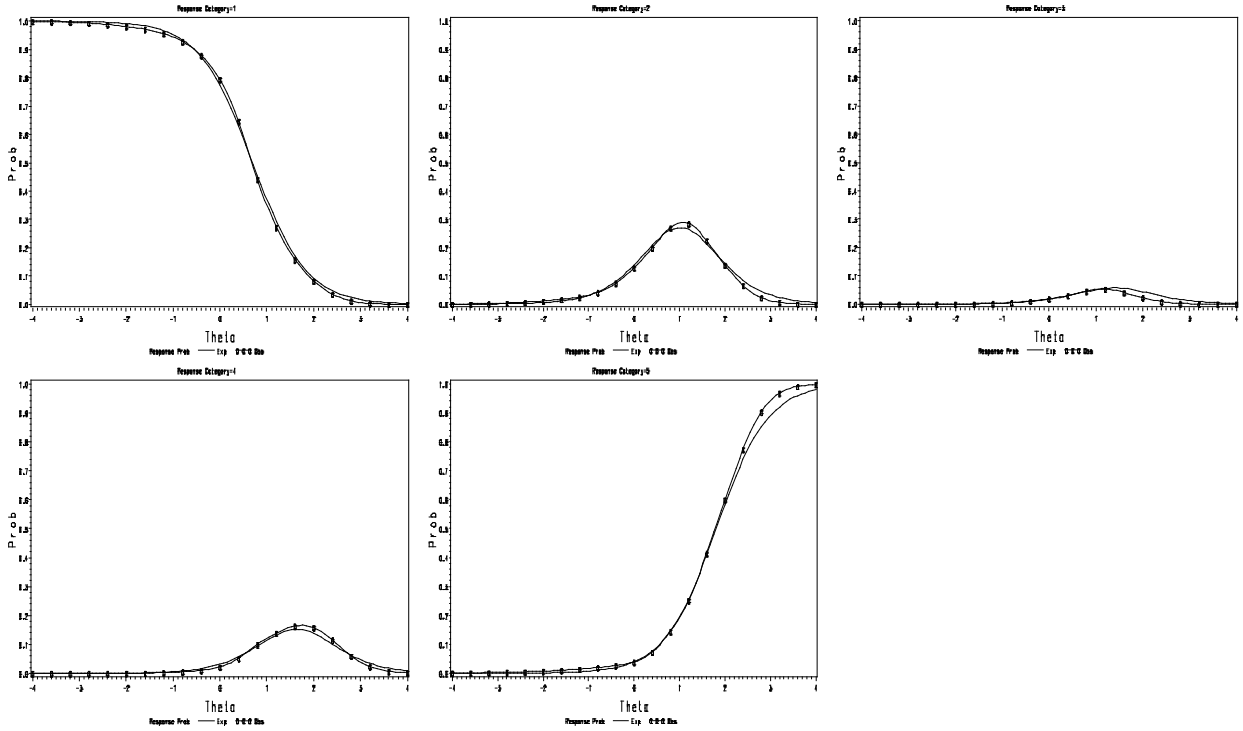
Response Probability Plots for Item 2

Item 3:



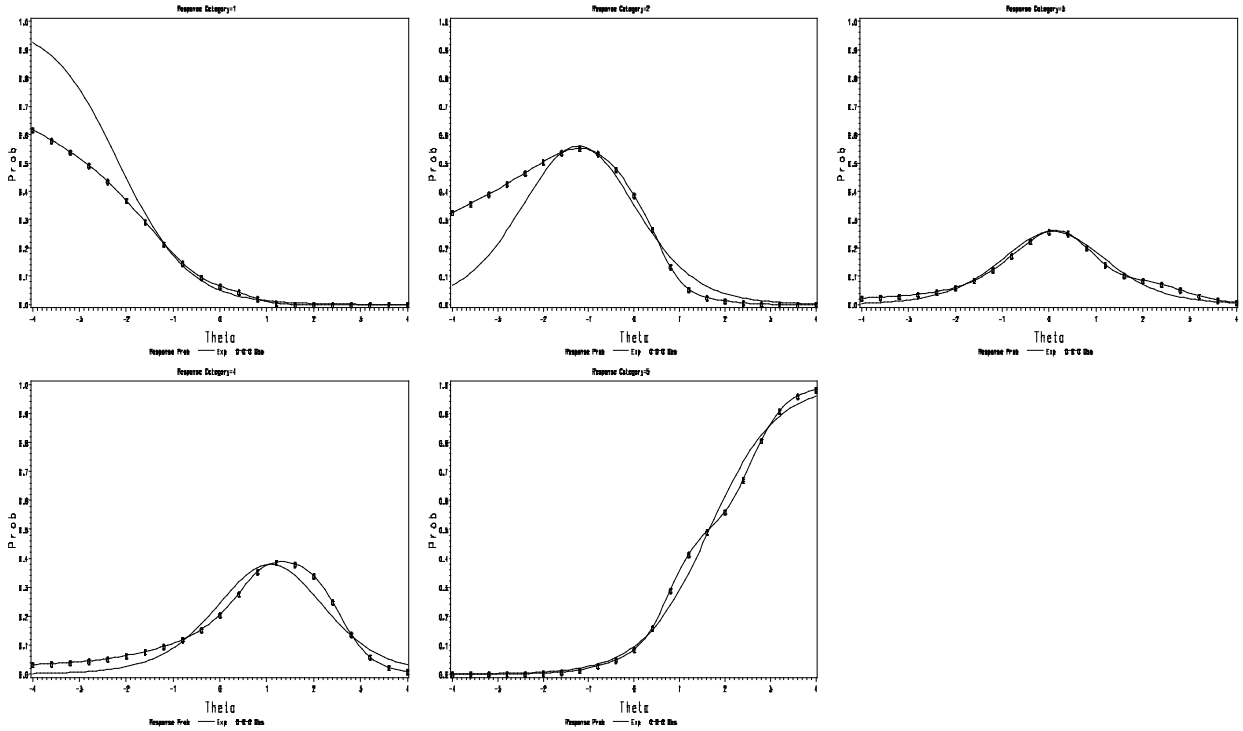
Response Probability Plots for Item 3

Item 4:



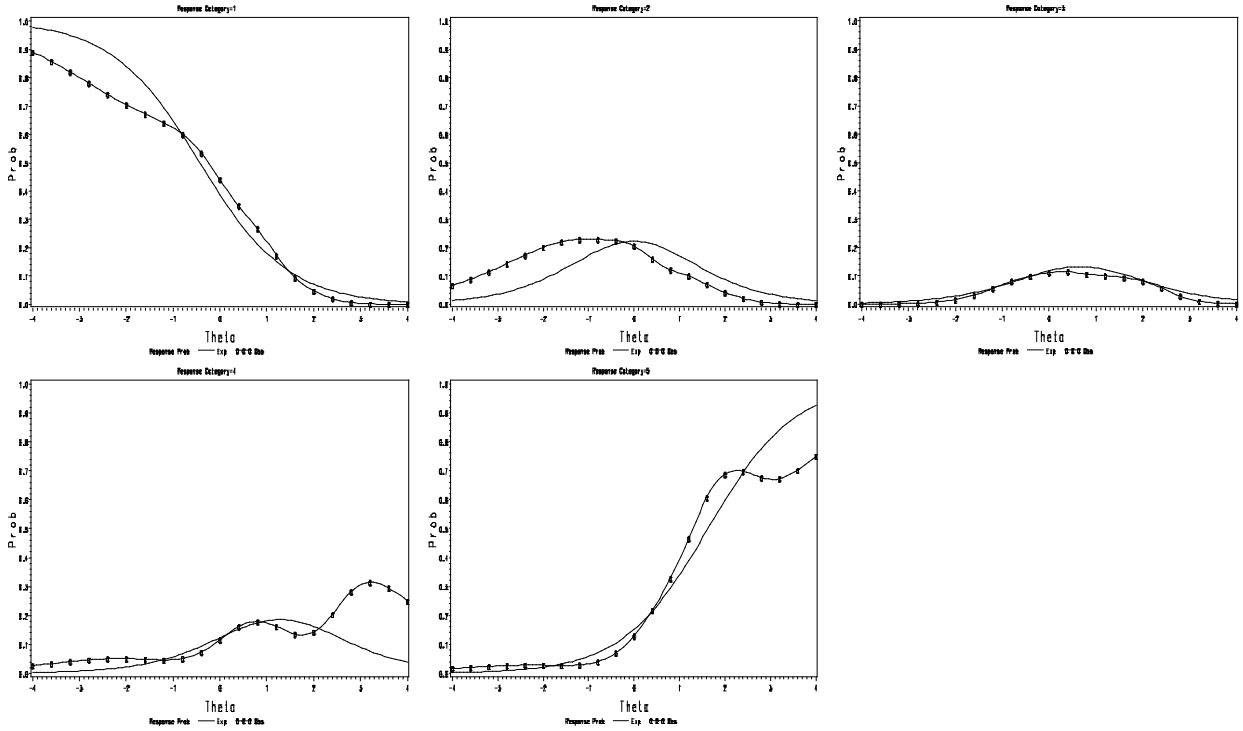
Response Probability Plots for Item 4

Item 5:



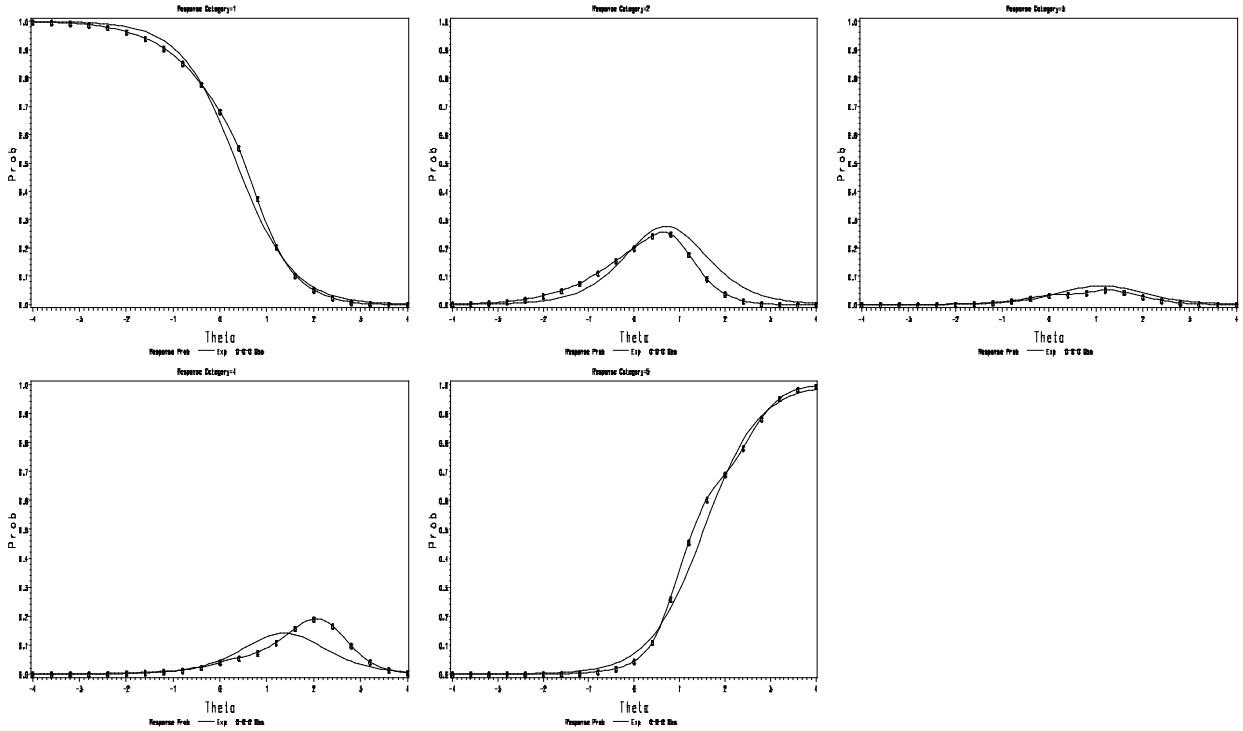
Response Probability Plots for Item 5

Item 6:



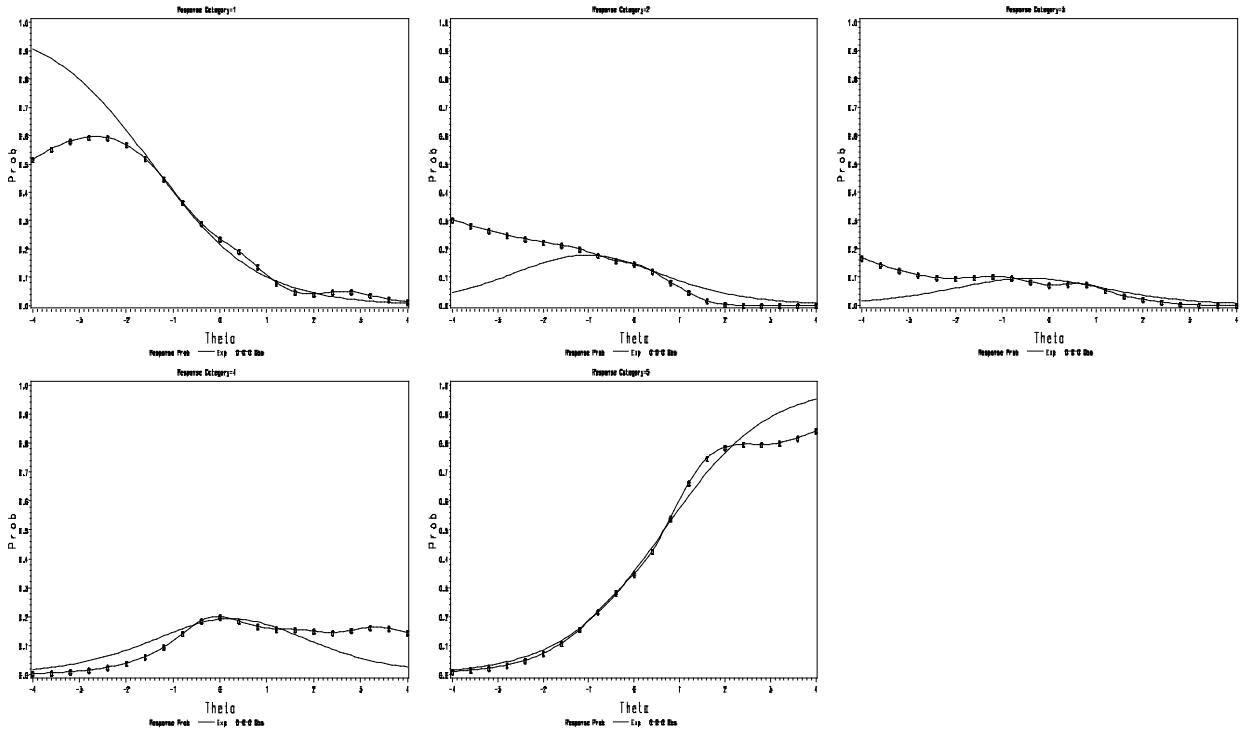
Response Probability Plots for Item 6

Item 7:



Response Probability Plots for Item 7

Item 8:



Response Probability Plots for Item 8

BIBLIOGRAPHY

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

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29(1),* 67-91.

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

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

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6),* 716-723.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood estimates for the graded model.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Services No ED. 347 208).

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

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4(4),* 305-318.

Bayarri, S., & Berger, J. (2000). P-values for composite null models. *Journal of the American Statistical Association, 95,* 1127-1142.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66,* 541-562.

Bjorner, J. B, Smith, K. J., Stone, C. A., & Sun, X. (2007). IRTFIT: A macro for item fit and local dependence tests under IRT models. Lincoln, RI: Quality Metric, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64(2),* 153-168.

Brooks, S., & Roberts, G. O. (1998). Convergence assessments of Markov chain Monte Carlo algorithms. *Statistics and Computing, 8,* 319–335.

Chen, W. (1998). IRTNEW [computer software]. Chapel Hill: University of North Carolina at Chapel Hill, L. L. Thurstone Psychometric Laboratory.

Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22(3),* 265-289.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91,* 883–904.

De Ayala, R.J. (1994). The influence of dimensionality on the graded response model. *Applied Psychological Measurement, 18,* 155-170.

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

Dollan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

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

Dresher, A. R. (2004). *The examination of local item dependency of NAEP assessments using the testlet model.* Unpublished dissertation. University of Pittsburgh.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey.

Ferrara, S., Huynh, H., & Bagli, H. (1997). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. *Applied Measurement in Education, 12,* 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36,* 119-140.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466-491.

Fu, J., Bolt, D. M., & Li, Y. (2005). *Evaluating item fit for a polytomous Fusion model using posterior predictive checks.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74,* 153-160.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.),*Bayesian statistics* (p. 147-167). Oxford: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York: Chapman & Hall.

Gelman, A., Meng, X., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 45,* 457–511.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27(3),* 217-233.

Guttman, I. (1976). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, 29,* 83-100.

Hansen, M. A. (2004) *Predicting the distribution of a goodness-of-fit statistics appropriate for use with performance-based assessments.* Unpublished dissertation. University of Pittsburgh.

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

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

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two parameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory* (pp. 109–130). New York: Springer.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347-387.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL (Version 8.8). Chicago: Scientific Software International.

Kang, T. & Chen T. T. (2008). *Performance of the generalized S-X^2 item fit index for the graded response model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Kim, J., & Bolt, D. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26,* 38-51.

Kim, S-H., Cohen, A. S., & Lin, Y-H. (2006). LDIP: a computer program for local dependence indices for polytomous items. *Applied Psychological Measurement, 30(6),* 509-510.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment. *Educational Measurement: Issues and Practice, 12,* 16-23.

Lane, S. & Stone, C.A. (2006). Performance Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Lane, S., Stone, C.A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response theory model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*(4), 313-340.

Levy, R. (2006) *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks.* Unpublished dissertation. University of Maryland.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30(1),* 3-21.

Liu, I-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometricis, 52,* 1223-1234.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8,* 453-461.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-122). New York: Springer.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer.

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23-40.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49-57.

Meng, X. (1994). Posterior predictive p-values.*The Annals of Statistics, 22(3),* 1142-1160.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73-90.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2006). Mplus: Statistical analysis with latent variables (Version 4.2). Los Angeles, CA: Muthén & Muthén.

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

Nandakumar, R., Yu, F., Li, H. H., Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22*, 99-115.

Orlando, M. (1997). Item fit in the context of item response theory. Doctoral dissertation, University of North Carolina**.** *Dissertation Abstracts International,* 58/04-B, 2175.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24(1),* 50-64.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item responses models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341-384.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (p. pp. 163-187). Washington DC: Chapman & Hall.

Raftery, A. E., & Lewis, S. M.(1992). One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science, 7*, 493-497.

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

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). The asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association, 95,* 1143–1172.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151-1172.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11(3),* 424-451.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17.*

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Silver, E. A. (1991). *Quantitative understanding: Amplifying student achievement and reasoning.* Pittsburgh, PA: Learning Research and Development Center.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29(4),* 461-488.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42(4),* 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical & Statistical Psychology, 59,* 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30(4),* 298-321.

Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64,* 583-640.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WINBUGS Version 1.4 User's manual* [Computer software manual]. Cambridge, UK: MRC Biostatistics Unit.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37(1),* 58-75.

Stone, C.A., Ankenmann, R. D., Lane, S., & Liu, M. (1993, April). *Scaling QUASAR's performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness of fit tests for IRT models. *Educational and Psychological Measurement, 60,* 974-991.

Stone, C. A., Mislevy, R. J., & Mazzeo, J. (1994, April). *Classification error and goodness-of-fit in IRT models.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331-352.

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

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimates. *Psychometrika, 55,* 293-326.

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.

Sung, H. J., & Kang, T. (2006). *Choosing a polytomous IRT model using Bayesian model selection methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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.

Tay-Lim, S. H., & Stone, C. A. (2000). *Assessing the Dimensionality of Constructed-Response Tests Using Hierarchical Cluster Analysis: A Monte Carlo Study.* Paper presented at the annual meeting of the American educational Research Association, New Orleans, LA.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Mooresville, IN: Scientific Software.

Thissen, D. J., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Boston, MA: Kluwer Academic Publishers.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339–368.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147-163.

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: theory and applications . *Applied Psychological Measurement, 26,* 109-128.

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,* 239-252.

Wu, M., Adams, R. J, & Wilson, M. (1998). *ACER ConQuest: Generalized item response modeling software.* Melbourne, Australia: The Australian Council for Educational Research.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8(2),* 125-145.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local Item dependence. *Journal of Educational Measurement, 30(3),* 187-213.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Yao, L. & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30(6),* 469-492.

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

Yu, F., & Nandakumar, R. (2001). Poly-Detect for quantifying the degree of multidimensionality of item response data. *Journal of Educational Measurement, 38 (2),* 99–120.

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

Zhang, B. (2003) *Goodness-of-fit statistics for compensatory multidimensional item response models using total scores.* Unpublished dissertation. University of Pittsburgh.