

UCLA

UCLA Electronic Theses and Dissertations

Title

Hierarchical Item Response Models for Cognitive Diagnosis

Permalink

<https://escholarship.org/uc/item/6ps9d3fd>

Author

Hansen, Mark

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Hierarchical Item Response Models for Cognitive Diagnosis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Mark Patrick Hansen

2013

© Copyright by
Mark Patrick Hansen
2013

ABSTRACT OF THE DISSERTATION

Hierarchical Item Response Models for Cognitive Diagnosis

by

Mark Patrick Hansen

Doctor of Philosophy in Education

University of California, Los Angeles, 2013

Professor Li Cai, Chair

Cognitive diagnosis models (see, e.g., Rupp, Templin, & Henson, 2010) have received increasing attention within educational and psychological measurement. The popularity of these models may be largely due to their perceived ability to provide useful information concerning both examinees (classifying them according to their attribute profiles) and test items (describing the particular attributes that are relevant to or required in order to achieve a certain response). However, the validity of such information may be undermined when diagnostic models are misspecified.

This research focuses on one aspect of model misspecification: violations of the local item independence assumption. Potential causes of dependence are examined, with a particular focus on those causes unrelated to the attributes a diagnostic test is intended to measure. Ignoring such dependencies, as is the standard practice in fitting traditional diagnostic models, may lead to biased estimates of model parameters and misclassification of examinees.

An alternative to traditional diagnostic models is presented, in which random effects are included in order to account for these nuisance dependencies. This approach is already well-established in item factor analysis, serving as the basis

for the testlet response model (Wainer, Bradlow, & Wang, 2007), random intercept item factor model (Maydeu-Olivares & Coffman, 2006), item bifactor model (Cai, Yang, & Hansen, 2011), and two-tier item factor model (Cai, 2010), among others.

The resulting hierarchical diagnostic item response model maintains the desirable properties of traditional diagnostic models (e.g., the classification of examinees with respect to fine-grained cognitive attributes), while allowing for greater complexity in the underlying response process. Importantly, the model may be estimated efficiently—even for models with a large number of nuisance variables—using an analytical dimension reduction technique described by Gibbons and Hedeker (1992).

The dissertation of Mark Patrick Hansen is approved.

José-Felipe Martínez-Fernández

Steven Reise

Noreen Webb

Ying Nian Wu

Li Cai, Committee Chair

University of California, Los Angeles

2013

for Millie, Lukas, and Jonah

TABLE OF CONTENTS

1	Introduction	1
1.1	On the Utility and Appeal of Diagnostic Models	1
1.2	A Limitation of Current Diagnostic Models	2
1.3	The Problem of Local Item Dependence	4
1.4	Strategies for Dealing with Local Dependence	5
1.5	Goals of this Research	6
1.6	Chapter Overview	7
2	The Proposed Modeling Framework	9
2.1	A Traditional Diagnostic Modeling Framework	10
2.1.1	Compensatory Diagnostic Models	12
2.1.2	Disjunctive Diagnostic Models	13
2.1.3	Conjunctive Diagnostic Models	14
2.1.4	Illustration of the Traditional Diagnostic Models	15
2.1.5	Application of the Traditional Framework to Ordinal Data	17
2.2	An Alternative Diagnostic Model to Account for Local Dependence	19
2.3	Analytical Dimension Reduction	21
2.4	Higher-order Traits	21
2.5	Model Estimation	23
2.6	Summary	25
3	Simulation Study Design	26
3.1	Data Generating Conditions for Simulation Studies	28

3.1.1	Q-matrices and Path Diagrams for Data Generation	29
3.1.2	Item Parameters for Data Generation	35
3.1.3	Distribution of Latent Attributes	38
3.2	Fitted Models and Evaluation Criteria	38
3.2.1	Evaluation of Model Estimation	40
3.2.2	Evaluation of the Impacts of Model Misspecification	42
3.2.3	Evaluation of Local Dependence Diagnostic Indices	44
4	Simulation Results: Parameter Recovery	48
4.1	Illustrative Results for one Set of Simulation Conditions	48
4.1.1	Graphical Presentation of Simulation Results: Parameter Recovery and Standard Errors of Measurement	52
4.1.2	Results for Compensatory (C-RUM) Models	53
4.1.3	Results for Conjunctive (DINA) Models	53
4.1.4	Results for Disjunctive (DINO) Models	58
4.1.5	Results for Simple Structure Models	58
4.2	Discussion	67
5	Simulation Results: Impact of Misspecification	73
5.1	Measures of Classification Performance	74
5.2	Receiver Operating Characteristics (ROC) Graphs and the Area Under the Curve (AUC)	75
5.3	Appraisal of Classification Certainty	78
5.4	Discussion	82
6	Simulation Results: Characterizing Model Misfit	84

6.1	Misspecification Due to Nuisance Dimensionality (Violations of Local Item Independence)	85
6.1.1	Calibration of LD X^2 for the Traditional DINA Model	87
6.1.2	Sensitivity of LD X^2 to Nuisance Dimensionality	92
6.1.3	LD X^2 for Correctly Specified Hierarchical Diagnostic Models	93
6.2	Diagnosis for Other Types of Model Misspecification	94
6.2.1	Models Specifying Extraneous Paths	96
6.2.2	Models with Paths Omitted	99
6.2.3	Models Specifying Extraneous Attributes	101
6.2.4	Models with Attributes Omitted	104
6.2.5	Models with Incorrect Specification of Item Type: C-RUM	107
6.2.6	Models with Incorrect Specification of Item Type: DINO	109
6.3	Discussion	112
7	Applications	113
7.1	A Model for a Testlet-based Reading Assessment	114
7.2	A Higher-order Diagnostic Model for a Mathematics Assessment	121
7.3	A Longitudinal Diagnostic Model for Assessing Attribute Stability	132
7.4	A Higher-order Diagnostic Model with a Random Intercept.	137
7.5	A Higher-order Diagnostic Model for Depression	141
7.6	Summary	147
8	Discussion	150
8.1	Review of Study Findings	150
8.1.1	Is Nuisance Dimensionality a Problem Deserving of Attention?	150

8.1.2	Can Anything be Done about Nuisance Dimensionality?	152
8.1.3	It is Possible to Tell When Nuisance Dimensionality is a Problem?	154
8.2	Directions for Future Research	155
8.2.1	Tests of Overall Model Fit	156
8.2.2	Validation of Diagnostic Models	157
8.3	Implications for Educational and Psychological Measurement	158
	Bibliography	160

LIST OF FIGURES

2.1	Path diagram for a traditional diagnostic model for items with compensatory (y_1), disjunctive (y_2), conjunctive (y_3), and simple (y_4) attribute influences.	15
2.2	Path diagram for a hierarchical diagnostic model for items with compensatory (y_1), disjunctive (y_2), conjunctive (y_3), and simple (y_4) attribute influences. Group-specific dimensions ξ_1 and ξ_2 account for item clustering.	20
2.3	Path diagram for a hierarchical diagnostic model with higher-order dimension θ	23
2.4	Conditional response probabilities for dichotomous items given attribute profiles and group-specific dimension ξ_s	24
3.1	Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with compensatory (i.e., C-RUM-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).	31
3.2	Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with disjunctive (i.e., DINO-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).	32
3.3	Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with conjunctive (i.e., DINA-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).	33

3.4	Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with “Simple” attribute influence (i.e., each item loads on exactly one attribute). Number of group-specific dimensions varies across models ($S = 1, 2, 4$).	34
4.1	Item parameter estimates and standard errors of measurement for C-RUM models, $I = 24, K = 2, N = 5000$	54
4.2	Item parameter estimates and standard errors of measurement for C-RUM models, $I = 24, K = 4, N = 5000$	55
4.3	Item parameter estimates and standard errors of measurement for C-RUM models, $I = 120, K = 2, N = 5000$	56
4.4	Item parameter estimates and standard errors of measurement for C-RUM models, $I = 120, K = 4, N = 5000$	57
4.5	Item parameter estimates and standard errors of measurement for DINA models, $I = 24, K = 2, N = 5000$	59
4.6	Item parameter estimates and standard errors of measurement for DINA models, $I = 24, K = 4, N = 5000$	60
4.7	Item parameter estimates and standard errors of measurement for DINA models, $I = 120, K = 2, N = 5000$	61
4.8	Item parameter estimates and standard errors of measurement for DINA models, $I = 120, K = 4, N = 5000$	62
4.9	Item parameter estimates and standard errors of measurement for DINO models, $I = 24, K = 2, N = 5000$	63
4.10	Item parameter estimates and standard errors of measurement for DINO models, $I = 24, K = 4, N = 5000$	64
4.11	Item parameter estimates and standard errors of measurement for DINO models, $I = 120, K = 2, N = 5000$	65

4.12	Item parameter estimates and standard errors of measurement for DINO models, $I = 120$, $K = 4$, $N = 5000$	66
4.13	Item parameter estimates and standard errors of measurement for Simple models, $I = 24$, $K = 2$, $N = 5000$	68
4.14	Item parameter estimates and standard errors of measurement for Simple models, $I = 24$, $K = 4$, $N = 5000$	69
4.15	Item parameter estimates and standard errors of measurement for Simple models, $I = 120$, $K = 2$, $N = 5000$	70
4.16	Item parameter estimates and standard errors of measurement for Simple models, $I = 120$, $K = 4$, $N = 5000$	71
5.1	Classification performance for C-RUM models with $I = 24$, $K = 2$, $N = 5000$. Results shown are for attribute x_2 using an EAP threshold of 0.5. Results obtained from the traditional diagnostic model are shown in gray; those obtained using the hierarchical diagnostic model are shown in black. OCC is overall correct classification, Sn is sensitivity, Sp is specificity, ϕ is the phi coefficient, r_{tc} is the tetrachoric correlation, and κ is Cohen's kappa.	76
5.2	Classification performance for C-RUM Models with $I = 120$, $K = 2$, $N = 5000$. Results shown are for attribute x_2 using an EAP threshold of 0.5. Results obtained from the traditional diagnostic model are shown in gray; those obtained using the hierarchical diagnostic model are shown in black. OCC is overall correct classification, Sn is sensitivity, Sp is specificity, ϕ is the phi coefficient, r_{tc} is the tetrachoric correlation, and κ is Cohen's kappa.	77

5.3	Estimated and true conditional attribute probabilities for the C-RUM Model, $I = 24$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model (in gray). Model 2 is the hierarchical diagnostic model (in black). Results shown are for attribute x_2 . Single point drawn on each curve corresponds to ROC coordinates for EAP threshold of 0.5.	79
5.4	Estimated and true conditional attribute probabilities for the C-RUM Model, $I = 120$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model (in gray). Model 2 is the hierarchical diagnostic model (in black). Results shown are for attribute x_2 . Single point drawn on each curve corresponds to ROC coordinates for EAP threshold of 0.5.	80
5.5	Estimated and true conditional attribute probabilities for the C-RUM Model with $I = 24$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model. Model 2 is the hierarchical diagnostic model. Results shown are for attribute x_2	81
5.6	Estimated and true conditional attribute probabilities: C-RUM Model with $I = 120$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model. Model 2 is the hierarchical diagnostic model. Results shown are for attribute x_2	83
6.1	LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 24$ items in $K = 2$ response categories and sample size of $N = 1000$	88
6.2	LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 24$ items in $K = 4$ ordered response categories and sample size of $N = 1000$	89

6.3	LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 120$ items in $K = 2$ response categories and sample size of $N = 1000$	90
6.4	LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 120$ items in $K = 4$ ordered response categories and sample size of $N = 1000$	91
6.5	LD X^2 -based RMSEA for models specifying extraneous paths. Tickmarks identify items 3 and 23, for which data generating Q-matrix elements $q_{3,1} = q_{23,1} = 0$ were replaced in the fitted models with $q_{3,1} = q_{23,1} = 1$	98
6.6	LD X^2 -based RMSEA for models with paths omitted. Tickmarks identify items 5 and 16, for which generating model Q-matrix elements $q_{5,1} = q_{16,1} = 1$ were replaced in the fitted models with $q_{5,1} = q_{16,1} = 0$	100
6.7	LD X^2 -based RMSEA for models specifying an extraneous attribute. Tickmarks identify items 2, 6, 8, and 13, which were incorrectly specified in the fitted models as loading on an attribute x_5 not present in the data generating model.	102
6.8	LD X^2 -based RMSEA for models with an omitted attribute. Tickmarks identify items loading on attribute x_4 in the data generating model. This attribute was omitted in the fitted models.	105
6.9	LD X^2 -based RMSEA for models with omitted attribute variable and fixed attribute distribution.	108
6.10	LD X^2 -based RMSEA for models with incorrect specification of item type: C-RUM	110

6.11	LD X^2 -based RMSEA for models with incorrect specification of item type: DINO	111
7.1	Illustration of EAP comparison used in analyses of empirical data. Model 1 (x -axis) is the traditional diagnostic model. Model 2 (y -axis) is the hierarchical model. Threshold $\text{logit}(\text{EAP})$ levels of 0 (corresponding to $\text{EAP}=0.5$) are shown as vertical and horizontal lines bisecting the scatterplot. Values at positions (A)–(H) represent proportions of examinees in (mutually exclusive) groups defined by the EAP values obtained from the alternative models: (A)–(C) are the proportions of examinees classified as possessing attribute x_j (i.e., “positive”) under both models for whom (A) EAP_1 is closer to 0.5 than EAP_2 , (B) $\text{EAP}_1=\text{EAP}_2$, or (C) EAP_2 is closer to 0.5 than EAP_1 . The values at positions (E)–(G) are the proportions classified as lacking attribute x_j (i.e., “negative”) under both models for whom (E) EAP_1 is closer to 0.5 than EAP_2 , (F) $\text{EAP}_1=\text{EAP}_2$, or (G) EAP_2 is closer to 0.5 than EAP_1 . The value at position (D) is the proportion classified positive under model 1 and negative under model 2, while (H) is the proportion classified negative under model 1 and positive under model 2.	115
7.2	Path diagram for a hierarchical diagnostic model fit to the PISA dataset.	117
7.3	LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the PISA dataset.	119
7.4	Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in the three PISA reading processes.	120

7.5	LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the TIMSS fourth grade mathematics assessment.	125
7.6	Item cluster M031242 A/B/C in TIMSS 2007 math booklet 4. . .	126
7.7	Path diagram for a hierarchical DINA model for TIMSS mathematics.	128
7.8	Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in TIMSS 4th grade mathematics attributes (x_1 – x_9). .	130
7.9	Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in TIMSS 4th grade mathematics attributes (x_{10} – x_{15}). .	131
7.10	LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the PROMIS Physical Functioning dataset.	135
7.11	Path diagram for a longitudinal diagnostic model of physical functioning.	136
7.12	Expected a posteriori (EAP) estimates of examinee physical functioning (in logit scale) at three points in time.	136
7.13	A higher-order, random intercept model for nicotine dependence. .	138
7.14	LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to PROMIS nicotine dependence data.	140
7.15	Expected a posteriori (EAP) estimates of three attributes related to nicotine dependence (in logit scale).	141
7.16	Path diagram for a diagnostic model of major depressive disorder.	143
7.17	LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to 19 items from the PDSQ measuring major depressive disorder.	144

7.18 Expected a posteriori (EAP) estimates of six attributes related to major depressive disorder (in logit scale).	146
--	-----

LIST OF TABLES

2.1	Probability of correct response given model, item parameters, and attribute profile.	16
2.2	Category response probabilities given model, item parameters, and attribute profile for ordinal data ($K = 3$).	18
3.1	Conditions manipulated in the simulation study design.	29
3.2	Q-matrices used in simulated tests of $J = 24$ Items.	30
3.3	Data generating parameters for one item, across model types. . .	37
3.4	Population distribution of attribute profiles for data generation. .	39
3.5	Misspecified models fit to simulated data.	41
4.1	Item slope parameters for higher order, hierarchical DINA model: Generating values, estimated bias, standard errors, and Monte Carlo standard deviations	50
4.2	Item intercept parameters for higher-order, hierarchical DINA model: Generating values, estimated bias, standard errors, and Monte Carlo standard deviations	51
6.1	Marginal distribution of attributes estimated from a fitted model with an extraneous attribute.	103
6.2	Marginal distribution of attributes estimated from a fitted model with an omitted attribute.	107
7.1	Description of latent variables specified in the traditional and hierarchical diagnostic models fit to the PISA dataset.	116

7.2	Q-matrix and group-specific slope parameters for the traditional and hierarchical diagnostic models fit to the PISA dataset.	118
7.3	Description of latent variables specified in the hierarchical model of TIMSS mathematics.	123
7.4	Q-matrix and group-specific slope parameters for a hierarchical model of TIMSS mathematics.	124
7.5	Description of latent variables specified in a longitudinal diagnostic model of physical functioning.	133
7.6	Q-matrix and group-specific slope parameters for a longitudinal diagnostic model of physical functioning.	134
7.7	Description of latent variables specified in a higher-order, random intercept model for nicotine dependence.	138
7.8	Q-matrix and group-specific slope parameters for a higher-order, random intercept model for nicotine dependence.	139
7.9	Description of latent variables specified in diagnostic model of major depressive disorder.	142
7.10	Q-matrix and group-specific slope parameters for a longitudinal diagnostic model of physical functioning.	149

ACKNOWLEDGMENTS

This work could not have been completed without the extensive support and guidance of my advisor and committee chair, Professor Li Cai. From the start of this project to its completion, Professor Cai provided crucial insights and suggestions that kept the project on-track, and I am indebted to him for his patient mentoring. In addition, the research presented here made extensive use of the flexMIRT item response modeling software, which Professor Cai developed. So I am also grateful for having had access to this tool.

The members of my committee—Professors José-Felipe Martinez-Fernández, Steve Reise, Noreen Webb, and Ying Nian Wu—also contributed greatly to this work through their helpful questions, feedback, and suggestions. Collectively, this group has taught me just about everything I know about educational and psychological assessment (and it was largely due to Professor Martinez-Fernández that I first became interested in this area of research). They are tremendous researchers and exceptional teachers. I am grateful to have had their input at crucial steps along the way.

The same can be said of the faculty in the social research methodology division, each of whom has been an important part of my learning these past few years and in various ways influenced this work. I feel incredibly fortunate to have received my training here. I wish to thank Professors Marvin Alkin, Tina Christie, Mike Rose, and Mike Seltzer—along with Professors Cai, Martinez-Fernández, and Webb—for all they have taught me, as well as for their kindness and generosity. I am particularly grateful to Professor Alkin, my first advisor and a supportive mentor throughout my time at UCLA.

Many fellow students have provided me with a great deal of support and assistance and have, in every way, enriched my experience of these past few years. I am especially grateful to Ji Seung Yang, Jordan Rickles, Anne Vo, Deirdre Kerr,

Larry Thomas, Scott Monroe, Moonsoo Lee, Megan Kuhfeld, Zhen Li, Lisa Dillman, Tim Ho, Rebecca Luskin, and Belinda Thompson. I also thank Drs. Taehun Lee and Carl Falk, past and current postdoctoral researchers in the psychometrics lab, who have been generous with their time and help.

I wish to acknowledge financial support for this research and my training that has been provided by the Institute of Education Sciences (through a predoctoral training grant awarded to UCLA, R305B080016), the National Science Foundation (through a doctoral dissertation improvement grant, 1260746), the National Conference of Bar Examiners (Covington Award for Research on Testing and Licensure), and the UCLA Department of Education. The views expressed here are my own and do not reflect the views or policies of these funding agencies.

Finally, I am grateful to my family—especially to my parents, parents-in-law, my sons Lukas and Jonah, and most of all, my wife Millie.

VITA

EDUCATION

- 1999 Bachelor of Arts, Boston University College of Arts and Sciences. Concentration in Biology. Minor concentrations in Mathematics and Visual Arts.
- 2005 Master of Public Health, Boston University School of Public Health. Concentration in Social and Behavior Sciences.

WORK

- 1999–2004 Research Technician, Molecular Diagnostics, Dana-Farber Cancer Institute.
- 2004–2006 Research Coordinator, Tieng Xanh-Voice, Inc.
- 2004–2007 Director of Evaluation and Research, National Center for Medical-Legal Partnership, Boston University Medical Center.
- 2007–Present Graduate Student Researcher, University of California, Los Angeles.

PUBLICATIONS

Cai, L., Yang, J., and Hansen, M. (2011). Generalized full-information bifactor analysis. *Psychological Methods*, 16, 221-248.

Yang, J., Hansen, M., and Cai, L. (2012). Characterizing sources of uncertainty

in IRT scale scores. *Educational and Psychological Measurement*, 72, 264-290.

Ryan, A. M., Kutob, R. M., Suther, E., Hansen, M., and Sandel, M. (2012). Pilot study of impact of medical-legal partnership services on patients' perceived stress and wellbeing. *Journal of Health Care for the Poor and Underserved*, 23, 1536-1546.

Hansen, M., Alkin, M. C., and Wallace, T. L. (2013). Depicting the logic of three evaluation theories. *Evaluation and Program Planning*, 38, 34-43.

Cai, L., and Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66, 245-276.

CHAPTER 1

Introduction

1.1 On the Utility and Appeal of Diagnostic Models

In recent years, cognitive diagnosis models (see, e.g., Rupp et al., 2010) have received increasing attention within educational and psychological measurement. The popularity of these models may be largely due to their focus on obtaining information not readily gained from alternative latent variable models, including those based on item response theory (IRT). Specifically, while IRT models have proven useful in ordering individuals on one or more continuous dimensions, they have not typically provided the sort of actionable information that might help inform, for instance, classroom instruction, treatment assignment, or remediation. In contrast, diagnostic models are “designed to measure specific knowledge structures and processing skills so as to provide information about [students’] cognitive strengths and weaknesses” (Leighton & Gierl, 2007, p. 3). As a result, these models may provide powerful insights concerning test takers, identifying skills or attributes not yet mastered that might be targeted in subsequent instruction. In addition, diagnostic models provide a new frame for examining the functioning of test items, allowing test developers to make explicit connections between the particular attributes that are hypothesized to be relevant to or required in order to achieve a certain response.

Given the benefits of diagnostic models, it is not surprising that there has been a tremendous eagerness to apply these models in test development and in

analyses of existing test data. Federal legislation (namely, the No Child Left Behind Act of 2001) requires that state assessment systems produce formative evaluations concerning students' specific academic needs. While any number of psychometric models might be used in the course of conducting such evaluations, it is notable that the still-dominant approaches seem somewhat ill-suited to fulfill this mandate. Many states' end-of-year proficiency tests are developed and scored using IRT or classical test theory. These testing programs provide rankings of students and classification into proficiency levels based on broad domain definitions. However, such information falls short of what is needed by teachers, schools, and families as they seek to address their students' academic needs. Given the enormous amounts of time and energy invested in testing programs, it is clear that test developers and the users of test data have a common interest in maximizing the instructional relevance of test results. Cognitive diagnosis models seem to fit well with this goal.

1.2 A Limitation of Current Diagnostic Models

A large number of cognitive diagnosis models have been developed in order to relate item responses to underlying cognitive attributes. A 2007 paper by Fu and Li identified 62 such models, and of course additional models have been proposed in the years since. Extensive reviews of these models, along with explanations of their origins and relationships to one another, have been provided by a number of authors, including Fu (2005); Fu and Li (2007); DiBello, Roussos, and Stout (2007); Rupp and Templin (2008b); and Rupp et al. (2010).

Most diagnostic models can be characterized as latent class models, in which examinee proficiencies are described in terms of discrete attributes. In other words, these are variables for which there is a finite number of possible states. Most typically, that number is two, such that an examinee might be classified as

either possessing or lacking a particular skill or knowledge structure. In addition, the sorts of attributes typically considered in diagnostic assessments are narrow in their conceptual breadth. Taking an example from later in this paper, a possible skill measured in a diagnostic test of math ability might be to “calculate and estimate perimeters, area, and volume” (see Table 7.3, adapted from Lee, Park, & Taylan, 2011).

This approach contrasts with traditional unidimensional and multidimensional IRT models, in which all latent variables are continuous (and typically somewhat broad or general in nature). As described above, it is the ability of diagnostic models to place examinees into classes or groups that contributes to these models’ perceived utility. At the same time, some researchers have noted that discrete variables alone may be insufficient to account for relationships among items and attributes. This observation has led to the development of hybrid models with both discrete and continuous latent variables. For example, de la Torre and Douglas (2004) proposed a model in which a higher-order continuous variable explains correlations among discrete attributes, while the attributes account for variations in item responses. Similarly, Choi (2010) specified a mixture IRT model in which item responses are regressed on both a continuous ability dimension and on discrete attribute variables. In this model, attribute profiles predict latent class membership, and item difficulty (relative to examinee standing on the continuous ability dimension) may vary across classes.

What is most important to note regarding the various cognitive diagnosis models currently in use—even those that incorporate continuous latent dimensions to better account for item dependence—is that the hypothesized latent variables specified in each diagnostic model are all assumed to be interpretable and of substantive interest. However, as discussed in the following section, there are frequently testing contexts in which substantively irrelevant, nuisance dimensions may be hypothesized.

1.3 The Problem of Local Item Dependence

Diagnostic models—like other item response models—assume local item independence, meaning that responses are independent, conditional on the latent variables modeled. In most current diagnostic models, those latent variables consist of the specific skills or cognitive attributes measured by test items. However, experience in fitting such models suggests that these attribute variables alone are frequently unable to fully account for associations among items (some examples will be provided in Chapter 7). Put another way, we often find that the assumption of local item independence is violated in practice. This problem is well-known in IRT modeling, and so it should perhaps be fully expected that local dependence would also arise in the context of diagnostic models. Indeed, measures of residual dependence have been suggested as possible tests of diagnostic model goodness-of-fit (de la Torre & Douglas, 2004; Templin & Henson, 2006; Rupp et al., 2010), following the use of such measures in evaluating the fit of IRT models (see, e.g., Yen, 1993; Chen & Thissen, 1997).

At the same time, evaluation of the local independence assumption has not yet become part of the routine practice in diagnostic modeling. Moreover, even in those cases where researchers have evidence of local dependence, there is currently little guidance to be found in the diagnostic modeling literature concerning what can be done to improve the situation. Much of the research on diagnostic model misspecification has focused on the validity of the mapping of attributes onto test items via the so-called “Q-matrix” (Tatsuoka, 1983). Given this focus, residual item dependence might be interpreted as a failure to include all relevant skills or attributes in the model or to properly match items onto the requisite skills or attributes. Accordingly, additional attributes might be proposed, or the pattern of item loadings onto attributes might be altered.

Certainly, there may be situations in which it would be correct to interpret lo-

cal dependence as evidence of Q-matrix misspecification. However, not all causes of item dependence are necessarily related to the attributes or skills of interest. For example, certain practices in test construction—such as the administration of item blocks following a common stimulus (an instance of testlet-based assessment; see, e.g., Wainer et al., 2007)—may result in entirely predictable patterns of local dependence. Similarly, idiosyncratic response styles (when using Likert-type scales) or within-test changes in the direction of item phrasing (e.g., positive versus negative), item type (e.g., multiple choice versus free response), or mode-of-administration may introduce associations between items, beyond what can be explained by the latent variables of substantive interest (Maydeu-Olivares & Coffman, 2006; Pomplum, 2007; Cai, 2010).

1.4 Strategies for Dealing with Local Dependence

Various approaches to account for these kinds of construct- or attribute-irrelevant dependencies have been developed in the context of IRT modeling, including the specification of hierarchical item factor models, which include item bifactor models (Gibbons & Hedeker, 1992), testlet response models (Wainer et al., 2007), and two-tier models (Cai, 2010), among others.

An important advantage of the hierarchical item factor model is that it may be efficiently estimated, despite potentially high overall dimensionality. This is due to a dimension reduction strategy discovered by Gibbons and Hedeker (1992) that takes advantage of the unique structure of the model, in which items load on at most one group-specific dimension, and the group-specific dimensions are independent of one another, conditional on the primary dimensions. When the fitted model adheres to this structure, it is possible to perform maximum marginal likelihood estimation by conducting a series of integrations over a number of dimensions equal to one more than the number of primary dimensions (rather than

the total number of dimensions, which can be much larger), without any loss in precision.

Prior research (e.g., DeMars, 2007) has demonstrated that hierarchical item factor models (including, e.g., bifactor and testlet response models) fit real data better than alternative models that ignore local item dependencies. Consequently, these models may provide less biased estimates of individuals' standing on the latent variables of interest (i.e., scores) and more accurate characterization of the uncertainty in those estimates (i.e., score reliability).

1.5 Goals of this Research

Despite the proven utility of hierarchical item factor models in IRT modeling, commonly utilized cognitive diagnosis models have generally not accounted for the influence of nuisance dimensions on item responses. As a consequence, constructs or attributes of substantive interest are likely to be confounded with irrelevant, nuisance dimensions. This creates some uncertainty concerning the extent to which model parameter estimates and examinee classification decisions should be trusted. This research seeks to extend developments in hierarchical item factor analysis to the context of cognitive diagnosis modeling, with the following goals:

1. Presentation of the proposed modeling framework;
2. Demonstration of model estimation under a wide range of simulated test conditions;
3. Examination of the potential consequences of local item independence violations within the context of diagnostic modeling;
4. Exploration of the potential utility of limited-information goodness-of-fit statistics for characterizing model misfit; and

5. Application of the framework to existing educational and psychological data.

Consequences of local item dependence have been relatively well-studied for unidimensional IRT models but not previously explored within the diagnostic modeling context. This research seeks to address that gap, examining the impacts of local dependence and evaluating a flexible modeling framework to account for systematic but construct-irrelevant influences on item responses. The parameterization and efficient estimation of a hierarchical item response model for diagnosis, as well as the approach used for checking model fit, builds on recent developments in item factor analysis.

There is growing interest in extracting diagnostic information from assessments of all kinds so that more meaningful feedback can be provided to all stakeholders of the assessments. However, up to this point, the question of whether traditional cognitive diagnostic models fit real test data has been somewhat neglected. To the extent that the proposed framework better fulfills modeling assumptions, its application will contribute to more valid model-based diagnostic inferences.

1.6 Chapter Overview

The remainder of this report is organized in the following manner. In Chapter 2, I describe an existing diagnostic modeling framework, then present the proposed hierarchical model as an extension of that framework. This model is offered as an alternative to traditional diagnostic models when nuisance dimensionality is present. The proposed model is described, along with an approach for its estimation.

In Chapter 3, I describe the design of a series of simulation studies, which were conducted in order to address three primary questions. The first question is whether the proposed model can be estimated accurately and efficiently, under various data generating conditions. The second question is whether extending

diagnostic models in order to account for local dependence is even necessary. To answer this question, a study was designed to examine the robustness of traditional models to mild and moderate local independence violations. The final question to be addressed through the simulation studies is whether a limited-information goodness-of-fit test already used in IRT model checking may be useful in examining the fit of diagnostic models. Of particular interest is whether this index might allow for characterization of the possible causes of misfit, not only informing ultimate judgments concerning the acceptance or rejection of a particular model.

After describing the design of the simulation studies, results are presented in the subsequent chapters. In Chapter 4, I examine effectiveness of the proposed estimation procedure—recently implemented in the flexMIRT[®] software (Cai, 2012)—in recovering model parameters. In Chapter 5, impacts of model misspecification are explored. Here, the focus is on accuracy of examinee classification with respect to the skills or cognitive attributes being measured. Chapter 6 presents results concerning the calibration and sensitivity of Chen and Thissen’s (1997) local dependence (LD) X^2 index to various types of model misspecification.

In Chapter 7, I present a series of real-data applications of the hierarchical diagnostic model. These illustrations were selected in order to demonstrate the flexibility of the modeling approach and to show how the limited-information goodness-of-fit statistics might be used to inform model specification. The applications in this chapter include a test of reading proficiency, a fourth grade mathematics assessment, a measure of physical functioning administered at three points in time, a measure of nicotine dependence, and a depression screening questionnaire.

Finally, in Chapter 8, I discuss the findings of this research and its relevance to educational and psychological testing. I also identify directions for further research.

CHAPTER 2

The Proposed Modeling Framework

In this chapter, I describe the proposed hierarchical item response model for cognitive diagnosis. This model may be best understood as an extension of traditional diagnostic models. To the basic structure of these existing models, random effects may be added in order to account for systematic sources of variability, unrelated to the attributes or skills of interest. Such nuisance dimensions might, for example, include testlet or other method-related influences.

Various frameworks or model families have been offered in recent years as a way of organizing the large number of diagnostic classification models in use. Such frameworks help to highlight the similarities and differences of particular models. Two such frameworks are von Davier's (2005) general diagnostic model (GDM) and Henson, Templin, and Willse's (2009) log-linear cognitive diagnosis model (LCDM). Although many of the traditional diagnostic models can be subsumed within either one of these frameworks, a recent implementation of hierarchical diagnostic modeling within the flexMIRT[®] computer software (Cai, 2012) more closely resembles the parameterization used in presentations of the LCDM. Thus, this framework is used as a starting point. After presenting this "traditional" model (or, more accurately, model framework) and showing some of its special cases, I then present a variation of the LCDM for use with ordinal (graded response data), as this will allow the model to accommodate a greater variety of response formats. Finally, I present the proposed hierarchical extension and describe its estimation.

2.1 A Traditional Diagnostic Modeling Framework

Suppose that a test is developed in order to measure a set of J underlying, latent attributes (or skills). An attribute profile may be defined as $\mathbf{x} = (x_1, \dots, x_j, \dots, x_J)'$, where each x_j takes on an unobserved value of 1 or 0, given the presence (i.e., possession or mastery of) or absence of skill j , respectively. This attribute profile is posited to influence how an individual responds to test items. For a test of length I (the number of items), the vector of observed responses is given by $\mathbf{y} = (y_1, \dots, y_i, \dots, y_I)'$. For dichotomous data, $y_i \in 0, 1$ (the more general case of ordinal data in K categories will be discussed later in this chapter), and the probability of a correct response (or item endorsement) is

$$P(y_i = 1 | \mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}))]}. \quad (2.1)$$

Here, α_i is the intercept parameter for item i , and $h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x})$ defines the manner in which the latent attributes enter the logit of the item response model. Within this mapping function, $\boldsymbol{\gamma}_i$ is a vector of the item's slope parameters. The second argument, \mathbf{q}_i , is the i th row of the so-called "Q-matrix" (Tatsuoka, 1983), which is an $I \times J$ pattern matrix of zeros and ones that appears in many common diagnostic models. The Q-matrix (\mathbf{Q}) identifies the attributes that influence or are relevant to each item. Alternatively, one could say this matrix identifies the items measuring each attribute. A possible Q-matrix for a four-item assessment measuring three latent attributes is given by

$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.2)$$

If $h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x})$ is taken as $\sum_{j=1}^J \gamma_{i,j} q_{i,j} x_j$, then the model in equation 2.1 is a discrete latent variable counterpart to the multidimensional extension of the two-parameter logistic model (M2PL; Reckase, 2009). It is also the structure of von Davier’s (2005) GDM. This model is compensatory in that possession of one attribute can make up for the absence of another. However, other mapping functions are possible. Within the LCDM framework (Henson et al., 2009), h is taken as the sum of the linear combinations of attributes and their interactions:

$$h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}) = \sum_{j=1}^J \gamma_{i,j} q_{i,j} x_j + \sum_{j=1}^{J-1} \sum_{j'>j}^J \gamma_{i,j \times j'} q_{i,j} x_j q_{i,j'} x_{j'} \dots, \quad (2.3)$$

where $\gamma_{i,j \times j'}$, is the effect of the 2-way interaction of attributes x_j and $x_{j'}$. For a diagnostic model with J latent attributes, this mapping would accommodate the J main effects, $J \times (J - 1)/2$ two-way interactions, and so on, up to the lone J -way interaction term. That said, for any given application, the number of main effects and interactions specified is typically a small subset of the possible combinations. For the Q-matrix shown in Equation 2.2, no item is influenced by many as three attributes, for example. Moreover, constraints may be placed on the slope parameters ($\gamma_{i,j}$), such that the parameter space is further reduced, and it is these constraints that reduce the LCDM in its general form to several of the most commonly utilized diagnostic models.

These special cases, and the relationship of each to the LCDM (and the GDM), have been described elsewhere (see, e.g., Rupp et al., 2010; Choi, Rupp, & Pan, 2013; von Davier, 2013). They are briefly reviewed here as examples of the “traditional” models that will eventually be extended through the incorporation of random effects to obtain the proposed hierarchical diagnostic model.

2.1.1 Compensatory Diagnostic Models

As described above, when the mapping function $h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}) = \sum_{j=1}^J \gamma_{i,j} q_{i,j} x_j$, the result is a compensatory diagnostic model. Within the LCDM framework, this mapping function can be obtained by fixing all slope parameters for interaction terms to zero, leaving only the main effects of the relevant attributes (those with non-zero elements in the \mathbf{Q} -matrix):

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + \sum_{j=1}^J \gamma_{i,j} q_{i,j} x_j)]}. \quad (2.4)$$

The GDM (von Davier, 2005) has this general form, as does the compensatory reparameterized unified model (C-RUM; Hartz, 2002). When an examinee possesses none of the attributes or skills relevant to item i , each product $q_{i,j} x_j$ is zero, and the probability of a correct response (or item endorsement) is

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp(-\alpha_i)}. \quad (2.5)$$

In the context of educational testing, this quantity is sometimes referred to as the “guessing” parameter. If all relevant attributes are possessed, then the probability of correct response is

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + \sum_{j=1}^J \gamma_{i,j})]}, \quad (2.6)$$

provided $\gamma_{i,j} = 0$ for all i and j with $q_{i,j} = 0$ (which highlights the fact that \mathbf{Q} is actually unnecessary once the appropriate constraints have been imposed on $\boldsymbol{\gamma}_i$). Since one minus the quantity in Equation 2.6 is the probability of an incorrect response despite possessing all relevant attributes, it is sometimes referred to as a “slipping” parameter.

Depending on the number of relevant attributes and their slope parameters,

there may be several levels of response probability between those associated with possessing none or all of the attributes. A defining feature of the compensatory models is that possession of one attribute could offset the absence of another (though how completely the effects offset would depend on the slope parameters for the possessed and absent attributes).

2.1.2 Disjunctive Diagnostic Models

As an alternative to an additive or compensatory model, one might imagine testing situations in which having a high probability of correct response depends on the possession of at least one the relevant attributes but possessing multiple relevant attributes provides no additional advantage. Such a model might be reasonable in situations in which there are multiple appropriate strategies (requiring different attributes or skills) that could be applied in order to obtain a correct answer. Within the LCDM framework, this disjunctive attribute behavior is modeled by constraining the slope parameters in such a way that interaction terms effectively cancel out lower-order effects whenever an examinee possesses multiple attributes.

Suppose that either attribute x_j or $x_{j'}$ is required by item y_i . To specify a disjunctive response model, the two main effects are constrained to be equal, and the two-way interaction of x_j and $x_{j'}$ is constrained to be the negative of the main effects ($\gamma_{i,j} = \gamma_{i,j'} = -\gamma_{i,j \times j'} = \gamma_i$):

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + \gamma_i x_j + \gamma_i x_{j'} - \gamma_i x_j x_{j'})]}. \quad (2.7)$$

For three attributes, the main effects and three-way interactions are constrained equal to each other and to the negative of the (equal) two-way interactions. Regardless of the number of attributes, the total number of free item parameters remains the same: one intercept, one slope.

When no relevant attributes are possessed, the “guessing” probability is once

again the inverse-logit of the intercept parameter (as shown above for the compensatory model in Equation 2.5). Meanwhile, for examinees possessing at least one of the relevant attributes, the probability of correct response is

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + \gamma_i)]}. \quad (2.8)$$

This disjunctive version of the LCDM is equivalent to a model termed the deterministic input, noisy “or” gate (DINO; Templin & Henson, 2006).

2.1.3 Conjunctive Diagnostic Models

In contrast to the compensatory and disjunctive models, where possession of one attribute may partially or fully offset the absence of another, conjunctive diagnostic models are applied to items for which it is believed that *all* relevant attributes are necessary in order for an examinee to have any increase in the probability of a correct response. Such a rule might be hypothesized when a test item requires successful completion of multiple steps or tasks, and each step represents a particular skill or attribute. The most common conjunctive diagnostic model is the deterministic input, noisy “and” gate model (DINA; Junker & Sijtsma, 2001). The LCDM parameterization of the DINA model is obtained by fixing all main effects (and lower-order interactions, if relevant) to zero and estimating only the highest-order interaction of all required attributes. Letting γ_i represent the slope parameter of that highest relevant interaction,

$$P(y_i = 1|\mathbf{x}) = \pi_i^{(1)}(\mathbf{x}) = \frac{1}{1 + \exp[-(\alpha_i + \gamma_i \prod_{j:(q_{i,j}=1)} x_{i,j})]}. \quad (2.9)$$

The probability of correct response when less than all relevant attributes are possessed is determined by the value of the item intercept (Equation 2.5), while the probability if all relevant attributes are possessed is determined by the sum

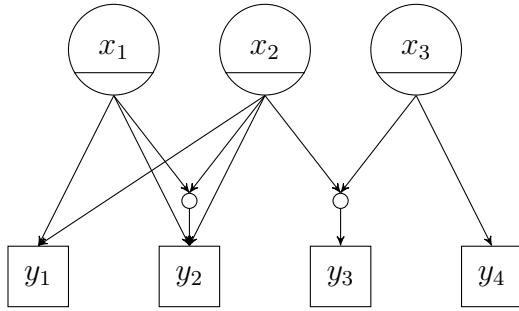


Figure 2.1: Path diagram for a traditional diagnostic model for items with compensatory (y_1), disjunctive (y_2), conjunctive (y_3), and simple (y_4) attribute influences.

of the intercept and slope parameters (Equation 2.8).

2.1.4 Illustration of the Traditional Diagnostic Models

Figure 2.1 presents a possible path diagram for the Q-matrix in Equation 2.2. Main effects of the attributes on the item responses are shown as the arrows drawn between latent attribute variables (x_p) and the observed item variables (y_i). Attribute interactions are shown by the arrows that join before connecting to an item. Horizontal lines are drawn through the attribute variables in order to emphasize that these are discrete variables (in contrast to the continuous variables in path diagrams for continuous factor analysis or IRT models) and will later be helpful in distinguishing between the modeled attributes and the random effects incorporated in the hierarchical model. The items are also discrete, of course, and thresholds could be drawn through their representative boxes. However, since all the manifest variables I consider here are discrete, there is no need to differentiate so the thresholds are not shown.

Each item in Figure 2.1 uses a different set of constraints. Note that the Q-matrix entries for items 1 and 2 are identical, with $\mathbf{q}_1 = \mathbf{q}_2 = (1, 1, 0)$. However, the path model implies a compensatory model with only main effects (e.g., C-RUM) for item 1 and, given appropriate constraints (specifically, $\gamma_{2,1} = \gamma_{2,2} =$

$-\gamma_{2,1 \times 2}$), a disjunctive model (e.g., DINO) for item 2. The Q-matrix entries for item 3, $\mathbf{q}_3 = (0, 1, 1)$, identify the influence of attributes 2 and 3. However, the path diagram implies that the slope parameters of the main effects for these attributes have been fixed to zero, leaving only the effect of the x_2x_3 interaction. This corresponds to the conjunctive case (e.g., DINA). Finally, item 4 is influenced only by attribute 3, since $\mathbf{q}_4 = (0, 0, 1)$. I'll refer to items having this structure as “simple” in that the influence of the attributes is limited to a single main effect.

Given the Q-matrix structure and the particular parameter constraints implied by the path diagram, probabilities of correct response may be obtained. This is illustrated in Table 2.1. In the upper portion of the table, values are assigned for each item parameter. The lower portion of the table shows the model-implied probabilities of incorrect and correct response—i.e., $\pi_i^{(0)}(\mathbf{x})$ and $\pi_i^{(1)}(\mathbf{x})$, respectively—given the item parameters and the attribute profile, $\mathbf{x} = (x_1, x_2, x_3)$.

Table 2.1: Probability of correct response given model, item parameters, and attribute profile.

Parameter	C-RUM Item 1	DINO Item 2	DINA Item 3	Simple Item 4
α_i	-1.3	-2.1	-1.8	-1.0
$\gamma_{i,1}$	1.5	3.6	0	0
$\gamma_{i,2}$	2.2	3.6	0	0
$\gamma_{i,3}$	0	0	0	2.4
$\gamma_{i,1 \times 2}$	0	-3.6	0	0
$\gamma_{i,1 \times 3}$	0	0	5.4	0
$\gamma_{i,2 \times 3}$	0	0	0	0
$\gamma_{i,1 \times 2 \times 3}$	0	0	0	0
\mathbf{x}	$\pi_i^{(0)}(\mathbf{x}), \pi_i^{(1)}(\mathbf{x})$			
0 0 0	(0.79,0.21)	(0.89,0.11)	(0.86,0.14)	(0.73,0.27)
0 0 1	(0.79,0.21)	(0.89,0.11)	(0.86,0.14)	(0.20,0.80)
0 1 0	(0.29,0.71)	(0.18,0.82)	(0.86,0.14)	(0.73,0.27)
0 1 1	(0.29,0.71)	(0.18,0.82)	(0.86,0.14)	(0.20,0.80)
1 0 0	(0.45,0.55)	(0.18,0.82)	(0.86,0.14)	(0.73,0.27)
1 0 1	(0.45,0.55)	(0.18,0.82)	(0.03,0.97)	(0.20,0.80)
1 1 0	(0.08,0.92)	(0.18,0.82)	(0.86,0.14)	(0.73,0.27)
1 1 1	(0.08,0.92)	(0.18,0.82)	(0.03,0.97)	(0.20,0.80)

2.1.5 Application of the Traditional Framework to Ordinal Data

Up until this point, the models have been limited to dichotomous response data. However, the LCDM framework may be adapted to handle polytomous data in ordinal response categories, which allows the application of these models to a broader range of assessments. For items with K ordered categories, let $y_i \in \{0, 1, \dots, K - 1\}$. Following the approach of Samejima's (1969) graded response IRT model (and subsequent multidimensional extensions, e.g., Muraki & Carlson, 1995; Gibbons et al., 2007), the probability of response in a particular category k may be obtained by taking the difference in adjacent *cumulative* response probabilities:

$$P(y_i = k|\mathbf{x}) = \pi_i^{(k)}(\mathbf{x}) = P(y_i \geq k|\mathbf{x}) - P(y_i \geq k + 1|\mathbf{x}). \quad (2.10)$$

These cumulative probabilities (describing the probability of response in category k or higher) are defined in the following manner:

$$\begin{aligned} P(y_i \geq 0|\mathbf{x}) &= 1, \\ P(y_i \geq 1|\mathbf{x}) &= \frac{1}{1 + \exp[-(\alpha_{i,1} + h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}))]} \\ &\vdots \\ P(y_i \geq k|\mathbf{x}) &= \frac{1}{1 + \exp[-(\alpha_{i,k} + h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}))]} \\ &\vdots \\ P(y_i \geq K - 1|\mathbf{x}) &= \frac{1}{1 + \exp[-(\alpha_{i,K-1} + h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x}))]} \\ P(y_i \geq K|\mathbf{x}) &= 0. \end{aligned} \quad (2.11)$$

Here, $\alpha_{i,1}, \dots, \alpha_{i,k}, \dots, \alpha_{i,K-1}$ are the $K - 1$ intercept parameters for item i . If each of the four items in Figure 2.1 have $K = 3$ ordered categories, the implied probabilities of response in each category—given the item parameters and attribute profile—can be obtained, from Equations 2.11 and 2.10, as illustrated in

Table 2.2.

Table 2.2: Category response probabilities given model, item parameters, and attribute profile for ordinal data ($K = 3$).

Parameter	C-RUM	DINO	DINA	Simple
	Item 1	Item 2	Item 3	Item 4
$\alpha_{i,1}$	0.2	0.6	0.4	-0.3
$\alpha_{i,2}$	-2.4	-1.6	-3.1	-2.2
$\gamma_{i,1}$	1.5	3.6	0	0
$\gamma_{i,2}$	2.2	3.6	0	0
$\gamma_{i,3}$	0	0	0	2.4
$\gamma_{i,1 \times 2}$	0	-3.6	0	0
$\gamma_{i,1 \times 3}$	0	0	5.4	0
$\gamma_{i,2 \times 3}$	0	0	0	0
$\gamma_{i,1 \times 2 \times 3}$	0	0	0	0
\mathbf{x}	$\pi_i^{(0)}(\mathbf{x}), \pi_i^{(1)}(\mathbf{x}), \pi_i^{(2)}(\mathbf{x})$			
0 0 0	(0.45,0.47,0.08)	(0.35,0.48,0.17)	(0.40,0.56,0.04)	(0.57,0.33,0.10)
0 0 1	(0.45,0.47,0.08)	(0.35,0.48,0.17)	(0.40,0.56,0.04)	(0.11,0.34,0.55)
0 1 0	(0.08,0.47,0.45)	(0.01,0.10,0.88)	(0.40,0.56,0.04)	(0.57,0.33,0.10)
0 1 1	(0.08,0.47,0.45)	(0.01,0.10,0.88)	(0.40,0.56,0.04)	(0.11,0.34,0.55)
1 0 0	(0.15,0.56,0.29)	(0.01,0.10,0.88)	(0.40,0.56,0.04)	(0.57,0.33,0.10)
1 0 1	(0.15,0.56,0.29)	(0.01,0.10,0.88)	(0.00,0.09,0.91)	(0.11,0.34,0.55)
1 1 0	(0.02,0.19,0.79)	(0.01,0.10,0.88)	(0.40,0.56,0.04)	(0.57,0.33,0.10)
1 1 1	(0.02,0.19,0.79)	(0.01,0.10,0.88)	(0.00,0.09,0.91)	(0.11,0.34,0.55)

The conditional density for the observed response y_i is given by

$$P(y_i|\mathbf{x}) = \prod_{k=0}^{K-1} [\pi_i^{(k)}(\mathbf{x})]^{\chi_k(y_i)}, \quad (2.12)$$

where $\chi_k(y_i)$ is an indicator function that takes on a value of 1 if $y_i = k$ and a value of 0 otherwise (including the case that y_i is missing).

Under the assumption of conditional item independence, the conditional probability of response pattern $\mathbf{y} = (y_1, \dots, y_I)'$ is given by

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^I \prod_{k=0}^{K-1} [\pi_i^{(k)}(\mathbf{x})]^{\chi_k(y_i)}. \quad (2.13)$$

However, there are situations in which the independence assumption is question-

able. For example, items are sometimes administered within blocks that follow a common stimulus (such as a reading passage or image). In other cases, subsets of items may share specific content or phrasing. Conditional independence implies that each item in an assessment provides independent information about the respondent. If an examinee’s true attribute profile were known, information about response to one item would provide no additional insights (beyond the attribute profile) concerning how he or she might be expected to respond to another item. However, the cases described above raise the possibility that items might be associated due to influences other than underlying attribute profiles. These would constitute violations of the local item independence assumption.

2.2 An Alternative Diagnostic Model to Account for Local Dependence

A standard strategy in item factor analysis is to include random effects to account for potential residual dependence due to the common source of variation shared by a set of items. Let there be S of these item clusters, indexed $s = 1, \dots, S$. In addition, let K be the number of ordered response categories, with the categories indexed $k = 0, \dots, K - 1$. If we assume that the clusters are mutually exclusive, we obtain the following diagnostic model with group-specific dimension ξ_s :

$$P(y_i = k | \mathbf{x}, \xi_s) = \pi_i^{(k)}(\mathbf{x}, \xi_s) = P(y_i \geq k | \mathbf{x}, \xi_s) - P(y_i \geq k + 1 | \mathbf{x}, \xi_s), \quad (2.14)$$

where the cumulative response probabilities are now given by

$$P(y_i \geq k | \mathbf{x}, \xi_s) = \frac{1}{1 + \exp [-(\alpha_{i,k} + h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x})) + \beta_{i,s} \xi_s]}, \quad (2.15)$$

with boundary conditions $P(y_i \geq 0 | \mathbf{x}, \xi_s) = 1$ and $P(y_i \geq K | \mathbf{x}, \xi_s) = 0$ defined as before (Equation 2.10). Here, $\beta_{i,s}$ is the slope of item i on the group-specific

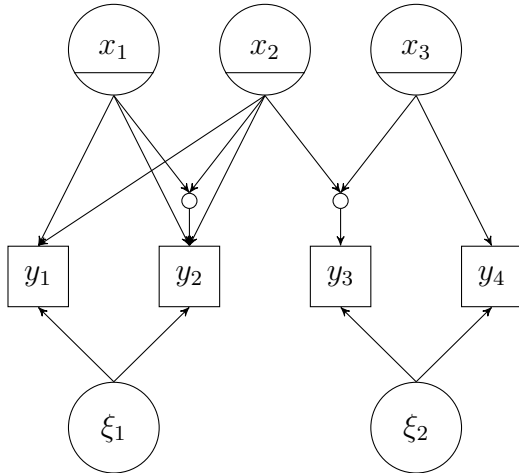


Figure 2.2: Path diagram for a hierarchical diagnostic model for items with compensatory (y_1), disjunctive (y_2), conjunctive (y_3), and simple (y_4) attribute influences. Group-specific dimensions ξ_1 and ξ_2 account for item clustering.

factor ξ_s , and each item is permitted to load on at most one group-specific dimension. The model resembles a two-tier item factor analysis model (Cai, 2010), with attributes replacing continuous factors in the primary tier. For models in which only a single attribute is measured, the model is closely related to an item bifactor model (Gibbons & Hedeker, 1992) or, with some constraints imposed, a testlet response theory model (Wainer et al., 2007). Equation 2.16 gives one possible item clustering for the earlier illustration (Figure 2.1):

$$\mathbf{B} = \begin{pmatrix} \beta_{1,1} & 0 \\ \beta_{2,1} & 0 \\ 0 & \beta_{3,2} \\ 0 & \beta_{4,2} \end{pmatrix}. \quad (2.16)$$

Figure 2.2 presents a path diagram for the resulting diagnostic model with random effects. There are now five latent variables in the model—three discrete x variables, and two continuous ξ variables.

2.3 Analytical Dimension Reduction

Having accounted for some of the dependence within item clusters, the conditional independence assumption may now be more realistic, leading to the following:

$$P(\mathbf{y}|\mathbf{x}, \xi_1, \dots, \xi_S) = \prod_{s=1}^S \prod_{i \in \mathfrak{S}_s} \prod_{k=0}^{K-1} [\pi_i^{(k)}(\mathbf{x})]^{X_k(y_i)}, \quad (2.17)$$

where \mathfrak{S}_s indicates the set of items that load on group-specific dimension s . Now, suppose the distribution of the ξ 's are given by $g(\xi_1)g(\xi_2) \dots g(\xi_S)$ —i.e., the group-specific dimensions are mutually orthogonal. In this case, it is possible to integrate out the S group-specific dimensions without a full S -dimensional integral. This is because we may utilize the familiar dimension reduction method (see Gibbons & Hedeker, 1992; Rijmen, 2009; Cai et al., 2011) developed for item bifactor analysis to transform the S -fold integral,

$$P(\mathbf{y}|\mathbf{x}) = \int \prod_{s=1}^S \prod_{i \in \mathfrak{S}_s} \prod_{k=0}^{K-1} [\pi_i^{(k)}(\mathbf{x})]^{X_k(y_i)} g(\xi_1) \dots g(\xi_S) d\xi_1 \dots d\xi_S, \quad (2.18)$$

into a series of products of one-dimensional integration:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{s=1}^S \int \prod_{i \in \mathfrak{S}_s} \prod_{k=0}^{K-1} [\pi_i^{(k)}(\mathbf{x})]^{X_k(y_i)} g(\xi_s) d\xi_s. \quad (2.19)$$

This rearrangement of terms can greatly reduce the amount of time needed for likelihood-based parameter estimation.

2.4 Higher-order Traits

The number of possible attribute profiles increases exponentially with the number of attributes (with the base determined by the number of categories or levels per attribute). For the example considered here (three dichotomous attributes), there

are $2^3 = 8$ possible profiles. In Chapter 7, an example is given of a diagnostic model with 15 dichotomous attributes and $2^{15} = 32768$ possible profiles. The proportions of examinees within each attribute class are parameters to be estimated (with the one constraint that they must sum to 1, bringing the number of these parameters to $2^J - 1$). As the profile space grows, it is increasingly difficult to model, unless some structure is imposed on the distribution of attributes.

As demonstrated by de la Torre and Douglas (2004), one possible approach is to regress the attributes on a higher-order, continuous latent trait, θ . Under this approach, the probability of possessing each attribute is assumed to depend on an examinee's standing on this higher-order dimension. Because the attributes are discrete variables, they may be treated in a similar manner to the items in an IRT model (with one important difference being that this higher-order model is fit to attribute profiles probabilities, rather than observed response patterns). For example, a two-parameter logistic model might be used to describe the relationship between attribute x_j and θ :

$$P(x_j = 1|\theta) = \pi_j^{(1)}(\theta) = \frac{1}{1 + \exp [-(c_j + a_j\theta)]}, \quad (2.20)$$

where c_j and a_j are intercept and slope parameters, respectively. In fact, any number of IRT models might be used to structure the multinomial distribution of attributes, though of course the number of “items” (the x variables) measuring the higher-order dimension(s) will necessarily be smaller than the number of actual test items (the y variables). This fact poses some practical limits on the range of higher-order structures that can be considered.

The full model—with random effects to account for nuisance dependencies and a higher-order factor modeling the attribute profiles—is illustrated in Figure 2.3. It is no longer quite so practical to present the category response probabilities within a table, since these probabilities now depend on the values of ξ_s , as well

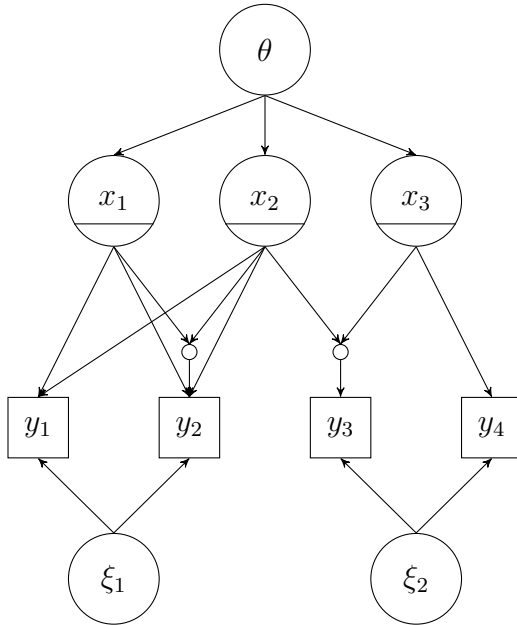


Figure 2.3: Path diagram for a hierarchical diagnostic model with higher-order dimension θ .

as the attribute profiles. Nor does it make as much sense to speak of “guessing” or “slipping” probabilities, as these would vary across individuals with identical attribute profiles (but differing in ξ_s). Figure 2.4 illustrates this variability. Each line represents a set of attribute profiles that produce identical response probabilities (conditional on ξ_s). Points are drawn at $\xi_s = 0$, since it is at this value of ξ_s that the model reduces to the traditional model (and the probabilities are identical to those reported in Table 2.1).

2.5 Model Estimation

Assuming conditional independence of the latent attributes given θ (which may be more reasonable due to the incorporation of the ξ variables into the model), we may write

$$P(\mathbf{x}|\theta) = \prod_{j=1}^J [\pi_j(\theta)]^{x_j} [1 - \pi_j(\theta)]^{1-x_j}. \quad (2.21)$$

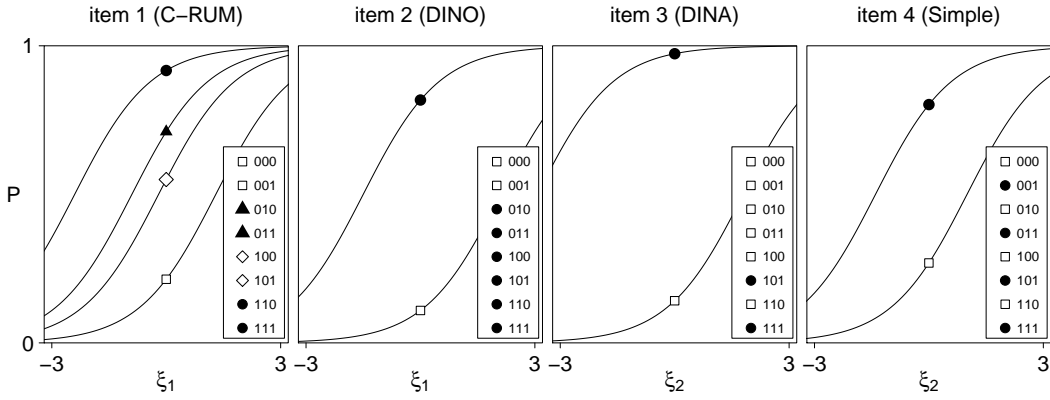


Figure 2.4: Conditional response probabilities for dichotomous items given attribute profiles and group-specific dimension ξ_s .

Combining $P(\mathbf{y}|\mathbf{x})$ from Equation 2.19 with $P(\mathbf{x}|\theta)$ from Equation 2.21, the marginal likelihood for \mathbf{y} can be obtained in two-steps. First,

$$P(\mathbf{y}|\theta) = \int P(\mathbf{y}|\mathbf{x})P(\mathbf{x}|\theta)d\mathbf{x}, \quad (2.22)$$

where the integration is a summation over the attribute profile probabilities, conditional on the values of θ . Second, θ is integrated against its distribution $g(\theta)$ to obtain the contribution to the marginal likelihood of response pattern \mathbf{y} :

$$P(\mathbf{y}) = \int P(\mathbf{y}|\theta)g(\theta)d\theta. \quad (2.23)$$

Standard numerical procedures such as the EM algorithm (Dempster, Laird, & Rubin, 1977) may be used to maximize the marginal likelihood efficiently, if dimension reduction is employed. This estimation strategy has recently been incorporated into the flexMIRT[®] item response modeling software (Cai, 2012), which was used to perform the analyses reported in this study. This software already had the capability to fit hierarchical IRT models (including item bifactor and two-tier item factor analysis models (Cai, 2010; Cai et al., 2011) using dimension reduction. The implementation of the hierarchical diagnostic modeling framework

described here provides an additional level of modeling flexibility.

2.6 Summary

In this chapter, I have described one of the existing modeling frameworks (the LCDM; Henson et al., 2009), from which many of the traditional and most frequently used diagnostic models may be derived. I noted that these models have generally ignored the problem of item clustering due to nuisance dimensionality. An alternative approach was proposed, in which such influences are explicitly modeled through incorporation of random variables. Due to the hierarchical structure of the proposed model, it is possible to utilize a dimension reduction strategy (Gibbons & Hedeker, 1992) to simplify the estimation computations. Subsequent chapters present efforts to evaluate the proposed model, first through a series of simulation studies, then by application to real educational and psychological assessment data.

CHAPTER 3

Simulation Study Design

In this chapter, I describe the design of a series of Monte Carlo studies. As mentioned previously, these were developed with three primary questions in mind. The first question deals with the implementation of the proposed hierarchical diagnostic model. I examine the estimation of model parameters under a range of data generating conditions. I also compare the variability in these estimates across replicated calibration samples with the standard errors of measurement in order to evaluate how well parameter uncertainty is characterized. In summary, the first question focuses on the viability of the model and the method of estimation.

The second question to be addressed through simulation study has a rather different focus. Whether or not the proposed model can be estimated well, it is reasonable to ask whether such a model is even needed. It is well-known that local independence violations in IRT models can impact the quality of item parameter estimates and, ultimately, induce biases in scoring. However, such effects have not been directly examined within the context of diagnostic models. Thus, this question deals with the extent to which unmodeled dimensionality impacts the fit of diagnostic models and, perhaps most importantly, the accuracy of examinee classifications. Put another way, how robust are various diagnostic models to local independence violations? The goal is to better understand if and when ignoring such violations could be detrimental.

The third question examined through simulation study concerns efforts to test model fit. Here, I evaluate the performance of an existing index of model

misspecification—one that specifically attends to violations of local item independence (Chen & Thissen, 1997). The study question can be stated as follows: Can indices of local dependence not only identify or diagnose misfit, but also characterize the nature of model misspecification? In this sense, my focus is more on the role that such goodness-of-fit tests might play in generating hypotheses concerning the underlying structure of test data (though the calibration and power of the indices are also examined). To explore the potential usefulness of the goodness-of-fit tests, a number of different types of misspecification are considered, including misassignment of Q-matrix elements, incorrect specification of the diagnostic rule (e.g., fitting a disjunctive or DINO-like model to an item with conjunctive or DINA-like structure), and failure to model nuisance dimensionality. Of course, it is this last form of misspecification that the proposed hierarchical model described is intended to address. Thus, I examine whether fit indices might provide some indication of when such a model is appropriate. A related question is whether local dependence might obscure other forms of misspecification and, if so, whether accounting for such dependencies might allow that misfit to be more readily identified.

In the following sections, I present the design of the simulation studies. Although the study questions above indicate somewhat divergent purposes, a common set of data generating conditions were utilized, so I begin with an description of these conditions. Then, for each study question, I discuss the particular approach that was taken, including explanation of the models that were fit to the data and various evaluative criteria that were used. Results from the simulation studies are presented in Chapters 4–6.

3.1 Data Generating Conditions for Simulation Studies

The hierarchical diagnostic model presented in Chapter 2 was used in data generation. Various aspects of this model were manipulated, in order that any conclusions made concerning the study questions might be based on a broad sampling of conditions. Table 3.1 summarizes the facets of the overall simulation design. Each of the constrained (special case) diagnostic response models presented in the previous chapter were used—C-RUM, DINA, DINO, and “Simple”. Unlike the earlier four-item illustration (Figure 2.3), however, all items within the simulated tests followed the same attribute-to-item mapping rule. The conditional attribute profile probabilities were structured by a higher-order continuous dimension (see Section 2.4).

Tests consisted of either $I = 24$ or $I = 120$ items. The number of group-specific dimensions S was manipulated across conditions ($S = 1, 2, 4$ when $I = 24$; $S = 1, 2, 4, 20$ when $I = 120$). The strength of these nuisance effects was manipulated through the values of the item slopes on the group-specific dimensions, $\beta_s = 0, 1, 2$. Of note, when $\beta_s = 0$, the data generating model is equivalent to a “traditional” diagnostic model, in which the items are independent, conditional on the latent attributes. The number of items loading on each group-specific dimension was equal with each condition (and was simply the total number of items divided by the number of clusters, I/S).

Although the existing diagnostic models (DINA and DINO, in particular) were developed for use with dichotomous data, the same mapping functions can be used with polytomous items, as shown previously. Accordingly, the number of response categories was manipulated within the simulation study ($K = 2, 4$). Finally, two different calibration sample sizes were used ($N = 1000, 5000$). The design factors in Table 3.1 were fully crossed, with one exception. The models with $S = 20$ group-specific dimensions were only used with the tests of $I = 120$ items. Overall,

384 conditions were examined.

Table 3.1: Conditions manipulated in the simulation study design.

factor	no. of levels	levels/values
model type (rule), $h(\boldsymbol{\gamma}_i, \mathbf{q}_i, \mathbf{x})$	4	C-RUM,DINA,DINO,Simple
no. of group-specific dimensions, S	3	1,2,4 (20, with $I = 120$ only)
group-specific slope parameter, β_s	3	0,1,2
no. of test items, I	2	24,120
no. of response categories, K	2	2,4
sample size, N	2	1000,5000

Note: All conditions fully crossed except for models with $S = 20$ group-specific dimensions, which were only included in the longer test conditions ($I = 120$ items).

3.1.1 Q-matrices and Path Diagrams for Data Generation

For all simulation conditions, the number of latent attributes was fixed to $K = 4$. Q-matrices for the C-RUM, DINA, and DINO models were obtained by randomly assigning each item to load on exactly two of the four attributes. The assignment was balanced, such that all two-attribute combinations were represented equally within the 24- and 120-item tests. The Q-matrix for the Simple model was also generated through random assignment. However, for this model, items were assigned to load on only one of the four attributes. The two Q-matrices used in the 24-item conditions are shown in Table 3.2. The first matrix (with two nonzero entries per row) was used for the C-RUM, DINA, and DINO conditions), and the second Q-matrix (with one nonzero entry per row) was used for the Simple model conditions.

Figures 3.1–3.4 present path diagrams for the 24-item tests. These diagrams are obtained by applying the attribute mapping rule (i.e., the diagnostic model, with its particular parameter constraints, as described in Chapter 2) to the appropriate Q-matrix. For the C-RUM and Simple models (Figures 3.1 and 3.4, respectively), all interaction effects are fixed to zero, such that attributes may influence items only through the main effects. In contrast, both the DINA and

Table 3.2: Q-matrices used in simulated tests of $J = 24$ Items.

Rows 1-12			Rows 13-24		
item	$\mathbf{Q}^{(1)}$	$\mathbf{Q}^{(2)}$	item	$\mathbf{Q}^{(1)}$	$\mathbf{Q}^{(2)}$
1	0 1 1 0	1 0 0 0	13	1 0 1 0	0 0 1 0
2	1 1 0 0	0 0 0 1	14	1 0 0 1	0 1 0 0
3	0 1 1 0	0 0 0 1	15	1 1 0 0	0 0 0 1
4	1 0 1 0	1 0 0 0	16	1 0 0 1	1 0 0 0
5	1 0 1 0	0 0 1 0	17	1 0 0 1	0 1 0 0
6	0 1 0 1	0 0 0 1	18	1 0 0 1	1 0 0 0
7	0 0 1 1	0 0 1 0	19	1 0 1 0	0 1 0 0
8	0 0 1 1	0 1 0 0	20	1 1 0 0	0 0 1 0
9	0 1 1 0	0 0 1 0	21	0 0 1 1	1 0 0 0
10	0 1 0 1	0 0 1 0	22	1 1 0 0	0 0 0 1
11	0 1 0 1	1 0 0 0	23	0 0 1 1	0 1 0 0
12	0 1 0 1	0 0 0 1	24	0 1 1 0	0 1 0 0

Note: $\mathbf{Q}^{(1)}$ matrix was used in C-RUM, DINA, and DINO models; $\mathbf{Q}^{(2)}$ was used in Simple structure models.

DINO models require nonzero interaction terms. The DINO model (Figure 3.2) appears rather cluttered due to the presence of both main effects and interaction terms. However, the number of parameters estimated in the DINA and DINO models is the same, due to the equality constraints placed on the DINO slope parameters (see Section 2.1.2).

For each model, three path diagrams are shown, corresponding to the manipulations of the number of group-specific dimensions ($S = 1, 2, 4$). The 120-item tests have the same basic structure as the 24-item tests, with Q-matrices being generated in exactly the same manner. However, the tests are different with respect to the number of items loading on each group-specific dimension. Since the item clusters are balanced within each test, for a given number of group-specific dimensions, the number of items per cluster for the 120-item tests is five times the number per cluster for the 24-item tests.

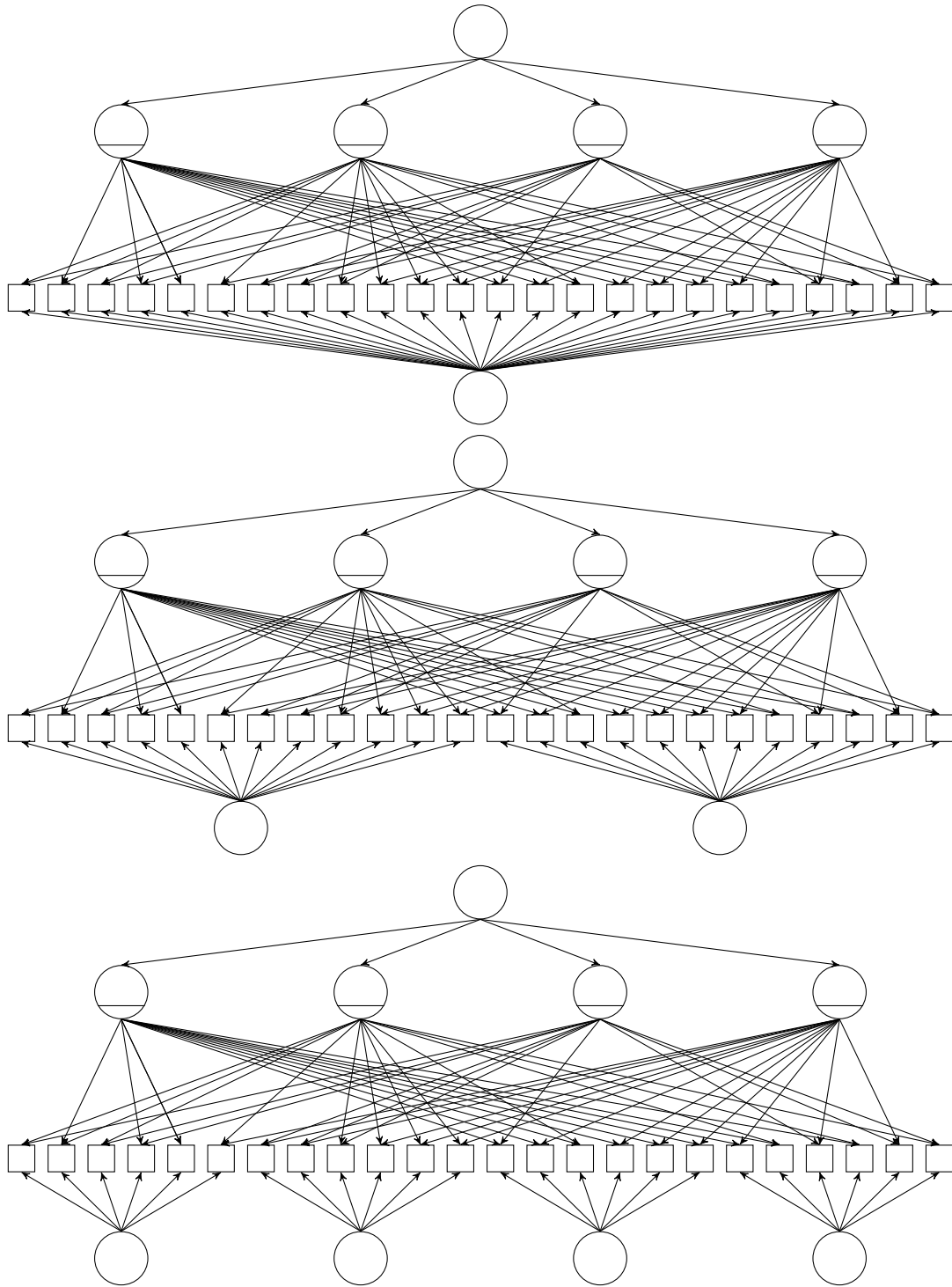


Figure 3.1: Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with compensatory (i.e., C-RUM-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).

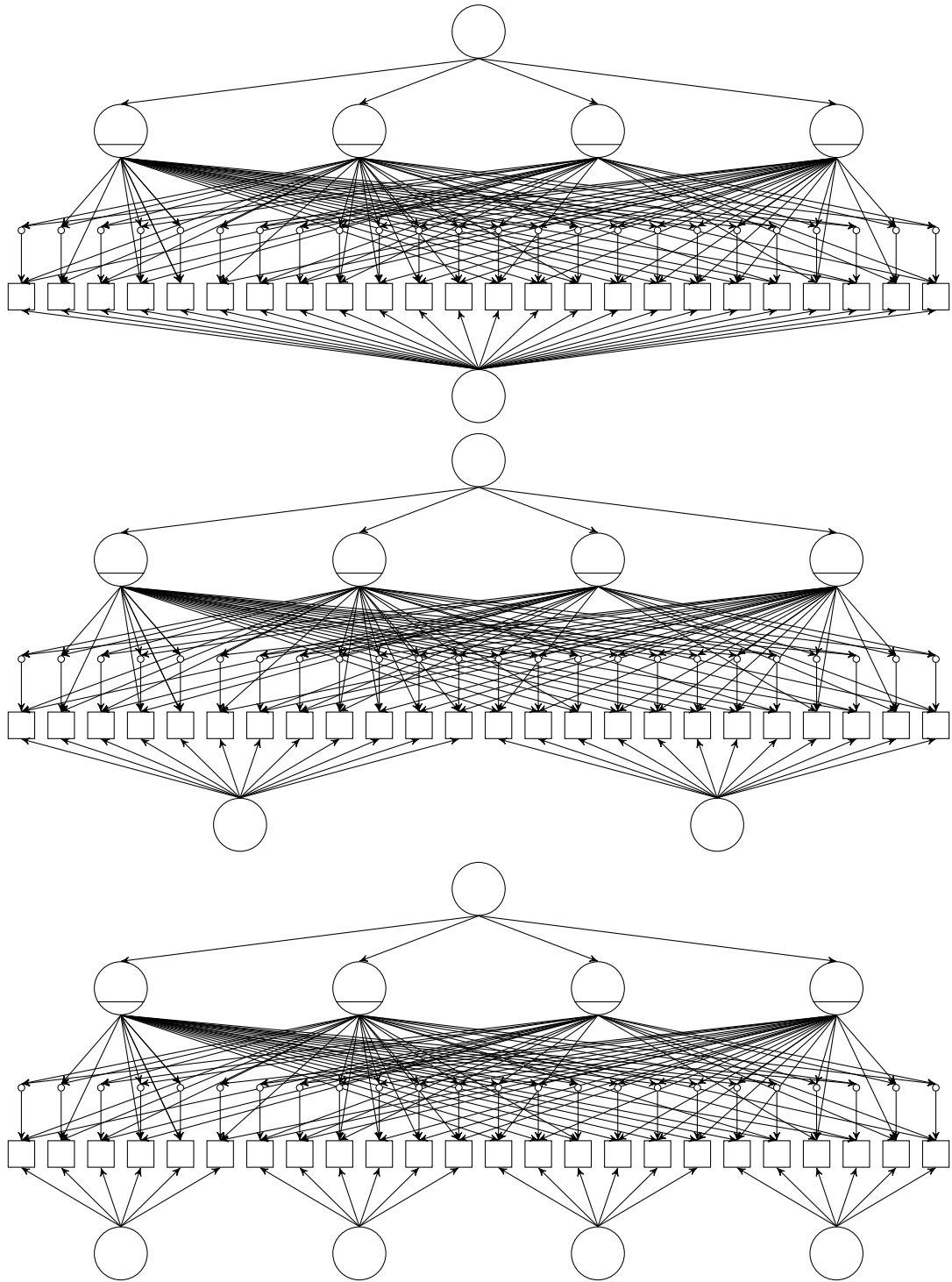


Figure 3.2: Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with disjunctive (i.e., DINO-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).

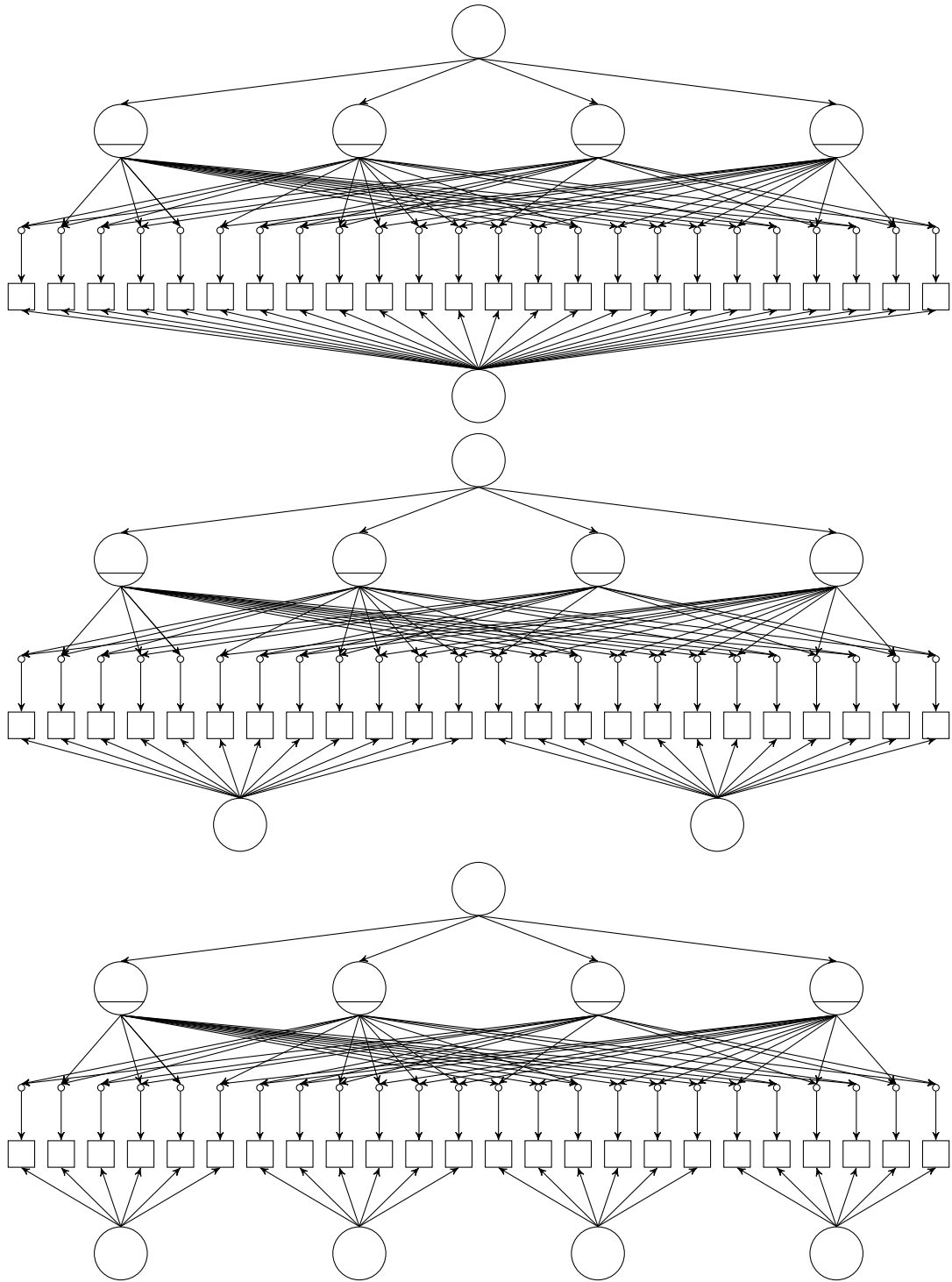


Figure 3.3: Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with conjunctive (i.e., DINA-like) attribute influences. Number of group-specific dimensions varies across models ($S = 1, 2, 4$).

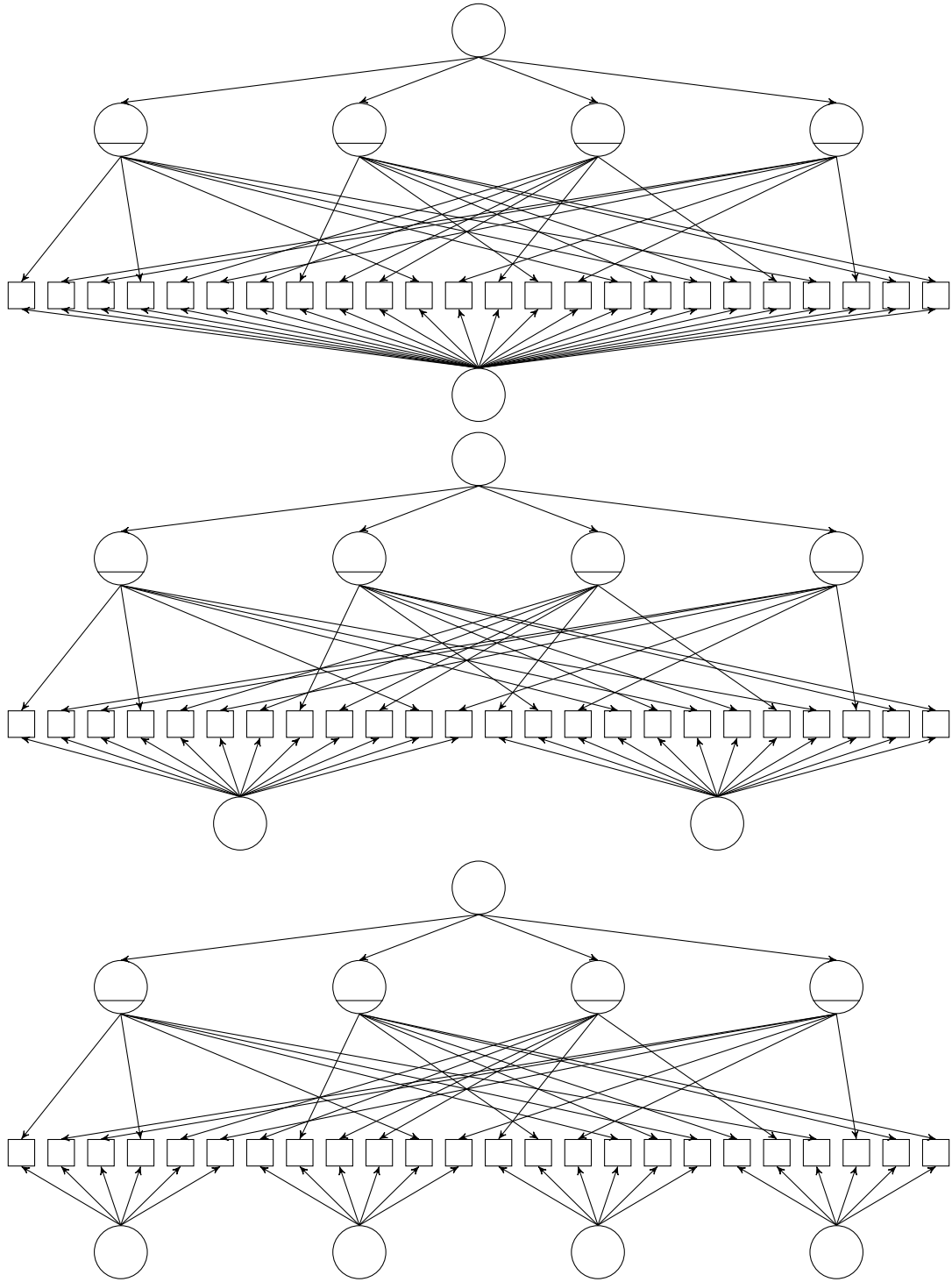


Figure 3.4: Path diagrams for higher-order, hierarchical diagnostic models for tests of $I = 24$ items with “Simple” attribute influence (i.e., each item loads on exactly one attribute). Number of group-specific dimensions varies across models ($S = 1, 2, 4$).

3.1.2 Item Parameters for Data Generation

Item parameters for data generation were randomly drawn from distributions specified on the basis of previous empirical analyses of educational assessment data. Rather than sampling slopes and intercepts directly, distributions were instead specified based on the plausible response probabilities for respondents that lack or possess requisite attributes (i.e., “guessing” and “slipping”, in the dichotomous case). These values were then transformed to match the LCDM parameterization (Henson et al., 2009).

For the dichotomous items, guessing parameters were drawn from a beta distribution with a mean of 0.2 and standard deviation of 0.05 ($g_i \sim \beta_{12.6,50.4}$). The sampling distribution for the slipping parameters had a mean of 0.10 and standard deviation of 0.05 ($s_i \sim \beta_{3.5,31.5}$). Item intercepts (α_i) were computed from the guessing parameters (g_i) in the following manner:

$$\alpha_i = -\log\left(\frac{1}{g_i - 1}\right). \quad (3.1)$$

This intercept, together with the slipping parameter (s_i), was then used to obtain the slope parameter:

$$\gamma_i = -\log\left(\frac{1}{1 - s_i} - 1\right) - \alpha_i. \quad (3.2)$$

For graded items, “guessing” and “slipping” probabilities don’t have quite the same meaning as in the dichotomous case. Still, the concepts may be readily applied to at least the highest response category ($y_i = K - 1$). In the dichotomous case, this is simply the probability of correct response. For examinees lacking the relevant attributes, the probability of response in the highest category ($y_i = 3$ for the simulation conditions with $K = 4$ ordered categories) was drawn from a distribution with a mean of 0.05 and standard deviation 0.01 ($\beta_{19,450.3}$). The probability of response in the highest category when relevant attributes are possessed

was drawn from a distribution with mean of 0.5 and standard deviation of 0.02 ($\beta_{312,312}$). For approximate balance in responses across categories, the probability of response in the lowest category for examinees lacking the relevant attributes ($y_i = 0$) was also drawn from a distribution with mean 0.5 and standard deviation of 0.02 ($\beta_{312,312}$).

From these three probabilities, the overall item slope (γ_i) and the first and third intercept parameters ($\gamma_{i,1}$, $\gamma_{i,3}$) follow from Equations 2.10 and 2.11. The second intercept parameter ($\alpha_{i,2}$) was randomly drawn from a uniform distribution over values between one-third and two-thirds of the distance between $\alpha_{i,3}$ and $\alpha_{i,1}$. Ensuring some distance between the intercepts helps to avoid the possibility of very small response probabilities for a particular category.

For the DINA, DINO, and Simple models, each response category has two possible probabilities (ignoring the influence of nuisance dimensions, for the moment). This is due to the fact that there is effectively (given the imposed constraints) only one slope parameter in each of these models. Thus, the two probabilities correspond to the cases in which the slope is or is not added to the logit, which depends on the particular mapping rule, as discussed previously. For the C-RUM model, however, there are additional levels of probability. Here, all items load on two attributes. If the slope parameters for the main attribute effects differ, then there are four possible probabilities, reflecting the four possible attribute profiles.

For the simulation study, slope parameters for the C-RUM model were obtained by splitting the slope parameter used in the other models, such that possessing both attributes would yield the same probability as possessing (a) both attributes in the DINA model, (b) at least one attribute in the DINO model, and (c) the lone required attribute in the Simple model. The splitting of the slope parameters for the C-RUM model was carried out by multiplying the single slope parameter used in the other models by a proportion (and one minus this proportion) drawn from a uniform distribution in the range 0.25–0.75. This allowed the

slope for one attribute to be no more than three times the slope for the other.

Table 3.3, shows the parameters for one item with $K = 4$ categories. This table illustrates how the values drawn for each item were applied, based on the elements of the Q-matrix (shown in the upper portion of the table) and the particular model type. Note that the C-RUM, DINO, and DINA models use the same Q-matrix (with all items loading on exactly two attributes), while a different Q-matrix was produced for the Simple models (in which items loaded on only one attribute).

Table 3.3: Data generating parameters for one item, across model types.

Parameter	C-RUM	DINO	DINA	Simple
$q_{1,1}$	0	0	0	1
$q_{1,2}$	1	1	1	0
$q_{1,3}$	1	1	1	0
$q_{1,4}$	0	0	0	0
$\alpha_{1,1}$	0.48	0.48	0.48	0.48
$\alpha_{1,2}$	-0.72	-0.72	-0.72	-0.72
$\alpha_{1,3}$	-2.83	-2.83	-2.83	-2.83
$\gamma_{1,1}$	0	0	0	2.40
$\gamma_{1,2}$	0.88	2.40	0	0
$\gamma_{1,3}$	1.52	2.40	0	0
$\gamma_{1,2 \times 3}$	0	-2.40	2.40	0

Notes: The values above are for item 1 from the 24-item test simulation conditions. Slope parameters for all attribute main and interaction effects not shown were fixed to zero for this item.

As evident from the table, the same intercept parameters were used across all models. The slopes were the also same across the DINO, DINA, and Simple models, applying the necessary constraints, as shown. Of course, the interpretation of these slopes differs across models. However, the use of common intercepts and slopes means that the two possible response probabilities of each category, given the possible attribute profiles (and conditional on the level of the group-specific dimension, ξ_s , if applicable), were the same. For C-RUM models, the same intercept parameters were also used. However, as described above, slope parameters for this model were obtained by multiplying the values used in the other models

by a splitting proportion (0.367 for the item in Table 3.3).

In addition to the latent attribute variables, items were influenced by the S group-specific dimensions ($S = 1, 2, 4$), as illustrated in the path diagrams (Figures 3.1–3.4). The slopes of the items on these dimensions were equal within a data generating condition ($\beta_s = 0, 1, 2$).

3.1.3 Distribution of Latent Attributes

A higher-order structure was used to generate the distribution of the latent attributes (the x variables). This model was a one-parameter logistic IRT model, with slope $a = 1$ for all items and intercepts of $c_1 = -0.75$, $c_2 = -0.25$, $c_3 = 0.25$, and $c_4 = 0.75$. This model is equivalent to that shown in Equation 2.20, with the added constraint that $a_j = 1$ for each attribute j . The higher-order dimension had a standard normal distribution, $\theta \sim \mathcal{N}(0, 1)$. Table 3.4 presents the resulting population distribution of attribute profiles (and the marginal probabilities for each individual attribute). Because the same higher-order model was used under all conditions, these attribute probabilities did not vary, regardless of changes to other aspects of the data generating model. Note that the model implies an ordering of the attributes along the underlying higher-order latent variable θ . The larger intercept for attribute 4 implies that this attribute is “less difficult” than the other attributes. This is reflected in the comparatively high marginal probability of possessing this attribute, as shown in the final row of Table 3.4.

3.2 Fitted Models and Evaluation Criteria

For each of the conditions described in Table 3.1, 100 datasets were generated. For each replication, two models were fit to the data. The first fitted model was the correctly specified hierarchical diagnostic model (i.e., the data generating model). The second fitted model was a traditional diagnostic model—correctly specified

Table 3.4: Population distribution of attribute profiles for data generation.

Profile	Attributes				$P(\mathbf{x})$
	x_1	x_2	x_3	x_4	
1	0	0	0	0	0.12
2	0	0	0	1	0.11
3	0	0	1	0	0.07
4	0	0	1	1	0.11
5	0	1	0	0	0.04
6	0	1	0	1	0.06
7	0	1	1	0	0.04
8	0	1	1	1	0.11
9	1	0	0	0	0.02
10	1	0	0	1	0.04
11	1	0	1	0	0.02
12	1	0	1	1	0.07
13	1	1	0	0	0.01
14	1	1	0	1	0.04
15	1	1	1	0	0.02
16	1	1	1	1	0.12
$P(x_j = 0)$	0.65	0.55	0.45	0.35	
$P(x_j = 1)$	0.35	0.45	0.55	0.65	

with respect to both the Q-matrix elements and the item model types, but ignoring the influences of the group-specific dimensions (the ξ variables). The traditional model is nested within the hierarchical model, since it may be obtained by fixing all group-specific slope parameters (the β_s 's) to zero. For each replication, model parameter estimates, their standard errors of measurement, marginal goodness-of-fit indices, and examinee attribute classifications were saved for both the correct and incorrect fitted models.

To complement the results obtained from the 100 Monte Carlo replications, a single large-sample dataset (with $N = 20000$ examinees) was generated for each of the simulation conditions. This was done in order to allow for a more in-depth examination of model misspecification. Once again, the correctly specified hierarchical diagnostic model was fit to the appropriate dataset, along with the

corresponding traditional diagnostic model that failed to account for the group-specific dimensions. Examinee expected a posteriori (EAP) scores from these alternative models were also obtained.

Finally, models representing various types of misspecification were fit to a subset of the large-sample datasets. This subset included the hierarchical DINA-type models for tests of $I = 24$ items with $K = 2$ categories. The particular types of misspecifications considered at this step are summarized in Table 3.5. The misspecifications were incorporated into fitted models that either did or did not account for the influence of group-specific dimensions. In this way, the goal was to conduct an exploratory assessment of the performance of goodness-of-fit indices for models either singly or doubly misspecified.

All data generation and model estimation were conducted with the flexMIRT[®] software (Cai, 2012), using maximum marginal likelihood estimation via the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981). The parameter error covariance matrices (which provide the standard errors of measurement) were computed in flexMIRT[®] using the Richardson extrapolation method (Jamshidian & Jennrich, 2000; Tian, Cai, Thissen, & Xin, 2012). All model estimations were checked for convergence. Unconverged replications were tallied, but results from these replications were excluded from subsequent analyses.

Having described the common data generating and model fitting procedures used across the simulation studies, I now return to the primary study questions. The approach used in addressing each of these questions is discussed, in turn.

3.2.1 Evaluation of Model Estimation

For each estimated model parameter, the mean of the point estimates across all replications within a condition was compared to the true (generating) parameter value. This provides a measure of whether parameter estimates, on average, were

Table 3.5: Misspecified models fit to simulated data.

i	change in \mathbf{Q}		resulting change in logit of the item response model	
	generating	(mis)fitted	generating	(mis)fitted
1. add paths ($x_1 \rightarrow y_2, x_1 \rightarrow y_{23}$)				
3	$\mathbf{0\ 1\ 1\ 0\ 0}$	$\mathbf{1\ 1\ 1\ 0\ 0}$	$z_3 = \alpha_3 + \gamma_{3,2 \times 3} x_2 x_3 + \beta_{3,s} \xi_s$	$z_3 = \alpha_3 + \gamma_{3,1 \times 2 \times 3} x_1 x_2 x_3 + \beta_{3,s} \xi_s$
23	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{1\ 0\ 1\ 1\ 0}$	$z_{23} = \alpha_{23} + \gamma_{23,3 \times 4} x_3 x_4 + \beta_{23,s} \xi_s$	$z_{23} = \alpha_{23} + \gamma_{23,1 \times 3 \times 4} x_1 x_3 x_4 + \beta_{23,s} \xi_s$
2. omit paths ($x_1 \rightarrow y_5, x_1 \rightarrow y_{16}$)				
5	$\mathbf{1\ 0\ 1\ 0\ 0}$	$\mathbf{0\ 0\ 1\ 0\ 0}$	$z_5 = \alpha_5 + \gamma_{5,1 \times 3} x_1 x_3 + \beta_{5,s} \xi_s$	$z_5 = \alpha_5 + \gamma_{5,3} x_3 + \beta_{5,s} \xi_s$
16	$\mathbf{1\ 0\ 0\ 1\ 0}$	$\mathbf{0\ 0\ 0\ 1\ 0}$	$z_{16} = \alpha_{16} + \gamma_{16,1 \times 4} x_1 x_4 + \beta_{16,s} \xi_s$	$z_{16} = \alpha_{16} + \gamma_{16,4} x_4 + \beta_{16,s} \xi_s$
3. add attribute (x_5)				
2	$\mathbf{0\ 1\ 1\ 0\ 0}$	$\mathbf{0\ 1\ 1\ 0\ 1}$	$z_2 = \alpha_2 + \gamma_{2,2 \times 3} x_2 x_3 + \beta_{2,s} \xi_s$	$z_2 = \alpha_2 + \gamma_{2,2 \times 3 \times 5} x_2 x_3 x_5 + \beta_{2,s} \xi_s$
6	$\mathbf{0\ 1\ 0\ 1\ 0}$	$\mathbf{0\ 1\ 0\ 1\ 1}$	$z_6 = \alpha_6 + \gamma_{6,2 \times 4} x_2 x_4 + \beta_{6,s} \xi_s$	$z_6 = \alpha_6 + \gamma_{6,2 \times 4 \times 5} x_2 x_4 x_5 + \beta_{6,s} \xi_s$
8	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{0\ 0\ 1\ 1\ 1}$	$z_8 = \alpha_8 + \gamma_{8,3 \times 4} x_3 x_4 + \beta_{8,s} \xi_s$	$z_8 = \alpha_8 + \gamma_{8,3 \times 4 \times 5} x_3 x_4 x_5 + \beta_{8,s} \xi_s$
13	$\mathbf{1\ 0\ 1\ 0\ 0}$	$\mathbf{1\ 0\ 1\ 0\ 1}$	$z_{13} = \alpha_{13} + \gamma_{13,1 \times 3} x_1 x_3 + \beta_{13,s} \xi_s$	$z_{13} = \alpha_{13} + \gamma_{13,1 \times 3 \times 5} x_1 x_3 x_5 + \beta_{13,s} \xi_s$
4. omit attribute (x_4)				
6	$\mathbf{0\ 1\ 0\ 1\ 0}$	$\mathbf{0\ 1\ 0\ 0\ 0}$	$z_6 = \alpha_6 + \gamma_{6,2 \times 4} x_2 x_4 + \beta_{6,s} \xi_s$	$z_6 = \alpha_6 + \gamma_{6,2} x_2 + \beta_{6,s} \xi_s$
7	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{0\ 0\ 1\ 0\ 0}$	$z_7 = \alpha_7 + \gamma_{7,3 \times 4} x_3 x_4 + \beta_{7,s} \xi_s$	$z_7 = \alpha_7 + \gamma_{7,3} x_3 + \beta_{7,s} \xi_s$
8	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{0\ 0\ 1\ 0\ 0}$	$z_8 = \alpha_8 + \gamma_{8,3 \times 4} x_3 x_4 + \beta_{8,s} \xi_s$	$z_8 = \alpha_8 + \gamma_{8,3} x_3 + \beta_{8,s} \xi_s$
10	$\mathbf{0\ 1\ 0\ 1\ 0}$	$\mathbf{0\ 1\ 0\ 0\ 0}$	$z_{10} = \alpha_{10} + \gamma_{10,2 \times 4} x_2 x_4 + \beta_{10,s} \xi_s$	$z_{10} = \alpha_{10} + \gamma_{10,2} x_2 + \beta_{10,s} \xi_s$
11	$\mathbf{0\ 1\ 0\ 1\ 0}$	$\mathbf{0\ 1\ 0\ 0\ 0}$	$z_{11} = \alpha_{11} + \gamma_{11,2 \times 4} x_2 x_4 + \beta_{11,s} \xi_s$	$z_{11} = \alpha_{11} + \gamma_{11,2} x_2 + \beta_{11,s} \xi_s$
12	$\mathbf{0\ 1\ 0\ 1\ 0}$	$\mathbf{0\ 1\ 0\ 0\ 0}$	$z_{12} = \alpha_{12} + \gamma_{12,2 \times 4} x_2 x_4 + \beta_{12,s} \xi_s$	$z_{12} = \alpha_{12} + \gamma_{12,2} x_2 + \beta_{12,s} \xi_s$
14	$\mathbf{1\ 0\ 0\ 1\ 0}$	$\mathbf{1\ 0\ 0\ 0\ 0}$	$z_{14} = \alpha_{14} + \gamma_{14,1 \times 4} x_1 x_4 + \beta_{14,s} \xi_s$	$z_{14} = \alpha_{14} + \gamma_{14,1} x_1 + \beta_{14,s} \xi_s$
16	$\mathbf{1\ 0\ 0\ 1\ 0}$	$\mathbf{1\ 0\ 0\ 0\ 0}$	$z_{16} = \alpha_{16} + \gamma_{16,1 \times 4} x_1 x_4 + \beta_{16,s} \xi_s$	$z_{16} = \alpha_{16} + \gamma_{16,1} x_1 + \beta_{16,s} \xi_s$
17	$\mathbf{1\ 0\ 0\ 1\ 0}$	$\mathbf{1\ 0\ 0\ 0\ 0}$	$z_{17} = \alpha_{17} + \gamma_{17,1 \times 4} x_1 x_4 + \beta_{17,s} \xi_s$	$z_{17} = \alpha_{17} + \gamma_{17,1} x_1 + \beta_{17,s} \xi_s$
18	$\mathbf{1\ 0\ 0\ 1\ 0}$	$\mathbf{1\ 0\ 0\ 0\ 0}$	$z_{18} = \alpha_{18} + \gamma_{18,1 \times 4} x_1 x_4 + \beta_{18,s} \xi_s$	$z_{18} = \alpha_{18} + \gamma_{18,1} x_1 + \beta_{18,s} \xi_s$
21	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{0\ 0\ 1\ 0\ 0}$	$z_{21} = \alpha_{21} + \gamma_{21,3 \times 4} x_3 x_4 + \beta_{21,s} \xi_s$	$z_{21} = \alpha_{21} + \gamma_{21,3} x_3 + \beta_{21,s} \xi_s$
23	$\mathbf{0\ 0\ 1\ 1\ 0}$	$\mathbf{0\ 0\ 1\ 0\ 0}$	$z_{23} = \alpha_{23} + \gamma_{23,3 \times 4} x_3 x_4 + \beta_{23,s} \xi_s$	$z_{23} = \alpha_{23} + \gamma_{23,3} x_3 + \beta_{23,s} \xi_s$
5. apply incorrect mapping/rule (C-RUM)				
8	$\mathbf{0\ 0\ 1\ 1\ 0}$	no change	$z_8 = \alpha_8 + \gamma_{8,3 \times 4} x_3 x_4 + \beta_{8,s} \xi_s$	$z_8 = \alpha_8 + \gamma_{8,3} x_3 + \gamma_{8,4} x_4 + \beta_{8,s} \xi_s$
6. apply incorrect mapping/rule (DINO)				
8	$\mathbf{0\ 0\ 1\ 1\ 0}$	no change	$z_8 = \alpha_8 + \gamma_{8,3 \times 4} x_3 x_4 + \beta_{8,s} \xi_s$	$z_8 = \alpha_8 + \gamma_{8,3} x_3 + \gamma_{8,4} x_4 - \gamma_{8,2} x_2 + \beta_{8,s} \xi_s$

Note: Altered elements of \mathbf{Q} -matrix (\mathbf{Q}) are shown in bold. All data generated from a higher-order DINA model.

biased. In addition, standard errors of measurement for each parameter were averaged across replications and compared to the standard deviation of the point estimates (the Monte Carlo standard deviation). This comparison provides a test of standard error estimation, evaluating whether the reported standard errors, on average, capture the uncertainty in estimates due to sampling variability.

3.2.2 Evaluation of the Impacts of Model Misspecification

The effects of ignoring the group-specific dimensions on classification decisions were studied by generating expected a posteriori (EAP) scores for each simulated examinee on the four latent attributes. These posterior scores are probabilities of class membership, given an examinee's item responses, the estimated item parameters, and the estimated population distribution of attribute profiles. Because the examinees were simulated, true attribute status was known for every examinee with perfect certainty, and EAP-based classifications could be compared to the true attribute profiles. Various measures of classification accuracy were examined, many of which are based the relative prevalence of each of four groups: true positives (those correctly classified as possessing the attribute), false positives (those incorrectly classified as having the attribute), false negatives (those incorrectly classified as lacking the attribute), and true negatives (those correctly classified as lacking the attribute). From the sizes of these groups (and the corresponding 2×2 contingency table), one can compute a large number of related indices: an overall correct classification (OCC) rate, sensitivity, specificity, Cohen's kappa (κ), and the phi coefficient (ϕ), among others (see, e.g., Streiner, 2003).

Of course, a good diagnostic test should provide a high rate of correct classification (i.e., high OCC), which can be defined as the sum of the proportions of true positives and true negatives. Indeed, this measure was used in a prior study examining the classification accuracy of cognitive diagnosis models (Kunina-Habenicht, Rupp, & Wilhelm, 2012). At the same time, various authors have noted that OCC

has some limitations as a measure of accuracy. For example, the true positive and true negative proportions depend on the population distribution of the attribute, not only the quality of the diagnostic test. In addition, OCC fails to account for the sometimes substantial proportion of correct classifications that would be expected simply due to chance. Cohen's kappa adjusts for this expected rate of correct classifications due to chance but remains dependent on the attribute distribution, as well as the frequencies with which a particular classification is applied (which may vary according the particular score cut-off or threshold used). Sensitivity and specificity are similarly dependent on these frequencies. For example, false negatives or false positives can be completely avoided if one respectively classifies *all* individuals as possessing or lacking the attribute, but of course such a test "result" does not provide any meaningful insight.

One approach to evaluating the usefulness of a diagnostic test that addresses these limitations is the construction of receiver operating characteristic (ROC) graphs. These graphs plot the relationship between the rates of false positives and true positives for a given test or classification method. Importantly, only cases classified as positives are considered. Consequently, ROC does not depend on the underlying population distribution of the attribute (Streiner, 2003; Fawcett, 2006). When classification is based on a underlying continuous variable or probability (as is the case with the EAP scores obtained from the diagnostic models), an ROC curve can demonstrate the trade-off between the true positive rate and false positive rate for all possible cut-off levels. In such cases, the area under the ROC curve (AUC) provides a measure of overall diagnostic usefulness, without requiring specification of a threshold score for classification.

Various classification accuracy measures were computed under correct and misspecified diagnostic models, including the indices described above. Of course, the focus of this study is on the differences in classifications between the alternative diagnostic models. However, the behavior of different measures of accuracy was

also explored. ROC curves were generated for the large-sample (single replication) datasets, and the corresponding AUC indices were also computed.

Changes in classification accuracy are one possible impact of model misspecification. However, of equal interest was the possibility that misspecification might impact estimates of classification certainty. Here, the focus is on the EAP score itself, not its subsequent uses, and the question concerns the correctness of the stated probabilities of possessing an attribute. In order to explore whether misspecification affects those probabilities, scores assigned to simulated examinees were compared to the model-implied attribute distribution among examinees with a given response pattern. This analysis was conducted within the large-sample datasets, which reduces the impact of sampling variability in the item parameter estimates.

3.2.3 Evaluation of Local Dependence Diagnostic Indices

The final questions addressed through simulation study concern the use of local dependence diagnostic indices for the evaluation of model fit (and the characterization of misfit). The analyses here utilize Chen and Thissen's (1997) local dependence (LD) X^2 statistic, which can be computed in the flexMIRT[®] software for most IRT models and for the various diagnostic models considered here. This test statistic is computed from the observed and expected bivariate response frequencies for a given item pair. In this way, it is a limited-information test (dealing with marginal, rather than overall model fit). Observed cell counts may be obtained simply by cross-tabulating responses. Expected frequencies are obtained by taking the product of the category response functions and integrating over the latent variables (the x and ξ variables for these models). Given the observed and expected probabilities, the LD X^2 test is computed as a Pearson X^2 statistic in

the following manner:

$$X^2 = N \sum_{k_i=0}^{K_i-1} \sum_{k_{i'}=0}^{K_{i'}-1} \frac{(p_{k_i, k_{i'}} - \pi_{k_i, k_{i'}})^2}{\pi_{k_i, k_{i'}}}, \quad (3.3)$$

where $p_{k_i, k_{i'}}$ and $\pi_{k_i, k_{i'}}$ are, respectively, the observed and expected bivariate response proportions for response in category k_i to item i and category $k_{i'}$ to item i' . The degrees of freedom for the test statistic is equal to the number of independent cells in the bivariate table, $(K_i - 1) \times (K_{i'} - 1)$, where K_i and $K_{i'}$ are the number of categories in items i and i' , respectively. In this study, all items within a simulation condition had the same number of response categories. Thus, for conditions in which $K = 2$, the degrees of freedom $df = (2 - 1) \times (2 - 1) = 1$. When $K = 4$, $df = (4 - 1) \times (4 - 1) = 9$.

As described by Chen and Thissen (1997), a signed version of the LD X^2 statistic may be obtained by comparing the observed and expected correlations between the two items. If the observed correlation is larger than the expected, then the model has under-explained the association between the two items. In such a case, the items are said to exhibit *positive* local dependence. If, on the other hand, the observed correlation is lower than the expected value, then the model has apparently overstated the relationship between the two items, an example of *negative* local dependence. The performance of the LD X^2 index for use with diagnostic models was examined in two contexts, as described in the following sections.

3.2.3.1 Performance of LD X^2 Under Violations of Local Item Independence

As a first evaluation of the utility of the goodness-of-fit index, values of LD X^2 were obtained for each item pair for models fit to the replicated datasets described in section 3.1. While the parameter recovery study (Section 3.2.1) deals only with

correctly specified models (those used to generate the data), here, misspecified models are also considered. Specifically, the traditional higher-order diagnostic models that ignore the influence of nuisance dimensionality (i.e., exclude the ξ variables) were fit to the data, along with the correct models. This provides an assessment of the extent to which the indices are sensitive to local independence violations, as well as the calibration of these indices when the model is correctly specified.

Within each replication, the LD X^2 obtained for each item pair was compared to its critical value for $\alpha = 0.05$. Rejection rates were calculated based on the total number of rejected tests across replications, providing information concerning type I error rates (under the correct models) and statistical power (under misspecification). It is often found that test statistics with reasonable power lead to high rates of rejection in any real data application. Given this reality, it is desirable to differentiate between reasonably good or acceptable (but wrong) models and those that are poor (and wrong). Thus, for the current study, the LD X^2 indices were also used to compute a root mean square error of approximation (RMSEA; Steiger & Lind, 1980):

$$\epsilon = \sqrt{\frac{\max[(X^2 - df), 0]}{df(N - 1)}}. \quad (3.4)$$

3.2.3.2 Performance of LD X^2 Under Q-matrix Misspecification or Incorrect Specification of Item Type

The types of misspecifications described above are limited to what might be considered as the second tier of the diagnostic model (with the latent attributes comprising the primary tier—and also being of primary measurement interest). In the final simulation-based analyses, I examine the use of the LD X^2 indices under conditions in which the primary model structure is misspecified. The particular types of model misfit I consider are summarized in Table 3.5. The models

were fit to the the large-sample ($N = 20000$) datasets generated from hierarchical DINA-type diagnostic models. Only the 24-item tests with dichotomous items were used in these analyses. The primary model misspecifications were combined with misspecifications of the secondary structure. The goal in conducting these analyses was to examine the performance of the LD X^2 indices given the various misspecifications with and without the added misspecifications related to the group-specific dimensions. A question this analysis sought to address was whether correct specification of the secondary structure might provide clearer characterization of misspecification in the primary structure.

CHAPTER 4

Simulation Results: Parameter Recovery

In this chapter I present results from the simulation study dealing with the estimation of the hierarchical diagnostic model. In these analyses, the fitted model was the same as the model used in generating the data. Thus, the focus here is on the recovery of the generating model parameters under correct model specification. Estimates of each parameter and their corresponding standard errors of measurement were obtained for each replication. Bias was evaluated by comparing the difference between the Monte Carlo average of the parameter estimates and the true parameter values. The Monte Carlo averages of the standard errors computed in flexMIRT[®] (Cai, 2012) using the Richardson extrapolation method (Jamshidian & Jennrich, 2000; Tian et al., 2012) were compared against the standard deviations of the item parameter estimates. Within this chapter, detailed results are provided for one set of data generating conditions. The results for additional conditions are then summarized graphically.

4.1 Illustrative Results for one Set of Simulation Conditions

Results from the Monte Carlo simulation study for one data generating model under the two calibration sample sizes ($N = 1000, 5000$) are presented in Tables 4.1 and 4.2. The particular condition shown here is for the higher-order, hierarchical DINA model with $S = 2$ group specific dimensions and $I = 24$ dichotomous

items (a path diagram for this model is shown in Figure 3.3). The slopes on the group-specific dimensions were $\beta_1 = \beta_2 = 1$. Of the 100 datasets analyzed for each sample size, results were obtained for 96 and 100 replications for the $N = 1000$ and $N = 5000$ conditions, respectively.

From the results in Table 4.1 it is evident that the slope parameters for this model were generally estimated with minimal bias. The one notable exception is item 2. In the sample of $N = 1000$ examinees, the difference between the true slope parameter and the average point estimate was 0.21. The average standard error measurement is also much larger than the standard deviation in the estimates. Although the magnitudes of these differences are somewhat concerning, it is worth noting that the difference in expected probabilities between the true parameter values and the average estimates for this item were actually quite small. When the required attributes are not possessed, $\pi^{(0)} = 0.208$ for the true item parameters and $\pi^{(0)} = 0.209$ for the average parameters. For examinees possessing the required attributes, the resulting expected probabilities are $\pi^{(1)} = 0.979$ and $\pi^{(1)} = 0.983$. These rather high $\pi^{(1)}$ probabilities may be contributing to the apparent instability in the estimates of this parameter. The bias is greatly reduced in the $N = 5000$ sample, and the average standard error of measurement is much closer to the standard deviation of the estimates.

The average estimates for the intercept parameters (shown in Table 4.2) are very close to the true parameter values. In addition, Monte Carlo averages of the standard errors appear to match the standard deviations of the parameter estimates, suggesting that the estimated standard errors provide an accurate characterization of sampling variability in the estimates of the intercept parameters.

Table 4.1: Item slope parameters for higher order, hierarchical DINA model: Generating values, estimated bias, standard errors, and Monte Carlo standard deviations

Parameter	Value	$N = 1000$ (96 reps)			$N = 5000$ (100 reps)		
		Bias	M(SE)	SD(Est)	Bias	M(SE)	SD(Est)
γ_1	3.66	0.00	0.38	0.31	0.01	0.16	0.14
γ_2	5.17	0.21	6.13	1.61	0.02	0.27	0.23
γ_3	3.34	-0.02	0.26	0.29	0.01	0.12	0.13
γ_4	3.85	0.02	0.32	0.30	0.00	0.15	0.14
γ_5	3.25	0.02	0.27	0.27	0.03	0.12	0.12
γ_6	3.97	-0.05	0.32	0.32	0.04	0.14	0.15
γ_7	3.46	0.00	0.26	0.25	0.00	0.12	0.12
γ_8	4.82	0.07	0.45	0.55	0.02	0.18	0.18
γ_9	3.35	0.05	0.27	0.26	0.02	0.12	0.12
γ_{10}	3.35	0.01	0.25	0.22	0.02	0.11	0.13
γ_{11}	3.98	0.07	0.30	0.34	0.01	0.13	0.13
γ_{12}	2.48	0.08	0.22	0.20	0.01	0.10	0.09
γ_{13}	3.73	-0.06	0.26	0.27	0.01	0.12	0.10
γ_{14}	3.06	-0.02	0.22	0.20	0.01	0.10	0.11
γ_{15}	3.30	0.00	0.23	0.23	0.02	0.10	0.10
γ_{16}	3.95	-0.02	0.27	0.24	-0.01	0.12	0.13
γ_{17}	3.64	0.07	0.31	0.28	-0.01	0.13	0.10
γ_{18}	3.28	-0.01	0.28	0.25	0.00	0.12	0.10
γ_{19}	3.42	0.06	0.29	0.23	-0.01	0.12	0.12
γ_{20}	3.61	0.01	0.30	0.26	0.01	0.13	0.10
γ_{21}	4.27	0.05	0.29	0.32	0.00	0.13	0.12
γ_{22}	3.19	0.02	0.23	0.22	0.01	0.10	0.10
γ_{23}	3.22	-0.01	0.23	0.23	0.00	0.10	0.10
γ_{24}	2.81	0.04	0.21	0.22	0.00	0.09	0.09
β_1	1.00	-0.01	0.06	0.05	0.00	0.03	0.02
β_2	1.00	0.00	0.06	0.05	0.00	0.02	0.02

Notes: Results are shown for higher-order, hierarchical DINA model with $S = 2$ group specific dimensions with slopes of $\beta_s = 1$. The test has $I = 24$ items with $K = 2$ response categories. Bias is the mean (across replications) of the parameter point estimates, minus the true parameter value. M(SE) is the Monte Carlo average of the estimated standard errors. SD(Est) is the Monte Carlo standard deviation of the parameter estimates.

Table 4.2: Item intercept parameters for higher-order, hierarchical DINA model: Generating values, estimated bias, standard errors, and Monte Carlo standard deviations

Parameter	Value	$N = 1000$ (96 reps)			$N = 5000$ (100 reps)		
		Bias	M(SE)	SD(Est)	Bias	M(SE)	SD(Est)
α_1	-1.57	0.01	0.11	0.12	0.00	0.05	0.05
α_2	-1.34	0.01	0.11	0.09	0.00	0.05	0.05
α_3	-1.42	0.00	0.11	0.11	0.00	0.05	0.05
α_4	-1.06	0.00	0.11	0.10	-0.01	0.05	0.04
α_5	-1.34	0.01	0.12	0.10	-0.01	0.05	0.05
α_6	-1.27	-0.03	0.19	0.11	0.00	0.08	0.05
α_7	-1.60	-0.01	0.14	0.16	0.00	0.06	0.06
α_8	-1.12	0.00	0.12	0.11	0.00	0.05	0.05
α_9	-1.69	-0.01	0.12	0.11	-0.01	0.05	0.05
α_{10}	-1.52	-0.01	0.19	0.12	0.01	0.08	0.04
α_{11}	-1.35	-0.02	0.19	0.10	0.00	0.08	0.06
α_{12}	-1.35	-0.01	0.19	0.11	0.00	0.08	0.05
α_{13}	-1.46	0.00	0.12	0.10	0.00	0.05	0.04
α_{14}	-1.11	-0.01	0.10	0.12	0.00	0.05	0.05
α_{15}	-1.34	0.00	0.10	0.11	0.00	0.04	0.05
α_{16}	-1.40	0.00	0.11	0.11	0.00	0.05	0.05
α_{17}	-1.53	-0.01	0.11	0.12	0.00	0.05	0.05
α_{18}	-1.20	-0.02	0.10	0.09	0.00	0.05	0.04
α_{19}	-1.73	0.00	0.14	0.11	0.00	0.06	0.05
α_{20}	-1.55	0.00	0.10	0.10	0.00	0.05	0.05
α_{21}	-1.11	0.01	0.12	0.12	0.01	0.05	0.05
α_{22}	-1.29	-0.02	0.10	0.10	0.01	0.04	0.04
α_{23}	-1.10	-0.02	0.12	0.12	0.01	0.05	0.05
α_{24}	-1.69	0.00	0.12	0.11	0.01	0.05	0.06

Notes: Results are shown for higher-order, hierarchical DINA model with $S = 2$ group specific dimensions with slopes of $\beta_s = 1$. The test has $I = 24$ items with $K = 2$ response categories. Bias is the mean (across replications) of the parameter point estimates, minus the true parameter value. M(SE) is the Monte Carlo average of the estimated standard errors. SD(Est) is the Monte Carlo standard deviation of the parameter estimates.

4.1.1 Graphical Presentation of Simulation Results: Parameter Recovery and Standard Errors of Measurement

In the following sections, results are shown from the range of simulation conditions. For each data generating model, the main scatterplots compare the Monte Carlo average of the parameter estimates (y -axis) to the true parameter values (x -axis). Points along the $y = x$ line are parameters that were estimated without bias. Points below or to the right of the line were estimated with negative bias (average estimates were smaller than the true parameters). Those above the line were estimated with positive bias (average estimates were larger than the true values).

In the upper left corner of each parameter scatterplot is a smaller inset showing the comparison of the standard deviations in the parameter estimates (x -axis) with the average standard errors of measurement (y -axis). Here, the focus is on the extent to which these averages match the observed variability in the estimates. Standard errors that are correctly estimated should fall along the $y = x$ line. When points fall below or to the right of this line, the estimated standard errors were, on average, too small. Points above the line indicate standard errors that, on average, overstated the uncertainty in parameter estimates.

Each figure presents the results for one model type (C-RUM, DINA, DINO, Simple), test length ($I = 24, 120$), and number of categories ($K = 2, 4$). The rows of each figure present results varying in the number of group-specific dimensions ($S = 1, 2, 4$ for 24-item tests; $S = 1, 2, 4, 20$ for 120-item tests). The columns present results with values of the group-specific slope parameters increasing from left to right ($\beta_s = 0, 1, 2$). Thus, the first column represents conditions in which the data were generated from a “traditional” diagnostic model, with no influence of group-specific dimensions (i.e., $\beta_s = 0$).

For simplicity, only results for the calibration sample size of $N = 5000$ are shown. For the conditions with $N = 1000$, there were a greater number of ap-

parently unstable estimates (as observed in Section 4.1). In addition, the average standard errors and the Monte Carlo standard deviations in the parameter estimates were of course consistently larger for the smaller sample size. That said, parameters and standard errors for which estimation bias was evident in the $N = 5000$ showed similar bias in the $N = 1000$ sample.

4.1.2 Results for Compensatory (C-RUM) Models

Results for the compensatory (C-RUM) diagnostic models are presented in Figures 4.1–4.4. Across these conditions, it appears that the item parameters are generally recovered quite well. The only apparent exception to this is the set of group-specific slope parameters (β_s), under some conditions. For tests of $I = 120$ items (Figures 4.3 and 4.4), these parameters are underestimated when $\beta_s = 2$. The extent of the bias appears to be related to the number of group-specific dimensions, which varies across the rows of the figures. For the conditions with $S = 1, 2, 4, 20$, the estimates—averaged across replications and dimensions (a separate slope parameter was estimated for each of the S dimensions)—were 1.12, 1.45, 1.67, and 1.98, respectively (the true value is $\beta_s = 2$).

For most conditions, the standard errors of measurement appear to have correct magnitude, generally falling close to the $y = x$ line. The only exceptions are for conditions with $\beta_s = 2$, for which the standard errors for the intercept parameters are underestimated. The amount of underestimation appears to be inversely related to the number of group-specific dimensions, with the largest bias for conditions with $S = 1$.

4.1.3 Results for Conjunctive (DINA) Models

Figures 4.5–4.8 show the results for DINA models. In general, these results are quite similar to those obtained for the C-RUM model. Across the conditions, it

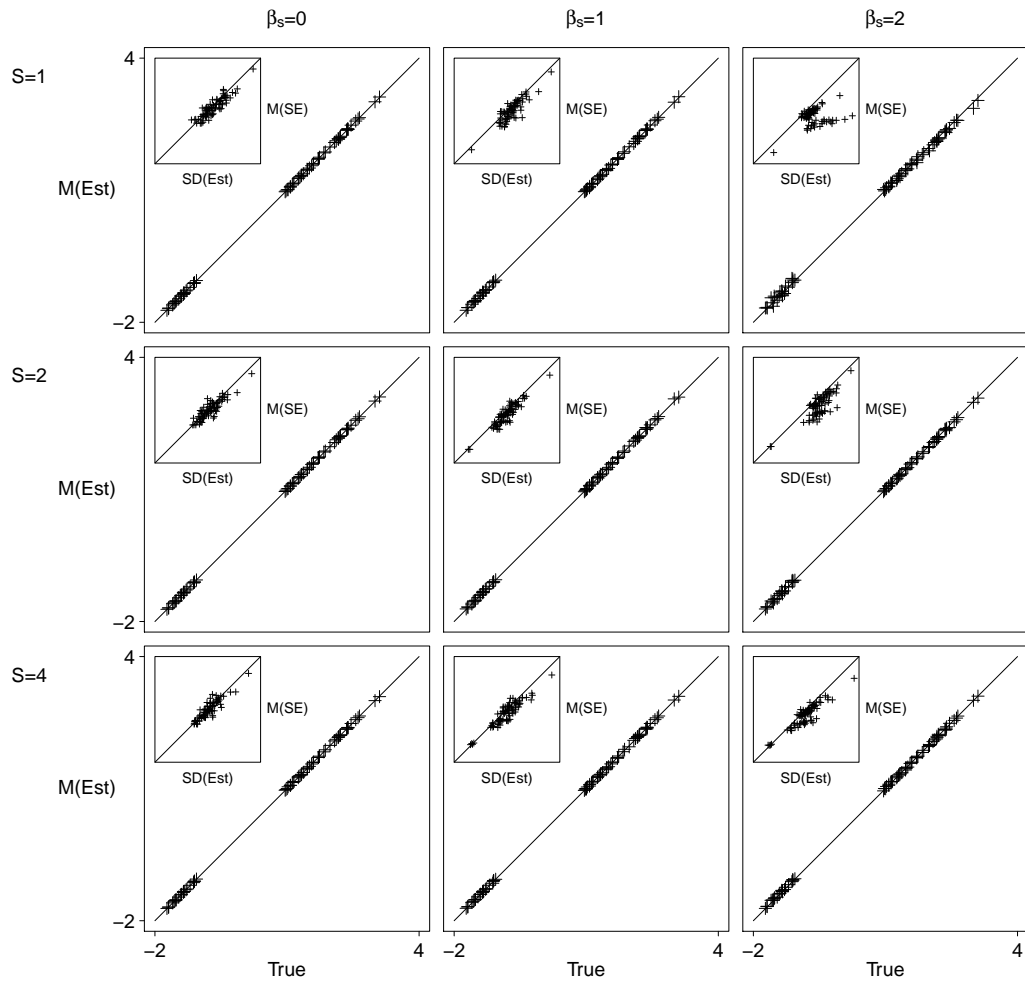


Figure 4.1: Item parameter estimates and standard errors of measurement for C-RUM models, $I = 24$, $K = 2$, $N = 5000$.

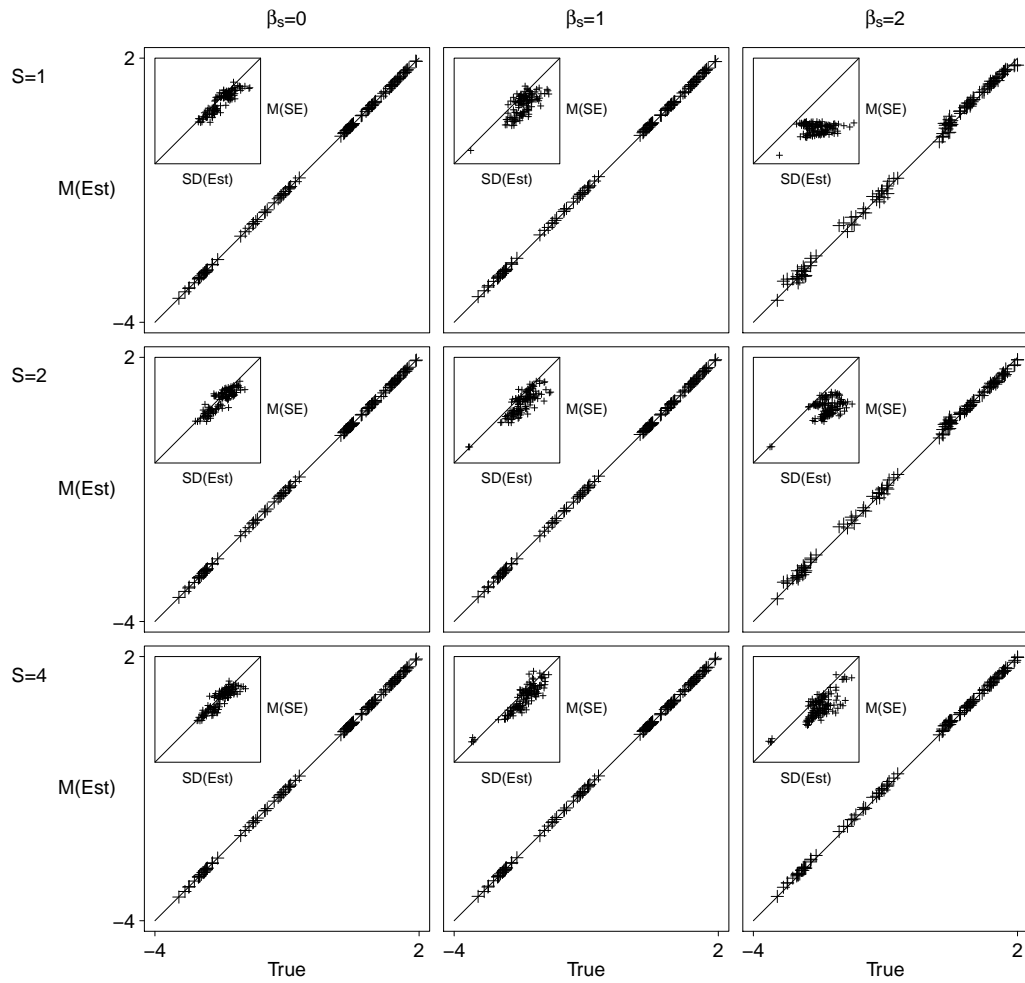


Figure 4.2: Item parameter estimates and standard errors of measurement for C-RUM models, $I = 24$, $K = 4$, $N = 5000$.

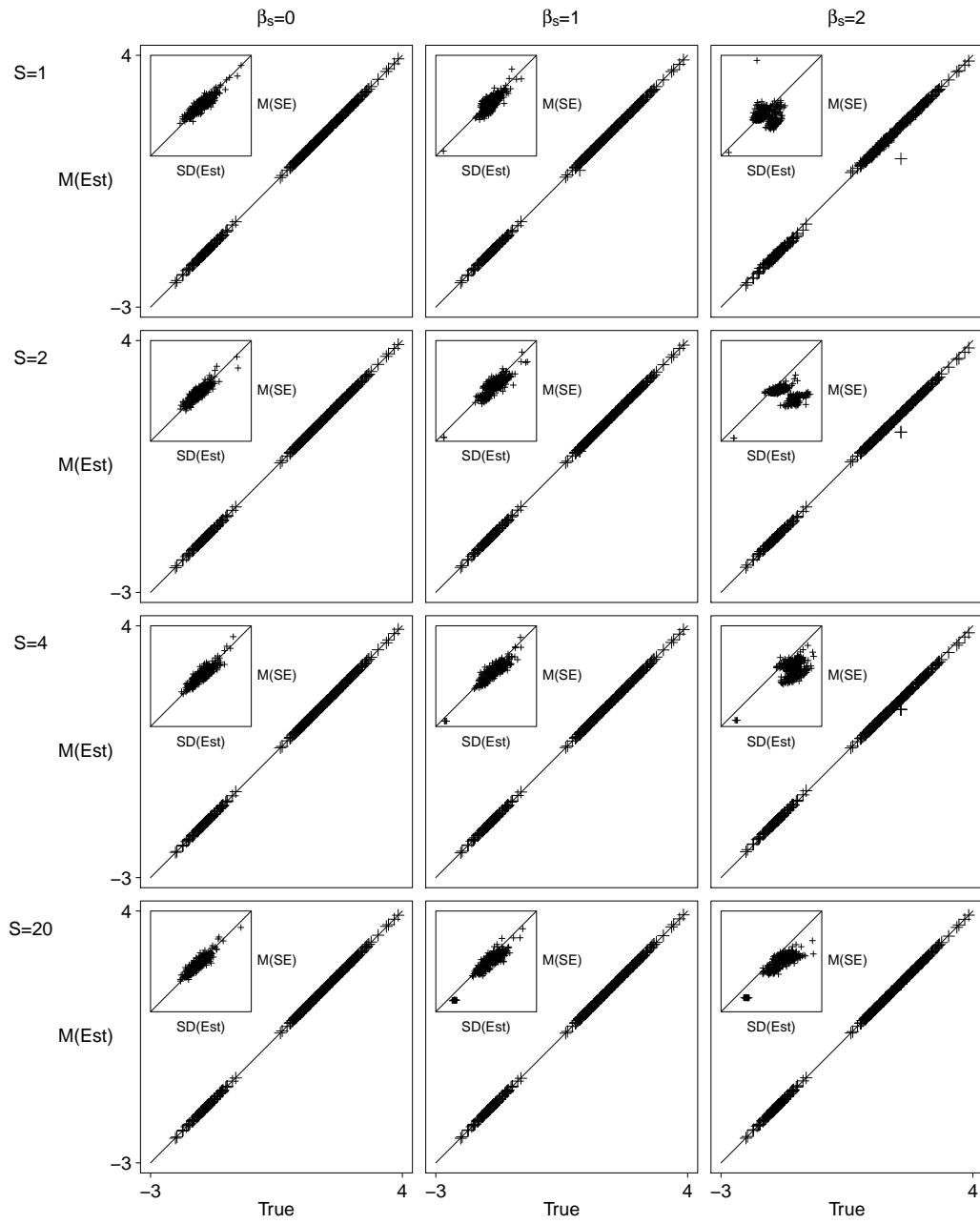


Figure 4.3: Item parameter estimates and standard errors of measurement for C-RUM models, $I = 120$, $K = 2$, $N = 5000$.

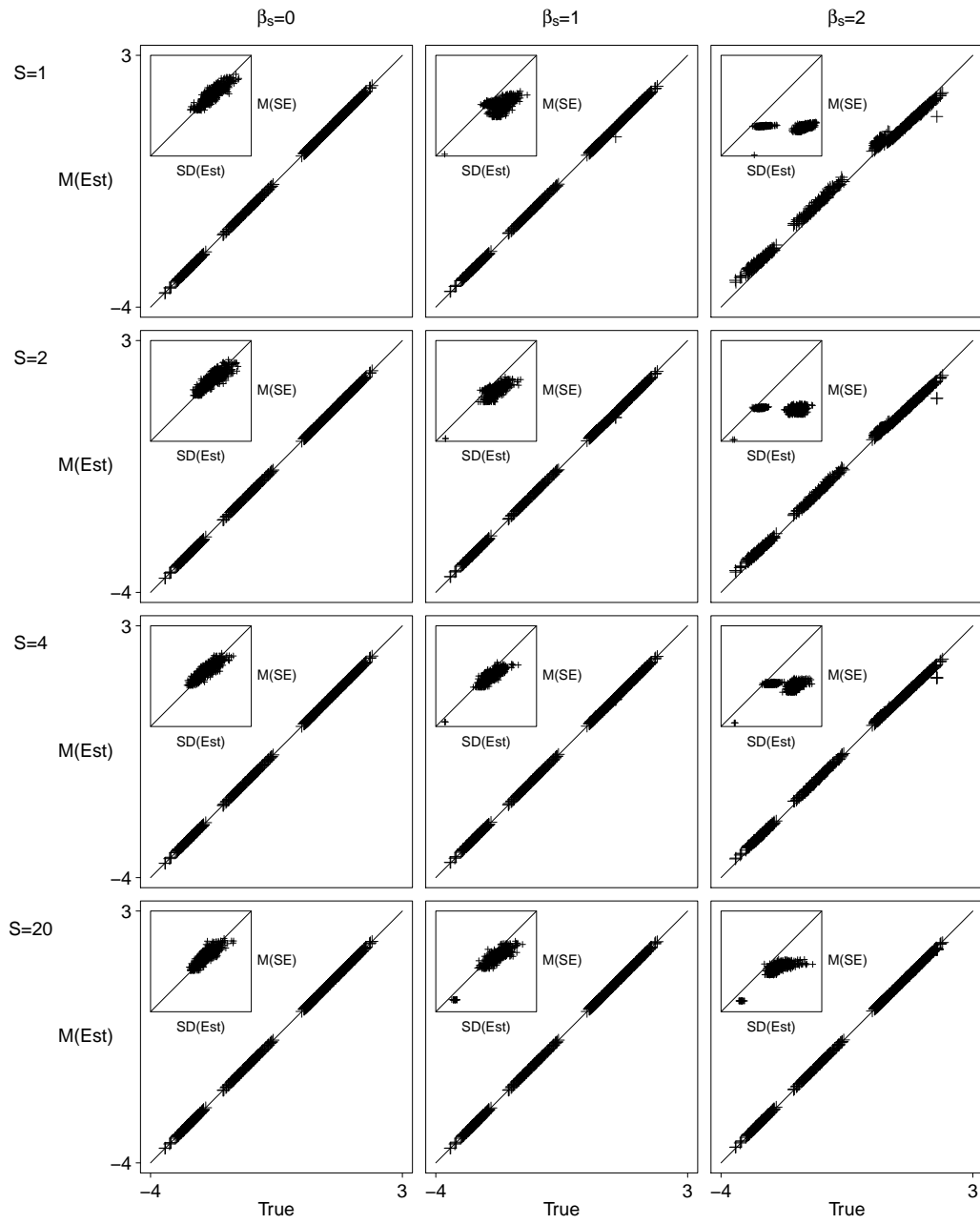


Figure 4.4: Item parameter estimates and standard errors of measurement for C-RUM models, $I = 120$, $K = 4$, $N = 5000$.

appears that the item parameters are estimated with little bias. Once again, the group-specific slope parameters are slightly underestimated when $\beta_s = 2$, and the amount of bias is reduced as the number of group-specific dimensions increases. The standard errors of measurement are also estimated well for all conditions except for $\beta_s = 2$, for which the standard errors for the intercept parameters are underestimated (most severely when $S = 1$).

4.1.4 Results for Disjunctive (DINO) Models

Simulation results for the DINO models are shown in Figures 4.9–4.12. Consistent with the findings for the other models, item parameters and their corresponding standard errors of measurement are estimated very well for the traditional diagnostic model ($\beta_s = 0$) and for hierarchical models in which the group-specific dimensions have somewhat mild influence on the item responses ($\beta_s = 1$). For conditions with stronger group-specific influence ($\beta_s = 2$), these slope parameters tend to be slightly underestimated. In addition, standard errors for the item intercepts are smaller than the Monte Carlo standard deviations of the estimates. This underestimation is most severe for conditions in which there is just one group-specific dimension with slope parameter $\beta_s = 2$. Compared to the other model types, there is slightly greater negative bias in the item slope parameters (the γ 's), which is accompanied by positive bias in the intercept parameters. These biases were greater for the conditions with ordinal data in $K = 4$ categories (see Figures 4.10 and 4.12). Perhaps not surprisingly, the conditions with $S = 1$ and $\beta_s = 2$ are the most problematic.

4.1.5 Results for Simple Structure Models

Results for the Simple structure models are presented in Figures 4.13–4.16. Across the conditions examined, the item parameters were generally estimated well, with

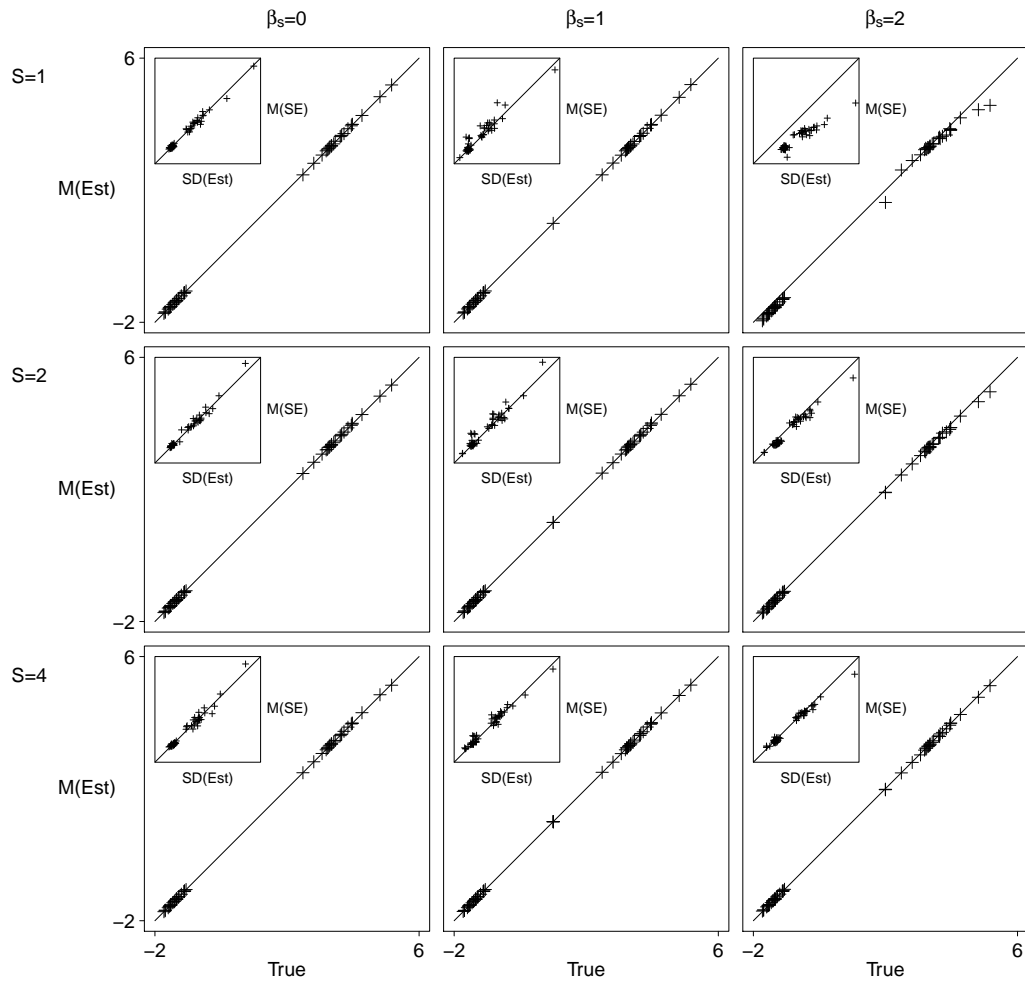


Figure 4.5: Item parameter estimates and standard errors of measurement for DINA models, $I = 24$, $K = 2$, $N = 5000$.

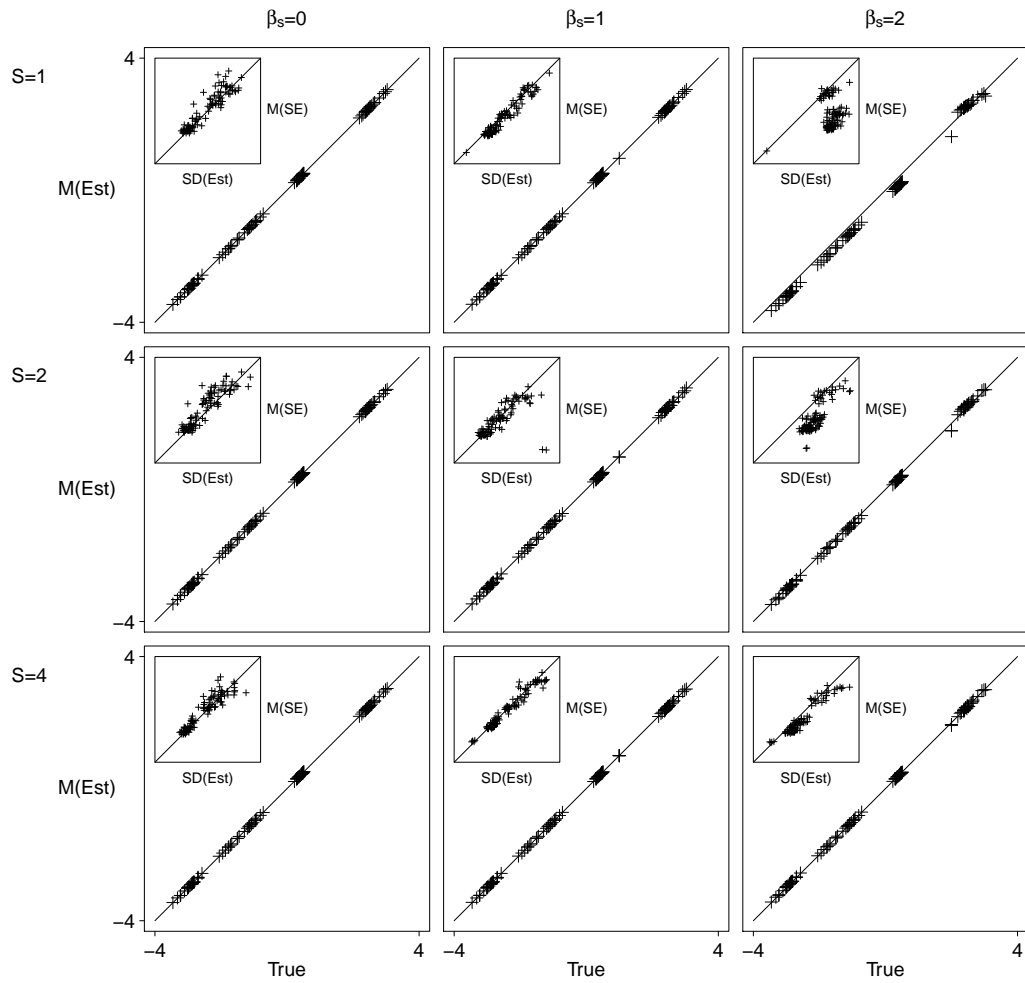


Figure 4.6: Item parameter estimates and standard errors of measurement for DINA models, $I = 24$, $K = 4$, $N = 5000$.

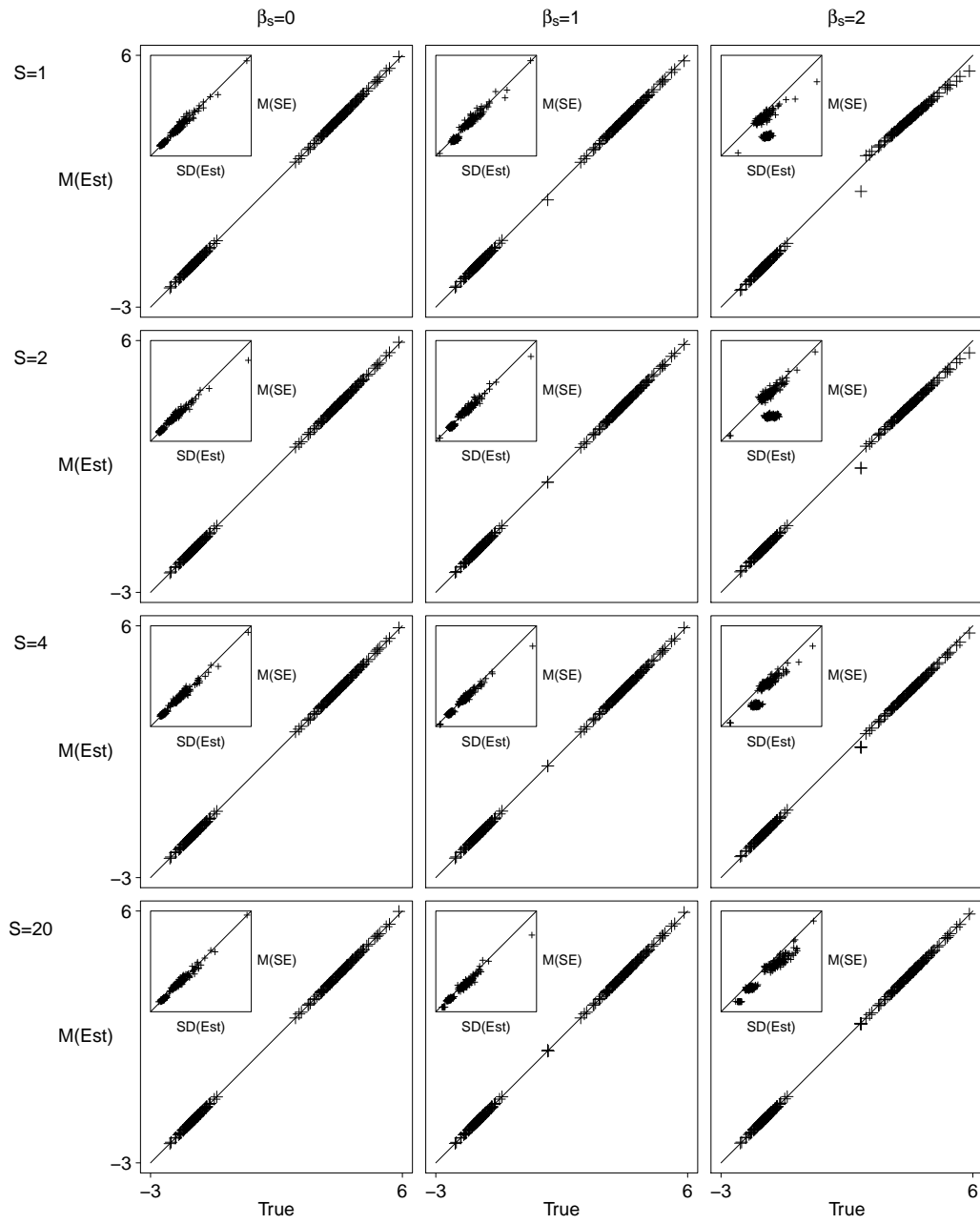


Figure 4.7: Item parameter estimates and standard errors of measurement for DINA models, $I = 120$, $K = 2$, $N = 5000$.

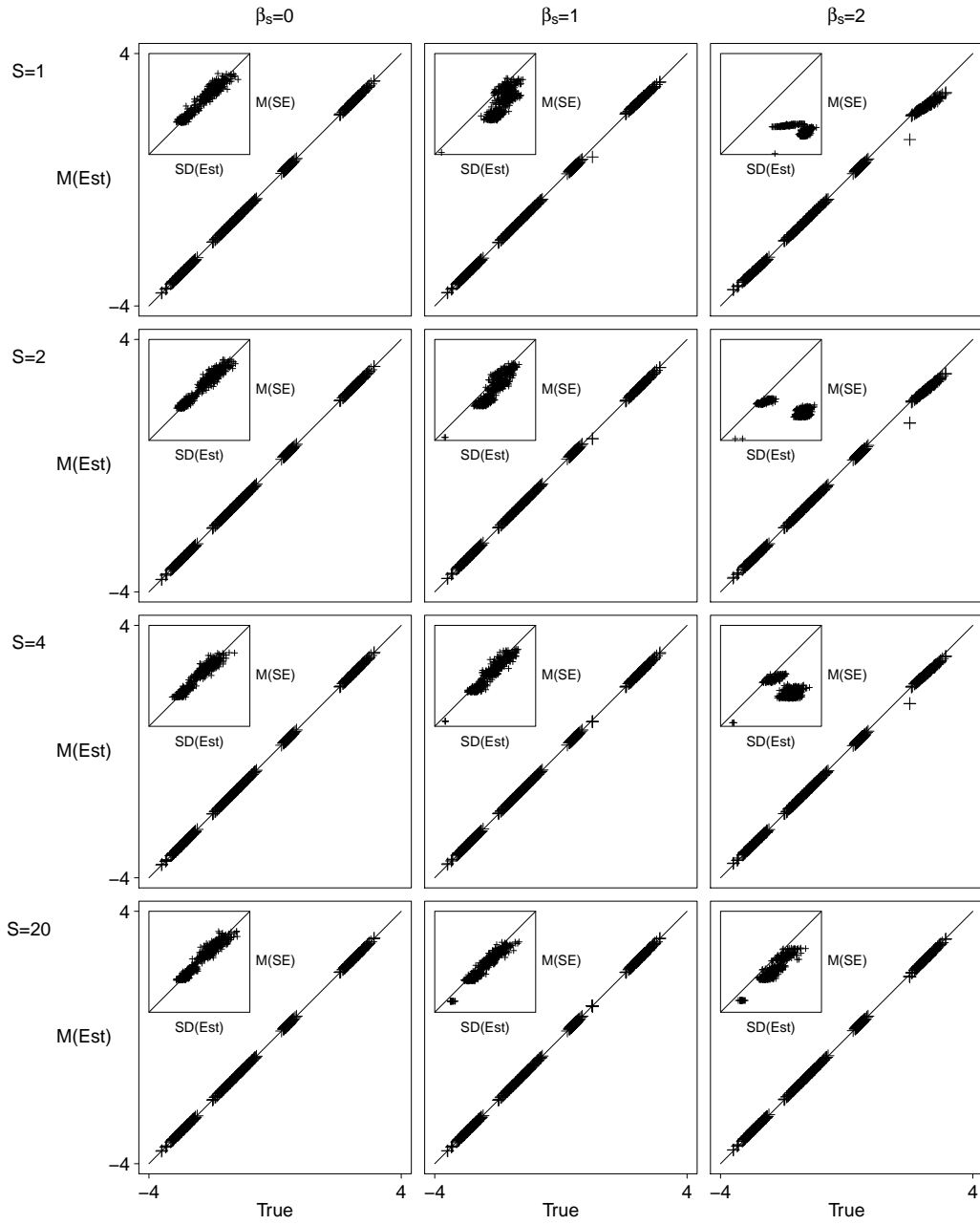


Figure 4.8: Item parameter estimates and standard errors of measurement for DINA models, $I = 120$, $K = 4$, $N = 5000$.

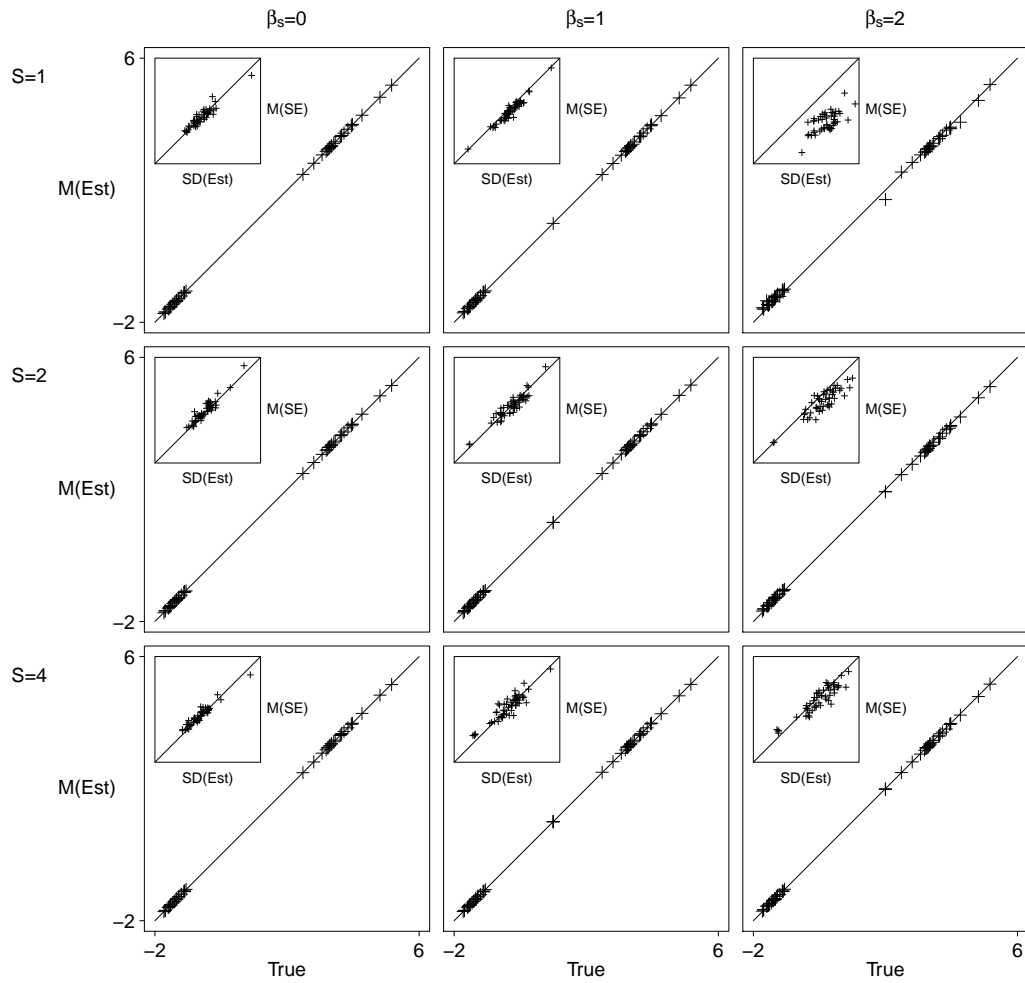


Figure 4.9: Item parameter estimates and standard errors of measurement for DINO models, $I = 24$, $K = 2$, $N = 5000$.

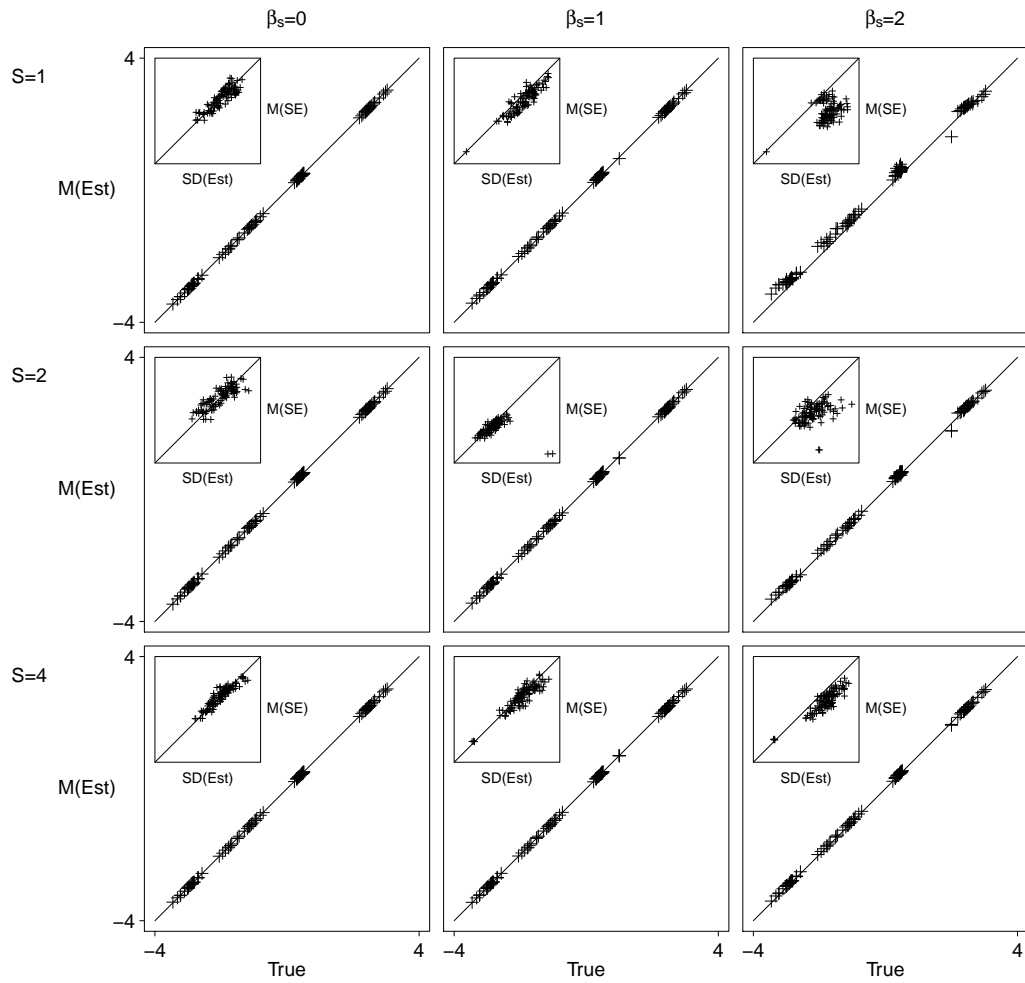


Figure 4.10: Item parameter estimates and standard errors of measurement for DINO models, $I = 24$, $K = 4$, $N = 5000$.

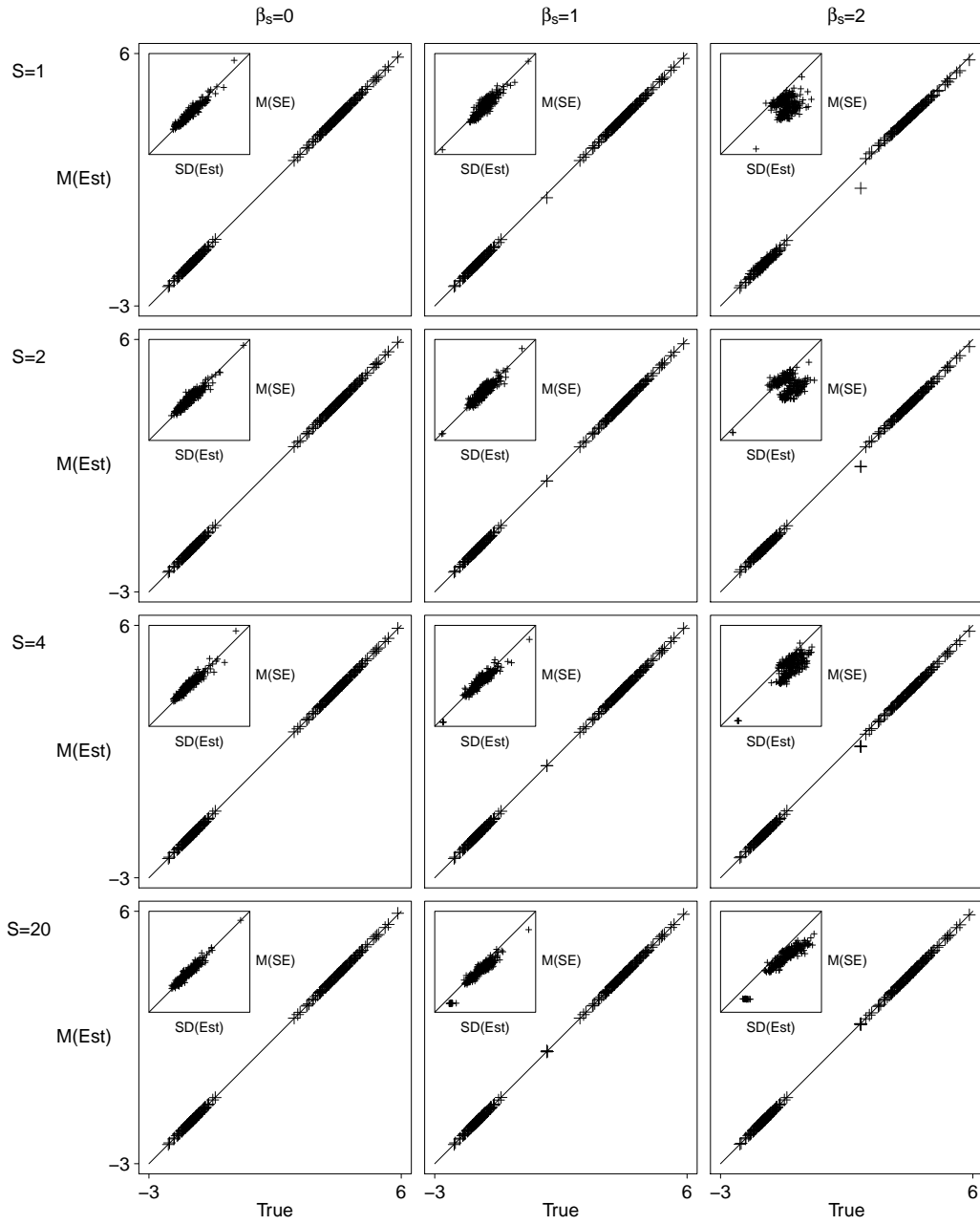


Figure 4.11: Item parameter estimates and standard errors of measurement for DINO models, $I = 120$, $K = 2$, $N = 5000$.

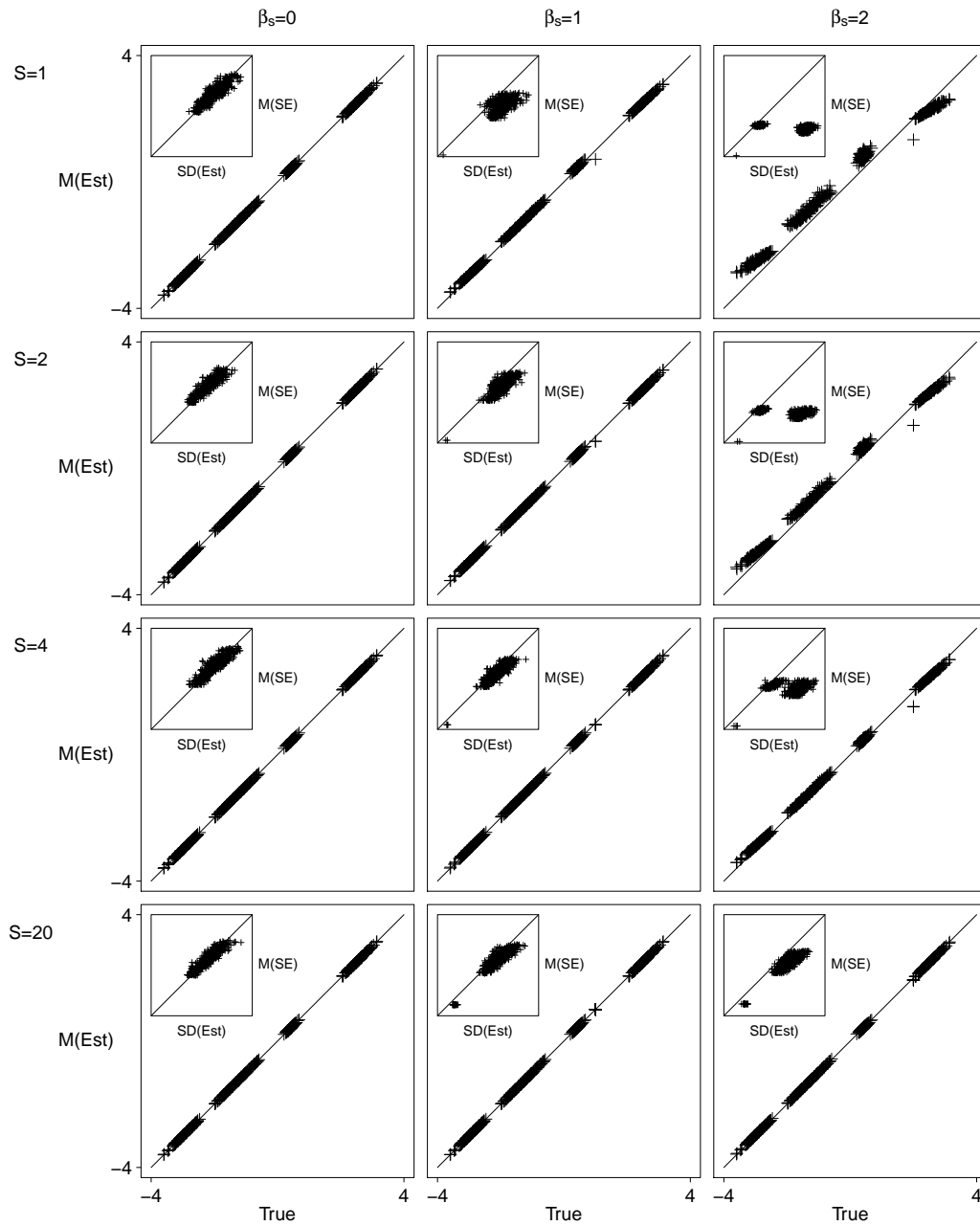


Figure 4.12: Item parameter estimates and standard errors of measurement for DINO models, $I = 120$, $K = 4$, $N = 5000$.

very little bias observed. As with the previous models, however, there was consistent negative bias in the estimation of the group-specific slope parameters (β_s), particularly the conditions with $\beta_s = 2$ and fewer group-specific dimensions ($S = 1, 2$). The standard errors of measurement followed similar trends. For the conditions $\beta_s = 0$ and $\beta_s = 1$, the Monte Carlo averages of the standard errors closely matched the standard deviations of the estimates.

4.2 Discussion

The results presented in this chapter indicate that the parameters of the proposed hierarchical diagnostic model can be estimated well under most conditions using the procedures described in Chapter 2 and implemented in the flexMIRT[®] software (Cai, 2012). In addition, the standard errors of measurement—which were computed in flexMIRT[®] using the Richardson extrapolation method to approximate the parameter error covariance matrix (Jamshidian & Jennrich, 2000; Tian et al., 2012)—closely matched (on average) the standard deviations of the parameter estimates, indicating that the reported standard errors accurately conveyed parameter uncertainty due sampling variability.

It is notable that the hierarchical model with $S = 20$ group-specific dimensions (for the tests with $I = 120$ items) was estimated without difficulty. This, of course, is a rather high-dimensional model. As such, this condition provides a good illustration of computational advantages achieved through analytical dimension reduction (Gibbons & Hedeker, 1992).

The generally positive results were consistent across the various model types (C-RUM, DINA, DINO, and Simple), test lengths ($I = 24, 120$ items), numbers of response categories ($K = 2, 4$), and sample sizes. The one set of conditions in which problems were repeatedly found was for models in which the group-specific dimensions had stronger influence (i.e., $\beta_s = 2$). Under these conditions, there was

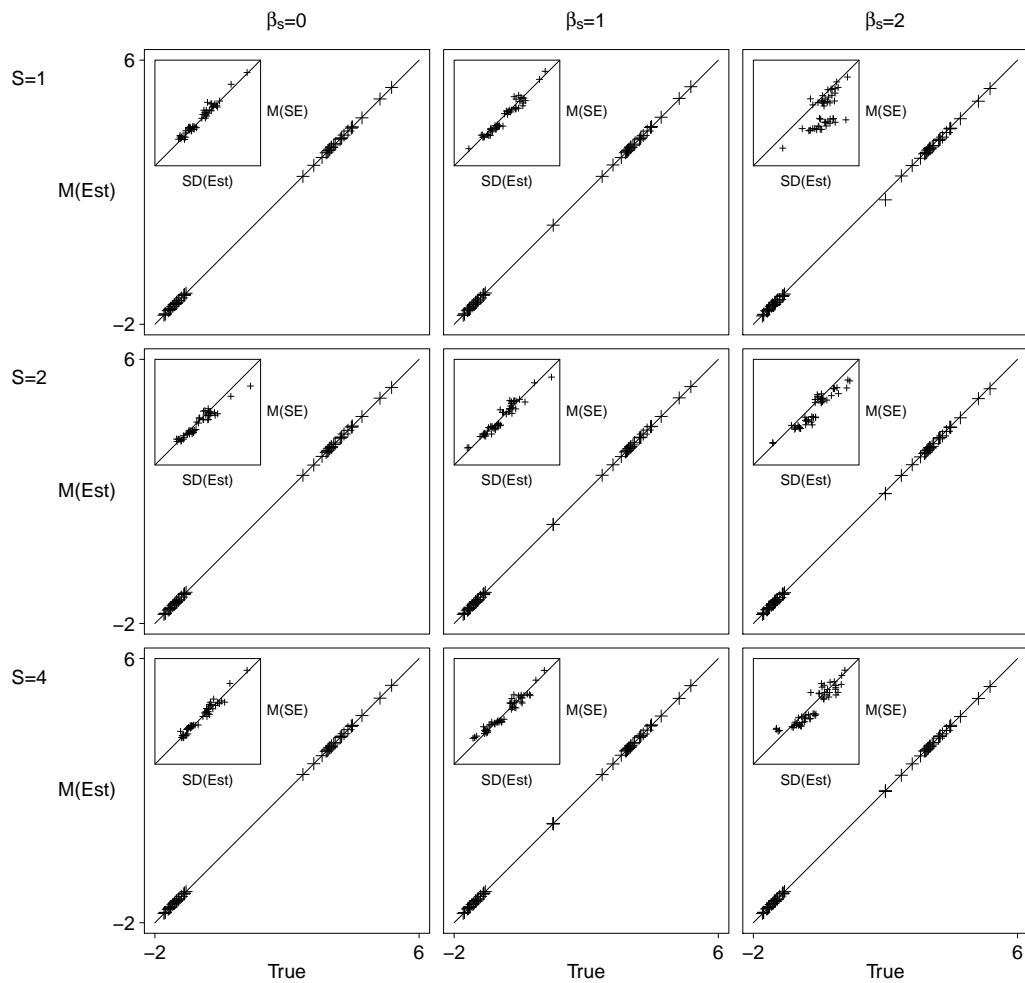


Figure 4.13: Item parameter estimates and standard errors of measurement for Simple models, $I = 24$, $K = 2$, $N = 5000$.

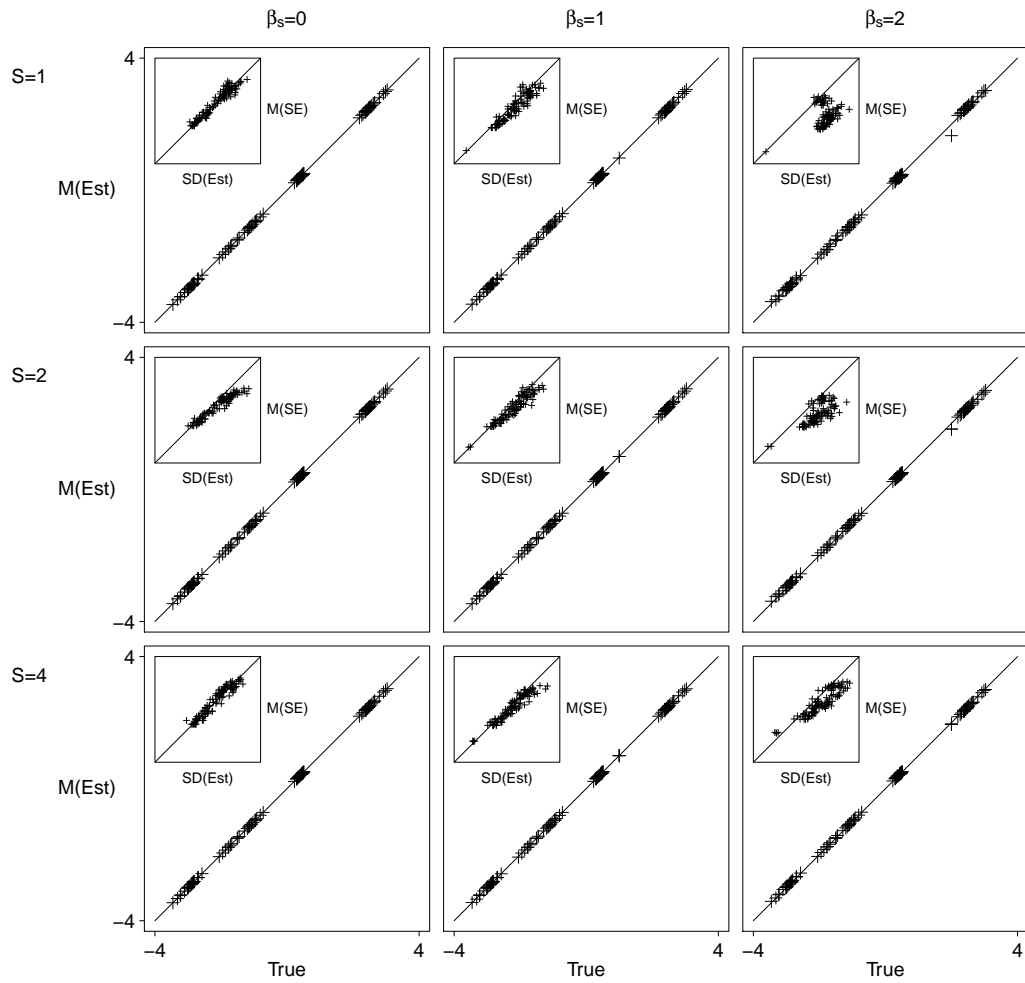


Figure 4.14: Item parameter estimates and standard errors of measurement for Simple models, $I = 24$, $K = 4$, $N = 5000$.

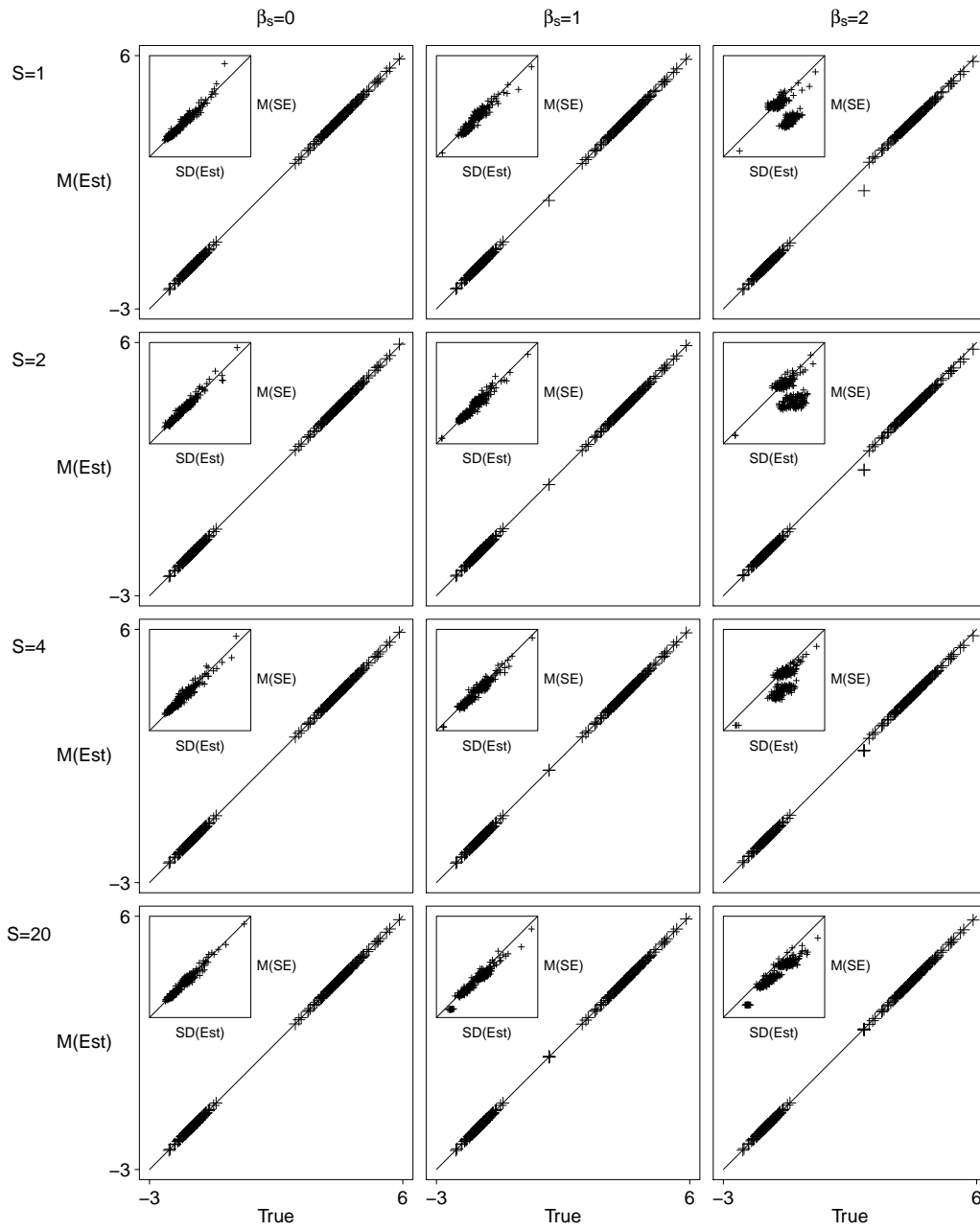


Figure 4.15: Item parameter estimates and standard errors of measurement for Simple models, $I = 120$, $K = 2$, $N = 5000$.

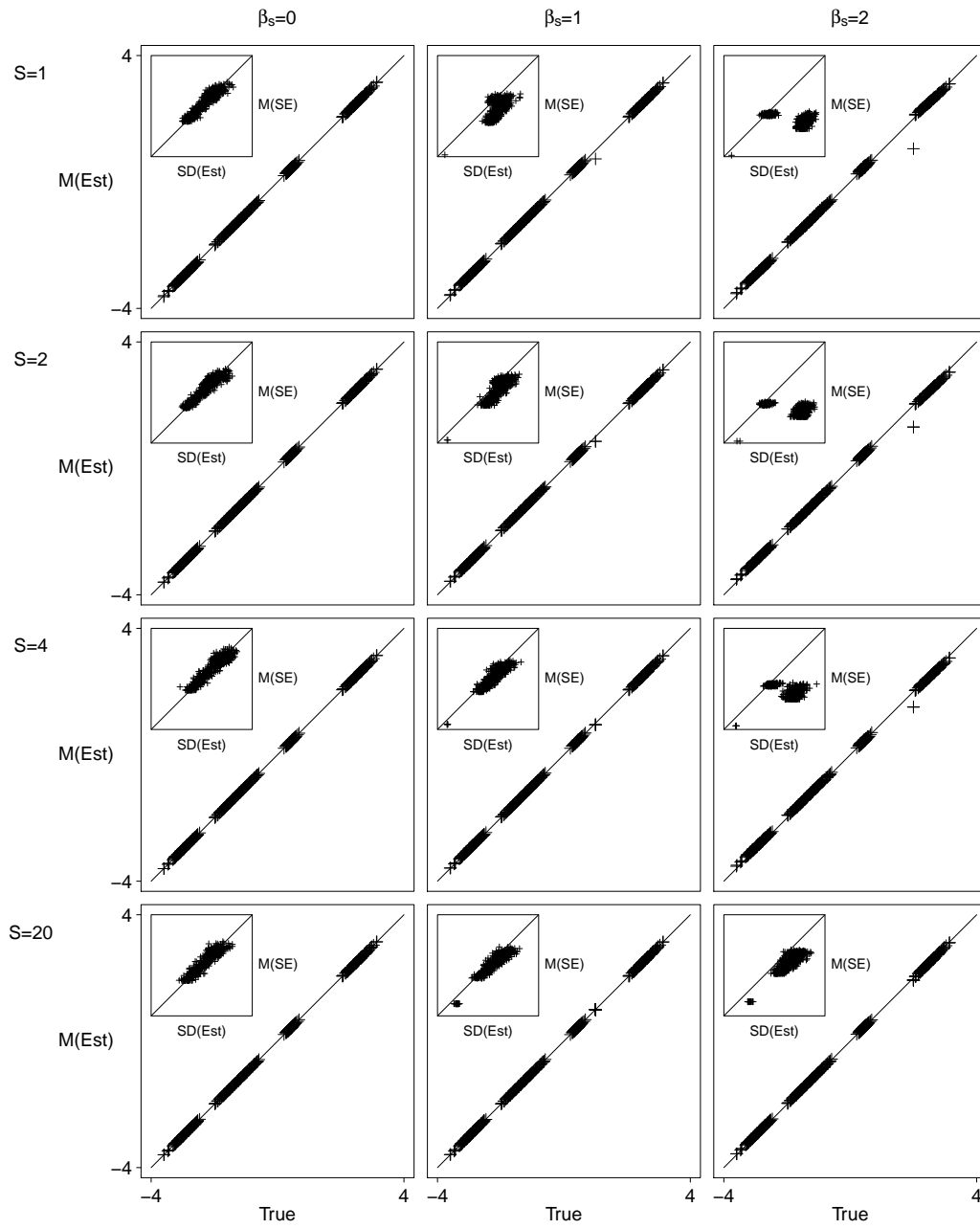


Figure 4.16: Item parameter estimates and standard errors of measurement for Simple models, $I = 120$, $K = 4$, $N = 5000$.

negative bias in the group-specific slope parameters, negative bias in the slopes on the attribute variables, and, in a smaller number of cases, some positive bias in the item intercepts. Standard errors were also underestimated under these conditions for many of the item parameters. Among the conditions with $\beta_s = 2$, magnitude of the bias was inversely related to the number of group-specific dimensions. For tests of $I = 120$ items, the negative bias in the slope parameters was negligible for models with $S = 20$ group-specific dimensions. Conditions with 4 or 2 group-specific dimensions exhibited some amounts of bias, while the model with $S = 1$ produced the most bias among the conditions examined.

Further study is needed to explore the sources of bias in these few problematic conditions and, if possible, to identify remedies. It should be noted that no prior distributions were placed on the item parameters when estimating these models. Specification of priors might resolve some of the difficulties, though it's also possible that the underlying problems go deeper. The fact that the most severe biases arose in the conditions with only one group-specific dimension raises the question of whether this model may be identified from the data. To date, there has been only limited discussion of how one might determine whether or not a given diagnostic model is identified (see, e.g., von Davier, 2013), but this will be an important matter to examine as the use of diagnostic models becomes more widespread.

CHAPTER 5

Simulation Results: Impact of Misspecification

The previous chapter demonstrated that the hierarchical diagnostic model can be estimated well for most conditions examined. However, there remains some uncertainty concerning the need for such a model. The purpose of this chapter, then, is to examine the consequences of ignoring the sorts of nuisance dimensions that the hierarchical diagnostic model is intended to address.

There are, of course, many different consequences that might be examined, including potential impacts on the estimates of item parameters, characterizations of the population distribution of attribute profiles, or various aspects of model fit. Some of these will be discussed in subsequent chapters. However, here the focus will be on the effects of model misspecification on the accuracy of examinee scores and classifications. There are two reasons for this particular focus. First, many of the other potential consequences—including each of those mentioned above—would be expected to be associated with some (not necessarily large) impact on score estimates. Thus, identification of any effect of misspecification on score estimates would imply some impact (i.e., bias or distortion) on the estimated model that mediates the effect on scoring. The second reason is related to the first: If there is no impact found on scores, then there is perhaps less reason to be concerned about other consequences. One might search for and find compelling evidence that ignoring nuisance dimensions results in biased item parameter estimates, for example. However, such a finding doesn't itself provide a clear answer to the question of whether one ought to be fitting a more complex model.

The results in this chapter will be presented in two parts. First, I report analyses of examinee scores obtained from the Monte Carlo study described in Chapter 3. For each replicated dataset, scores were obtained under the correctly specified hierarchical model and under a traditional diagnostic model that ignores the influence of the data generating (hierarchical) model’s group-specific dimensions. Attribute classification decisions were made based on these scores, and a number of accuracy measures—including overall correct classification rate (OCC), sensitivity, and specificity, among others—were computed for each study replication. The mean values of these measures over the replications are reported for the various study conditions.

To complement the Monte Carlo study results, a second analysis is presented that utilizes the large-sample datasets ($N = 20000$) generated from the same models used in the simulation study. Here, however, only single replications from each condition were analyzed. The first purpose of this in-depth analysis was to examine an alternative measure of classifier accuracy or utility—the area under the receiver operating characteristic (ROC) curve. Second, I examine the extent to which examinee EAP estimates accurately convey the uncertainty in classification, given observed item response data.

5.1 Measures of Classification Performance

In this section, examinee classifications made on the basis of EAP scores from the traditional and hierarchical models are evaluated on the basis of several traditional measures of diagnostic performance. The measures presented in this section are derived from the contingency tables obtained by cross-classifying examinees by their true attribute status and their test-based classification. The relative strengths and weaknesses of each have been discussed extensively (see, e.g., Streiner, 2003). They are presented here to illustrate the variation in results

that are may be obtained across the measures.

The differences in results across models and sample sizes were minimal. Accordingly, only results for the C-RUM models are shown here (and in the subsequent sections). Figures 5.1 and 5.2 show the performance of the tests of $I = 24$ and $I = 120$ items, respectively. Within each plot, values of one performance measure are given for the values of group-specific slope parameters ($\beta_s = 0, 1, 2$). Thus, the influence of nuisance dimensions increases from left to right. The number of group-specific dimensions also varies across the columns (from left to right, $S = 1, 2, 4$).

Across all conditions and performance measures, the models perform equivalently when $\beta_s = 0$, as is expected. However, as the group-specific dimensions exert greater influence, the levels of performance for the two models diverge. Across the various measures, the hierarchical diagnostic model consistently demonstrated better classification, even in the presence of strong nuisance dimensions. Differences among performance measures are noteworthy. The overall correct classification rate (OCC) tended to be quite high across conditions. In contrast, Cohen's kappa statistic, which adjusts for the amount of correct classification that would be expected due to chance, provides a much less favorable assessment.

5.2 Receiver Operating Characteristics (ROC) Graphs and the Area Under the Curve (AUC)

In this section, I examine the performance of the diagnostic tests with respect to an alternative measure, the receiver operating characteristic (ROC) graph, which plots the true positive rate (TPR) of classification against the false positives rate (FPR). As such, it can be viewed as a comparison of diagnostic test benefits against costs. The advantage of this approach to describing classification performance is that, unlike the measures considered in the previous section, it is

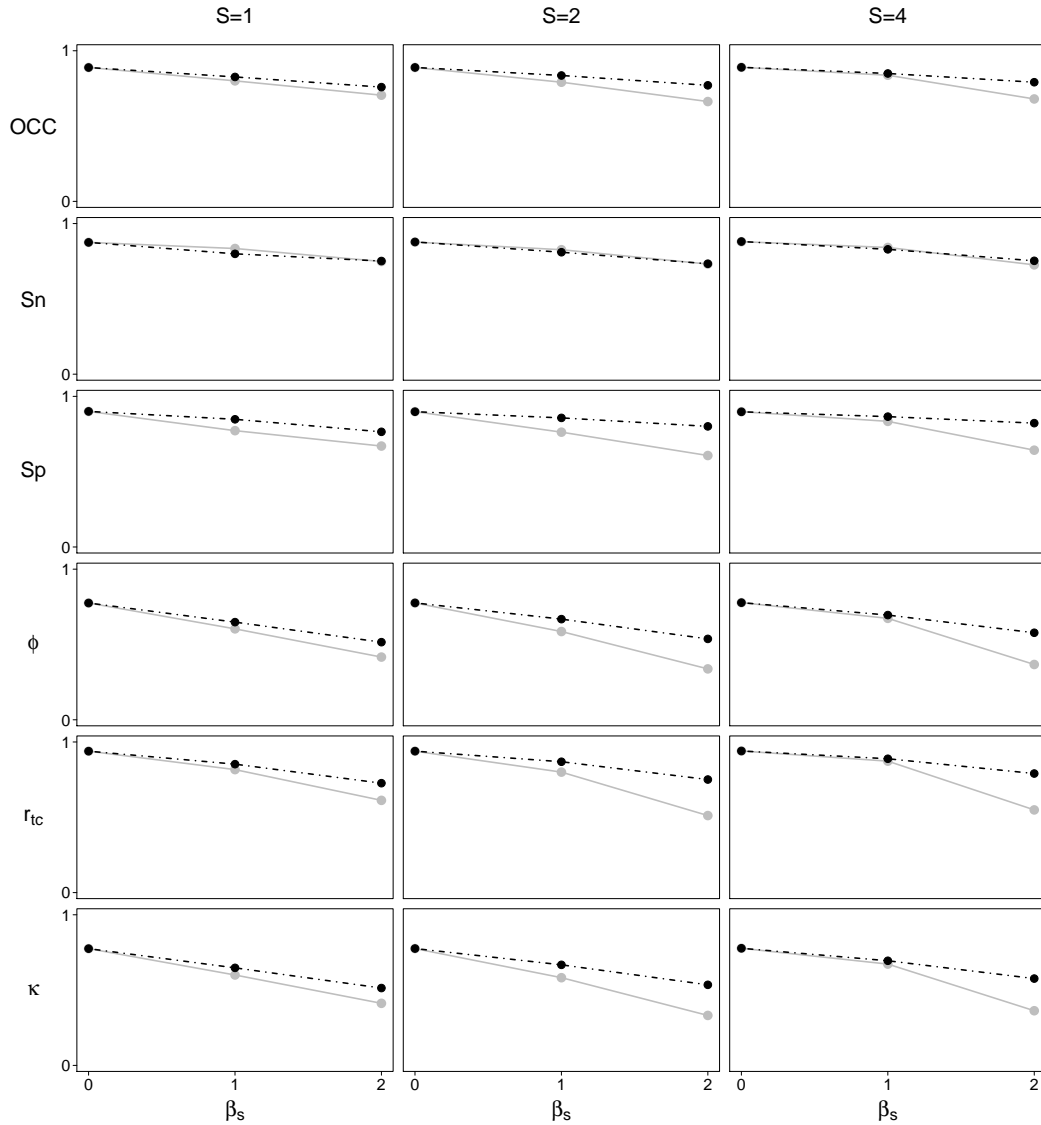


Figure 5.1: Classification performance for C-RUM models with $I = 24$, $K = 2$, $N = 5000$. Results shown are for attribute x_2 using an EAP threshold of 0.5. Results obtained from the traditional diagnostic model are shown in gray; those obtained using the hierarchical diagnostic model are shown in black. OCC is overall correct classification, Sn is sensitivity, Sp is specificity, ϕ is the phi coefficient, r_{tc} is the tetrachoric correlation, and κ is Cohen's kappa.

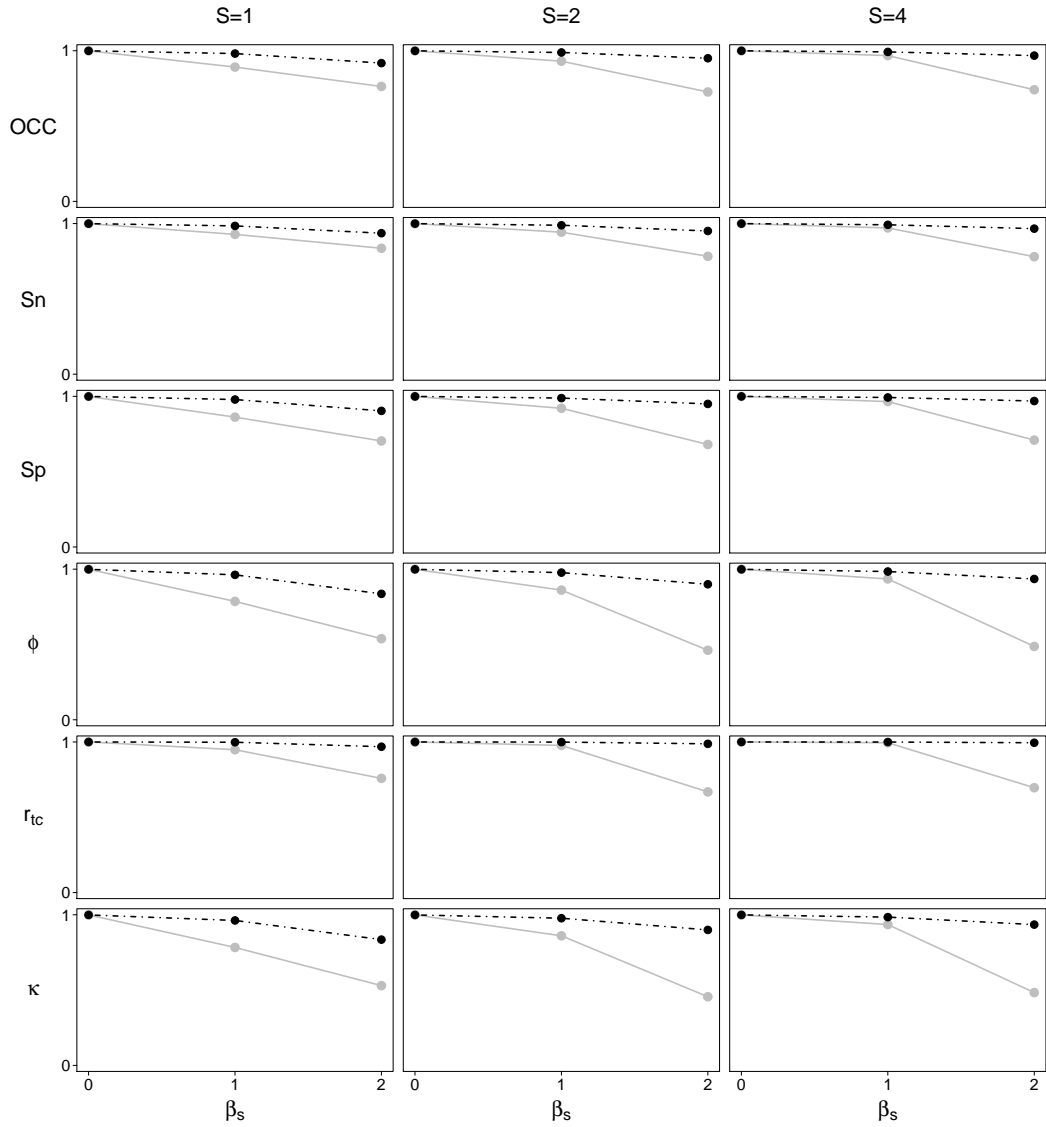


Figure 5.2: Classification performance for C-RUM Models with $I = 120$, $K = 2$, $N = 5000$. Results shown are for attribute x_2 using an EAP threshold of 0.5. Results obtained from the traditional diagnostic model are shown in gray; those obtained using the hierarchical diagnostic model are shown in black. OCC is overall correct classification, Sn is sensitivity, Sp is specificity, ϕ is the phi coefficient, r_{tc} is the tetrachoric correlation, and κ is Cohen's kappa.

unnecessary to specify a threshold for classification; the graph shows trade-offs between TPR and FPR over the entire range of possible classifications. A second benefit of using the ROC is that TPR and FPR are independent of the sample examines and the distribution (or prevalence) of the attributes being measured.

The area under the ROC curve (AUC) may be used as a single-number summary of classification performance. The value of AUC ranges from 0.5 to 1.0. Values close to 0.5 indicate that the classification does little better than would be expected by chance.

ROC curves were generated (and corresponding AUC values were computed) from examinee EAP scores and true attribute status for the large ($N = 20000$), single-replication simulated datasets under both the traditional and hierarchical diagnostic models. Results from these analyses were similar across the model types and number of response categories. Thus, as in the previous section, I focus on the results obtained for the C-RUM models for dichotomous data. Figure 5.5 presents the ROC curves for tests of $I = 24$ items. Results for the tests of $I = 120$ items are shown in Figure 5.6.

For the condition where $\beta_s = 0$ (in the first column) the two models produce identical results, as expected. For conditions with $\beta_s = 1$, the curves show some minor divergence, and the AUC values of the traditional models are slightly lower than those of the hierarchical model. Differences between the models are substantially greater for the conditions with $\beta_s = 2$. Notably, for tests of $I = 120$ items, the hierarchical model maintains very large AUC values (>0.97)—indicating excellent classification performance, despite the presence of the nuisance dimensions.

5.3 Appraisal of Classification Certainty

As a final evaluation of the impact of nuisance dimensionality, EAP scores (which describe posterior probabilities of possessing an attribute) were compared to the

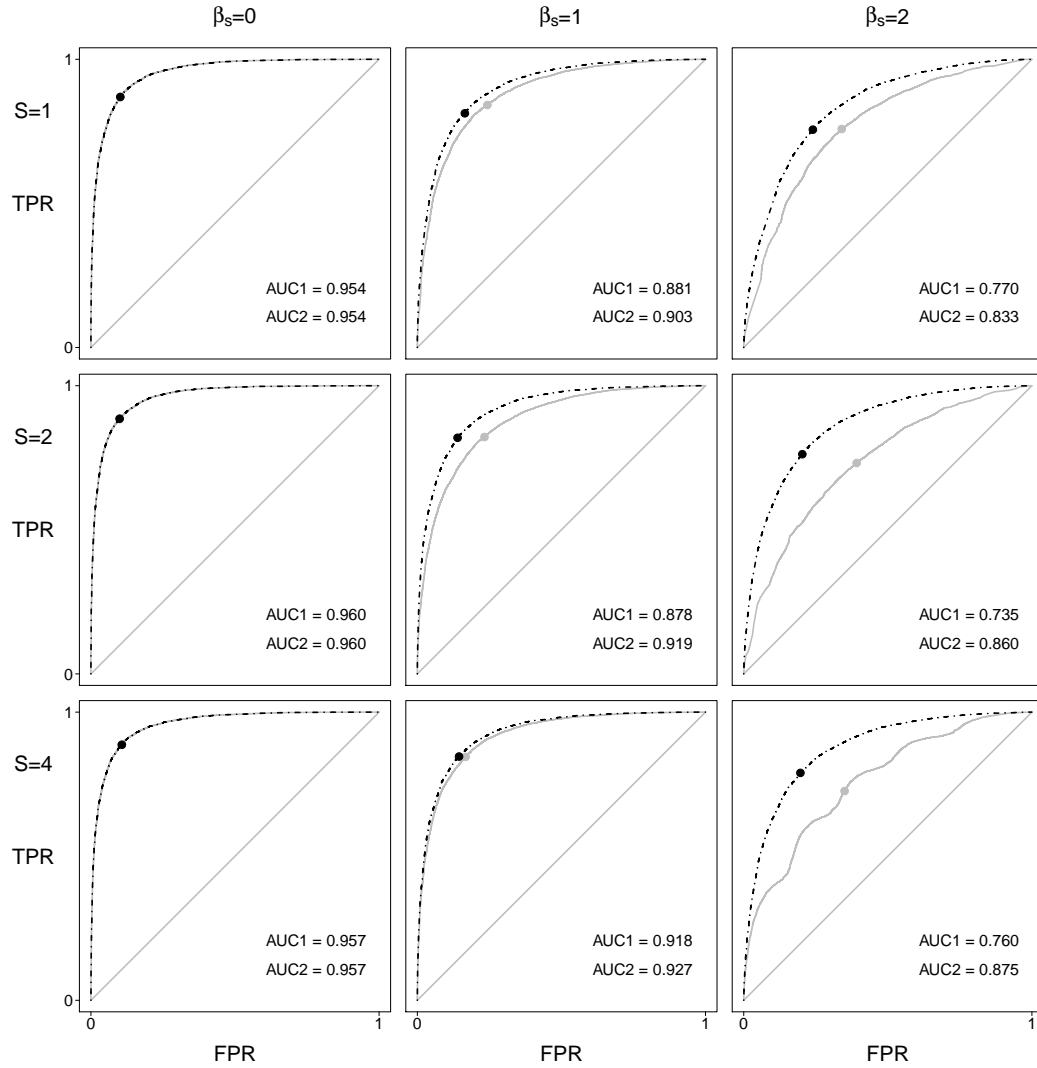


Figure 5.3: Estimated and true conditional attribute probabilities for the C-RUM Model, $I = 24$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model (in gray). Model 2 is the hierarchical diagnostic model (in black). Results shown are for attribute x_2 . Single point drawn on each curve corresponds to ROC coordinates for EAP threshold of 0.5.

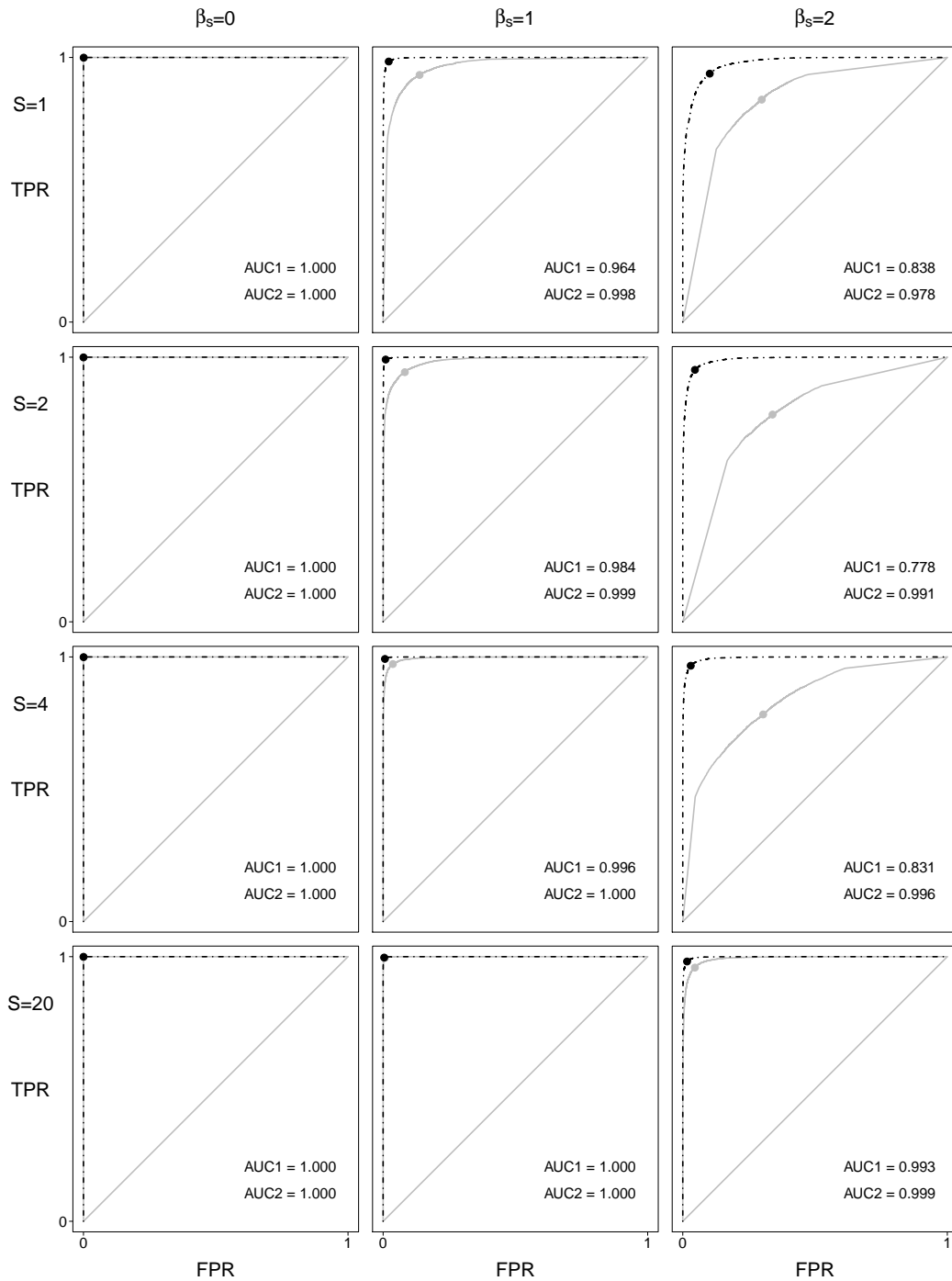


Figure 5.4: Estimated and true conditional attribute probabilities for the C-RUM Model, $I = 120$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model (in gray). Model 2 is the hierarchical diagnostic model (in black). Results shown are for attribute x_2 . Single point drawn on each curve corresponds to ROC coordinates for EAP threshold of 0.5.

true (data generating) model-implied attribute distributions for each pattern of observed item responses. Once again, the large-sample ($N = 20000$), single replication simulated datasets were used. Figures 5.5 and 5.6 present scatterplots for the C-RUM models with dichotomous data. In these plots, the EAP score (shown as $\hat{P}(x = 1|\mathbf{y})$) is plotted against the average attribute status for the response pattern, $M(x|\mathbf{y})$. Also shown on the plots are a root means square difference between the estimate and population value and Kendall's tau (τ) coefficient.

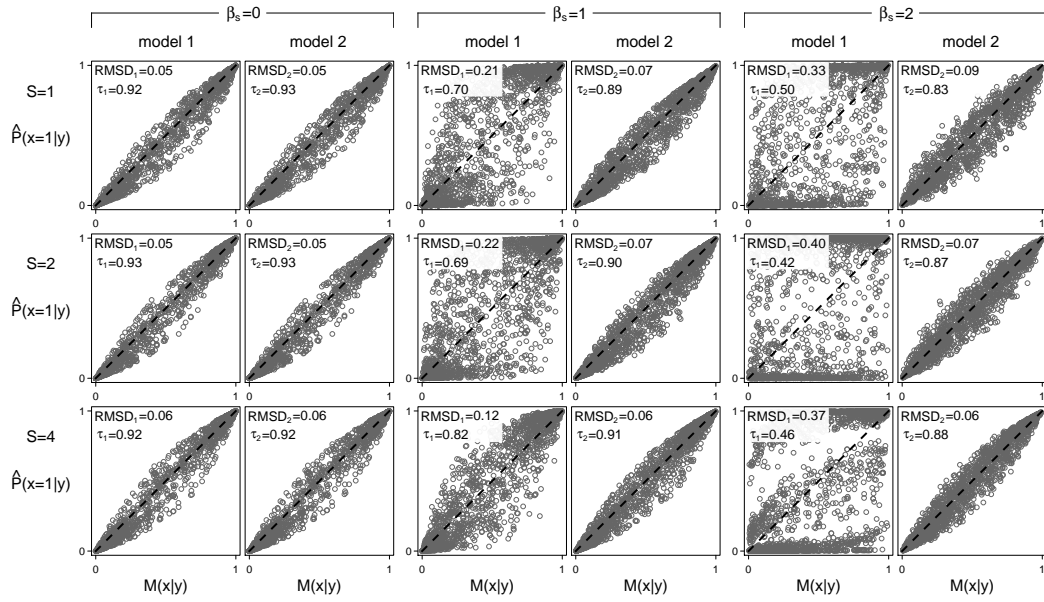


Figure 5.5: Estimated and true conditional attribute probabilities for the C-RUM Model with $I = 24$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model. Model 2 is the hierarchical diagnostic model. Results shown are for attribute x_2 .

The results of these analyses indicate that the EAP scores under the traditional models are not well-calibrated in the presence of nuisance dimensionality. This means that the level of certainty implied by the estimated posterior probability is often not justified, given the actual attribute probability for an individual response pattern. Perhaps most concerning are cases with true probabilities near 0.5 that are assigned EAP scores close to 0 or 1.

5.4 Discussion

The analyses presented in this chapter were conducted in order to address the question of whether hierarchical diagnostic models are necessary. The approach taken was to examine the effects of nuisance dimensionality on the scores obtained using traditional diagnostic models. It is clear that the presence of nuisance dimensions has a detrimental effect on the ability to accurately classify individuals on the latent attributes of interest. Importantly, the hierarchical models were able to maintain substantially better levels of classification performance. In addition, scores obtained from the hierarchical diagnostic models provide much better characterization of the uncertainty in attribute status, given observed response pattern, than the traditional models. These results suggest that there may be contexts in which the traditional models would fail to provide good classification but that the hierarchical models may represent an effective alternative.

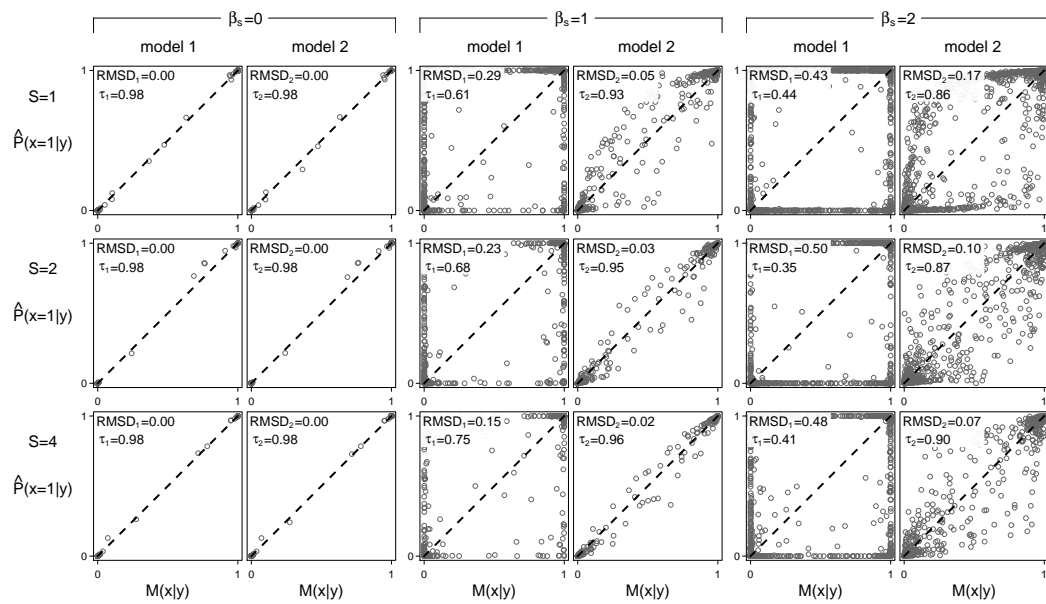


Figure 5.6: Estimated and true conditional attribute probabilities: C-RUM Model with $I = 120$, $K = 2$, $N = 20000$. Model 1 is the traditional diagnostic model. Model 2 is the hierarchical diagnostic model. Results shown are for attribute x_2 .

CHAPTER 6

Simulation Results: Characterizing Model Misfit

In this chapter, I present results from simulation studies examining the performance of the Chen and Thissen LD X^2 statistic in detecting and characterizing violations of local independence due to various types of model misspecification.

In the first section, I focus on the types of misspecification that motivate the development of the hierarchical diagnostic model: unmodeled dimensions that are typically unrelated to the attributes of primary interest. Such dimensions create inter-item dependencies that cannot be fully explained by the attributes. If the LD X^2 index is sensitive to these types of misspecification, then it could perhaps be used to (a) identify contexts in which the hierarchical model is needed, (b) inform the specification of that hierarchical model (by characterizing the particular items exhibiting residual dependence), and (c) evaluate whether the alternative model has achieved the desired improvement in model fit.

In the second section, I explore the utility of the LD X^2 index for detection of a broad range of types of model misspecification, including incorrect specification of the Q-matrix and application of the wrong diagnostic model (e.g., fitting a DINO model to data generated according to a DINA model). The fit indices are evaluated in cases where these misspecifications are present either in isolation or in combination with unmodeled nuisance dependencies. Related to this second case, one question this study seeks to address is whether the presence of nuisance dimensions could obscure other types of model misspecification (and, further, whether modeling these nuisance dimensions—even as the model remains

misspecified some other way—might help to identify that additional misfit).

6.1 Misspecification Due to Nuisance Dimensionality (Violations of Local Item Independence)

In order to investigate the performance of the LD X^2 index in detecting local independence violations due to nuisance dimensionality, the values of the index for every item pair were obtained from each replication of the Monte Carlo study described previously (Chapter 3). Two models were fit to each simulated dataset. One model was a traditional diagnostic model, the other a hierarchical model that includes random effects to account for latent dimensions unrelated to the cognitive attributes of interest. The data generating conditions varied in diagnostic model type, number of nuisance dimensions, strength (or level of influence) of those dimensions, number of item response categories, test length, and sample size (see Table 3.1). For each simulation condition, 100 datasets were generated. Here, the performance of the LD X^2 index across those replications is examined.

Due to the similarity of findings across model types and sample sizes, I focus here on the results obtained for the DINA model (and the hierarchical DINA) with calibration sample sizes of $N = 1000$.

Figures 6.1–6.4 show the Monte Carlo averages of the RMSEA for each item pair under the various simulation conditions through the shading of boxes corresponding to the matrix coordinates of each item pair. A large value for the test statistic (and corresponding RMSEA) conveys that the association between two items is not explained well by the model. However, when a large value is observed, it is often also useful to know in what way the explanation is wrong. As described in Chapter 3, a signed version of LD X^2 may be obtained by comparing the observed and expected correlations between two items, which provides an indication of whether the model has over- or under-explained the association.

The information concerning the direction of dependence is represented in the figures by the color of shading. Item pairs for which greater than 75% of replications exhibited positive local dependence are shown in shades of red that increase in intensity with the average RMSEA value. Item pairs with negative dependence in more than 75% of replications are represented in shades of blue, with darker blues thus indicating larger RMSEA values and consistent negative local dependence. The remaining item pairs, which displayed positive dependence in some replications and negative dependence in others (but without one result being substantially more frequent than the other) are shaded gray. In fact, these were generally cases with very small LD X^2 values, such that the shading is very light anyways. The RMSEA values obtained from the traditional DINA model (ignoring group-specific dimensions) are shown below the main diagonal, while results from the hierarchical model are shown above the diagonal. The figures also report the mean RMSEA values across all replications and item pairs, $M(\epsilon_1)$ and $M(\epsilon_2)$ for the traditional and hierarchical models, respectively.

For each condition, a mean rejection rate was computed for each model. The rejection rates are the number of times the LD X^2 test statistic exceeded the critical value for $\alpha = 0.05$ for a given item pair, over the total number of Monte Carlo replications. The mean rate is the average of the rejection rates across all item pairs, of which there are $I(I - 1)/2$ (276 for the 24-item test and 7140 for the 120-item test). The mean rejection rates—reported in the figures as $M(\text{rej}_1)$ for the traditional model and $M(\text{rej}_2)$ —are primarily useful in judging the overall calibration of the test statistic under the conditions of correct model misspecification. Of course, the individual rejection rates also indicate the level of power to detect misspecification for each item pair, when such misspecification is present. However, the *mean* rejection rates cannot be interpreted in this way. The one possible exception would be those conditions in which all items load on a single group-specific dimension ($S = 1$) with nonzero slope ($\beta_s = 1, 2$), since that would

be a case in which the traditional model's specification of the latent variables with common influence on any two items would be incomplete for every single item pair.

In the following sections, I discuss the results presented in Figures 6.1–6.4. I first examine the calibration of the LD X^2 statistic when the data generating model is a traditional diagnostic model (with no influence of group-specific dimensions on item responses). Then, I discuss the extent to which the index provides characterization of misspecification when the traditional model is fit to data generated from the hierarchical model. Finally, I examine the extent to which fitting the correctly specified hierarchical model resolves the dependence identified under the traditional model.

6.1.1 Calibration of LD X^2 for the Traditional DINA Model

As an initial assessment of the potential usefulness of the LD X^2 index for diagnosing model misspecification, the calibration of the statistic under the traditional model was examined. Of interest here is whether the empirical rejection rate is close to the nominal value used to specify critical values for the test. This analysis focuses on the first column of Figures 6.1–6.4 (those conditions with $\beta_s = 0$). Results for the traditional model are shown below the diagonal within each panel. Across the conditions shown, the average rejection rate for dichotomous items (Figures 6.1 and 6.3) ranged from 0.023 to 0.027. For the items with $K = 4$ ordered categories (Figures 6.2 and 6.4), the average rejection rate ranges from 0.044 to 0.047. This result is consistent with previous findings for IRT models, that LD X^2 is somewhat conservative for dichotomous items (Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2012) but that its calibration improves as the number of categories increases (Hansen & Cai, 2012). The average RMSEA values under these conditions were consistently quite small (0.008 for these conditions), indicating the good fit of these correctly specified models.

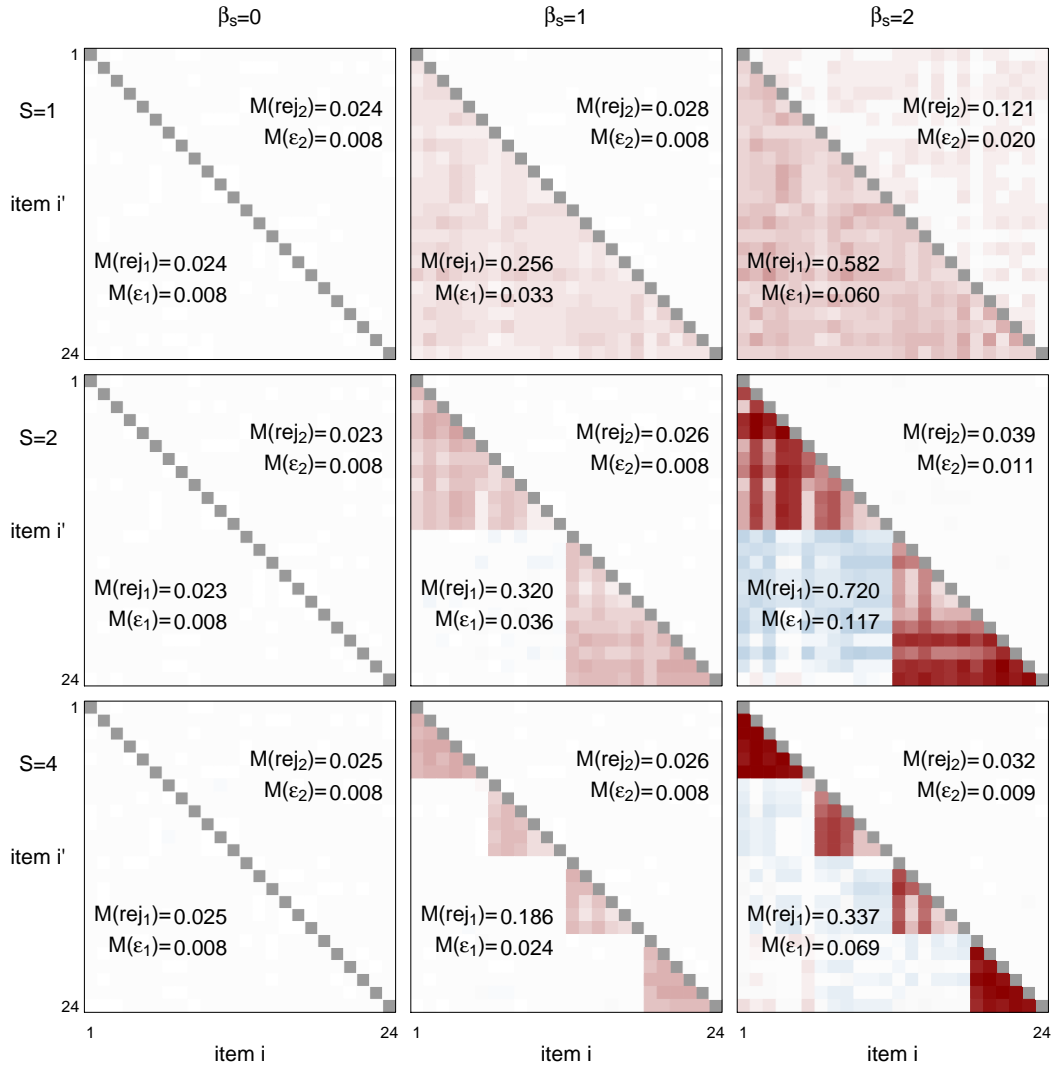


Figure 6.1: LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 24$ items in $K = 2$ response categories and sample size of $N = 1000$.

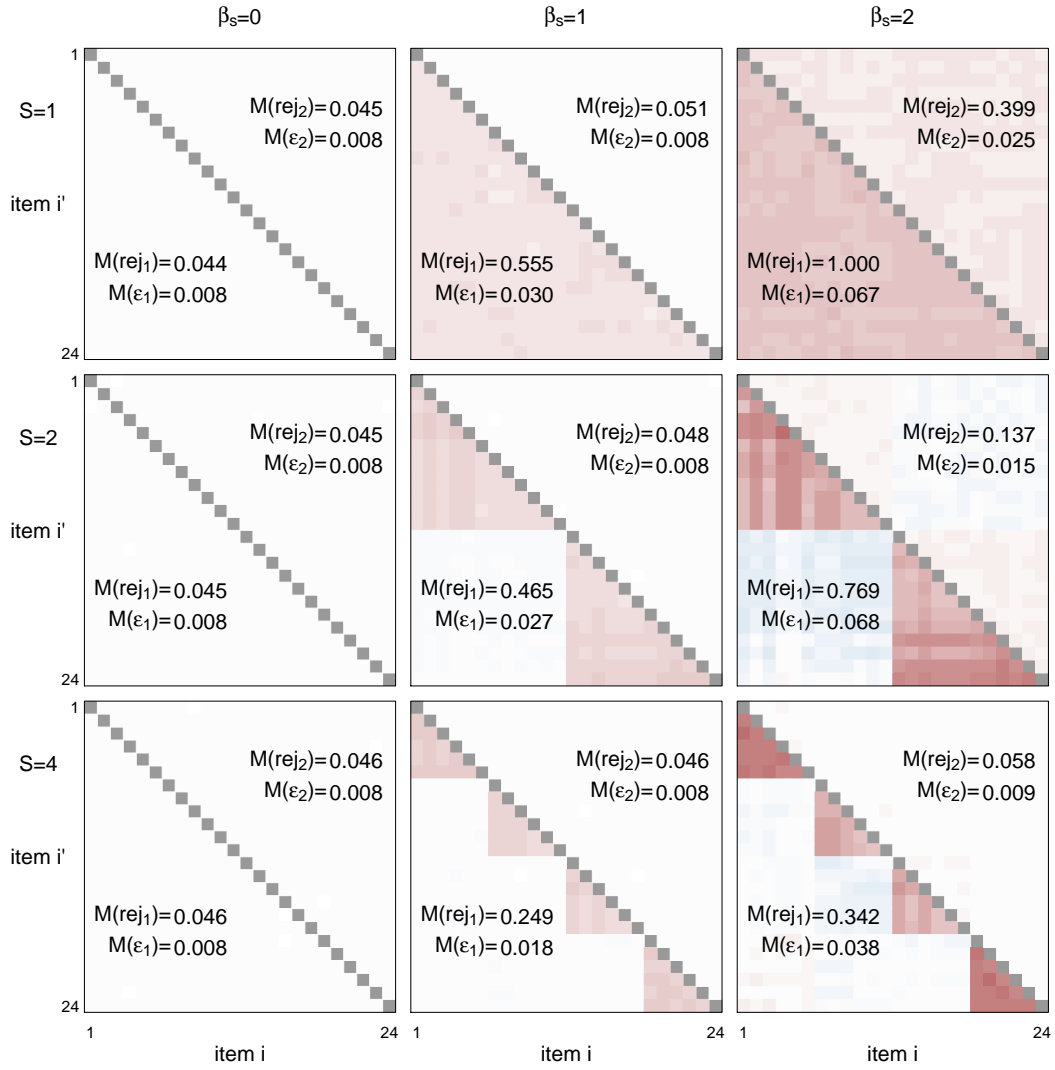


Figure 6.2: LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 24$ items in $K = 4$ ordered response categories and sample size of $N = 1000$.

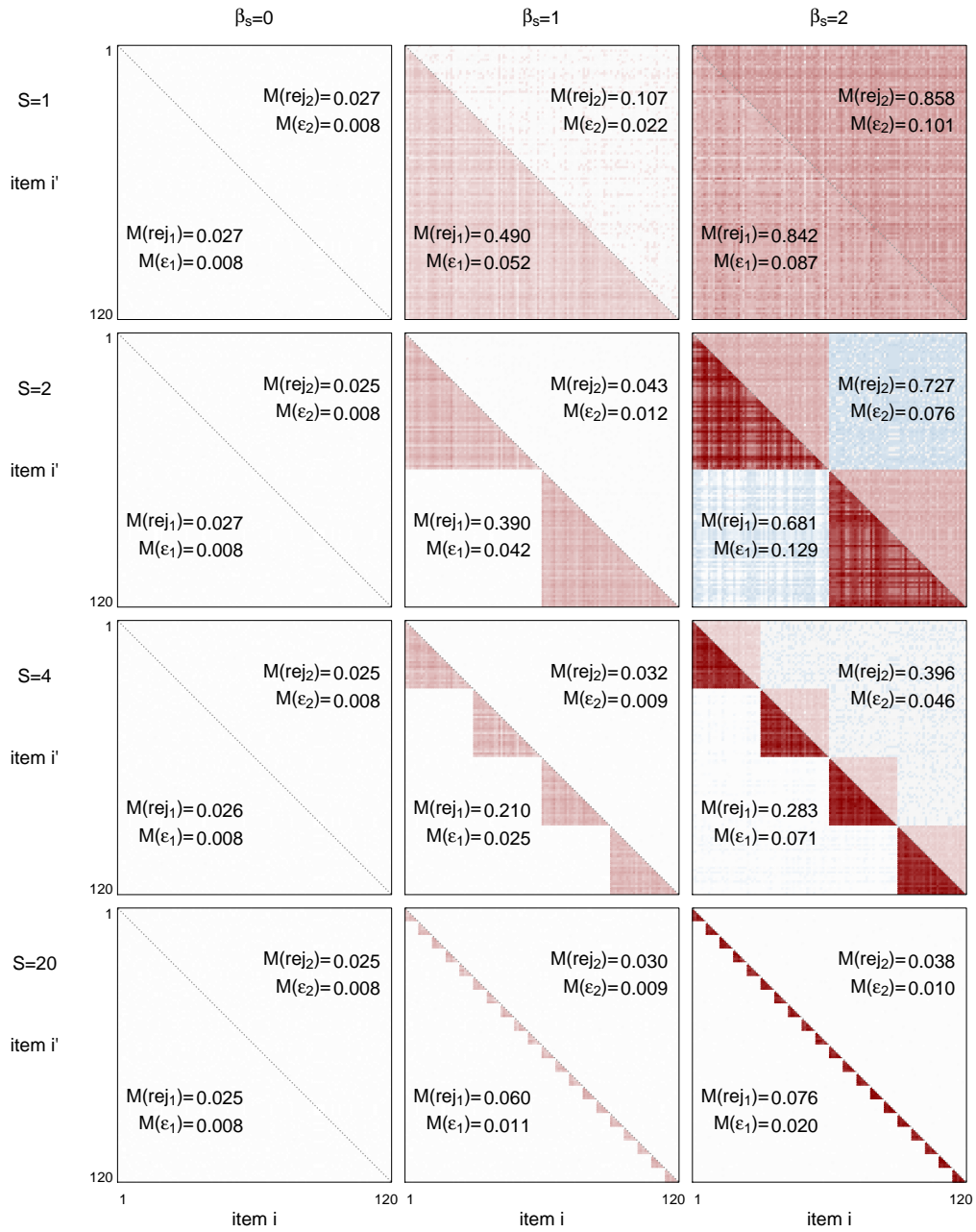


Figure 6.3: LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 120$ items in $K = 2$ response categories and sample size of $N = 1000$.

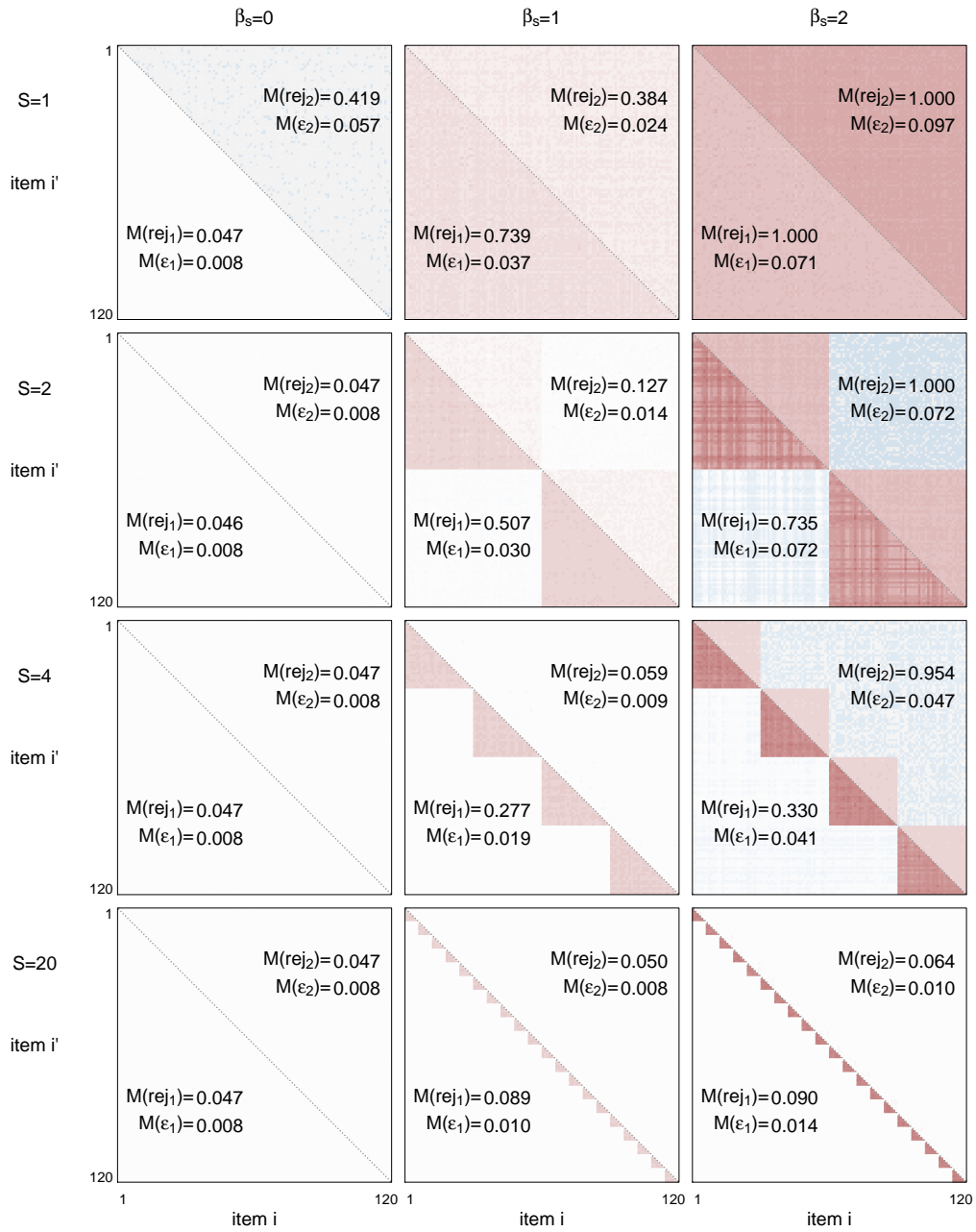


Figure 6.4: LD X^2 rejection rates and RMSEA (ϵ) for traditional (below diagonal) and hierarchical (above diagonal) DINA models with $I = 120$ items in $K = 4$ ordered response categories and sample size of $N = 1000$.

6.1.2 Sensitivity of LD X^2 to Nuisance Dimensionality

Having established that LD X^2 has good type I error rate control (and is, in fact, a bit conservative in this regard—particularly for dichotomous data), I now examine the sensitivity of the index to model misspecifications. Here, the relevant results are those presented in the lower portions of each plot (which show the results obtained fitting the traditional diagnostic model) for models with a hierarchical data generating model (i.e., those with nonzero group-specific slope parameters $\beta_s = 1$ and $\beta_s = 2$, which are presented in the second and third columns of each figure).

Across the conditions shown (and for the models not shown), the LD X^2 statistics provide a very clear and accurate characterization of the unmodeled dimensions. Positive local dependence is consistently identified among those items sharing the common influence of a group-specific dimension. Since those dimensions influence clusters of adjacent items (see, e.g., the path diagrams shown in Figure 3.2), this positive dependence appears as blocks of red on the diagonal. In general, the item pairs off the diagonal (by which I mean those item pairs not loading on the same group-specific dimension) are unaffected, since the fitted model includes the attribute variables that explain the common variability for these items. Of course, for the conditions with only one group-specific dimension (the first row of each figure), all items belong to that locally dependent block. Thus, there is no “off”-diagonal, and all pairs exhibit positive local dependence.

There are, however, some conditions for which off-diagonal elements exhibit local dependence. When there are $S = 2$ group-specific dimensions (i.e., the second row of each figure), the LD X^2 index identifies some negative dependence. This is particularly evident for the larger group-specific slope conditions ($\beta_s = 2$, in the rightmost column). The finding of negative dependence means that the model

is under-explaining the association between the item pairs, despite the fact that—as noted above—all true sources of common variance for these items were included in the fitted models. This result is an indication that the failure to account for the group-specific dimensions has altered the relationship between the attributes and the items and is consistent with the impacts on examinee classifications that were observed in Chapter 5. In the context of continuous underlying dimensions, this phenomenon of unmodeled dimensions causing construct distortions has been termed “ θ theft” (Thissen & Steinberg, 2010).

6.1.3 LD X^2 for Correctly Specified Hierarchical Diagnostic Models

The previous analyses suggest that LD X^2 may be used to detect local item dependencies resulting from the failure to account for group-specific dimensions unrelated to the attributes of interest. Thus the index might be a useful tool for identifying contexts in which a hierarchical model is appropriate. If such a model were fit, it would be useful to determine whether it has the intended result of more fully accounting for the associations among the test items. For the data generating conditions presented in Figures 6.1–6.4, the RMSEA values based on the LD X^2 for the hierarchical models are presented above the diagonal of each plot. This allows for a side-by-side “before” and “after” comparison of results obtained by fitting two alternative models to the same simulated datasets (with results for the hierarchical model representing the assessment of local dependence “after” accounting for the influence of the group-specific dimensions).

The leftmost columns (for which $\beta_s = 1$) illustrate the case in which the hierarchical model is fit to data for which the model is unnecessary. For most conditions, there is little cost for this over-fitting. Results when the number of group-specific dimensions in the fitted model is 2 or more are nearly identical to those obtained with the more parsimonious traditional model. However, for the longer test length ($I = 120$) with items having $K = 4$ ordered response

categories, fitting the hierarchical model results in an over-explanation of item covariation (i.e., negative dependence). This makes sense, since any estimates of the group-specific slope parameters other than zero would overstate the influence of the group-specific dimensions for this condition (the average estimate for this condition was $\beta_1=0.34$). In fact, it seems that the condition $S = 1$ is generally problematic for the tests of 120 items. There is also some evidence of model misfit for the condition with $\beta_s = 2$. Although the severity of the local dependence was generally reduced under the hierarchical model, some dependence remains.

To be clear, these results should not be interpreted as flaws in the test statistic. Rather, it is useful to recall that the conditions in which local dependence is now observed (despite fitting the correct model) are the same as those conditions in which there was evidence of bias in parameter estimates (Chapter 4). The failure to recover item parameters for these conditions limits the ability of the model to fully explain the covariation among the test items, resulting in the patterns of local dependence now observed.

Despite the problems found for the $S = 1$ conditions (particularly for the tests of $I = 120$ items), the more frequent result was that the hierarchical model provided improvements in model fit (over the traditional model) that were reflected in reductions in local item dependence, as measured by LD X^2 . Type I error rates were quite low (below the nominal alpha level for dichotomous items and approaching the correct level for items with $K = 4$ categories, as seen previously for the traditional model), and the average RMSEA was generally quite small (less than 0.01).

6.2 Diagnosis for Other Types of Model Misspecification

The Chen and Thissen (1997) LD X^2 statistic performs well in detecting violations of local independence due to the presence of unmodeled group-specific dimensions.

In this section, the performance of the index is examined within the context of a broader range of misspecification types, as described in Table 3.5. These include the sorts of modeling errors that have received greater attention in discussions of diagnostic model fit: various types of Q-matrix misspecifications (see, e.g., de la Torre, 2008; Rupp & Templin, 2008a; Kunina-Habenicht et al., 2012) and application of the incorrect diagnostic model or mapping rule (von Davier, 2013).

These sorts of misspecifications are not a central focus of this research but are examined here for three reasons. First, it is expected that the sensitivity of LD X^2 may vary according to the particular type of misspecification error. Thus, this analysis provides an initial assessment of the types of errors one might expect to manifest as local item independence violations. Second, it can be safely assumed that models fit to real data are wrong in multiple ways (see, e.g., Tucker, Koopman, & Linn, 1969; MacCallum & Tucker, 1991). Conditions in which, for example, the fitted diagnostic model has an incorrectly specified Q-matrix and *also* fails to account for nuisance dimensions (i.e. doubly misspecified) may more closely resemble real modeling contexts and, thus, provide more realistic assessments of the performance of the fit index. Finally, it is perhaps the case that some users of diagnostic models would be more concerned about the misspecifications discussed in this section than those arising due to the sorts of unmodeled nuisance dimensions that I have discussed up to this point (and which the hierarchical diagnostic model is intended to accommodate). LD X^2 might be a good tool for identifying these misspecifications and making improvements to the model. However, because this index is also sensitive to the presence of nuisance dimensions, it is possible that these dimensions could hinder the detection of the misspecifications that are deemed to be of greater importance. In that case, the hierarchical model might be needed in order to obtain an unobscured assessment of the fit of the attribute model (i.e., controlling for the influence of the nuisance dimensions on item responses).

The first two cases examined in this section are examples of incorrect specification of individual elements of the Q-matrix (ones replacing zeros and vice versa). The third and fourth cases deal with errors in the number of latent attributes (adding columns and deleting columns from the Q-matrix, respectively). In the fifth and sixth cases, an incorrect model type (C-RUM and DINO, respectively, instead of DINA) is specified to one of the test items. For each case, RSMEA values based on LD X^2 are depicted for both the traditional (doubly misspecified) and hierarchical (singly misspecified) models.

For each example presented below, data were generated from the higher-order, hierarchical DINA model (see Figure 3.3). Then, two kinds of models were fit to the data. The first type was a traditional higher-order DINA model that does not include the group-specific dimensions in its specification. This model was further misspecified according to one of the six types of errors described. The second model type was a hierarchical, higher-order diagnostic model. These models also included one of the six misspecifications but was otherwise correctly specified (i.e., with respect to the presence of the group-specific dimensions).

LD X^2 -based RMSEA values were obtained with the traditional (below the diagonal) and hierarchical (above the diagonal) models for a single replication ($N = 20000$) under various nuisance dimension conditions (number of group factors $S = 1, 2, 4$; group-specific slope parameters $\beta_s = 0, 1, 2$).

6.2.1 Models Specifying Extraneous Paths

The first case considered is the specification of an extraneous path from an attribute to an item. This error is obtained by replacing a zero-valued Q-matrix element with an entry of one. Two items in the simulated example were misspecified in this way, items 3 and 23. Neither item depends on attribute x_1 in the data generating model. Thus, the true Q-matrix entries are $q_{3,1} = q_{23,1} = 0$. In

the fitted model, these entries were incorrectly specified as $q_{3,1} = q_{23,1} = 1$. The resulting changes in the logit of the item response model were previously described in Table 3.5.

The plots in the first column of Figure 6.5 show the patterns of local item dependence unobscured by any further model misspecification since, for these conditions, the group-specific slope parameters were $\beta_s = 0$. Bands of color identify the strong local dependence (based on the LD X^2 RMSEA values) between the two misspecified items (which are indicated by the tickmarks on the x - and y -axes) and the other items on the test. Item pairs shown in blue are those with negative local dependence. Through comparison with the Q-matrix (Figure 3.2), it is apparent that items 3 and 23 exhibit negative dependence with the items loading on x_1 . This is to be expected, since adding paths from x_1 to y_3 and y_{23} implies a common source of variability that was not present in the data generating models. Thus, the fitted models overstate the dependence among these items, and this dependence is detected by the LD X^2 statistic. The items with greatest positive dependence are those that in the generating model did not co-depend (with items 3 and 23) on any common attributes.

Positive local dependence (shown in red) is observed between the misspecified items and those items that do not load on x_1 in the generating models. This is due to the fact that specifying an extraneous dependence on x_1 implies a weaker association of item responses than is observed. The restrictions of the DINA model contribute to this effect. Since this model requires all attributes to be present in order for the response probability to change, any variability in x_1 results in a weaker implied dependence of the misspecified variables on their underlying attributes. The items for which the negative dependence is greatest are those that, in the generating model, have the same two underlying attributes as either item 3 or item 23.

In the absence of further model misspecification (i.e., when $\beta_s = 0$), it seems

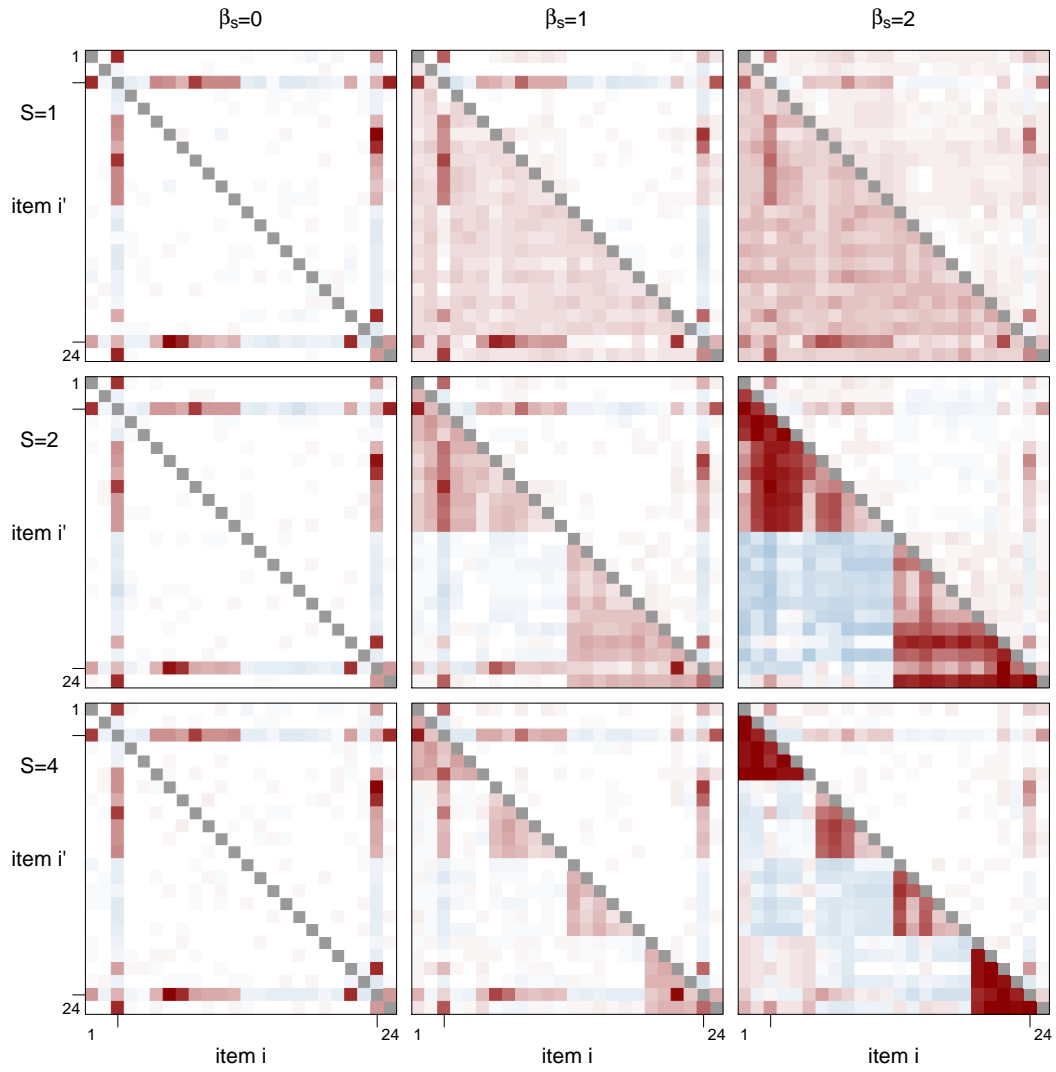


Figure 6.5: LD X^2 -based RMSEA for models specifying extraneous paths. Tick-marks identify items 3 and 23, for which data generating Q-matrix elements $q_{3,1} = q_{23,1} = 0$ were replaced in the fitted models with $q_{3,1} = q_{23,1} = 1$.

the LD X^2 index identifies a pattern of local dependence that is consistent with the expected effects of adding an extraneous path. The results for the traditional and the hierarchical models are similar for these conditions. When group-specific dimensions are added to the data generating models (as in the second and third columns ($\beta_s = 1, 2$), the number of locally dependent item pairs increases. For the traditional models (shown below the diagonal), blocks of positive dependence appear among the items loading on each group-specific dimension (compare to Figure 4.5). The presence of additional dependencies complicates the task of identifying the Q-matrix misspecification. On the other hand, fitting the hierarchical models—which are still misspecified with respect to the the extraneous paths—helps to resolve a good deal of the dependence. As a consequence, the patterns of dependence under the hierarchical models closely resemble the results presented in the first column (for which $\beta_s = 0$). The one exception is the condition with $S = 1$ group-specific dimension and a larger group-specific slope parameter ($\beta_s = 2$). This, of course, is a condition for which the group-specific slope parameter was underestimated (See Section 4.1.3). As a consequence, the model fails to account fully for the positive dependence of that group-specific dimension. For all other conditions, however, the hierarchical provides an improved characterization of the local dependence arising due to Q-matrix misspecification.

6.2.2 Models with Paths Omitted

The second type of misspecification considered was the omission of paths. This error is made by replacing positive elements of the Q-matrix with zeros. For this analysis, the paths between attribute x_1 and items 5 and 16 were omitted. Q-matrix entries $q_{5,1} = q_{16,1} = 1$ were replaced with $q_{5,1} = q_{16,1} = 0$. Changes to the logit of the item response model are shown in Table 3.5. The impact of these errors is examined in Figure 6.6.

As was true of the previous condition (in which extraneous paths were added),

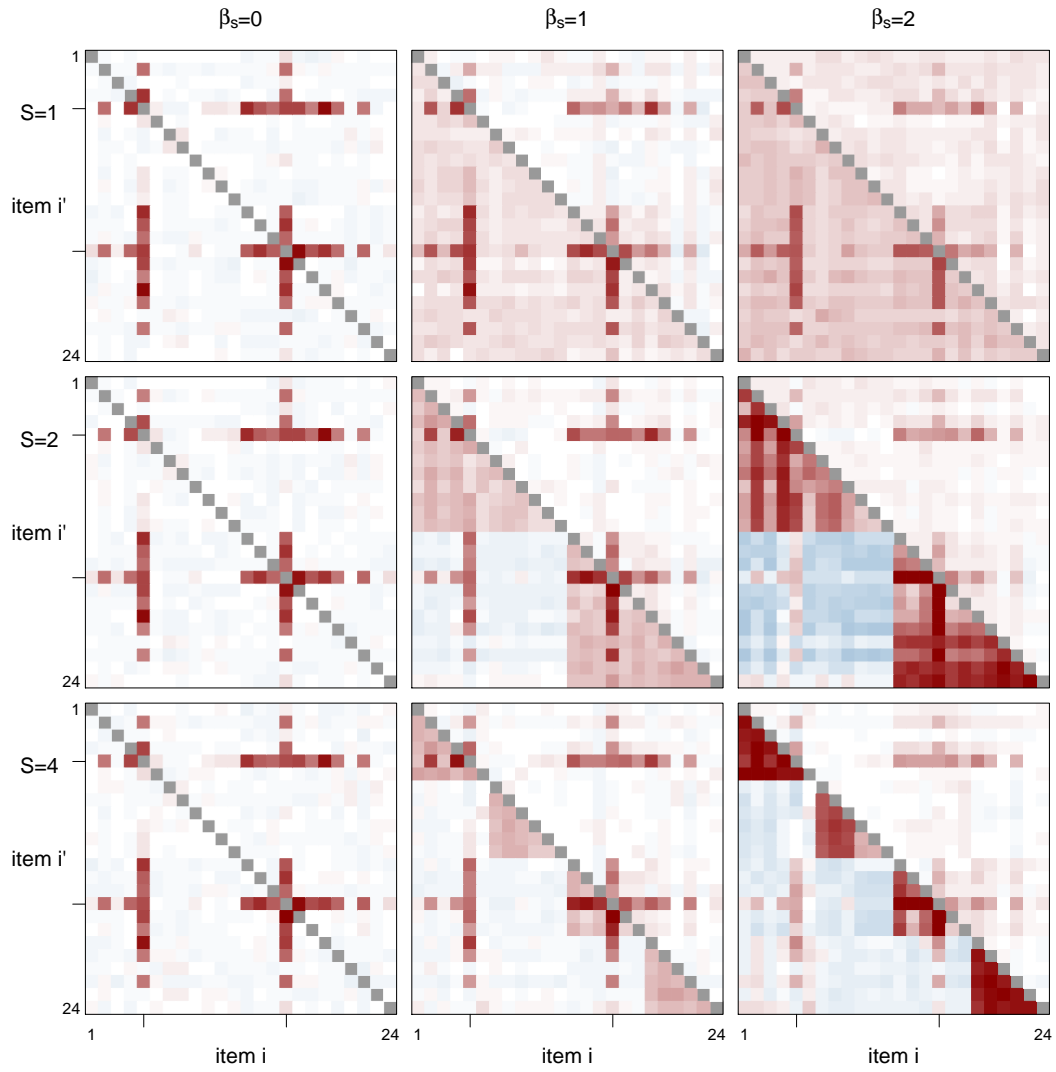


Figure 6.6: LD X^2 -based RMSEA for models with paths omitted. Tickmarks identify items 5 and 16, for which generating model Q-matrix elements $q_{5,1} = q_{16,1} = 1$ were replaced in the fitted models with $q_{5,1} = q_{16,1} = 0$.

the omitted paths manifest as bands of strong dependence between the misspecified items and the other items on the test. While there is some evidence of weak negative local dependence, the largest LD X^2 -based RMSEA values are for positive dependencies between the misspecified items (5 and 16) and those items loading on x_1 . Since these items share the common influence of x_1 in the data generating models, the fitted models ignoring the paths from x_1 to items 5 and 16 greatly understate these associations. The particular pairs with greatest positive dependence (i.e., those pairs with the darkest red shading: 5–4, 5–13, 5–19, 16–14, 16–17, 16–18) are those that share two common underlying attributes in the data generating models (x_1 and x_3 for items 5, 4, 13, and 19; x_1 and x_4 for items 16, 14, 17, and 18; see Table 3.2).

When group-specific dimensions are present in the data generating models (i.e., for $\beta_s = 1, 2$), fitting the traditional model results in the now familiar patterns of positive dependence among the items loading on the unmodeled dimensions. The dependence patterns due to the omitted paths—clearly observed in the first column ($\beta_s = 0$)—are still somewhat discernible. That said, fitting the hierarchical model (as shown in the results presented above the diagonal) greatly reduces the amount of “noise” and improves the possibility of detecting the Q-matrix error.

6.2.3 Models Specifying Extraneous Attributes

The third type of misspecification considered here is the inclusion in the fitted models of an attribute that was not present in the data generating models. This can be thought of as another sort of Q-matrix misspecification, as the error amounts to adding a column to this matrix and assigning some nonzero elements. In this analysis, four items—2, 6, 8, and 13—were assigned to load on the extraneous attribute x_5 (i.e., $q_{2,5} = q_{6,5} = q_{8,5} = q_{12,5} = 1$, and $q_{i,5} = 0$ for all other items i). The LD X^2 -based RMSEA values are shown in Figure 6.7.

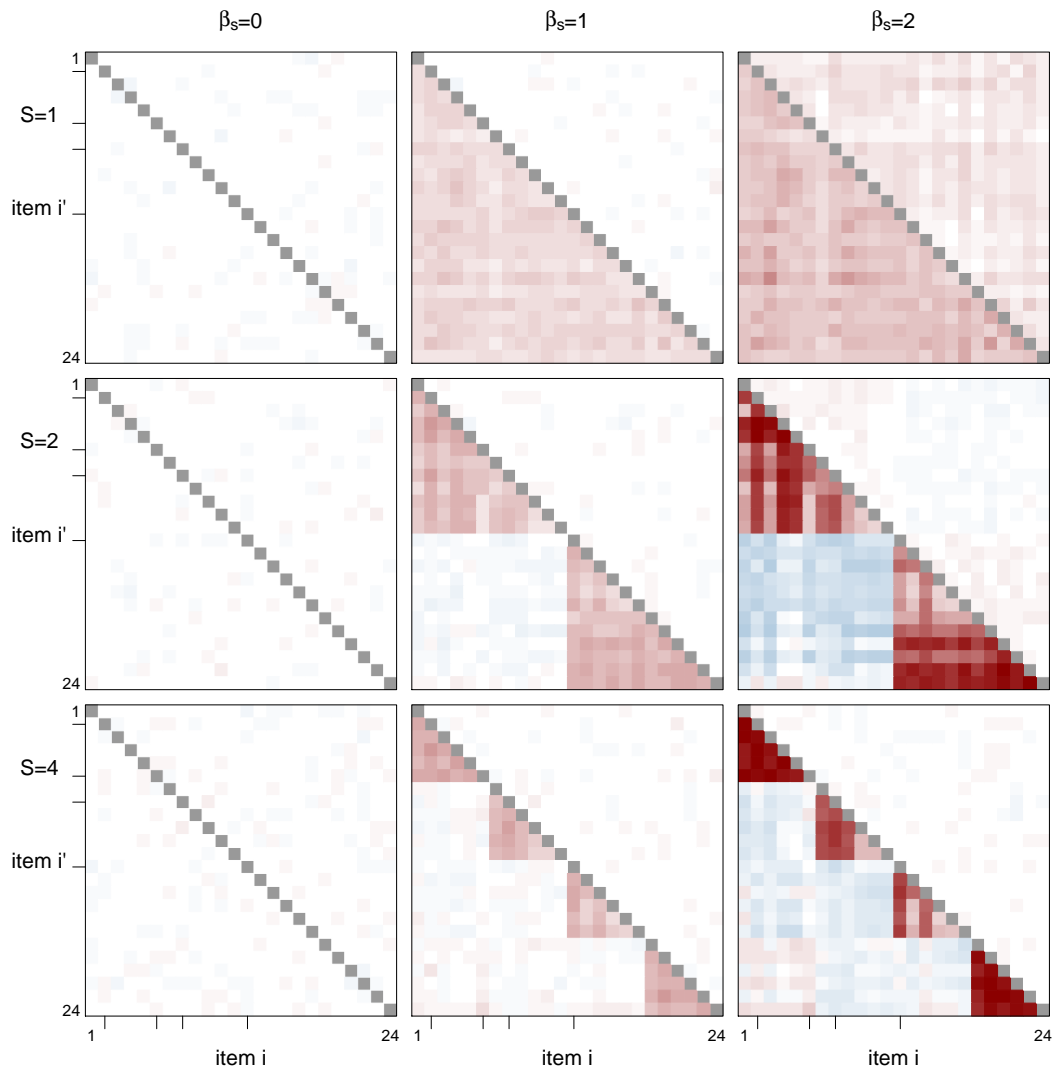


Figure 6.7: LD X^2 -based RMSEA for models specifying an extraneous attribute. Tickmarks identify items 2, 6, 8, and 13, which were incorrectly specified in the fitted models as loading on an attribute x_5 not present in the data generating model.

Here, what is noteworthy is the absence of any substantial local independence violations. No clear patterns of positive or negative dependence are apparent in the first column (for which the added attribute is the only misspecification), and the dependence observed in the remaining columns can be attributed to the influence of the group-specific dimensions (which largely resolves upon fitting the hierarchical model). Put another way, LD X^2 does not provide any suggestion of Q-matrix misspecification.

Focusing on the case with $\beta_s = 0$ (i.e., generating from the traditional, rather than hierarchical, model), it is interesting to note that the overall fit of the misspecified, five-attribute model is very similar to the correctly specified four-attribute model. The five attribute model has a slightly lower $-2 \times \log$ -likelihood, though the four-attribute model (requiring estimation of one less parameter) is favored by AIC: 509797 vs. 509718. That additional estimated parameter is the intercept for the higher-order model (i.e., in the regression of the attribute variables onto the continuous dimension θ , as described in Section 2.4). The estimate of this parameter is $\hat{c}_5=6.20$. Taken together with the estimated slope parameter ($\hat{a} = 1.01$), it is apparent that this is an attribute that is an extremely “easy” to possess for the population of (simulated) examinees. Exactly how easy is made clear in Table 6.1, shows the marginal attribute probabilities for the true (data-generating) model and for the estimated, five-attribute model.

Table 6.1: Marginal distribution of attributes estimated from a fitted model with an extraneous attribute.

	x_1	x_2	x_3	x_4	x_5
Population (Data Generating Model)					
$P(x_j = 0)$	0.651	0.551	0.449	0.349	N/A
$P(x_j = 1)$	0.349	0.449	0.551	0.651	N/A
Fitted Model with Extraneous Attribute					
$P(x_j = 0)$	0.647	0.553	0.449	0.349	0.003
$P(x_j = 1)$	0.353	0.447	0.551	0.651	0.997

Introduction of the extraneous attribute had virtually no impact on the dis-

tributions of the remaining attributes, and there are only a very small number of subjects expected to lack attribute 5. This is noteworthy because one effect of introducing x_5 into a conjunctive (DINA) model such as this is to increase the order of the interactions (as shown in Table 3.5). For example, the second term in the logit for item 2 in the true model is $\gamma_{2,2 \times 3} x_2 x_3$. In the misspecified model, the term involves a three-way interaction, $\gamma_{2,2 \times 3 \times 5} x_2 x_3 x_5$. However, if x_5 always (or nearly always) has a value of 1, then the slope parameters $\gamma_{2,2 \times 3 \times 5}$ and $\gamma_{2,2 \times 3}$ are essentially estimating the same quantity (the change in the logit due to possessing attributes 2 and 3), which is consistent with the resulting parameter estimates: $\hat{\gamma}_{2,2 \times 3} = 3.76$ for the correct model, $\hat{\gamma}_{2,2 \times 3 \times 5} = 3.77$ for the misspecified model.

To summarize, it appears LD X^2 may provide little assistance in identifying the specification of an extraneous attribute, since it appears that the models may absorb such attributes with little disruption of the relationships among the other attributes and the items. For this situation, examination of attribute probabilities (and the related parameters in a higher-order diagnostic model) may provide more insight.

6.2.4 Models with Attributes Omitted

The fourth type of misspecification considered here is the omission of a relevant attribute. Instead of adding a column to the Q-matrix (as in the previous section), the error here amounts to removing a column (or, equivalently, fixing all the elements in a column to zero). For this analysis, diagnostic models missing x_4 were fit to the data generated from models with four-attributes. The resulting LD X^2 -based RMSEA values are presented in Figure 6.8.

In the data generating models, twelve items load on x_4 (see Table 3.2): 6, 7, 8, 10, 11, 12, 14, 16, 17, 18, 21, and 23. These items are identified by tickmarks in Figure 6.8. This misspecification involves a larger number of items that have been

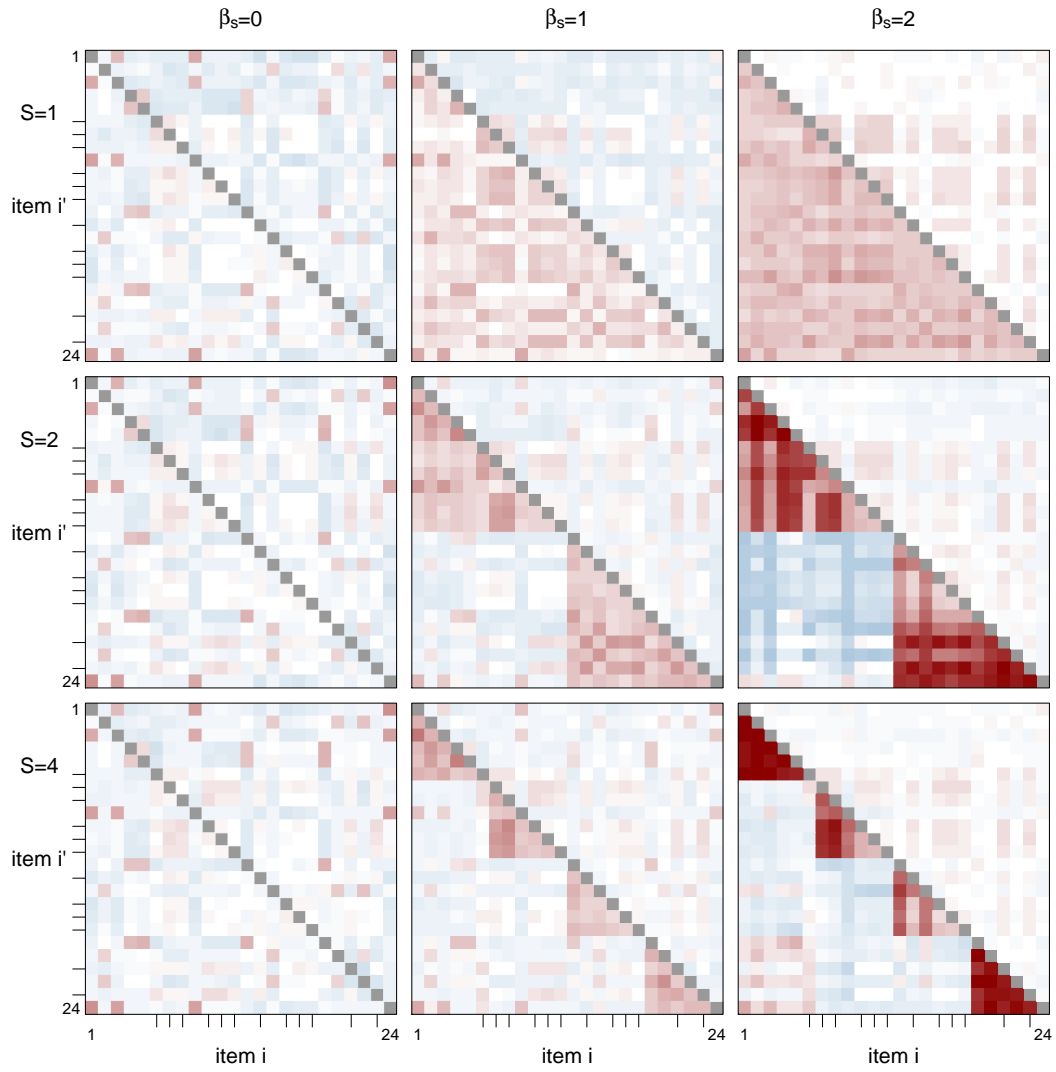


Figure 6.8: LD X^2 -based RMSEA for models with an omitted attribute. Tick-marks identify items loading on attribute x_4 in the data generating model. This attribute was omitted in the fitted models.

considered in the previous examples, and so it is perhaps to be expected that the patterns of local dependence would be somewhat more complex.

The plots in the first column—in which the data were generated from the traditional model (i.e., with $\beta_s = 0$)—reveal positive dependence among those items loading on x_4 in the generating model. However, this dependence is not particularly strong. Instead, for this condition, the item pairs with strongest positive dependence are those that depend in the true model on pairs of attributes that do not include x_4 (2–15, 2–20, 2–22, 15–20, 15–22, 20–22, which load on both x_1 and x_2 ; 4–5, 4–13, 2–19, 5–13, 5–19, 13–19, which load on x_1 and x_3 ; and 1–3, 1–9, 1–24, 3–9, 3–24, and 9–24, which load on x_2 and x_3).

Thus, it seems that the fitted model explains associations among those items depending on latent variables that are present in the model *less completely* than it explains the associations among items depending on the latent variable (x_4) that is omitted. Once again, this may be due to the particular model type in use. Because of the conjunctive nature of the DINA, values of x_4 can only maintain or decrease the probability of correct response in the true model, conditional on the remaining relevant variables. Thus, the observed proportion of correct responses on an item depending on x_4 would be less than what one would expect if the item depended only on x_1 – x_3 .

One possible explanation for lower than expected observed proportions of items correct would be that the required attributes are possessed by a lower proportion of the population. Table 6.2 presents the marginal attribute distributions from the population and based on the misspecified model (from the conditions with $\beta_s = 0$ fitting a traditional model with x_4 omitted. The prevalence of each attribute specified in the fitted model is greatly underestimated. Thus, it appears that the misspecification has resulted in an altered attribute distribution.

These results may be compared with a case in which the population distribution is unaltered. This may be accomplished by fixing the higher-order model

Table 6.2: Marginal distribution of attributes estimated from a fitted model with an omitted attribute.

	x_1	x_2	x_3	x_4
Population (Data Generating Model)				
$P(x_j = 0)$	0.651	0.551	0.449	0.349
$P(x_j = 1)$	0.349	0.449	0.551	0.651
Fitted Model with Omitted Attribute				
$P(x_j = 0)$	0.718	0.655	0.581	N/A
$P(x_j = 1)$	0.282	0.345	0.419	N/A

parameters (i.e., the attribute intercepts and their slopes on the higher-order trait) to their true values. The LD X^2 -based RMSEA values for the same data generating conditions—but with the attribute distribution fixed—are shown in Figure 6.9. Now that the model misspecification can no longer be absorbed by changes in the attribute distribute, a much clearer pattern of very strong local dependence is visible, with the items loading on the omitted attribute displaying the strongest positive dependencies.

These two examples of changes in the number of variables modeled—through either adding or omitting attributes—identify some important limitations of the LD X^2 index. In such cases, examination of the estimated attribute distributions may provide helpful insights concerning the possibility of misspecification.

6.2.5 Models with Incorrect Specification of Item Type: C-RUM

The fifth type of model misspecification I consider is the application of an incorrect item model. In the data generating models used for these misspecification analyses, all items were generated from a conjunctive (DINA) model. In the section, I examine the results of incorrectly fitting a compensatory model (C-RUM) to an item. Here, the error is made for item 8.

The changes in the logit of the item response model resulting from the altered mapping rule of the compensatory model are described in Table 3.5. Instead of

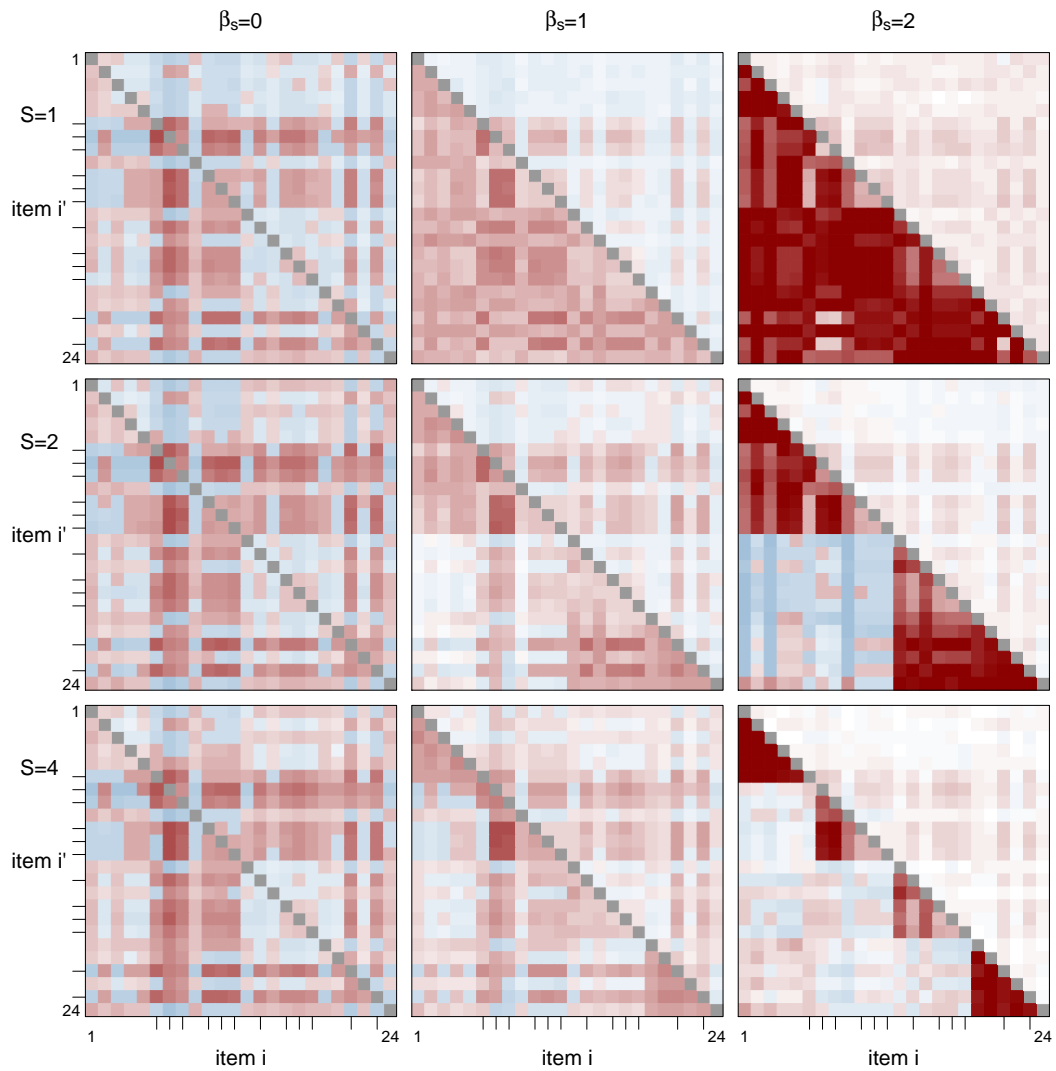


Figure 6.9: LD X^2 -based RMSEA for models with omitted attribute variable and fixed attribute distribution.

a single two-way interaction term, the misspecified model has two main effects—one for each underlying attribute. Figure 6.10 presents the LD X^2 -based RMSEA values for this type of model misspecification.

There is very little evidence of misspecification involving item 8 across the conditions examined. The result is consistent with recent analyses by von Davier (2013), who noted that compensatory models may often turn out to fit data from a conjunctive model quite well and, moreover, that a compensatory model may be specified that is entirely equivalent to the DINA model. Thus, it may be the case that empirical criteria alone—including the sort of diagnostic or goodness-of-fit test examined here—will not provide a strong indication of the superiority of one model or another.

6.2.6 Models with Incorrect Specification of Item Type: DINO

The final type of misspecification considered is another application of the wrong item response model. Here, a DINO model is specified for an item that was generated from a DINA model. Once again, item 8 is used for the illustration. As with the previous examples, the resulting changes in the logit of the item response model are described in Table 3.5. This error amounts to a relaxation of the item’s cognitive requirements. Whereas the DINA model requires that *all* attributes be present in order to have a high probability of correct response, the DINO model simply requires that the examinee possess one or more of the attributes. The results are shown in Figure 6.11.

The first column (for which $\beta_s = 0$) provides a clear look at the local dependence arising due to this misspecification. Item 8 exhibits positive dependence with all the items in the test, particularly with items 7, 21, and 23. These are items that load on the same attribute variables (x_2 and x_3) as item 8 in the data generating model. The DINO specification implies a weaker association than what

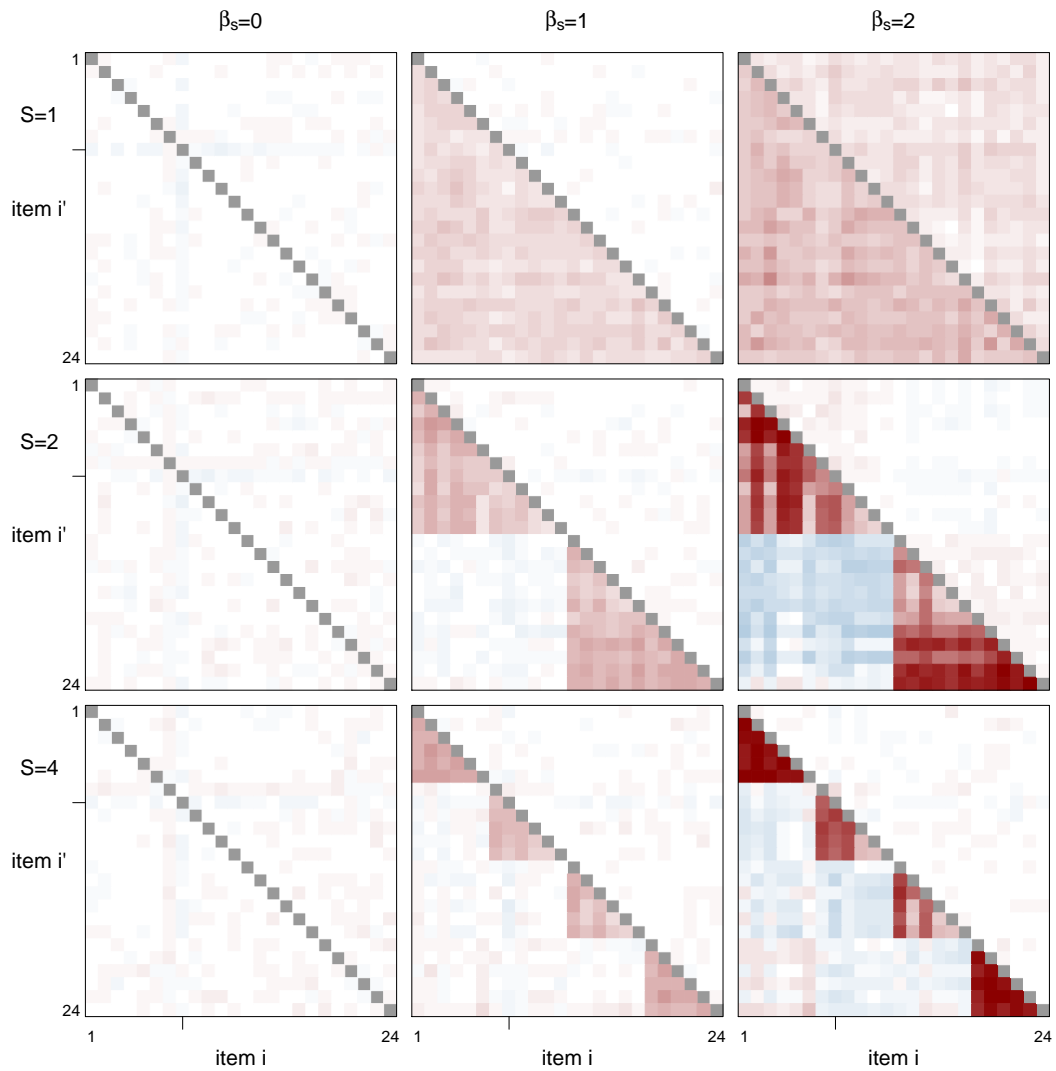


Figure 6.10: LD X^2 -based RMSEA for models with incorrect specification of item type: C-RUM

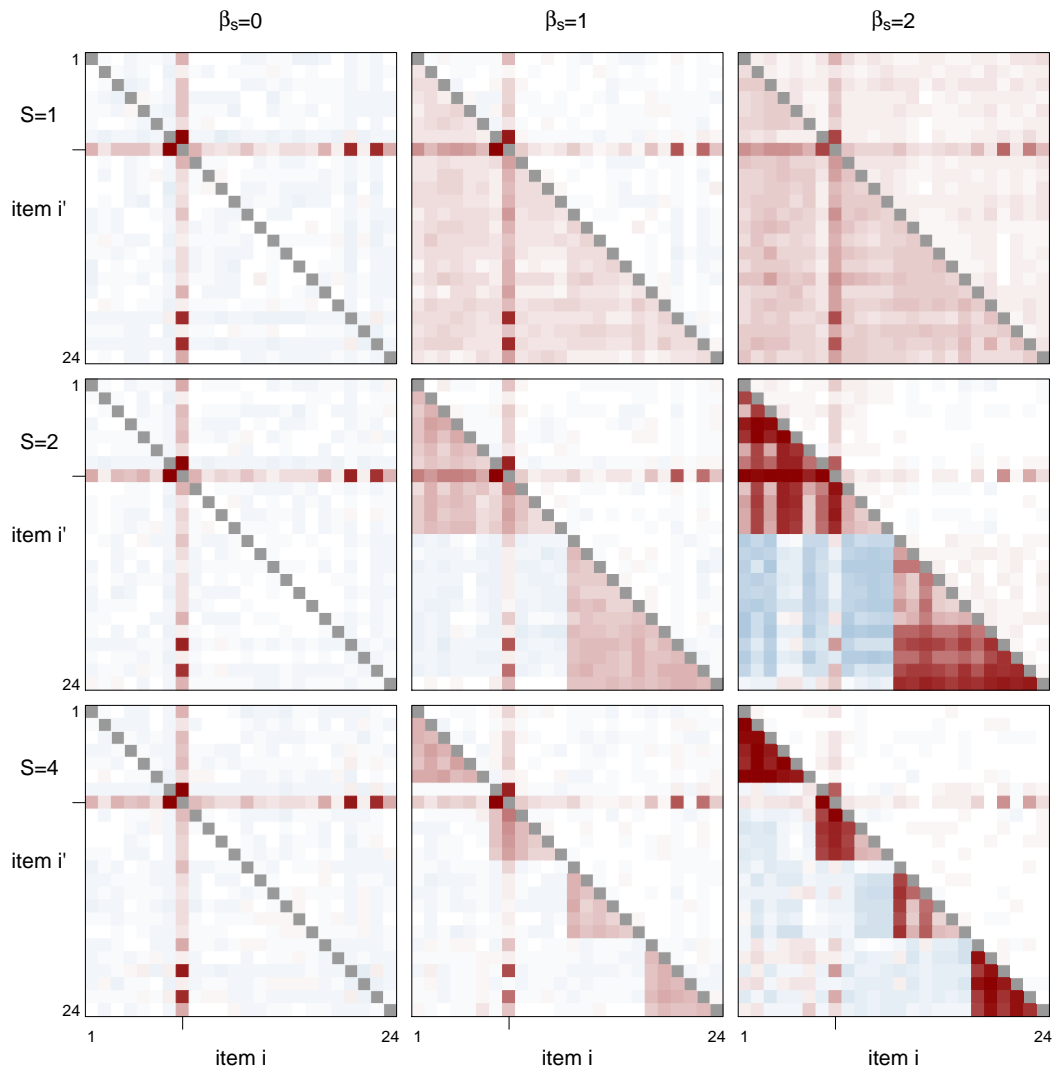


Figure 6.11: LD X^2 -based RMSEA for models with incorrect specification of item type: DINO

is actually observed between item 8 and these items.

The pattern of positive dependence remains evident even in the presence of additional nuisance dimensions. That said, for the conditions with large group-specific slope parameters ($\beta_s = 2$), the blocks of positive dependence do obscure the results to a small degree. Fitting the hierarchical diagnostic model appears to consistently improve the clarity of the characterization.

6.3 Discussion

In this chapter, I have examined the potential utility of the Chen and Thissen (1997) LD X^2 index as a tool for detecting and characterizing diagnostic model misspecification. The calibration of the test statistic was evaluated through a Monte Carlo study. The index demonstrated good type I error control. It was a bit conservative—particularly for dichotomous items. However, its calibration improved with a larger number of response categories ($K = 4$).

LD X^2 was found to be quite sensitive to a broad range of misspecifications, including failure to model nuisance dimensions (as is standard practice with the traditional models), incorrect specification of the Q-matrix, and application of an incorrect item model or mapping rule. At the same time, there were some conditions in which the test statistic provided little evidence of misspecification. The local dependence indices are obtained by comparing the observed and expected bivariate response probabilities. When the distributions of the underlying latent variables are altered—which, of course, is one of the reasons to be concerned about misspecification—results from LD X^2 are less definitive.

Finally it was found that for models doubly misspecified (e.g., with respect to both the Q-matrix and nuisance dimensions), fitting a singly misspecified hierarchical model can be a useful step in diagnosing some forms of model misspecification.

CHAPTER 7

Applications

Previous chapters demonstrated that the proposed hierarchical diagnostic model can be estimated well (Chapter 4), that this model provides an approach for reducing the potentially serious impacts of nuisance dimensions (Chapter 5), and that an existing goodness-of-fit index may be used to identify various types of model misspecification, including errors that might be addressed by utilizing the hierarchical model (Chapter 6). In this chapter, I present a series of empirical applications, demonstrating the use of the hierarchical model in a variety of educational and psychological testing contexts. My goal in presenting these examples is to illustrate the flexibility of the modeling framework and to provide further illustration of the use of the Chen and Thissen (1997) LD X^2 index for examining model fit.

For each example, two alternative fitted models are presented. The first is a traditional diagnostic model, with item responses influenced by attributes only (and attribute distributions either estimated freely or structured by a higher-order dimension). The second fitted model is a hierarchical diagnostic model, with one or more latent dimensions (random effects) included to account for some form of nuisance dimensionality. In one example (a test of fourth grade mathematics proficiency; see Section 7.2), the alternative model also includes changes to the Q-matrix, based on analyses of the fit indices obtained under the initial model and a review of the item content.

Chen and Thissen (1997) LD X^2 -based RMSEA values are presented as in

the previous chapters, with red and blue shading used to identify the intensity of positive and negative local independence violations. Likelihood-based fit indices, including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also provided. Side-by-side comparisons of the traditional and hierarchical models are provided for each example.

Finally, EAP scores obtained from the two models are compared through scatterplots. These identify cases of re-classification (given a fixed cut-off or threshold EAP score), as well as differences in the distributions of scores for each attribute measured. Figure 7.1 describes the various pieces of information that are presented in these scatterplots, including the proportions of examinees re-classified under the alternative model. The EAP scatterplots presented in the subsequent sections follow this same layout.

7.1 A Model for a Testlet-based Reading Assessment

The first empirical application is an analysis of 31 items from booklet 8 of the 2000 Programme for International Student Assessment (PISA; Adams & Wu, 2002) reading test. A sample of 3000 students from the United States was used. Although all the items in this assessment are intended to measure general reading literacy, test developers also targeted three distinct sub-domains or reading processes: “interpreting text”, “reflection and evaluation”, and “retrieving information” (Adams & Wu, 2002, p. 199). Each item in the assessment is intended to require one of these processes. For this analysis, I treated these reading processes as the discrete attributes on which examinees are to be classified.

The PISA reading assessment is testlet-based, and the 31 items considered in this analysis were nested within nine clusters of two to five items each. To account for nuisance dependencies arising from the testlet structure, a hierarchical diagnostic model was fit to the data. This model specified a total of 12 latent

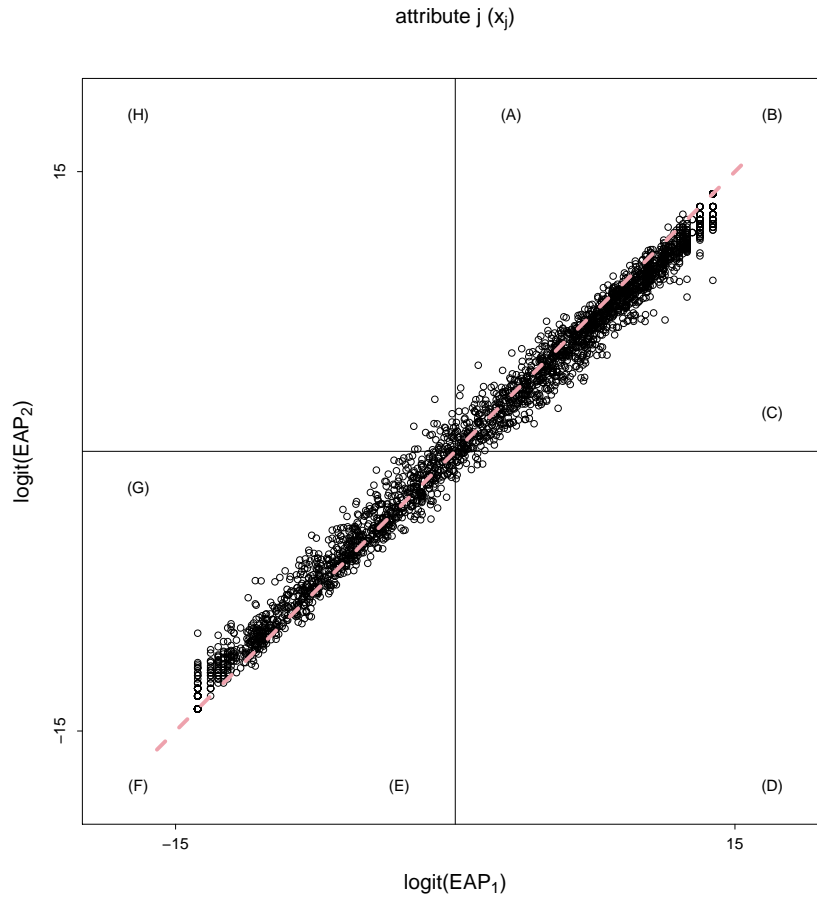


Figure 7.1: Illustration of EAP comparison used in analyses of empirical data. Model 1 (x -axis) is the traditional diagnostic model. Model 2 (y -axis) is the hierarchical model. Threshold $\text{logit}(\text{EAP})$ levels of 0 (corresponding to $\text{EAP}=0.5$) are shown as vertical and horizontal lines bisecting the scatterplot. Values at positions (A)–(H) represent proportions of examinees in (mutually exclusive) groups defined by the EAP values obtained from the alternative models: (A)–(C) are the proportions of examinees classified as possessing attribute x_j (i.e., “positive”) under both models for whom (A) EAP_1 is closer to 0.5 than EAP_2 , (B) $\text{EAP}_1=\text{EAP}_2$, or (C) EAP_2 is closer to 0.5 than EAP_1 . The values at positions (E)–(G) are the proportions classified as lacking attribute x_j (i.e., “negative”) under both models for whom (E) EAP_1 is closer to 0.5 than EAP_2 , (F) $\text{EAP}_1=\text{EAP}_2$, or (G) EAP_2 is closer to 0.5 than EAP_1 . The value at position (D) is the proportion classified positive under model 1 and negative under model 2, while (H) is the proportion classified negative under model 1 and positive under model 2.

variables (see Table 7.1): three attribute variables (the reading processes) and nine testlet dimensions.

Table 7.1: Description of latent variables specified in the traditional and hierarchical diagnostic models fit to the PISA dataset.

Var.	Description	No.
Attribute variables (x_j)		
x_1	Retrieving information: “locating one or more pieces of information in a text”	14
x_2	Interpreting: “constructing meaning and drawing inferences from one or more parts of a text”	6
x_3	Reflecting/evaluating: “relating a text to one’s experience, knowledge and ideas”	11
Group-specific dimensions (ξ_s)		
ξ_1	Testlet dimension 1 (items 1–5)	5
ξ_2	Testlet dimension 2 (items 6–9)	4
ξ_3	Testlet dimension 3 (items 10–14)	5
ξ_4	Testlet dimension 4 (items 15–18)	4
ξ_5	Testlet dimension 5 (items 19–23)	5
ξ_6	Testlet dimension 6 (items 24–25)	2
ξ_7	Testlet dimension 7 (items 26–27)	2
ξ_8	Testlet dimension 8 (items 28–29)	2
ξ_9	Testlet dimension 9 (items 30–31)	2

Note: “Var.” indicates the latent variable. “No.” indicates the number of items loading on the latent variable. Group-specific dimensions (ξ ’s) are omitted in the traditional diagnostic model.

The hypothesized relationships between the latent variables and the test items are summarized in Table 7.2. The Q-matrix is given in the first three columns and is based on the PISA manual (Adams & Wu, 2002), which identifies the targeted reading process (the x variable) for each item. Because one and only one process is identified with each item, the Q-matrix has an independent cluster structure that is comparable to the “Simple” models examined in the earlier simulation studies (see, e.g., Figure 3.4). The final nine columns in Table 7.2 show the β parameters, which relate the test items to the nine testlet dimensions (the ξ variables). For this model, group-specific slope parameters were estimated without any constraints for the five testlets of three or more items. For two-item testlets

(clusters 6–9), the group-specific slope parameters were constrained to be equal (for model identification).

Two alternative diagnostic models were fit to the PISA data. The first fitted model is a traditional diagnostic model, as given by Equation 2.1. This model ignores the clustering of items within testlets. The second model, given by Equation 2.14 and shown in Figure 7.2, specifies nine testlet dimensions. The traditional model is nested within this hierarchical model and is obtained by fixing the β_s parameters in the hierarchical model to zero.

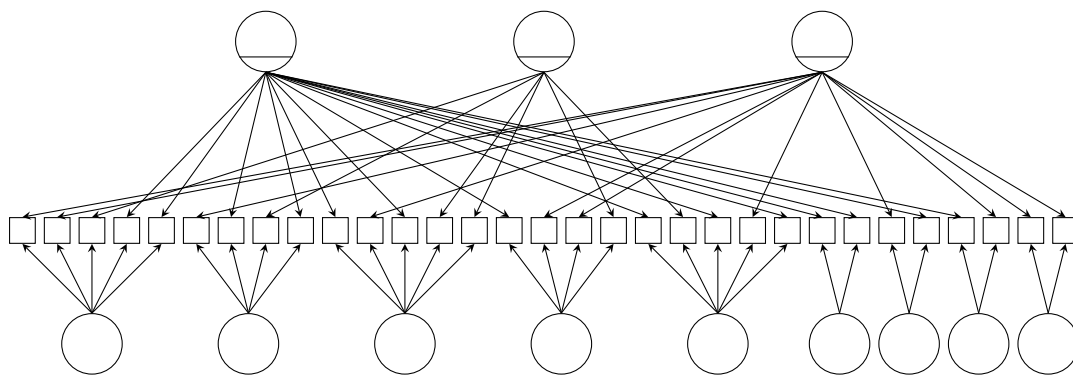


Figure 7.2: Path diagram for a hierarchical diagnostic model fit to the PISA dataset.

Maximum marginal likelihood estimation was used in fitting both models. In the case of the rather high-dimensional hierarchical model, the estimation was facilitated by the the dimension reduction technique discovered by Gibbons and Hedeker (1992) and described earlier (see Equations 2.18 and 2.19).

For both models, local item independence was evaluated using the Chen and Thissen (1997) LD X^2 statistic, from which an RMSEA was computed (see Equation 3.4). The RMSEA values for each item pair under the traditional and hierarchical diagnostic models are shown in Figure 7.3. It is important to note that many of the most severe residual dependencies occur among items belonging to the same testlet. This result is consistent with the notion that the testlet-based design creates inter-item associations that are not fully explained by the read-

Table 7.2: Q-matrix and group-specific slope parameters for the traditional and hierarchical diagnostic models fit to the PISA dataset.

Item	y	Q			B								
		x_1	x_2	x_3	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	ξ_9
R040Q02	y_1	0	0	1	$\beta_{1,1}$	0	0	0	0	0	0	0	0
R040Q03a	y_2	0	0	1	$\beta_{2,1}$	0	0	0	0	0	0	0	0
R040Q03b	y_3	0	1	0	$\beta_{3,1}$	0	0	0	0	0	0	0	0
R040Q04	y_4	1	0	0	$\beta_{4,1}$	0	0	0	0	0	0	0	0
R040Q06	y_5	1	0	0	$\beta_{5,1}$	0	0	0	0	0	0	0	0
R077Q02	y_6	0	0	1	0	$\beta_{6,2}$	0	0	0	0	0	0	0
R077Q04	y_7	1	0	0	0	$\beta_{7,2}$	0	0	0	0	0	0	0
R077Q05	y_8	0	1	0	0	$\beta_{8,2}$	0	0	0	0	0	0	0
R077Q06	y_9	1	0	0	0	$\beta_{9,2}$	0	0	0	0	0	0	0
R088Q01	y_{10}	1	0	0	0	0	$\beta_{10,3}$	0	0	0	0	0	0
R088Q03	y_{11}	0	0	1	0	0	$\beta_{11,3}$	0	0	0	0	0	0
R088Q04	y_{12}	1	0	0	0	0	$\beta_{12,3}$	0	0	0	0	0	0
R088Q05	y_{13}	0	1	0	0	0	$\beta_{13,3}$	0	0	0	0	0	0
R088Q07	y_{14}	0	1	0	0	0	$\beta_{14,3}$	0	0	0	0	0	0
R110Q01	y_{15}	1	0	0	0	0	0	$\beta_{15,4}$	0	0	0	0	0
R110Q04	y_{16}	0	0	1	0	0	0	$\beta_{16,4}$	0	0	0	0	0
R110Q05	y_{17}	0	0	1	0	0	0	$\beta_{17,4}$	0	0	0	0	0
R110Q06	y_{18}	0	1	0	0	0	0	$\beta_{18,4}$	0	0	0	0	0
R216Q01	y_{19}	1	0	0	0	0	0	0	$\beta_{19,5}$	0	0	0	0
R216Q02	y_{20}	0	1	0	0	0	0	0	$\beta_{20,5}$	0	0	0	0
R216Q03	y_{21}	1	0	0	0	0	0	0	$\beta_{21,5}$	0	0	0	0
R216Q04	y_{22}	0	0	1	0	0	0	0	$\beta_{22,5}$	0	0	0	0
R216Q06	y_{23}	1	0	0	0	0	0	0	$\beta_{23,5}$	0	0	0	0
R236Q01	y_{24}	1	0	0	0	0	0	0	0	β_6	0	0	0
R236Q02	y_{25}	1	0	0	0	0	0	0	0	β_6	0	0	0
R237Q01	y_{26}	0	0	1	0	0	0	0	0	0	β_7	0	0
R237Q03	y_{27}	1	0	0	0	0	0	0	0	0	β_7	0	0
R239Q01	y_{28}	1	0	0	0	0	0	0	0	0	0	β_8	0
R239Q02	y_{29}	0	0	1	0	0	0	0	0	0	0	β_8	0
R246Q01	y_{30}	0	0	1	0	0	0	0	0	0	0	0	β_9
R246Q02	y_{31}	0	0	1	0	0	0	0	0	0	0	0	β_9

Note: “Item” indicates the PISA variable label for each item. The Q-matrix is given by **Q**. The matrix **B** gives the pattern of group-specific (i.e., testlet) slope parameters. Item subscripts for group-specific dimensions 6–9 were dropped to indicate that these parameters were constrained equal across items within the doublet (e.g., $\beta_{24,6}=\beta_{25,6}=\beta_6$).

ing processes alone. When these testlet effects are explicitly modeled (as in the hierarchical model), model fit with respect to the bivariate margins appears to improve substantially.

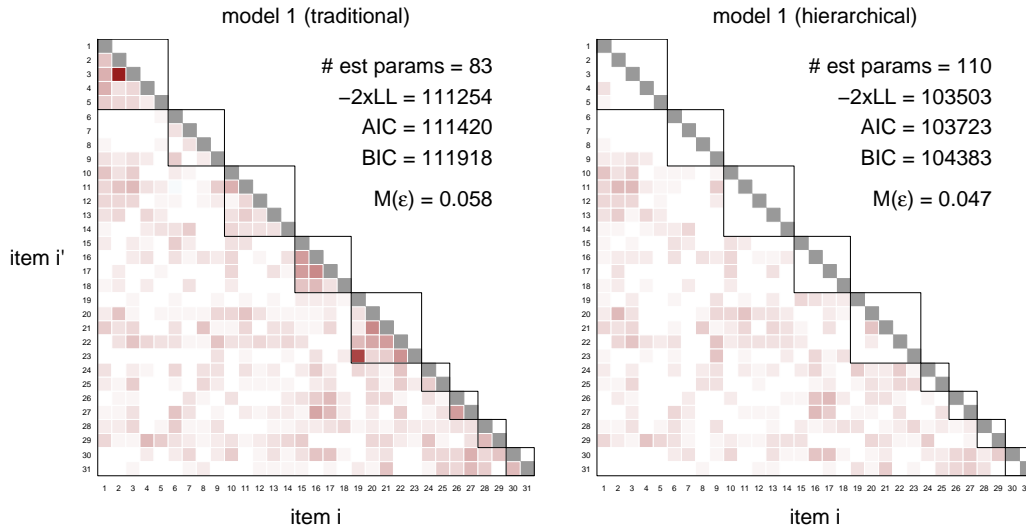


Figure 7.3: LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the PISA dataset.

The models were also compared on the basis of likelihood-based fit indices. The $-2 \times \log$ -likelihood for the hierarchical model was 103503, versus 111254 for the traditional model. Of course, some improvement in fit is guaranteed, since the first model is nested within the second. However, the large difference in likelihood that is observed with the estimation of twenty-seven additional parameters indicates that the model without testlet effects fits significantly worse (likelihood ratio test $X^2 = 7751$; with 27 degrees of freedom, $p < 0.0001$). AIC (103723 versus 111420) and BIC (104383 versus 111918) also favored the hierarchical (testlet) diagnostic model.

Figure 7.4 compares the logit of the expected a posteriori (EAP) probability (\hat{P}) of being proficient in the attribute, $\ln[\hat{P}/(1 - \hat{P})]$, under the traditional model (x -axis) and hierarchical model (y -axis) for each of the three reading process variables.

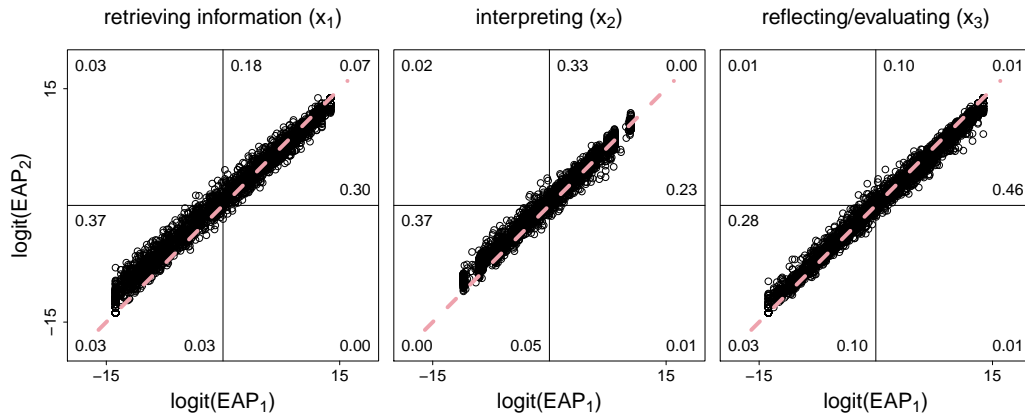


Figure 7.4: Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in the three PISA reading processes.

The EAP estimates obtained from the two models are generally quite similar. Very few examinees (about 2–4% for any single attribute) would be classified differently under the hierarchical model, given an EAP cut-off of 0.5. It does appear that EAP estimates from the hierarchical model are generally closer to 0.5 (and, thus, logits of the EAP estimates are closer to 0) than the estimates from the traditional model. This means that the hierarchical model provides a more conservative assessment of the probability of attribute classifications. Note, for example, that many more examinees receive EAP estimates of zero or one under the traditional (rather than hierarchical) model. Although classification certainty is seemingly desirable, it was found in Chapter 5 that there are cases in which high levels of certainty (as implied by an EAP near zero or one) are not justified. It is arguably much better to have an accurate characterization of this certainty, even if being more accurate necessarily means being less certain.

Of course, in this analysis we don’t know students’ “true” attribute classifications. Nor do we have the benefit of alternative measures of proficiency compare against. As a result, it is not possible to determine which of the two models provided more accurate classification (or, for that matter, whether *either* model is reasonable). Given the overall similarity in classification decisions, it is likely that

the two models would provide nearly identical overall rates of agreement with true attribute status. There might be differences, however, in the extent to which the EAP scores capture the distributions of the attribute profiles within each item response pattern.

The analysis of the PISA reading test illustrates a diagnostic modeling context in which local independence violations may be expected to arise—an assessment in which items are administered within a series of blocks or testlets. Consistent with this test structure, the LD X^2 indices suggested that a traditional diagnostic model did not fully account for covariation among the test items, particularly among those items belonging to the same testlet. A hierarchical diagnostic model fit to the same data seemed to better account for those nuisance (method-related) dependencies and, overall, was found to be a better fitting model. That said, incorporating random effects (i.e., testlet dimensions) into the diagnostic model seemed to make only small differences in examinee classification decisions.

7.2 A Higher-order Diagnostic Model for a Mathematics Assessment

The second illustration is an analysis of data from the 2007 Trends in Mathematics and Science Study (TIMSS). Here, a traditional diagnostic model is contrasted with an alternative that includes a random effect to model the dependence between a pair of strongly related items, along with some modifications to the attribute model. This example builds on prior work by Lee et al. (2011), who analyzed data from booklets 4 and 5 from the TIMSS 2007 fourth grade mathematics test. As part of their study, several teachers and content experts reviewed the TIMSS test items and coded them according to specific testing objectives described in the TIMSS 2007 framework. For the 25 items considered in the study, 15 unique testing objectives were identified (out of the 32 total objectives in the test frame-

work).

Brief descriptions of these 15 objectives—which were treated as the diagnostic attributes in Lee et al.’s (2011) study—are provided in Table 7.3. It is apparent that these objectives are much more conceptually narrow than, for example, the three reading processes identified in the PISA framework, as described in Section 7.1). Such fine-grained attribute specification may enhance the usefulness of the diagnostic model. Classification of examinees on the basis of these attributes would provide a fairly rich and useful profile of their strengths and weaknesses, which could inform and focus instructional efforts. That said, there is a good deal of variation in the number of items measuring each attribute. Ten of the 15 objectives are measured by only two or three items. In contrast, attributes x_2 and x_3 are measured by 16 and 11 items, respectively. Thus, it is expected that the test as a whole will not provide equally accurate classifications for all attributes.

Expert judgements about the attributes required for each test item produced a 25×15 dimensional Q-matrix, reported in Lee et al. (2011) and reproduced in the first 15 columns of Table 7.4. There is a good deal of complexity in the attribute loading pattern (compared to the independent cluster structure of reading processes for the PISA test considered earlier, for example); among the 25 items, the number of underlying attributes ranged from 1 (items 2, 9, 24) to 6 (item 14).

Lee et al. (2011) specified a conjunctive (DINA) model for each item in the analysis. For the current study, I fit the higher-order version of the DINA model (de la Torre & Douglas, 2004) using the Q-matrix exactly as reported in the earlier study (and shown in Table 7.4) to a sample of 564 students from the United States. The Chen and Thissen (1997) LD X^2 -based RMSEA values for this initial model are presented on the left side of Figure 7.5. Compared to the previous example, very few item pairs display evidence of local independence violations, which suggests that the qualitative work involved in developing the Q-matrix was rather effective.

Table 7.3: Description of latent variables specified in the hierarchical model of TIMSS mathematics.

Var.	Description	No.
Higher-order factor (θ)		
θ	Fourth Grade Mathematics	N/A
Attribute variables (x_j)		
x_1	Represent, compare, and order whole numbers; demonstrate knowledge of place value	6
x_2	Recognize multiples, computing with whole numbers using the four operations; estimating computations	16
x_3	Solve problems, including those set in real life contexts (for example, measurement and money problems)	11
x_4	Solve problems involving proportions	3
x_5	Recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents	3
x_6	Solve problems involving simple fractions and decimals including their addition and subtraction	2
x_7	Find missing number or operation; model simple situations involving unknowns	2
x_8	Describe relationships in patterns; generate pairs of whole numbers given rule and identify rule given pairs	3
x_9	Measure, estimate, and understand properties of lines and angles and be able to draw them	3
x_{10}	Classify, compare, recognize geometric figures and shapes and their relationships and elementary properties	7
x_{11}	Calculate and estimate perimeters, area, and volume	2
x_{12}	Locate points in an informal coordinate to recognize and draw figures and their movement	3
x_{13}	Read data from tables, pictographs, bar graphs, and pie charts	4
x_{14}	Comparing and understanding how to use information from data	3
x_{15}	Understanding different representations and organizing data using tables, pictographs, and bar graphs	2
Group-specific dimensions (ξ_s)		
ξ_1	Testlet dimension (items 18–19)	2

Note: “Var.” indicates the latent variable. “No.” indicates the number of items loading on the latent variable.

Table 7.4: Q-matrix and group-specific slope parameters for a hierarchical model of TIMSS mathematics.

Item	Q															B	
	y_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	ξ_1
M041052	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M041056	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
M041069	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
M041076	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
M041281	0	1	1	1	0	0	0	(1)	0	0	0	0	0	0	0	0	0
M041164	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
M041146	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
M041152	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0
M041258A	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
M041258B	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
M041131	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0
M041275	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
M041186	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
M041336	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0
M031303	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M031309	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M031245	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M031242A	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	β_1
M031242B	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	β_1
M031242C	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0
M031247	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
M031219	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
M031173	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M031085	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
M031172	1	1	1	0	0	0	0	0	0	0	0	0	0	(1)	0	1	0

Note: The Q-matrix (reproduced from Lee et al., 2011) is given by Q. Q-matrix elements in parentheses ($q_{5,8}$ and $q_{25,13}$) had values of one in the initial models and zero in the alternative model. The column B gives the pattern of group-specific slope parameters for the alternative model. The slopes of items 18 and 19 on the group-specific (testlet) dimension ξ_1 were constrained to be equal, so the item subscript is omitted.

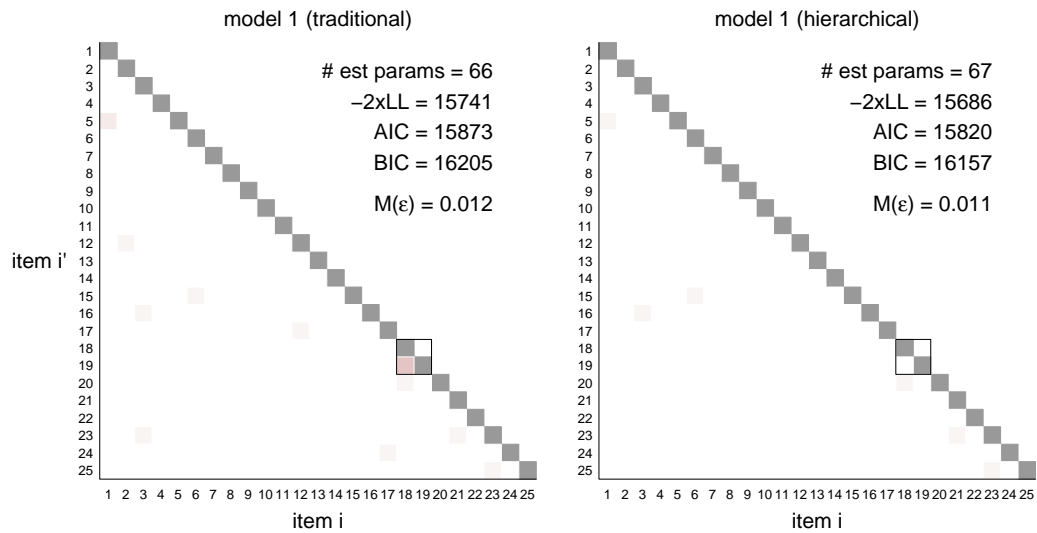


Figure 7.5: LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the TIMSS fourth grade mathematics assessment.

Because there are relatively few locally dependent item pairs, it is possible to individually examine the pairs with large RMSEA values in order to consider possible causes of dependence. Items 18 and 19 form the pair with the strongest residual dependence (LD X^2 -based RMSEA, $\epsilon=0.095$ in the initial model). These items appear are shown in Figure 7.6 (Part A is item 18; Part B is item 19).


In this case, a correct response to item 18 would seem to greatly simplify the task of answering item 19, since the answer to the latter may be read from the table completed by the examinee in responding to the former. This may explain why the attribute model doesn't fully explain the covariation in responses between these two items. Although it would be possible to arrive at the correct answer to item 19 by applying the skills identified in the Q-matrix as being relevant, those skills are less necessary once the examinee answered item 18. In order to model this lack of independence between these items, a testlet effect could be specified. Note that there are actually three parts to the "testlet". However, Part C (item 20) doesn't rely quite as directly on information from Parts A and B, so

Mathematics
Fourth Grade

Posters for two sports clubs that rent bikes are shown below.


Mountain Bike Rentals

8 zeds for 1st hour
3 zeds for each additional hour



Roadrace Bike Rentals

10 zeds for 1st hour
2 zeds for each additional hour



A. Use the information in the posters to complete the tables.

Mountain Bike Rentals	
Hours	Cost (zeds)
1	8
2	11
3	
4	
5	
6	

Roadrace Bike Rentals	
Hours	Cost (zeds)
1	10
2	12
3	
4	
5	
6	

B. For what number of hours are the rental costs the same at the two clubs?

Answer: _____

C. From which club does it cost less to rent a bike for 12 hours?

- (A) Mountain Bike Rentals
- (B) Roadrace Bike Rentals
- (C) They are both the same
- (D) It cannot be worked out

M031242

Copyright © 2008 International Association for the Evaluation of Educational Achievement (IEA). All rights reserved.

Content Domain
Number

Cognitive Domain
Applying

Maximum Points
1

Key
See scoring guide

Figure 7.6: Item cluster M031242 A/B/C in TIMSS 2007 math booklet 4.

for this illustration only items 18 and 19 load on the testlet dimension. Similarly, items 9 and 10 are administered as a testlet but show very little evidence of local dependence, so no random effect is specified for these items.

Two additional item pairs demonstrate fairly strong positive dependence: 1–5 ($\epsilon=0.057$) and 23–25 ($\epsilon=0.051$). The content of these items were examined for evidence of possible misspecification.

For item 1, it seemed that the relevant attributes (x_1 and x_2 , as shown in Table 7.4) represent alternative approaches for solving the problem, rather than necessary steps. Thus, a disjunctive (e.g., DINO) item response model was deemed a possible alternative to the DINA model used in the initial model. The study by Lee et al. (2011) identified three attributes expected to influence responses to item 5: x_2 , x_3 , and x_8 . However, in reviewing the content of this item, the relevance of x_8 was unclear. Thus, an alternative model for item 5 was considered, in which the response depends only on attributes x_2 and x_3 .

No clear misspecification was found for item 23, so the model for this item was unchanged in the alternative model. Four attributes had been identified by Lee et al. (2011) as being relevant to item 25: x_1 , x_2 , x_{13} , and x_{15} . However, upon review, it was unclear that x_{13} represents a *required* skill (in the way that the DINA rule would imply). This item presents the examinee with a pictograph, and attribute x_{13} deals with reading data from graphs of various kinds. So there is some apparent relevance of this attribute. On the other hand, the question asked in item 25 does not actually require using any of the information from the pictograph. Thus, an alternative model for item 25 was considered in which response no longer depends on x_{13} .

In sum, the alternative diagnostic model incorporates four changes, based on evaluation of the review of the goodness-of-fit indices obtained from the initial and review of the item content. First, a testlet effect (ξ_1) was added to account for the dependence between items 18 and 19. Second, the attribute model/rule used

for item 1 was changed from conjunctive (DINA) to disjunctive (DINO). Third, item 5 was specified to depend only on x_2 and x_3 (and not x_8). Fourth, item 25 was specified to depend only on x_1 , x_2 , and x_{15} (and not x_{13}). A path diagram for the alternative model is shown in Figure 7.7.

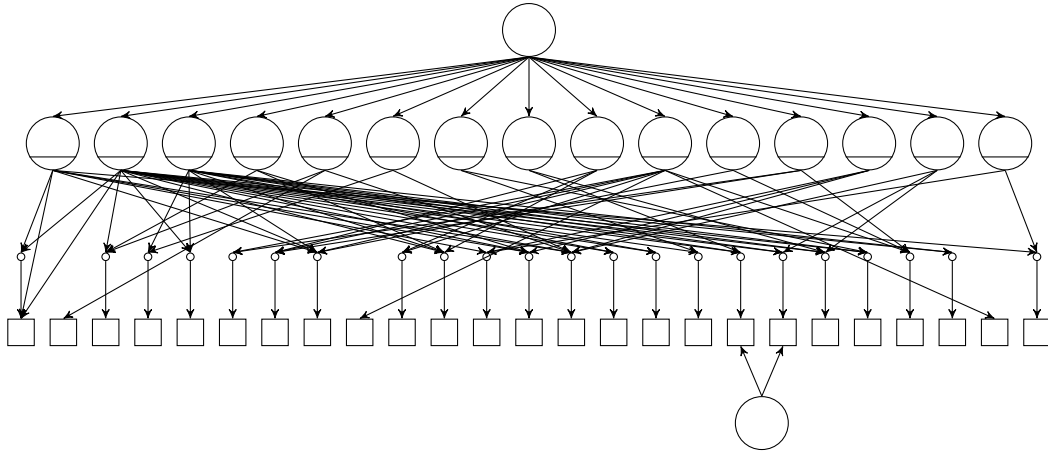


Figure 7.7: Path diagram for a hierarchical DINA model for TIMSS mathematics.

The total number of estimated parameters is 67 for the alternative model—one more than required for the initial model. Due to the particular equality constraints imposed for the DINO model, the same number of item parameters are estimated for item 1 as in the initial model (DINA). Similarly, changes to the models for items 5 and 25 do not affect the number of free parameters. The highest-order interaction slope parameter estimated in the initial model is fixed, and a lower-order interaction slope is instead estimated ($\gamma_{5,2 \times 3}$ instead of $\gamma_{5,2 \times 3 \times 8}$ for item 5; $\gamma_{25,1 \times 2 \times 15}$ instead of $\gamma_{25,1 \times 2 \times 13 \times 15}$ for item 25). The one additional parameter is the slope parameter for the testlet effect (for identification, the slopes of items 18 and 19 on ξ_1 were constrained to be equal—i.e., $\beta_{18,1} = \beta_{19,1} = \beta_1$).

The fitted hierarchical provided improved fit over the traditional model. Importantly, this model accounts for the strong positive local dependence between items 18 and 19 that was not explained by the traditional model (LD X^2 -based RMSEA, $\epsilon_2 = 0$ under the alternative model, from $\epsilon_1 = 0.095$ under the initial model; also see Figure 7.6). The dependence between items 1 and 5 was also reduced,

but only very slightly ($\epsilon_2=0.049$ versus $\epsilon_1=0.057$). Dependence between items 23 and 25 was actually greater under the alternative model, though the difference was also quite small ($\epsilon_2=0.052$ versus $\epsilon_1=0.051$). Likelihood-based fit statistics also favor the alternative model, which had lower AIC (15820 versus 15872) and BIC (16157 versus 16204) than the initial model.

Figures 7.8 and 7.9 compare the EAP scores obtained under the two models. Given the relatively small number of modifications specified in the alternative model, the effects on score estimates are surprisingly substantial. For example, the EAP scores for attribute 1 were consistently smaller under the alternative model. In the initial model, all examinees were estimated to have greater than 0.5 probability of possessing this attribute. Under the alternative model, about 78% had greater than 0.5 probability. Thus about 22% of this sample would be given different classifications for x_1 under the two models (assuming an EAP threshold of 0.5). In fact, ten of the fifteen attributes displayed re-classification rates of at least 5%: x_1 (22%), x_2 (8%), x_4 (13%), x_5 (6%), x_6 (6%), x_8 (9%), x_{13} (6%), x_{14} (7%), and x_{15} (6%).

Notably, differences in classification were not limited to those attributes on which the items with modifications loaded (e.g., x_4 , which is not measured by any of the items with changed item response functions in the alternative model), illustrating how changes in model specifications may have far-reaching consequences on the model-based interpretations (compare with, e.g., Steiger, 2002).

In this analysis of a fourth grade mathematics assessment, the hierarchical diagnostic model was used as an alternative to a traditional model described in a previous study (Lee et al., 2011). The model from the previous study seemed to provide rather good fit to the bivariate margins (as evident from the small number of items with strong local dependence). That said, there were a handful of item pairs with some evidence of misfit (local independence violations). The alternative diagnostic model—with a single group-specific dimension to model

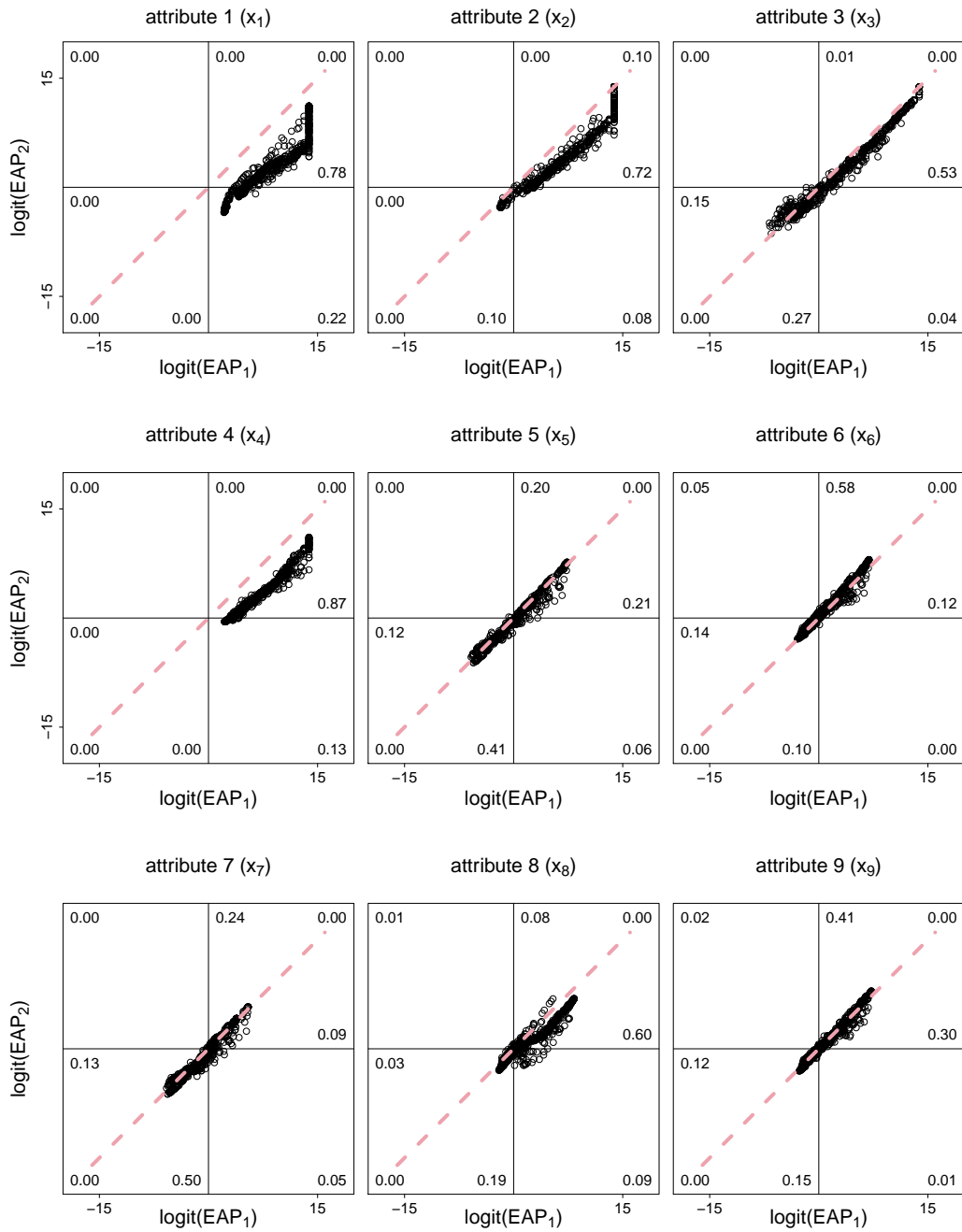


Figure 7.8: Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in TIMSS 4th grade mathematics attributes (x_1 – x_9).

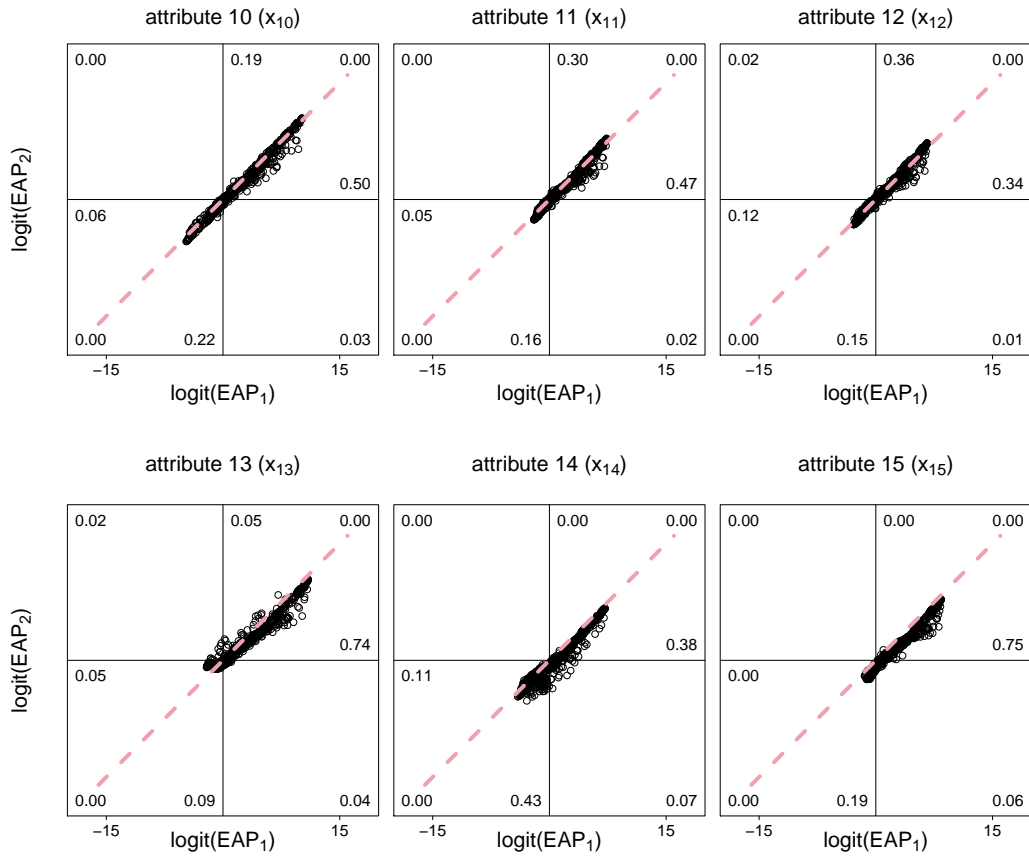


Figure 7.9: Expected a posteriori (EAP) estimates of examinee proficiency (in logit scale) in TIMSS 4th grade mathematics attributes (x_{10} – x_{15}).

dependence between a pair of items in a testlet, as well as some changes to the response functions of a few other items—provided improved fit to the data, both in terms of the LD X^2 indices and information-based criteria. EAP scores obtained from the two models were strongly associated but would nonetheless produce a substantial number of differences in classification decisions.

7.3 A Longitudinal Diagnostic Model for Assessing Attribute Stability

The third empirical application utilizes data from a study related to the development of measures of physical functioning (Fries, Cella, Rose, Krishnan, & Bruce, 2009; Rose, Bjorner, Becker, Fries, & Ware, 2008). This study was conducted as part of the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS; Reeve et al., 2007). For this illustration, I focus on five physical functioning items administered to 401 subjects at three points in time. Each of the items has three response categories. The discrete attributes of diagnostic interest (i.e., the x variables) are physical functioning measured on each of the three testing occasions. The attributes are regressed onto a higher-order dimension, and the attributes' standardized factor loadings on this dimension may be used to estimate the correlations between the attributes and, thus, the stability of attribute status over time.

In this example, the hierarchical diagnostic model is used to account for the dependencies among repeated items. The model is analogous to the longitudinal IRT models proposed by Hill (2006) and Cai (2010). Descriptions of the latent variables are provided in Table 7.5, and the corresponding Q-matrix and pattern of group-specific slope parameters are given in Table 7.6. A path diagram for the hierarchical model is presented in Figure 7.11. Slopes on the residual dependence factors are equal for each item across the three time points, so the item subscript

is dropped (e.g., $\beta_{1,1} = \beta_{6,1} = \beta_{11,1} = \beta_1$). The traditional model is nested within the hierarchical model, obtained by fixing the group-specific slope parameters to zero.

Table 7.5: Description of latent variables specified in a longitudinal diagnostic model of physical functioning.

Var.	Description	No.
Attribute variables (x_j)		
x_1	Physical functioning, time 1	5
x_2	Physical functioning, time 2	5
x_3	Physical functioning, time 3	5
Group-specific dimensions (ξ_s)		
ξ_1	Repeated item 1, residual dependence factor	3
ξ_2	Repeated item 2, residual dependence factor	3
ξ_3	Repeated item 3, residual dependence factor	3
ξ_4	Repeated item 4, residual dependence factor	3
ξ_5	Repeated item 5, residual dependence factor	3

Note: “Var.” indicates the latent variable. “No.” indicates the number of items loading on the latent variable.

Figure 7.10 shows the Chen and Thissen (1997) LD X^2 -based RMSEA values for the traditional (left) and hierarchical (right) diagnostic models. The traditional model fails to account for the very strong dependence among responses to the multiple administrations of each item. For example, the first physical functioning item (dealing with lifting or carrying groceries) appears in the model as items 1 (time 1), 6 (time 2), and 11 (time 3). The RMSEA values for the item pairs 1–6, 1–11, and 6–11 were 0.239, 0.217, and 0.246, respectively. Similarly strong dependencies were observed among the multiple instances of the other four items.

Figure 7.11 shows a path diagram for the proposed hierarchical model. Five group-specific dimensions were specified in order to account for the strong associations of responses to the same item across time.

Local item dependence is substantially reduced (even as some misspecification clearly remains) under the hierarchical model, as evident from the right side of Figure 7.10. For item pairs 1–6, 1–11, and 6–11 (all found to have LD X^2 -based

Table 7.6: Q-matrix and group-specific slope parameters for a longitudinal diagnostic model of physical functioning.

Time point (x_k)		Q					B			
Item Text (does health limit...)	y	x_1	x_2	x_3	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	
Physical Functioning, Time 1 (x_1)										
Lifting or carrying groceries.	y_1	1	0	0	β_1	0	0	0	0	
Climbing several flights of stairs.	y_2	1	0	0	0	β_2	0	0	0	
Bending, kneeling, or stooping.	y_3	1	0	0	0	0	β_3	0	0	
Walking one hundred yards.	y_4	1	0	0	0	0	0	β_4	0	
Bathing or dressing yourself.	y_5	1	0	0	0	0	0	0	β_5	
Physical Functioning, Time 2 (x_2)										
Lifting or carrying groceries.	y_6	0	1	0	β_1	0	0	0	0	
Climbing several flights of stairs.	y_7	0	1	0	0	β_2	0	0	0	
Bending, kneeling, or stooping.	y_8	0	1	0	0	0	β_3	0	0	
Walking one hundred yards.	y_9	0	1	0	0	0	0	β_4	0	
Bathing or dressing yourself.	y_{10}	0	1	0	0	0	0	0	β_5	
Physical Functioning, Time 3 (x_3)										
Lifting or carrying groceries.	y_{11}	0	0	1	β_1	0	0	0	0	
Climbing several flights of stairs.	y_{12}	0	0	1	0	β_2	0	0	0	
Bending, kneeling, or stooping.	y_{13}	0	0	1	0	0	β_3	0	0	
Walking one hundred yards.	y_{14}	0	0	1	0	0	0	β_4	0	
Bathing or dressing yourself.	y_{15}	0	0	1	0	0	0	0	β_5	

Note: Items are abridged for space. The Q-matrix is given by **Q**. The matrix **B** gives the pattern of group-specific slope parameters.

RMSEA values greater than 0.20 under the traditional model), the RMSEA values under the hierarchical model were 0.152, 0.135, and 0.108, respectively. Results were similar for the other sets of repeated items.

A likelihood ratio test was performed to examine the plausibility of the parameter constraints imposed on the hierarchical model in order to obtain the traditional model (i.e., $\beta_s=0$ for all the group-specific dimensions $s = 1, \dots, 5$). The test statistic, obtained by multiplying -2 times the difference in log-likelihood for the two models, was $X^2=1006$. With five degrees of freedom (the difference in the number of parameters estimated in the two models), $p < 0.0001$. This result suggests that the constraints implicit in the traditional model are well-supported by the data. The hierarchical model was also favored in comparisons of AIC (4418

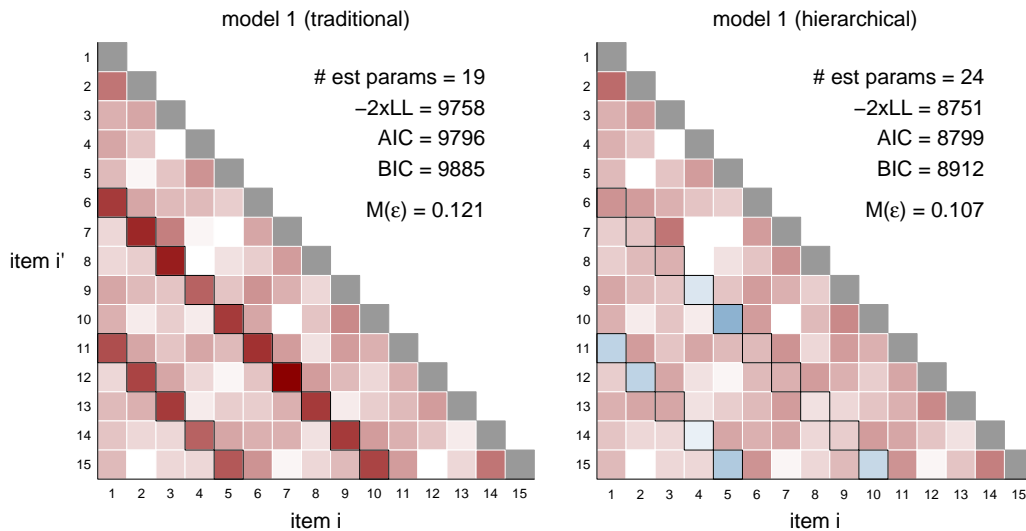


Figure 7.10: LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to the PROMIS Physical Functioning dataset.

versus 5414), and BIC (4507 versus 5480).

Figure 7.12 shows the logit of the EAP scores obtained from the two models for the three time points. For this example, incorporation of the group-specific dimensions has little effect on classification decisions (less than 4% difference in positive classification for each attribute, given a fixed EAP threshold of 0.5). Similar to the earlier example using the PISA reading test (Section 7.1), the primary impact of modeling the dependencies across time points seems to be shrinkage of the EAP estimates towards 0.5. For example, among examinees classified negative for attribute x_1 under either model (i.e., with $EAP_1 < 0.5$ and $EAP_2 < 0.5$), 92% had EAP values closer to 0.5 (and, thus further from zero) under the hierarchical. For those classified positive under both models, 44% had EAP values closer to 0.5 under the hierarchical model than under the traditional model, 35% had identical scores under the two models, and 20% had higher EAP scores under the hierarchical model. The fact that scores tend to be closer to 0.5 under the hierarchical model means that this model provides more

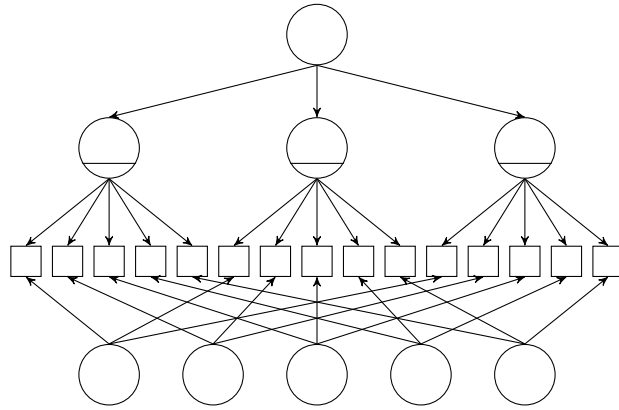


Figure 7.11: Path diagram for a longitudinal diagnostic model of physical functioning.

conservative assessments of the probability of latent class membership, which *may* be a more accurate characterization of the level of certainty than is provided by the traditional model (which, under some circumstances, may overstate the level of confidence; see Section 5.3).

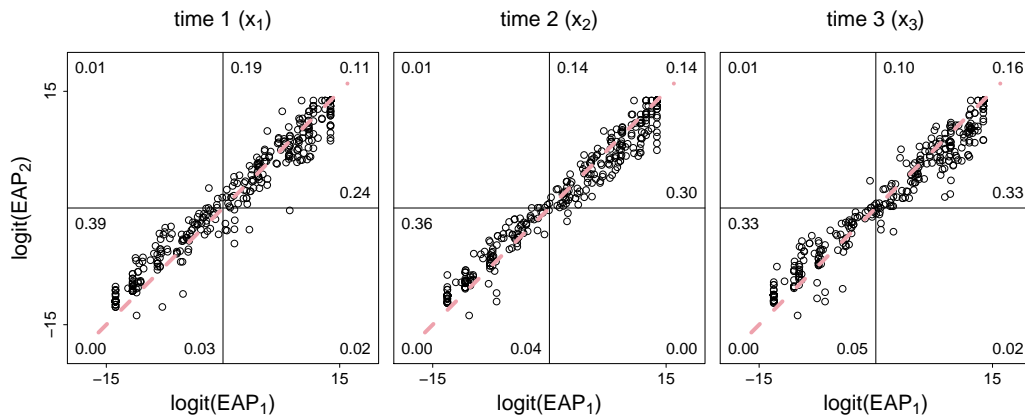


Figure 7.12: Expected a posteriori (EAP) estimates of examinee physical functioning (in logit scale) at three points in time.

In this example, the hierarchical model was used to account for residual dependencies among responses to the three administrations of each test item. This model compared favorably to the traditional model with respect to marginal goodness-of-fit. Impacts of model misspecification on EAP-based classification

decisions were minimal. However, the two models did provide somewhat different assessments of probable class membership, with the hierarchical models generally being somewhat conservative with respect to the level of classification certainty.

7.4 A Higher-order Diagnostic Model with a Random Intercept.

The fourth application also utilizes data from the PROMIS initiative, this time from a study involving the development of a measure of nicotine dependence among adult smokers (Shadel, Edelen, & Tucker, 2011; Edelen, Tucker, Shadel, Stucky, & Cai, 2012). Dependence is usually conceived of as a rather multi-dimensional construct, with several constituent symptoms or criteria (American Psychiatric Association, 2001). Here, I consider a set of 19 items measuring three possible dependence criteria: withdrawal, persistence, and abuse. Responses were obtained for 1365 adult smokers living in the United States.

Presence or absence of the three discrete symptoms (i.e., the latent attributes) could be modeled as a function of an overall dependence construct, as in a higher-order diagnostic model. However, one might be concerned about individual differences in the interpretation of response options as a source of nuisance dependence among items. Subjects responded to each dependence item in one of five response categories (e.g., *never, . . . , always*). Random intercept models have been proposed in such contexts (Maydeu-Olivares & Coffman, 2006; Cai, 2010). Adding an orthogonal random variable to the higher-order diagnostic model results in a hierarchical diagnostic model with a single group-specific dimension. The latent variables in the model are described in Table 7.7. The Q-matrix and patterns of slope parameters are presented in Table 7.8 and illustrated in Figure 7.13. The random intercept is the product $\beta_1 \xi_1$, where β_1 is the common slope parameter for all items (i.e., $\beta_{1.1} = \dots = \beta_{15.1} = \beta_1$).

Table 7.7: Description of latent variables specified in a higher-order, random intercept model for nicotine dependence.

Var.	Description	No.
Higher-order factor (θ)		
θ	Cigarette dependence	N/A
Attribute variables (x_j)		
x_1	Withdrawal	8
x_2	Persistent desire	6
x_3	Abuse	5
Group-specific dimensions (ξ_s)		
ξ_1	Random intercept (response style)	15

Note: “Var.” indicates the latent variable. “No.” indicates the number of items loading on the latent variable.

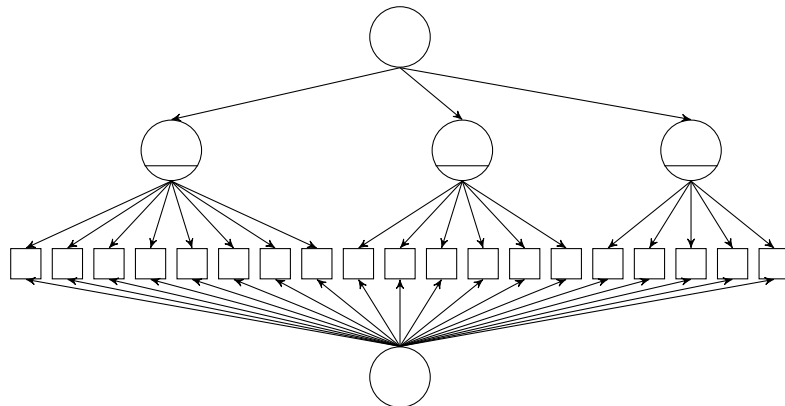


Figure 7.13: A higher-order, random intercept model for nicotine dependence.

Figure 7.14 presents the Chen and Thissen (1997) LD X^2 -based RMSEA values for the traditional model and the hierarchical model (with random intercept). The pattern of dependence observed under the traditional model was consistent with the notion of a single underlying continuous dimension influencing most or all of the items in the instrument. Specifically, the vast majority of item pairs displayed strong positive local dependence (across all pairs, the average RMSEA was 0.080).

The addition of the random intercept seems to provide improved fit, compared to the traditional model. Average RMSEA was 0.054. For some item pairs, the hierarchical model seems to over-correct the positive dependence, inducing

Table 7.8: Q-matrix and group-specific slope parameters for a higher-order, random intercept model for nicotine dependence.

Dependence Attribute (x_k)		Q			B
Item	y	x_1	x_2	x_3	ξ_1
Withdrawal (x_1)					
Withdrawal item 1	y_1	1	0	0	β_1
Withdrawal item 2	y_2	1	0	0	β_1
Withdrawal item 3	y_3	1	0	0	β_1
Withdrawal item 4	y_4	1	0	0	β_1
Withdrawal item 5	y_5	1	0	0	β_1
Withdrawal item 6	y_6	1	0	0	β_1
Withdrawal item 7	y_7	1	0	0	β_1
Withdrawal item 8	y_8	1	0	0	β_1
Persistent desire (x_2)					
Persistent desire item 1	y_9	0	1	0	β_1
Persistent desire item 2	y_{10}	0	1	0	β_1
Persistent desire item 3	y_{11}	0	1	0	β_1
Persistent desire item 4	y_{12}	0	1	0	β_1
Persistent desire item 5	y_{13}	0	1	0	β_1
Persistent desire item 6	y_{14}	0	1	0	β_1
Abuse (x_3)					
Abuse item 1	y_{15}	0	0	1	β_1
Abuse item 2	y_{16}	0	0	1	β_1
Abuse item 3	y_{17}	0	0	1	β_1
Abuse item 4	y_{18}	0	0	1	β_1
Abuse item 5	y_{19}	0	0	1	β_1

Note: The model's underlying Q-matrix is given by **Q**. All items load on ξ_1 with a common slope of β_1 .

negative local dependence (indicated in Figure 7.14 by blue shading). This was especially true for those pairs including items 15, 17, 18, or 19, all of which load on the abuse attribute/criterion, x_3). That said, the overall trend was a reduction in the RMSEA values, suggesting less severe local independence violations under the hierarchical model. This model was also preferred by the likelihood-based fit indices. AIC was lower for the hierarchical model than for the traditional model (59332 versus 69056), as was BIC (59858 versus 69577). A likelihood ratio test was performed to examine the plausibility of fixing the group-specific slope parameter in the hierarchical model to zero (i.e., $\beta_1=0$), which gives the traditional

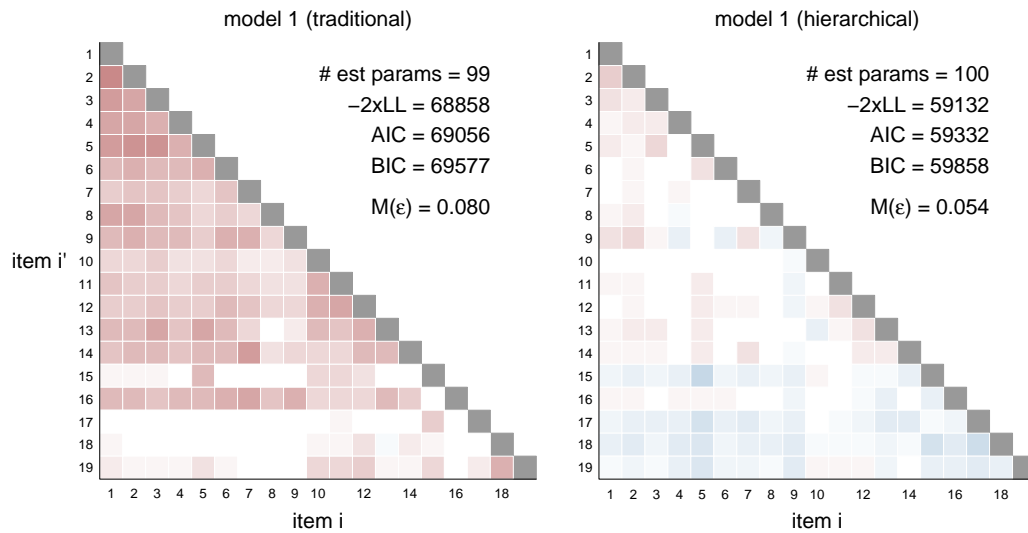


Figure 7.14: LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to PROMIS nicotine dependence data.

model. The test statistic, -2 times the difference in log-likelihoods, was $X^2=9726$. With degree of freedom (the difference in the number of parameters estimated), $p < 0.0001$. Thus, it seems the traditional model is overly restrictive.

EAP estimates from the two models are presented in Figure 7.15. There are substantial differences in these scores, with large percentages ($> 25\%$ for each of the three attributes) of subjects receiving different classifications under the two models, given a threshold score of 0.5. For attributes x_1 and x_2 (withdrawal and persistent desire, respectively), the EAP estimates under the hierarchical model are shrunken toward values of 0.5 (or logit of zero), which means that there is less certainty in classifications (i.e., fewer EAP values close to zero or one). Scores for attribute x_3 (abuse) show no particular bias or shrinkage under one model versus the other, just a good deal of dispersion; the correlation between the logit of the EAP scores is only 0.37 (compared to 0.61 and 0.62 for attributes x_1 and x_2). This result raises the possibility that the constructs measured by x_3 in the two models are not the same.

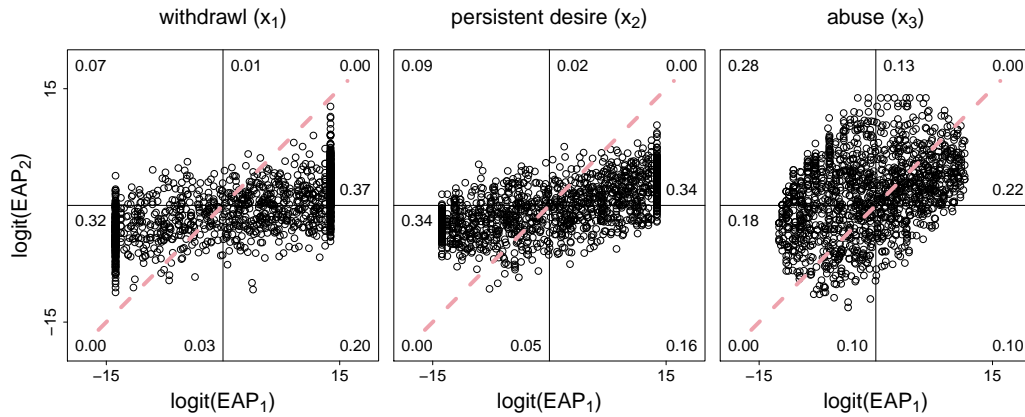


Figure 7.15: Expected a posteriori (EAP) estimates of three attributes related to nicotine dependence (in logit scale).

This section illustrated the use of a hierarchical diagnostic model in which a single random effect was used to model differences among subjects in their use of ordinal response categories. Because a common slope parameter (β_1) was used for all items, the hierarchical model required estimation of just one additional parameter, yet resulted in substantial improvements in model fit. EAP scores varied enough across the models to produce different classification decisions for large proportions of subjects. Thus, decisions to include a random intercept or not can be quite consequential.

7.5 A Higher-order Diagnostic Model for Depression

The final empirical application utilizes data from 3999 respondents to the Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001). The PDSQ consists of 125 items measuring fifteen common psychiatric disorders. For this analysis, I examined responses to a subset of the items measuring major depressive disorder (MDD). All subjects participating in the study had a clinical diagnosis of MDD, obtained using guidelines from the fourth edition of the Diagnostic Statistical Manual (DSM-IV; American Psychiatric Association, 2001).

Although there was no variability in the MDD diagnosis (since such a diagnosis was an eligibility criterion), subjects did vary in the particular combination of symptoms or aspects that led to the positive diagnosis. PDSQ items matching these symptoms were identified and used in the analysis. In order to obtain attribute variables measured by more than one item, some DSM criteria were combined. In the end, six attribute variables were specified, measured by a total of 19 items.

As in the previous examples, two alternative diagnostic models were fit to the data. Here, the traditional model had a simple loading structure, and the distribution of attribute profiles was estimated freely (i.e., without the specification of a higher-order latent trait).

The alternative, hierarchical model included random effects to account for two sets of locally dependent items. The latent variables for this model are summarized in Table 7.9. Item descriptions, the Q-matrix, and the \mathbf{B} matrix with group-specific slope parameters are shown in Table 7.10.

Table 7.9: Description of latent variables specified in diagnostic model of major depressive disorder.

Var.	Description	No.
Attribute variables (x_j)		
x_1	Depressed mood	2
x_2	Diminished interest or pleasure in daily activities	2
x_3	Physical impacts	4
x_4	Feelings of worthlessness, excessive or inappropriate guilt	3
x_5	Diminished ability to think or concentrate, indecisiveness	2
x_6	Recurrent thoughts of death	6
Group-specific dimensions (ξ_s)		
ξ_1	Opposing item pair (more or less sleep)	2
ξ_2	Item cluster (suicide ideation)	3

Note: “Var.” indicates the latent variable. “No.” indicates the number of items loading on the latent variable.

A path diagram for the hierarchical model is provided in Figure ???. The two group-specific dimensions were specified on the basis of results obtained from

fitting the initial model and subsequent review of item content. Figure 7.17 shows the Chen and Thissen (1997) LD X^2 -based RMSEA values obtained under the two models. Strong positive dependence is observed among items 17, 18, and 19. These items deal with suicide ideation, and this common focus appear to create a stronger association than is explained by the more general attribute (recurrent thoughts of death). Thus, a group-specific dimension was specified to account for this dependence.

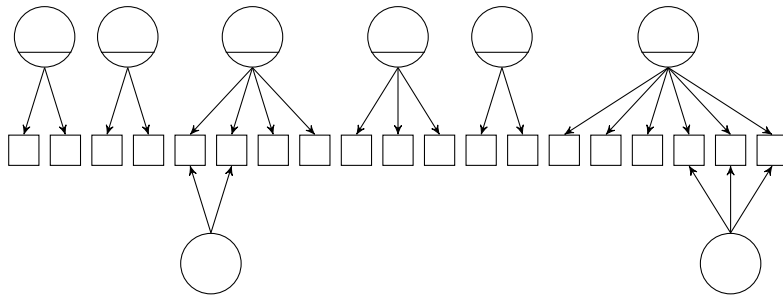


Figure 7.16: Path diagram for a diagnostic model of major depressive disorder.

A second group-specific dimension was added to model dependence between items 5 and 6. The reason for including this dimension is not readily apparent from the RMSEA tables. It was actually the slopes of these items on the attributes, along with knowledge of the item wording, that informed the modifications used in the alternative model. When the slope parameters of these items are estimated in the traditional model without any constraints, the slope for item 5 was estimated to be 12.6, while the slope for item 6 was -2.6. This slope parameter estimate for item 5 is much larger than the estimates for the other items in this cluster. Meanwhile, the negative estimate for the item 6 slope means that endorsement of the item will result in lower estimated probability of possessing the attribute, which is inconsistent with the DSM guidelines.

The apparent problem is that items 5 and 6 form a mutually exclusive pair. Item 5 describes getting less sleep than usual, while item 6 describes getting more sleep. Thus, respondents should not endorse both items (and when they

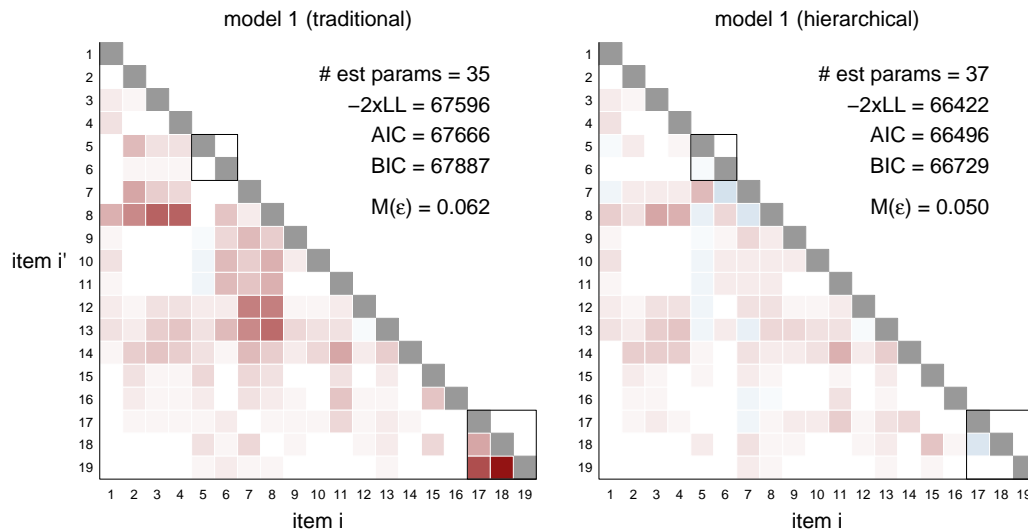


Figure 7.17: LD X^2 -based RMSEA values for the traditional (model 1, left) and hierarchical (model 2, right) diagnostic models fit to 19 items from the PDSQ measuring major depressive disorder.

do, it's likely a mistake). This creates a strong negative association between the two items, which, in turn, produces the unusual parameter estimates in the traditional model. A number of strategies could be taken, including dropping one of the items or combining the two items into a single item. The first strategy is not desirable, since it omits information that is relevant to the diagnosis. The second strategy (combining the items) might be reasonable but would make it harder to evaluate the performance of the individual items (which might be useful in future form or test assembly. In any case, the particular approach adopted here was to introduce a group-specific dimension on which items 5 and 6 would load. In order to identify the model and use the random effect to account for the *negative* dependence between these two items, the group-specific slope parameter for item 6 was constrained to be equal to the negative of the slope parameter for item 5 (i.e., $\beta_{6,1} = -\beta_{5,1}$). Because of the equality constraint, the item subscript is dropped (i.e., $\beta_{5,1} = \beta_1$ and $\beta_{6,1} = -\beta_1$).

Thus, the hierarchical model differs from the traditional model in the addition

of two group-specific dimensions (random effects)—one accounting for negative dependence between items 5 and 6, the other accounting for positive dependence among items 17–19. The hierarchical model appears to fit the bivariate margins better than the traditional model, as seen in Figure 7.17). The average LD X^2 -based RMSEA was 0.062 for the traditional model and 0.050 for the hierarchical model.

The group-specific dimension influencing items 5 and 6 allows both of these items to now have positive slopes on attribute x_3 , reducing the severity of local independence violations between these items and others in the instrument. The group-specific dimension accounting for the clustering of items 17–19 also seems to have served its intended purpose, as there is little remaining dependence among these items under the hierarchical diagnostic model. The likelihood ratio test indicates that the traditional model fits significantly worse ($X^2=1174$; with two degrees of freedom, $p < 0.0001$) than the hierarchical model. The hierarchical model is also favored on the basis of AIC (66496 versus 67666) and BIC (66729 versus 67887).

EAP scores for each of the six attributes were obtained for the traditional and hierarchical diagnostic models. These scores are compared in Figure 7.18. In general, the scores obtained from the two models were quite similar for attributes x_1 , x_2 , x_4 , and x_5 . For these attributes, the EAP scores provide similar assessments of classification uncertainty and almost identical classification decisions (given EAP threshold of 0.5). However, some differences are apparent in the scores for attributes x_3 and x_6 . These, of course, are the attributes for which a subset of items were deemed to load on a group-specific dimension. Most striking are the results for x_3 . Under the traditional model, subjects appear to be split into two groups (i.e., those with EAP scores either close to zero or close to one) with very little overlap. In contrast, more intermediate EAP scores are observed under the hierarchical model. The correlation between the EAP scores under these two models

is about 0.41. Meanwhile, for all other attributes, the correlations were at least 0.97. Thus, it seems that the particular hierarchical model leaves most attributes unaffected but significantly changes the estimated scores for x_3 in particular.

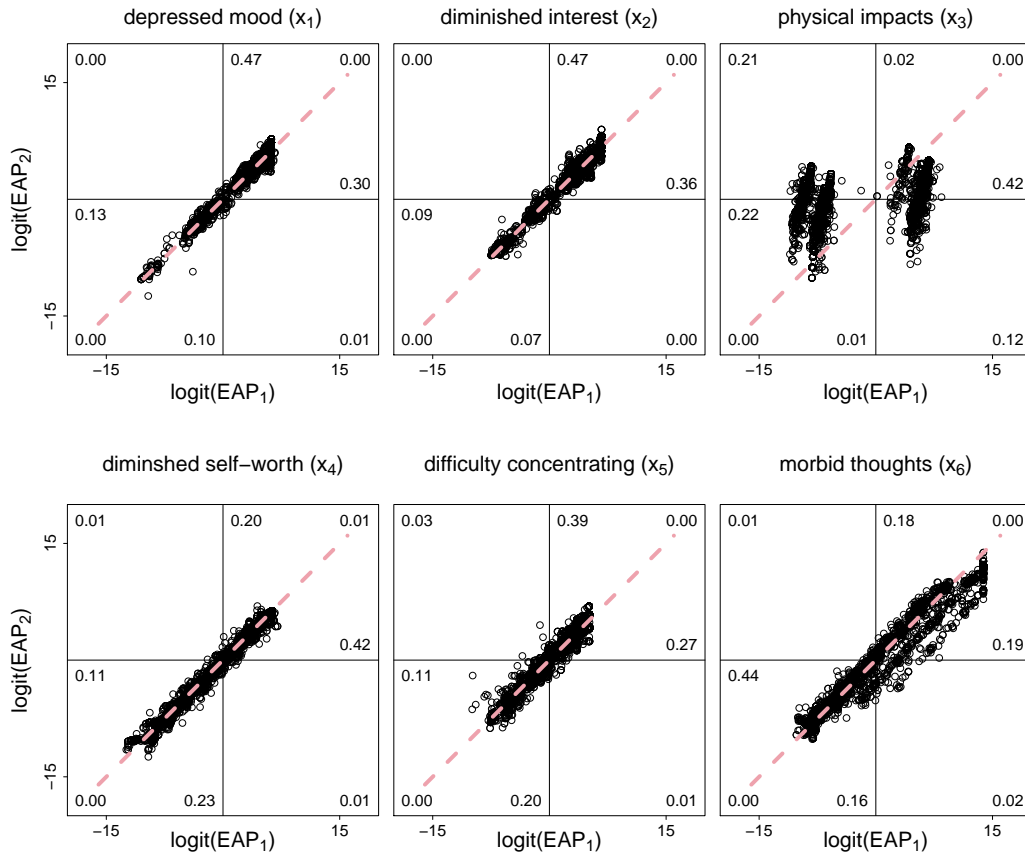


Figure 7.18: Expected a posteriori (EAP) estimates of six attributes related to major depressive disorder (in logit scale).

Because the subjects in this study were also classified according to the DSM-IV criteria for MDD, there is opportunity to compare EAP-based classifications under the two models to the clinical diagnosis. With six dichotomous attributes, the number of possible attribute profiles is $2^6 = 64$. Of the 3999 subjects in the study, 913 (23%) received an EAP-based classification that matched their clinical diagnosis. Under the hierarchical model 1328 subjects (33%) received a matching classification. Although both rates of correct classification are somewhat

low, it should be noted that some attributes were measured with as few as two items. In addition, the reliability of the clinical diagnoses are unknown. That said, it is perhaps uncommon to have any means of checking the validity of the classifications based on the diagnostic item response models. The fact that the hierarchical model would increase by nearly 50% the number of subjects correctly classified lends further support—in addition to the earlier evidence of improved fit—for the use of this model.

7.6 Summary

In this chapter, various hierarchical diagnostic models were applied to a series of real data examples. The results obtained under these models were compared against those from traditional diagnostic models, which do not explicitly model nuisance dimensionality. The hierarchical models were consistently found to fit the data better than the traditional models—both in terms of the bivariate margins (as evaluated by the LD X^2 indices) and the likelihood-based measures of overall fit (AIC and BIC). The magnitude of the differences in EAP estimates varied across the examples. However, in each case, some differences were evident—either changes in the relative certainty of classification or proportions of students receiving different classifications under the two models.

The examples were offered in order to illustrate the flexibility of the hierarchical model and to explore some of the issues arising in its application. Of particular interest may be the difficulty in determining when the hierarchical model should be favored. Clearly, fit indices are generally going to support its use (even acknowledging the increase in the number of parameters estimated). On the other hand, the impacts on scores and their interpretations must be considered. Unfortunately, classifications based on a standard or already accepted method were only available for the final example presented. Future studies would do well to

incorporate into their data collection design some alternative measures that would allow for the validity of classifications under alternative diagnostic models to be evaluated.

Table 7.10: Q-matrix and group-specific slope parameters for a longitudinal diagnostic model of physical functioning.

Time point (x_k)		Q						B	
Item Text (during past two weeks. . .)	y	x_1	x_2	x_3	x_4	x_5	x_6	ξ_1	ξ_2
Depressed mood (x_1)									
Felt sad or depressed.	y_1	1	0	0	0	0	0	0	0
Sad or depressed most of the time.	y_2	1	0	0	0	0	0	0	0
Diminished interest or pleasure in daily activities (x_2)									
Less joy or pleasure.	y_3	0	1	0	0	0	0	0	0
Less interest.	y_4	0	1	0	0	0	0	0	0
Physical impacts (x_3)									
Slept less than usual.	y_5	0	0	1	0	0	0	β_1	0
Slept more than usual.	y_6	0	0	1	0	0	0	$-\beta_1$	0
Felt jumpy and restless.	y_7	0	0	1	0	0	0	0	0
Felt tired.	y_8	0	0	1	0	0	0	0	0
Feelings of worthlessness, excessive or inappropriate guilt (x_4)									
Frequently felt guilty.	y_9	0	0	0	1	0	0	0	0
Negative thoughts about self.	y_{10}	0	0	0	1	0	0	0	0
Felt like a failure.	y_{11}	0	0	0	1	0	0	0	0
Diminished ability to think or concentrate, indecisiveness (x_5)									
Problems concentrating.	y_3	0	0	0	0	1	0	0	0
Difficultly making decisions.	y_3	0	0	0	0	1	0	0	0
Recurrent thoughts of death (x_6)									
Thought of dying in passive way.	y_3	0	0	0	0	0	1	0	0
Wished were dead.	y_3	0	0	0	0	0	1	0	0
Thought better off dead.	y_3	0	0	0	0	0	1	0	0
Had thoughts of suicide.	y_3	0	0	0	0	0	1	0	β_2
Seriously considered taking own life.	y_3	0	0	0	0	0	1	0	β_2
Thought specific way of taking life.	y_3	0	0	0	0	0	1	0	β_2

Note: Items are abridged for space. The Q-matrix is given by **Q**. The matrix **B** gives the pattern of group-specific slope parameters.

CHAPTER 8

Discussion

In this final chapter, I briefly summarize the major findings presented in this paper, discuss the implications of this work for educational and psychological measurement, and offer some possible directions for future research.

8.1 Review of Study Findings

As outlined previously, this study was organized around a handful of research questions, which will be used to here to structure my review of what was learned. The major questions could be stated as follow. First, is nuisance dimensionality a problem for diagnostic models that deserves attention? Second, even if nuisance dimensionality is a problem, is there anything that can be done about it? Third (and finally), supposing that nuisance dimensionality could be a problem in some cases (but not others, perhaps), how might one determine whether nuisance dimensionality must be attended to in the one's present testing context? Study findings relevant to each of these questions are discussed in the following sections.

8.1.1 Is Nuisance Dimensionality a Problem Deserving of Attention?

One of the primary findings of this study is that, indeed, failure to account for the influence of nuisance dimensions on item responses can undermine the usefulness of diagnostic models. Now, to be sure, the extent to which these nuisance dimensions adversely impact the test and test-based inferences depends on the

number, type, and strength of the nuisance dimensions, as well as other properties of the test. Thus, there are conditions in which there may be little or even no harm or loss due to ignoring these dimensions. On the other hand, there clearly exist some conditions in which the effects of misspecification are quite substantial. This variability in impact means that researchers must constantly be aware of the *possibility* of negative impacts and, perhaps, view them as explanations of variation in item responses that must be ruled out (rather than simply assuming such impacts to be ignorable).

The various simulation-based results presented in Chapters 4–6 identified several reasons why nuisance dimensions are a problem deserving of attention. First, failure to model such dimensions—as is standard practice in the current applications of traditional diagnostic models—can result in biased score estimates. When score estimates are inaccurate, examinees are more likely to be misclassified. Chapter 5 presented results showing that, for every measure of classification accuracy examined, the performance of traditional diagnostic models declined as the influence of nuisance dimensions increased.

A second reason why the problem is deserving of attention is that the biases in score estimates can result in completely misleading assessments of the level of certainty concerning the probability that an examinee possesses or lacks a particular attribute. This problem is, of course, related to the first (misclassifications). However, it is a slightly more subtle effect. It was demonstrated in Chapter 5 that one effect of ignoring nuisance dimensions is the over-estimation of classification certainty. As shown in Figures 5.5 and 5.6, EAP scores obtained from traditional models are often closer to zero or one than is warranted, given the theoretical (population-wide) prevalence of the attribute among individuals with a particular item response patterns. Such shifts in EAP scores could result in levels of confidence or certainty that are either higher or lower than warranted. In an extreme case, the EAP scores (which are the posterior probabilities of attribute possession)

obtained from the traditional model might be entirely uninterpretable. Interestingly, longer test length does not provide any protection against such bias. In fact, EAP scores from longer tests are even more likely to be close to values of zero or one

A third reason why the problem of nuisance dimensionality is deserving of attention is that the presence of such dimensions can have the effect of obscuring other types of model misspecification. In making this argument, I am assuming that one would already consider other types of model misspecification—such as problems with the Q-matrix, or use of the wrong type of item response function—as deserving of attention, since some effects of such errors have been examined (see, e.g., Kunina-Habenicht et al., 2012; von Davier, 2013). The results in Chapter 6 demonstrated that many of these sorts of misspecifications may be detected using the Chen and Thissen (1997) LD X^2 statistic. However, when nuisance dimensions are present, the task is greatly complicated, since the number of items with apparent local independence violations may be quite large.

8.1.2 Can Anything be Done about Nuisance Dimensionality?

The development of the hierarchical item response model provides a partial answer to this question. In essence, the approach is motivated by previous developments in factor analysis and IRT modeling. Specifically, I noted in Chapter 2 that a well-established strategy for dealing with nuisance dimensions in those modeling contexts is the specification of random effects, as applied in the item bifactor model (Gibbons & Hedeker, 1992) or two-tier item factor analysis model (Cai, 2010). Thus, a possible answer to the question of whether anything can be done about nuisance dimensions in the context of diagnostic modeling is that one might try applying that same strategy—which is the basis for the proposed hierarchical diagnostic model.

The development of this model does not provide a complete answer, however. In order to be useful, it must be possible for the model to be estimated efficiently and accurately. Thus, an important aspect of this research was the evaluation of an implementation (in the flexMIRT[®] software; Cai, 2012) of the hierarchical model. Results for this evaluation were presented in Chapter 4.

Through simulation study, it was found that under most conditions, model parameters were recovered with very little bias for each of the diagnostic models examined. Moreover, the estimated standard errors of measurement for the parameter estimates were, on average, quite similar in magnitude to the Monte Carlo standard deviations of the parameter estimates, indicating that these errors are estimated correctly.

There were a few cases in which some amount of bias was evident. This was primarily observed in conditions with strong nuisance dimensions (i.e., with slope parameters $\beta_s = 2$) and only one or two group-specific dimensions. The biases were more pronounced for the conditions with $K = 4$ response categories. Standard errors under these same conditions also demonstrated some amount of bias. It is possible that the stability or accuracy of the parameter and standard error estimates could be improved with specification of prior distributions or through the use of a larger number of quadrature points along the continuous latent dimensions in the fitted models.

Despite a small number of problematic conditions, however, the overall finding was that the implementation of the hierarchical diagnostic model allows for accurate estimation of model parameters. In addition, the estimation is accomplished with a great deal of efficiency, though the use of analytical dimension reduction (Gibbons & Hedeker, 1992), as described in Chapter 2. Very high-dimensional models (with as many as 20 group-specific dimensions) were able to be estimated, with computing time increasing linearly rather than exponentially in the number of group-specific dimensions.

Finally, results in Chapter 5 revealed that use of the hierarchical diagnostic model may alleviate many of the biases arising when nuisance dimensions are simply ignored (as in a traditional diagnostic model). Classification accuracy was better for the hierarchical model than for the traditional model, across all the measures examined. In addition, the EAP scores appears to be much better calibrated—i.e., provided posterior probabilities that more closely match the true (population) attribute prevalence for a given response pattern.

In summary, the hierarchical diagnostic model does seem to offer an approach for dealing with nuisance dimensionality. An implementation of this model is available, allowing researchers to estimate the model efficiently and accurately across a broad range of conditions and diagnostic model types. Perhaps most importantly, the model seems to be effective in dealing with nuisance dimensions, allowing for unbiased estimates of model parameters and good characterization of test examinees.

8.1.3 It is Possible to Tell When Nuisance Dimensionality is a Problem?

Given the findings that nuisance dimensions are a potentially serious problem in diagnostic modeling contexts and that the hierarchical model may provide an effective alternative to traditional models in cases where nuisance dimensions are present, a final primary study question was whether a model fit index might provide some insights concerning the need for the hierarchical model. The particular index examined was Chen and Thissen's (1997) LD X^2 statistic, which quantifies local item dependence between two items.

The index was computed for all possible item pairs in a test after fitting both the traditional and hierarchical models over the range of simulation conditions and data replications. Results from these analyses were presented in Chapter

6 Across all conditions, the index demonstrated good type I error control (i.e., maintained a low rate of rejection when the model was correctly specified). In addition, the index was sensitive to the presence of unmodeled nuisance dimensions in the traditional model, correctly identifying items pairs sharing the influence of these dimensions. When hierarchical models were fit to the data with nuisance dimensions in the generating model, the LD X^2 indices again demonstrated good type I error control.

Although detection and characterization of unmodeled nuisance dimensions was a primary focus, the LD X^2 indices were also evaluated as a possible tool for detecting other forms of misspecification (including modifications to the Q-matrix or use of the wrong item response function). The results of these analyses were mixed—not due to a limitation of the index but, instead, the apparent flexibility of the diagnostic model to absorb various types of misspecification (by adjusting to near-unity the prevalence of an extraneous attribute, for example).

Despite some limitations in the ability to detect some types of model misspecification, simulation results demonstrated that the LD X^2 index could provide very consistent and accurate characterization of underlying dimensions omitted from a diagnostic model. Use of the index would allow researchers to identify when a traditional diagnostic model might be inadequate in accounting for variability in item responses. Patterns of index values might even identify the specific clusters of items that might load on group-specific dimensions. Finally, the indices provide a tools for evaluating the effectiveness of the hierarchical model in accounting for the local dependence observed under the traditional model.

8.2 Directions for Future Research

Despite the current interest in diagnostic models, there continue to be many questions related to their use and interpretation. This study has attempted to explore

some of those questions—in particular, the ones outlined earlier in this chapter. However, there is certainly room to improve on quality of the answers obtained here, as well as to examine new questions. In the following sections, I identify some possible directions for future research.

8.2.1 Tests of Overall Model Fit

Evaluations of model fit in the current study were based exclusively on the Chen and Thissen (1997) LD X^2 statistic and on the overall log-likelihood (and related indices). The former constitutes a test of absolute fit but only for a very specific aspect of the model (how well the model explained associations between two items). The latter describes the overall fit of the model, but not in absolute terms. In other words, it was possible to verify in each of the real data examples that the hierarchical model fit the data better than the traditional model. However, it was not possible to verify that the hierarchical model was a good fitting model, in any absolute sense. Full-information goodness-of-fit statistics, such as Pearson's X^2 and the likelihood ratio statistic G^2 might serve such a purpose. However, it is a well-known problem in evaluating models fit to multinomial data (including the sorts of models considered here) that the full underlying contingency table will inevitably be sparse for tests of reasonable length and realistic sample sizes (Bartholomew & Tzamourani, 1999). For example, the full table for a test of 20 dichotomous items has 2^{20} (> 1 million) cells. Under such conditions, X^2 and G^2 do not follow their asymptotic distributions and, thus, are poorly calibrated. Limited-information fit statistics (e.g., Bartholomew & Leung, 2002) have been suggested as a possible alternative. Perhaps most notable is the M_r family of statistics (Maydeu-Olivares & Joe, 2005; Joe & Maydeu-Olivares, 2010), which are computed from marginal subtables, rather than the full contingency table. These marginal subtables are also the basis for the Chen and Thissen (1997) LD X^2 statistic. The fact that this index demonstrated good calibration and sensitivity

to various types of model misfit suggests that aggregations of the marginal subtables may also prove to be effective. Thus, extension of the limited-information tests such as M_2 to the case of diagnostic models should be explored.

8.2.2 Validation of Diagnostic Models

A limitation of the empirical analyses presented in the previous chapter was that an alternative method of examinee classification was only available in one case (and even in that case, the reliability of the alternative method—which was a clinician’s identification of depression symptoms—was unknown). Results from simulation study demonstrate the *potential* for rather severe impacts of classification accuracy when traditional diagnostic models are fit to data generated from models with nuisance dimensions. However, additional empirical analyses are needed to verify that this is actually the case in practice. However, to be most informative, these analyses would need to include some criterion measures through which the validity of classifications could be evaluated. Such measures would provide a more realistic test of the performance of the hierarchical model. This model worked well when matched exactly to the data generating model. However, it is unknown how well it would perform with the sorts of misspecifications that would be more typical of real data. Having alternative measures of the latent attributes would, of course, be useful. However, validity of the classifications might also be established by examining the relationships between these attributes—as measured by the diagnostic model—and other variables or processes. Of particular interest would be the extent to which classifications are influenced by prior treatment (e.g., instruction). Prospective studies might also look at the extent to which classifications predict response to treatment.

8.3 Implications for Educational and Psychological Measurement

Diagnostic models are growing in popularity and use. These models are well-suited to the sorts of formative purposes that might be used to improve instruction or clinical treatment. However, up to this point, the question of whether traditional cognitive diagnosis models fit real test data has been largely neglected. The purpose of this research has been to explore the potential consequences of ignoring nuisance dimensions and to develop a modeling framework that acknowledges the influence of these dimensions more explicitly.

In analyses of both simulated and real data, it was found that failure to account for nuisance dimensions can be detrimental, leading to significant rates of examinee misclassification. This undermines the primary function of diagnostic models. Given that nuisance dimensions are likely to be ubiquitous in real data analyses, there is reason to be skeptical of the validity of inferences made on the basis of traditional diagnostic models.

The alternative framework presented in this study seeks to model the influences of nuisance dimensions. In so doing, the hierarchical model was found to retain good classification accuracy in conditions where traditional models had greatly diminished utility. It also provided more correct estimates of classification certainty, which is significant given decisions about where to set threshold probability levels might be based on beliefs about the impact of such decisions on the rates of misclassifications. If the estimated probabilities are skewed (as they were under the traditional model in the presence of nuisance dimensions), there can no longer be any basis for predicting the costs or benefits of particular threshold levels.

To the extent that the proposed framework better fulfills modeling assumptions, its application will contribute to improved test development. Models that

more accurately account for the true underlying data structure will allow for more realistic statements concerning the reliability of inferences that may be supported by a diagnostic test. Clarifying the influence of nuisance dimensions will allow developers to better understand test design requirements, such as overall test length (number of items) and item heterogeneity (with respect to both diagnostic attributes and nuisance dimensions). This, in turn, is expected to enhance decision-making based on the results of diagnostic assessments, such as improved instructional strategies. Consequences of—and methods for dealing with—violations of local item independence have been relatively well-characterized for IRT models but not previously explored within the diagnosis modeling context. By exploring this issue, this study begins to fill that gap.

BIBLIOGRAPHY

- Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development.
- American Psychiatric Association. (2001). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15.
- Bartholomew, D. J., & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research*, *27*, 525–546.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.
- Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.
- Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.
- Cai, L., & Hansen, M. (2012). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/j.2044-8317.2012.02050.x
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.

- Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.
- Choi, H.-J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models*. Unpublished doctoral dissertation, Department of Psychology, University of Georgia.
- Choi, H.-J., Rupp, A. A., & Pan, M. (2013). Standardized diagnostic assessment design and analysis: Key ideas from modern measurement theory. In M. M. C. Mok (Ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific, Education in the Asia-Pacific region: Issues, Concerns and Prospects 18* (pp. 61–85). Dordrecht: Springer.
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.
- DeMars, C. E. (2007). “Guessing” parameter estimates for multidimensional item response theory models. *Educational and Psychological Measurement*, *67*, 433–446.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, *39*, 1–38.
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics, Vol. 26, Psychometrics* (pp. 979–1030). New York: Elsevier.
- Edelen, M. O., Tucker, J. S., Shadel, W. G., Stucky, B. D., & Cai, L. (2012). Toward a more systematic assessment of smoking: Development of a smoking

- module for PROMIS. *Addictive Behaviors*, *37*, 1278–1284.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.
- Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *The Journal of Rheumatology*, *36*, 2061–2066.
- Fu, J. (2005). *A polytomous extension of the fusion model and its Bayesian parameter estimation*. Unpublished doctoral dissertation, University of Wisconsin–Madison.
- Fu, J., & Li, Y. (2007, April). Cognitively diagnostic psychometric models: An integrative review. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Grochocinski, V. J. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.
- Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423–436.
- Hansen, M., & Cai, L. (2012, April). The potential of local dependence diagnostics to inform or mislead. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, BC.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.
- Hill, C. D. (2006). *Two models for longitudinal item response data*. Unpublished doctoral dissertation, Department of Psychology, University of North

Carolina at Chapel Hill.

- Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*, 257–270.
- Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393–419.
- Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.
- Lee, Y., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic models of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.
- Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3–18). New York: Cambridge University Press.
- Liu, Y., & Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*. doi: 10.1177/0013164412453841
- MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*, 502–511.
- Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362.

- Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.
- Maydeu-Olivares, A., & Montaña, R. (2012). How should we assess the fit of Rasch-type models? approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, 1-18. doi: 10.1007/s11336-012-9293-1
- Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73–90.
- Pomplum, M. (2007). A bifactor analysis for a mode-of-administration effect. *Applied Measurement in Education*, *20*, 137–152.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22–31.
- Rijmen, F. (2009). *Efficient full information maximum likelihood estimation for multidimensional IRT models* (Tech. Rep. No. RR-09-03). Princeton, NJ: Educational Testing Service.
- Rose, M., Bjorner, J., Becker, J., Fries, J., & Ware, J. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the patient-reported outcomes measurement information system (PROMIS). *Journal of Clinical Epidemiology*, *61*, 17–33.
- Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96.
- Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic

- classification models: a comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, *17*.
- Shadel, W. G., Edelen, M., & Tucker, J. S. (2011). A unified framework for smoking assessment: The PROMIS Smoking Initiative. *Nicotine & Tobacco Research*, *13*, 399–400.
- Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, *7*, 210–227.
- Steiger, J. H., & Lind, J. M. (1980, June). Statistically based tests for the number of common factors.
- Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment*, *81*, 209–219.
- Tatsuoka, K. K. (1983). An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.
- Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.
- Thissen, D., & Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 123–144). Washington, DC: American Psychological Association.
- Tian, W., Cai, L., Thissen, D., & Xin, T. (2012). Numerical differentiation methods for computing covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement*. doi: 10.1177/0013164412465875

- Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic procedures by means of simulated correlation matrices. *Psychometrika*, *34*, 421–459.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS No. RR-05-16). Princeton, NJ: ETS.
- von Davier, M. (2013). The DINA model as a constrained general diagnostic model: two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*. Retrieved from <http://dx.doi.org/10.1111/bmsp.12003> doi: 10.1111/bmsp.12003
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.
- Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire (PDSQ). *Archives of General Psychiatry*, *58*, 787–794.