Posterior Predictive Model Checking for the Diagnostic Input Noisy and Gate Model

by

Cigdem Alagoz Ekici

(Under the Direction of Seock-Ho Kim and Allan S. Cohen)

Abstract

This study presents a posterior predictive model checking (PPMC) method for the deterministic inputs, noisy "and" gate (DINA) model. The potential of the PPMC method is examined for detecting problems with the DINA model. χ^2 statistics are calculated based on latent class and raw score groups to evaluate model fit and item fit. Then PPP-values are calculated using these χ^2 values as discrepancy measures for both item fit and model fit evaluation. Two problem conditions were simulated to study these fit indices. The first problem situation occurs, when the higher order structure among the attributes are ignored, when analyzing the data. The second problem situation occurs, when the Q–matrix is misspecified. The performance of the fit indices was evaluated under the presence of these two problem situations. Type I error rates and power were calculated. χ^2 is calculated based on latent classes. PPP-values based on this χ^2 produced small Type I error rates and very good power. On the other hand, Type I error rates and power from χ^2 calculated based on raw score groups and PPP-values based on this χ^2 were not in the acceptable range. Item fit indices successfully detected problems with the Q–matrix misspecification. This helped identify which items were misspecified. However, neither item fit nor model fit indices detected problems with the modeling of the attribute relationship structure. When the Q–matrix misspecification was small, model fit indices did not reject the model. When 5% or more of the

Q–matrix were misspecified, the overall χ^2 calculated based on latent classes successfully rejected the model. A real data analysis was presented to demonstrate the application of these model and item fit indices for the DINA model.

INDEX WORDS: PPMC, Bayesian model fit, Model evaluation, Diagnostic classification, Higher order DINA model

Posterior Predictive Model Checking for the Diagnostic Input Noisy and Gate Model

by

Cigdem Alagoz Ekici

B.A., Gazi University, 2000

M.A., University of Georgia, 2005

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2012

Posterior Predictive Model Checking for the Diagnostic Input Noisy and Gate Model

by

Cigdem Alagoz Ekici

Major Professors: Seock-Ho Kim
Allan S. Cohen

Committee: Jonathan Templin
Gauri Datta

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

Dedication

To Hazel and Celil

Acknowledgments

I feel deeply indebted to many people for their assistance and encouragement during my doctorate, which has been demanding but also exciting part of my life.

First of all, I wish to thank the University of Georgia for financial support during my doctoral studies. I sincerely thank Dr. Seock-Ho Kim for setting a great example of an advisor that I hope to proudly emulate someday to the best of my abilities. I deeply appreciated Dr. Allan S. Cohen's genuine interest and support to ensure my continuing professional development as a strong scholar in the field of educational measurement and research. I am grateful to the faculty in the REMS program at the University of Georgia, who always had time to talk, asked stimulating questions, and encouraged me to investigate my ideas. Special thanks to those who agreed to serve on my committee: Dr. Seock-Ho Kim, Dr. Allan S. Cohen, Dr. Jonathan Templin, and Dr Gauri Datta. Their questions and comments during my program of study and the development of my research helped me to focus and critique my work. I deeply appreciated Dr. Jonathan Templin's invaluable feedbacks about diagnostic classification models.

Dr. Steve Cramer has been one supervisor that I cherished the opportunity to work with. His support facilitated me to gain empowering professional experiences as a psychometrician in the Georgia Center for Assessment. I would like to extend my deepest gratitude to Dr. Dorothy Harnish from the Educational Policy Evaluation Center at UGA. It was a great delight to learn from Dr. Dorothy Harnish while working under her supervision in many projects.

I wish to thank all my departmental colleagues, in particular Feiming Lee, Hye-Jeong Choi and Youn-Jeng Choi. Intellectual and supportive discussions with you have been helpful for me during my doctoral studies.

I am grateful to my parents, Hatice and Zeki Alagoz, for fully supporting me and my decisions. I also thank my daughter Hazel who has been a huge inspiration to me by showing what a person is capable of doing. Last, but by no means the least, I wish to express my deepest gratitude to my husband. His support and patience carried me along, especially with his ability to see possibilities rather than problems during this journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A number of diagnostic classification models (DCMs) have been proposed recently (de la Torre & Douglas, 2004; DiBello, Stout, & Roussos, 1995; Haertel, 1989; Hartz, 2002; Henson, Templin, Willse, 2009; Junker & Sijtsma, 2001; Templin, 2004).

DCMs are used to provide fine-grained information about examinees' strengths and weaknesses in responding to test items. This is different from score-based information that is provided by the usual classical test and item response theory models. These DCMs have the potential to be a new way of providing feedback to teachers and students alike. These developments in psychometric theory are both exciting and promising.

Concurrently, computational tools have been steadily improving and making feasible calculations that were previously not feasible. For example, it is relatively easy to find an estimator for a statistical model with Bayesian estimation methods using algorithms such as Markov chain Monte Carlo (MCMC).

Being able to find an estimator does not directly guarantee the correctness of the interpretations of estimated parameters. Prior to interpreting the results, one should investigate the model fit. Several fit indices have been proposed with DCMs, some at the item level, some at the test level, and some to compare alternative models (de la Torre & Douglas, 2004, 2008; Sinharay, 2006). As yet, however, fit indices have not been studied with respect to their properties such as Type I error rate and power.

Purpose of the Study

In this study, model-data fit methods are examined in the context of Bayesian estimation for one of the basic DCMs, the Deterministic Inputs, Noisy "And" gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977). Model fit is examined using the Bayesian posterior predictive model checking index (PPMC, Guttman, 1967; Rubin, 1984). Diagnostics obtained based on the PPMC are examined to see if they can be used to identify the problems with the DINA model and if they suggest ways to improve it. These diagnostics are applied to a real data example and results are reported. A simulation study is presented to demonstrate the effectiveness of the PPMC for performing model evaluation in the presence of a Q–matrix misspecification and under two higher order structure conditions.

Research Questions

The research questions are:

- How do the PPMC discrepancy statistics perform for the DINA model?
- What are the empirical Type-I error rates and power for suggested discrepancy measures?
- To what extent is the PPMC method useful in detecting the Q–matrix misspecification?
- To what extent is the PPMC method useful when the relationship among the attributes is hypothesized incorrectly?
- How do the PPMC discrepancy measures based on raw scores perform in comparison to the discrepancy measures based on latent classes?

Chapter 2

Theoretical Framework

Item response theory (IRT) models used for most testing programs typically model a single continuous unidimensional latent variable. Recent research, however, is focusing on modeling continuous multidimensional latent variables (e.g., Reckase, 2010). Another approach receiving attention in the educational measurement field is modeling of categorical multidimensional latent variables. One family of these models are known as diagnostic classification models (DCMs, Rupp, Templin, & Henson, 2010). These models are receiving attention because they offer the possibility of providing more fine-grained information than is available with standard item response models. DCMs can provide information that can be used to diagnose students' strengths and weaknesses based on their responses to test items. DCMs employ a vector of binary latent variables to which of these variables are required for each individual test item. The term *attribute* is used in the DCM literature to refer to these latent variables. A skill or knowledge state could be considered as an attribute in the diagnostic classification modeling context. The term attribute sometimes also refers to rules, subgoals, skills, or knowledge in various studies required for answering test items (Junker, 1999).

Multiple classification latent class models are a class of models used for person-by-items data. Every person is placed into a latent class by these models (Maris, 1999). DCMs are specific cases of these multiple classification latent class models. In DCMs, latent classes are created by the vector of binary latent variables indicating the attributes required for correctly answering each item.

Even though it is desirable to search for more detailed information about students' knowledge states from an educational assessment, it is also highly likely that these attributes are

related to a more general latent trait. Math ability, for example, is a general latent trait, and addition, subtraction, multiplication, and division are specific skills associated with this general ability. A math test could be designed to specifically measure these skills. It is highly likely that the dependence among the items are explained by the general math ability coupled with these additional more specific skills. The objective of the DCMs is to provide information about students' mastery of the specific skills. That is, a DCM could explain the dependence among the items due to the general math ability along with the specific attribute-level information.

Higher-order latent trait models aim to provide multivariate latent trait modeling when item response models appear to be an appropriatealternative. These higher-order latent trait models specify the attributes or knowledge states as arising from a latent trait. This is similar to the continuous latent ability in IRT models. A higher-order latent trait model defines the relationship between the general latent ability and specific knowledge as measured by the joint distribution of the attributes. Attribute classification and estimation of a latent trait are achieved by the same analysis (de la Torre & Douglas, 2004).

Let $\mathbf{Y}$ denote an item response vector with dichotomous elements for J items. Item responses are modeled as statistically independent given the attribute vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)'$. The kth element, α_k of α, is a binary indicator of a subject's mastery classification for the kth attribute. For example, α_k might be an indication of a student's mastery of addition where k represent the skill of addition. To achieve a complete latent variable model for $\mathbf{Y}$, conditional distribution of $\mathbf{Y}$ given an attribute pattern α needs to be formulated as well as the joint distribution of α (de la Torre & Douglas, 2004). A number of models have been presented for defining the conditional distribution of $\mathbf{Y}$ given α in the context of diagnostic classification models. For a response pattern $\mathbf{y}$, the conditional distribution is given by

$$P(\mathbf{Y}|\boldsymbol{\alpha}) = \prod_{j=1}^{J} P(y_i|\boldsymbol{\alpha}), \tag{2.1}$$

Item response functions require construction of a $\mathbf{Q}$–matrix. A $\mathbf{Q}$–matrix is defined in which attributes are indicated that are required to solve each item on the test. $\mathbf{Q}$ is a $J \times$

K matrix of zeros and ones, its element on the jth row and kth column is q_{jk}, which is the indicator of whether skill k is required to produce a correct response for the jth item.

The DINA model is one of the DCMs that defines the conditional distribution of $\mathbf{Y}$ given α.

DINA Model

The DINA model is a stochastic conjunctive model. A correct response is produced when an examinee masters all the attributes indicated in the $\mathbf{Q}$–matrix. The probability of a correct response for an examinee is the same if he/she has only some of the required attributes or none of the required attributes. In this sense, the DINA model is a conjunctive model. DINA is a stochastic model in that mastering all the attributes indicated in the $\mathbf{Q}$–matrix does not guarantee a correct response and missing one or more of these attributes does not result in an incorrect response. For each examinee and for each item an indicator, η_{ij}, is calculated for mastering required attributes for that item. This process provides the deterministic aspect of the model, and η_{ij} is determined by the attribute vector $\boldsymbol{\alpha}_i$ and by q_j (the row of $\mathbf{Q}$ that corresponds to the jth item for the ith subject) and is calculated as

$$\eta_{ij} = \prod_{k=1}^{K} \boldsymbol{\alpha}_{ik}^{q_{jk}}, \tag{2.2}$$

where η_{ij} equals 1, when an examinee possesses all the required attributes for a correct response and 0 when an examinee does not posses one or more of the required attributes. Some attribute patterns may result in calculation of the same η_{ij} and the same ideal item score patterns. Attribute patterns create equivalence classes with respect to their corresponding single ideal item score pattern. These are called *equivalent attribute patterns* (Tatsuoka, 2002).

The latent response variables η_{ij} are related to observed task performances X_{ij} according to the probabilities $s_j = P(X_{ij} = 0|\eta_{ij} = 1))$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0))$. s_j refers to the probability of slipping and missing the item when an examinee possesses all the required attributes. g_j refers to the probability of guessing and giving a correct response when an

examinee does not possess all the required attributes. Both are error probabilities, that is, false negative and false positive rates. When an examinee has all the required attributes to solve an item, the probability of correct response equals $1 - s_j$. When an examinee misses any one of the attributes necessary to solve the item, probability of correct response is equal to g_j. The IRF (item response function) for a single task is defined as

$$P(X_{ij} = 1|\boldsymbol{\alpha}, s, g) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} = P_j(\boldsymbol{\alpha}_i). \tag{2.3}$$

Individuals' responses are assumed to be independent of each other and responses are conditional on attributes indicated as $\boldsymbol{\alpha}$. The joint likelihood function of the DINA model is given as

$$L(s, g; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{J} [s_j^{1-y_{ij}} (1 - s_j)^{y_{ij}}]^{\eta_{ij}} [g_j^{y_{ij}} (1 - g_j)^{1-y_{ij}}]^{1-\eta_{ij}}. \tag{2.4}$$

The DINA model is appropriate for tests for which the conjunction of several attributes is required to achieve a correct response and lacking one required attribute results in incorrect response. It is easy to interpret and the model and the algorithms are available as well as a number of applications of it.

Higher-order Structure Among the Attributes

The DINA model defines the distribution of $\mathbf{Y}$ conditional on $\boldsymbol{\alpha}$. Modeling $\mathbf{Y}$ with the DINA model provides diagnostic feedback from a measurement. The objective of higher-order latent trait models is to combine this diagnostic information with a summative assessment by modeling $\boldsymbol{\alpha}$ in addition to item response data, $\mathbf{Y}$. If a test has been designed to measure cognitive abilities with higher-order latent trait models in mind, then it is possible for DCMs, such as the DINA model, to provide more specific cognitive diagnostic information in the test results. In addition, these models also have the potential to provide summative information about a general latent ability, such as from being combined with IRT models.

Latent attributes, $\boldsymbol{\alpha}$, are distributed independently conditional on a general latent trait, $\boldsymbol{\alpha}$. Dichotomous responses of items, $\mathbf{Y}$, are independently distributed conditional on $\boldsymbol{\alpha}$. The

probability distribution of $\boldsymbol{\alpha}$ defines the higher-order structure among the attributes. Several models for the probability distribution of $\boldsymbol{\alpha}$ have been developed (e.g., de la Torre & Douglas, 2004; Hartz, 2002; Marris, 1999; Li, 2008).

One way to define the probability distribution of $\boldsymbol{\alpha}$ is through the independence model (Maris, 1999). Independence model assumes that the latent classes, which are based on $\boldsymbol{\alpha}$ are independently distributed. Population proportions for each attribute are estimated for the probability distribution of $\boldsymbol{\alpha}$. The independence model might not be realistic in educational measurement because the attributes or cognitive skills are almost never independent of each other. Attributes are related to each other through a general latent trait.

Another model suggested for $\boldsymbol{\alpha}$ is a loglinear model (Henson, Templin, & Willse, 2009; Maris, 1999). A loglinear model defines the attribute main effects through the use of a log link function. This could be extended to model the interactions among the attributes. In another approach Hartz (2002) assumed that normally distributed latent variables generated $\boldsymbol{\alpha}$, dichotomous attributes.

The latent traits are modeled according to the complexity or simplicity of attributes. Based on the assessment, unidimensionality or multidimensionality of the attributes are assumed. The probability model for $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\theta}$ is

$$P(\boldsymbol{\alpha}|\boldsymbol{\theta}) = \prod_{k=1}^{K} P(\boldsymbol{\alpha}_k|\boldsymbol{\theta}). \tag{2.5}$$

General latent trait, $\boldsymbol{\theta}$, is assumed to be unidimensional in this study and it is normally distributed with $\mu = 0$ and $\sigma = 1$. In this study, attributes, $\boldsymbol{\alpha}_k$, are modeled with a 2 parameter logistic model with latent covariate, $\boldsymbol{\theta}$, and are defined as

$$\text{logit}[P(\boldsymbol{\alpha}_{jk}|\boldsymbol{\theta}_j)] = a_k\boldsymbol{\theta}_j - \boldsymbol{\beta}_k, \tag{2.6}$$

where $\boldsymbol{\alpha}_{jk}$ is a binary indicator of a mastery status of the jth subject for the kth attribute, θ_j is the ability parameter of the jth subject, and a_k and β_k are, respectively, the discrimination and difficulty parameters for the kth attribute.

The higher-order DINA model is used to classify the subjects according to specific attributes and concurrently provide estimates of general ability or latent traits in the same analysis. The 2 parameter logistic model for modeling the attributes and DINA for modeling the dichotomous item responses facilitated classification of examinees into latent classes and at the same time provided an estimate of their general ability.

Model Fit in DCMs

The model fit for DCMs is still an underdeveloped area (Habenicht-Kunina, 2010; Sinharay & Almond, 2007). A widely known model fit index for use with categorical data is a chi–square statistic based on response patterns. Such a chi–square statistic is expected to be a popular model fit index in an analysis of DCMs. The number of possible response patterns, however, generally is too large to provide a chi-square that functions well. Too many response patterns create contingency tables that are too sparse and a chi-square which is almost always 0 (Rupp, Templin, & Henson, 2010).

Utilizing the mixed number subtraction data from Tatsuoka (1990) and the higher-order DINA model, de la Torre and Douglas (2004) estimated item pair relationships as part of the model fit evaluation. This relationship took the form of the log-odds ratio which is a measure of association between binary random variables. The log-odds ratio for items j and j' is defined as

$$\log\left[\frac{(P(Y_j=1,Y_{j'}=1)P(Y_j=0,Y_{j'}=0))}{(P(Y_j=1,Y_{j'}=0)P(Y_j=0,Y_{j'}=1))}\right]. \tag{2.7}$$

de la Torre and Douglas calculated the difference between the observed and expected log-odds ratios for each item. The mean absolute difference, which is the average of observed and expected log-odds ratio differences over the items, were obtained. Competing models were compared in terms of the mean absolute difference. Results suggested that the higher order model produced smaller mean absolute difference compared to the independence model. Three other indices at the test level, Bayes factor (Kass & Raftery, 1995), Akaike information criterion (AIC; Akaike, 1973), and Bayesian information criterion (BIC, Schwarz, 1978), were

computed to provide global measures of the relative fit of the models. It should be noted each of these indices can be used as measures of evidence for a model with respect to another even when models are not nested. Even though the indices offered could provide insight into item fit as well as model fit, they are relative measures and need to be compared with other candidate models. In addition to log-odds ratio, de la Torre and Douglas (2008) proposed use of the proportion of examinees correctly responding to each item and product-moment correlations to check model fit. These statistics were computed for each item across several models and were employed to compare these models.

Sinharay (2006) and Sinharay and Almond (2007) suggested creating item fit plots and calculating χ^2-like and G^2 item fit statistic and overall fit statistic for assessing the model fit with Bayes network models. Item fit statistics were calculated based on both latent classes and raw scores. In the context of the Bayesian networks, the PPMC method presented by Sinharay (2006) and Sinharay and Almond (2007) produced promising results. In this study, methods recommended by Sinharay (2006) and Sinharay and Almond (2007) were investigated further and extended to the DINA model. Further explanation of these methods is presented in the following section.

Sinharay. Johnson, and Stern (2006) used the biserial correlation coefficient as a measure of inadequacy of IRT models. Sinharay (2004) also presented the biserial correlation coefficient in the context of Bayesian Network as a measure for item fit.

Templin and Henson (2006) calculated posterior predictive p-values based on the discrepancy measures comparing the RMSE of the item pair association, using Pearson correlations and Cohens kappa.

2.1 Posterior Predictive Model Checking

The posterior predictive model checking idea was first proposed by Guttman (1967). Guttman proposed a comparison of the observed and theoretical frequencies based on the

posterior predictive distribution of a future observation. This was described as a χ^2-type procedure and model fit was represented based on p-values (Bayarri & Berger, 1997).

PPMC uses the posterior predictive distribution as a reference distribution for comparing to the observed data. Similarity of observed data and the replicated data generated under the model becomes the objective of the PPMC methods. Replicated values, y^{rep}, can be generated from the posterior predictive distribution by adding a step in MCMC sampling using the likelihood function $f(y^{rep|\theta^{(t)}})$ that is evaluated at parameter values $\theta^{(t)}$ of the current state of the algorithm. The predictive data y^{rep} reflect the expected observations after replicating the experiment in the future, given observed y and assuming the particular model is true. If the model fits, then observed data should be plausible under the posterior predictive distribution. To check model fit, these samples of simulated values from the posterior predictive distribution of replicated data are compared to the observed data. In practice, the vectors y and y^{rep} are expected to be close to each other, if the adopted model is appropriate for describing the observed data. Discrepancies between the simulations and the data indicate possible problems with the analysis. Such a comparison can be facilitated by considering summary functions $T(y, \theta)$, which play the role of a test statistic for checking the assumption under investigation and measure discrepancies between the data and the model (Gelman, Carlin, Stern, & Rubin, 1994; Ntzoufras, 2009).

Let $p(y|\theta)$ denote the likelihood distribution for a statistical model where y denotes the data and θ denotes the parameters in the model. Let $p(\theta)$ denote the prior distribution of the parameters. Then the posterior density of θ is given as

$$P(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta}. \tag{2.8}$$

Let y^{rep} denote the replicated data that could have been observed, or the data that would be observed in the future if the experiment that produced y today were replicated with the same model and same value of θ that produced the observed data. The posterior predictive distribution is given as

$$P(y^{rep}|y) = \int_\theta p(y^{rep}|\theta)p(\theta|y)d\theta. \tag{2.9}$$

The discrepancy between model and data is measured by test quantities. A test quantity, $T(y, \theta)$, is defined as a scalar to summarize parameters and data to compare data to predictive simulations.

Posterior predictive p-value

Rubin (1984) presented the posterior p-values as the tail area probabilities which were calculated by utilizing the posterior predictive distribution that correspond to some test statistic, $T(Y)$. Meng (1994) and Gelman, Meng, and Stern (1996) defined the posterior predictive p-values by calculating test quantity of $T(Y)$ depending on the parameter θ and conditioning on the value of some auxiliary statistic $A(y)$ (Bayarri & Berger, 1997). The posterior predictive p–value (PPP–value) is defined as the probability that replicated data could be more extreme than observed data. PPP-value is measured by the test quantity p_B which is given as

$$p_B = P(T(y^{rep}, \theta) \geq T(y, \theta)|y) = \int \int I_{[T(y^{rep},\theta) \geq T(y,\theta)]} p(y^{rep}|\theta) p(\theta|y) dy^{rep} d\theta, \qquad (2.10)$$

where $I_{[A]}$ is the indicator function for the event A.

Posterior predictive p–values are easy to compute with the usual outputs from Bayesian numerical analysis. The posterior predictive distribution $p(y|\theta)$ of θ is computed by simulating replicated data sets. MCMC is one of the popular computing techniques in Bayesian analysis. Adding a new step in the recursive MCMC simulation in which y is generated from $f(y|\theta)$ provides a sample from the posterior predictive distribution from which the PPP–value is estimated. N simulations, $\theta^1, \theta^2, ..., \theta^N$, are drawn from the posterior density of θ, $p(\theta|y)$. Then one $y^{rep,n}$ from the likelihood distribution $p(y|\theta^n)$ for each simulated θ is drawn. This process results in N samples from the joint posterior distribution, $p(y^{rep}, \theta|y)$. The posterior predictive check compares the realized test quantities $T(y, \theta^l)$ and the predicted test quantities $T(y^{repl}, \theta^l)$. The proportion of N simulations for which the test quantity, $T(y^{rep,n}, \theta^n)$ equals or exceeds its realized value, $T(y, \theta^n)$, provides the estimated PPP–value. Extreme PPP–values indicate problems with the model fit (Gelman, Carlin, Stern, & Rubin, 2003).

Some researchers remarked on problematic aspects of the application of the PPP–value. For example, Bayarri and Berger (1997) pointed out that "too much" use of the observed data causes the inadequate behavior of the PPP–value. Furthermore, Bayarri and Berger (1997) indicated the similarity of the PPP–value to classical p–values and resulting inheritance of inadequacies of classical p–values.

2.2 The Graphical and Numerical Test Quantities and Summaries Utilized

The PPMC method utilizes a number of graphical summaries as well as some numerical test statistics to facilitate the model fit evaluation. Some possible graphical summaries are presented in this study in the real data analysis section. Test statistics that were employed are presented below.

Direct Data Display

One way to provide an overall idea about model fit is to draw the real data and a number of randomly selected replicated data. Then the image of the real data is visually compared to the images of the replicated data sets. When the model is fit to the data, it is expected to observe similar patterns from both the real and replicated data sets.

Residual Analysis

Residuals convey information about errors caused by the model. A visual inspection of a graphic display of the residuals is an easy way to check problems with the model. Suppose X_i denotes a scalar response for an individual i. The realized residual is $R_i = X_i - E(X_i|\theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$ denotes the vector of parameters in the model. X_i is the observed response to an item. Residuals are calculated at each MCMC iteration with the current values of correct response estimates and observed responses. At each iteration, residuals are calculated as the discrepancy between the observed total number of correct responses and the sum of the correct response probabilities for each item. A total score is considered outlying

if the posterior distribution for its residual is located too far from zero. Too many outliers indicate poor model fit (Sinharay & Almond, 2007).

Item Fit Analysis

Two discrepancy measures are used in this study to evaluate item fit which are χ^2_{lt} which is calculated based on latent class groups and χ^2_{raw} which is calculated based on raw score groups. Two posterior predictive p-values based on these two discrepancy measures, χ^2_{lt} and χ^2_{raw}, are used for item fit evaluation.

Item Fit Measures Based on Latent Class Membership

For IRT models, one way to assess item fit is to compare the average item performance for various proficiency groups to the performance levels predicted by the model (Hambleton & Swaminathan, 1985). This idea is extended to cognitive diagnostic models (Sinharay & Almond, 2007) with the use of proficiency classes. These classes are created by examinees' knowledge states $\boldsymbol{\alpha}_i = \alpha_{i1}, \alpha_{i2}, ..., \alpha_{ik}$. For these proficiency groups, expected and observed proportion of correct responses are compared. The observed and expected proportion of correct responses are obtained at each iteration of the MCMC algorithm for the examinees' knowledge states at that MCMC iteration. These knowledge states are updated by the MCMC algorithm at each step. When the sample size of the latent class is small, latent classes might be combined to calculate a more stable χ^2 statistic. Too many poorly fitting items indicate a problem with the fit of the model to the data whereas few poorly fitting items indicate problems with the items. Large values of this discrepancy measure indicate poor fit.

A χ^2 discrepancy measure is given by

$$\chi^2 = \sum_{k=1}^{K} N_k \frac{(O_{jk} - E_{jk})^2}{E_{jk}(N_k - E_{jk})}, \tag{2.11}$$

where O_{jk} is the proportion of examinees in proficiency class k, $k = 1, 2, ..., K$ who answer item j correctly, $j = 1, ..., J$. E_{jk} is the expected probability of an examinee in proficiency class k which is also defined as $P(X = 1|\alpha_{js}, \theta_j)$, where α_{js} is a binary indicator of a mastery status of the jth subject for the sth attribute. It is noted that O_{jk} is not truly observed since it is calculated based on the estimated proficiency class membership. O_{jk} would have been known if the proficiency class membership is known.

A comparison between the posterior distribution of χ^2_j and the posterior predictive distribution of χ^2_j provides a summary for the fit of the item j. This summary could take the form of a graph. Another measure to summarize this discrepancy for item fit is the posterior predictive p-value (PPP value). A PPP value close to 0 or 1 indicates that the variability in the data set compared to that predicted by the model is large and suggests poor fit. Summing these χ^2 discrepancy measures across items provides an overall discrepancy measure for assessing the fit of the model and is given by

$$\chi^2_{overall} = \sum_j \chi^2_j. \tag{2.12}$$

The posterior predictive p-value corresponding to the overall χ^2 provides an index for the overall goodness of fit.

Item Fit Measures Based on Raw Scores

Item fit measures presented above based on the equivalence classes include error due to latent class estimation. Similar goodness of fit indices can be calculated based on groups created based on their raw scores (Sinharay, 2003, 2004). The suggested χ^2–type measure is given by

$$\chi^2 = \sum_{k=1}^{J-1} N_k \frac{(O_{jk} - E_{jk})^2}{E_{jk}(1 - E_{jk})}, \tag{2.13}$$

where O_{jk} represents the observed proportion of examinees who answer item j correctly from among the examinees who answered k items correctly in total. E_{jk} is the expected probability of an examinee who obtained the raw score k.

Chapter 3

Design of Simulation Study

This simulation study had two motivations. First, it was designed to determine whether the PPMC discrepancy measures could detect problems with model fit, when the structural relationship among the attributes were modeled incorrectly. Thus, two versions of the DINA model were involved in the simulations: The independence DINA model and the higher–order DINA model. The second motivation was to investigate the effect of Q-matrix misspecification on the performance of the PPMC procedures. Simulations used one true Q–matrix and six incorrectly specified Q-matrices for this purpose. These conditions produced 28 combinations of two models and seven Q–matrices to investigate. These 28 conditions are presented in Table 3.1.

3.1 Number of Attributes, Sample Size, and Number of Items

Only the most relevant factors were varied to properly address the research questions and maintain the manageability of the simulation design. Number of attributes, sample size, and number of items were fixed for this study. Most applications on multiple classification models use about four to eight attributes (de la Torre & Douglas, 2004; Rupp & Templin, 2008). Although current research studies are undertaken with these numbers, it must be noted that the choice of these numbers is mostly due to the long calculation times required for larger numbers of attributes. In this study, the number of attributes was fixed at 5. The choice of an attribute number as 5 is not uncommon (de la Torre & Douglas, 2004; Li, 2008; Sinharay, 2006). This choice provided an opportunity to apply and extend previous research in this context.

Since each attribute means additional latent variables to estimate, assessments need a sufficient number of items to provide reliable information about each latent variable in the model. The number of items was fixed at 20 in this study. This choice with the choice of Q–matrix structure enabled each attribute to be measured by the same number of items in the test. Ten items were simulated as measured by single attributes, and ten items were simulated as measured by a pair of attributes. Complexity of the Q–matrix is the ratio of the number of items to the number of attributes. This proportion was fixed at $20/5 = 4$ for this study. The Q–matrix used in this study could be considered as moderately dense. When the Q–matrix is dense, misspecification errors tend to be smaller. Baker (1993) found that all levels of misspecifications (from 1% to 10 %) yielded larger errors in the parameter estimates for sparse Q–matrix conditions than in the dense Q–matrix. Therefore, in this study, it was expected that the choice of a moderately dense Q–matrix would accurately reflect errors caused by misspecified Q–matrix. Next, the sample sizes were fixed at 1,000 for each simulated data set under investigation. Previous research with similar sample sizes provided accurate and stable estimates, such as 1000 examinees with 30 items (de la Torre & Douglas, 2004), five hundred and 1000 examinees and 25 items (Li, 2008) and 1,500 examinees with 40 items (Hartz, 2002).

The structural parameters s, slip, g, guessing, $\boldsymbol{\alpha}$, attribute pattern and β, attribute difficulty were fixed across replications. Slip and guessing parameters were generated from a uniform (.1, .3) distribution (see Table 3.2). Attribute difficulty parameters and attribute discrimination parameters were fixed at 0 and 1 respectively for all five attributes.

Table 3.1
True and Related Fitting Models for Simulation

	True	Fitting
1	Q_0+Higher order DINA	Q_0+Higher order DINA
2	Q_0+Higher order DINA	Q_{O1}+Higher order DINA
3	Q_0+Higher order DINA	Q_{O2}+Higher order DINA
4	Q_0+Higher order DINA	Q_{U1}+Higher order DINA
5	Q_0+Higher order DINA	Q_{U2}+Higher order DINA
6	Q_0+Higher order DINA	Q_{B1}+Higher order DINA
7	Q_0+Higher order DINA	Q_{B2}+Higher order DINA
8	Q_0+Higher order DINA	Q_0+DINA
9	Q_0+Higher order DINA	Q_{O1}+DINA
10	Q_0+Higher order DINA	Q_{O2}+DINA
11	Q_0+Higher order DINA	Q_{U1}+DINA
12	Q_0+Higher order DINA	Q_{U2}+DINA
13	Q_0+Higher order DINA	Q_{B1}+DINA
14	Q_0+Higher order DINA	Q_{B2}+DINA

3.2 True and Misspecified Q–matrices

The Q–matrix conveys the attribute-item relationship information of the test being analyzed. It is created by the content experts. It is possible, however, that there may not be a consensus among experts. For even the simplest tasks, such as the fraction-subtraction, content experts have created more than one Q–matrix (de la Torre, 2004; Tatsuoka, 1990). The Q–matrix is central to the estimation of parameters. Even a small change in the correct specification of the Q–matrix can potentially result in large root mean squares of the difference between the parameter estimates obtained under correct Q–matrix specification conditions and those obtained under misspecification conditions (Baker, 1993).

Q–matrix misspecification has been studied from several perspectives. Rupp and Templin (2008) examined three types of Q-matrix misspecification, underspecification of the Q-matrix, overspecification of the Q-matrix, and a balanced misspecification of the Q-matrix. Underspecification of the Q-matrix occurs when 0's are specified in the Q-matrix where there should have been 1s. Overspecification of the Q-matrix is observed when 1s are specified in

place of the 0s in the Q-matrix. When the same number of 0s and 1s in the Q-matrix are exchanged, balanced misspecification of the Q-matrix takes place. Q-matrix misspecification can potentially create attribute patterns that are already being measured or could create new attribute patterns given that the all possible attribute patterns are not measured initially. Baker (1993) investigated Q–matrix misspecification by the percentage of misspecified Q–matrix indices by randomly changing 0s to 1s and 1s to 0s. Misspecification percentages of 1, 2, 3, 5, 7.5, and 10 were studied. Im and Corter (2011) examined two types of Q–matrix misspecification: Exclusion of an essential attribute that was needed to produce a correct response, and inclusion of a superfluous attribute that was not necessary to produce a correct response. Each of these misspecification scenarios are possible in real data. The Q–matrix that was used to generate the item response data in this study is presented in Table 3.2.

Table 3.2
Data Generating Q–matrix and Assessment Characteristics

Item no	α_1	α_2	α_3	α_4	α_5	g	s
1	1	0	0	0	0	.128	.141
2	0	1	0	0	0	.201	.267
3	0	0	1	0	0	.111	.140
4	0	0	0	1	0	.145	.260
5	0	0	0	0	1	.278	.191
6	1	0	0	0	0	.254	.214
7	0	1	0	0	0	.227	.226
8	0	0	1	0	0	.252	.275
9	0	0	0	1	0	.240	.184
10	0	0	0	0	1	.161	.181
11	1	1	0	0	0	.265	.236
12	1	0	1	0	0	.256	.206
13	1	0	0	1	0	.247	.125
14	1	0	0	0	1	.151	.194
15	0	1	1	0	0	.112	.111
16	0	1	0	1	0	.280	.131
17	0	1	0	0	1	.268	.139
18	0	0	1	1	0	.211	.238
19	0	0	1	0	1	.106	.278
20	0	0	0	1	1	.243	.208

To create the Q–matrix misspecification, the true Q–matrix was modified by changing several indicators of attributes required to produce a correct response to an item. Six misspecified Q-matrices were created, two were overfitting, Q_{O1}, Q_{O5}, two were underfitting, Q_{U1}, Q_{U5}, and two were balanced Q_{B2}, Q_{B6}. One percent of the indices were randomly selected and modified for the Q_{O1} and Q_{U1} matrices. Five percent of the indices were modified for Q_{O5} and Q_{U5}. This level of misspecification could be considered a high level of misspecification. Two percent of indices were modified from to create Q_{B2} and 6% of the indices were modified to create Q_{B6}. Balanced misspecifications modified even numbers of indices to yield equal numbers of overfitting and underfitting misspecifications. These matrices are presented below. Indices that were modified from the true Q–matrix are underlined.

To create overfitting Q–matrices, specified number of 0s were randomly selected and changed to 1s. For Q_{O1}, 1% of 100 indices were selected from among the 0s of the true Q–matrix and changed to 1s. For Q_{O5}, 5% of 100 the indices were randomly selected from among the 0s of true Q–matrix and changed to 1s. First, 0s of Q–matrix were assigned integers from left top of the Q–matrix to the right and bottom to randomly select the indices to be changed. Then, using a random integer generator (website http://www.random.org/integers/), the required number of integers were selected randomly between the 1s and 0s: 1 integer for 1% and 5 integers for 5%.

$$Q_{O1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & \underline{\mathbf{1}} & 1 & 1 \end{pmatrix} \quad Q_{O5} = \begin{pmatrix} 1 & 0 & 0 & \underline{\mathbf{1}} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \underline{\mathbf{1}} & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & \underline{\mathbf{1}} & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & \underline{\mathbf{1}} & \underline{\mathbf{1}} & 1 & 1 \end{pmatrix} \quad Q_{U1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & \underline{\mathbf{0}} & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$Q_{U5} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & \underline{\mathbf{0}} & 0 & 0 & 0 \\ 1 & 0 & \underline{\mathbf{0}} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & \underline{\mathbf{0}} & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & \underline{\mathbf{0}} & 0 \\ 0 & 0 & \underline{\mathbf{0}} & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad Q_{B2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \underline{\mathbf{1}} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \underline{\mathbf{0}} & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad Q_{B6} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & \underline{\mathbf{1}} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \underline{\mathbf{1}} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ \underline{\mathbf{1}} & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & \underline{\mathbf{0}} & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & \underline{\mathbf{0}} & 0 & 1 \\ 0 & 0 & 0 & \underline{\mathbf{0}} & 1 \end{pmatrix}$$

3.3 Data Generation Procedures

Data were generated in accordance with the higher order DINA model. The Q_0 matrix was used to generate the data. The following steps were taken to simulate item response data with higher order DINA model.

1. Attribute difficulty parameters and attribute discrimination parameters are fixed to 0 and 1 respectively for all five attributes.

2. General ability parameters are randomly generated from $N \sim (0, 1)$ distribution

3. Attribute patterns are generated for all examinees based on 1PL model, utilizing the general ability distribution of $N \sim (0, 1)$ and attribute discrimination and difficulty parameters of 0 and 1.

4. Slip and guessing parameters are generated from uniform $\sim$(.1, .3) distribution only once.

5. Item response data are generated based on DINA model, by using the generated attribute profiles, slip and guessing parameters.

3.4 Convergence

Simulated draws from $p(\theta|y)$ were used to summarize the posterior density and to compute means, quantiles, and descriptive statistics. Posterior predictive simulations of unobserved outcomes $\tilde{y}$ were obtained by conditioning on the drawn values of θ. In this process, iterative simulation should proceed long enough to be representative of the target distribution. Even when the iterative simulations converge and the subsequent string becomes representative of the target distribution, the first iterations still do not represent the target distribution. To diminish the effect of early iterations, some number of iterations are discarded. This process is called as *burn-in*. The number of burn-in iterations depends on the context (Gelman, Carlin, Stern, & Rubin, 2003). One way to investigate convergence is to visually inspect the trace plots. Irregularities or failure to reach a consistent pattern typically indicates problems with convergence.

One measure to monitor convergence is the convergence diagnostic $\hat{R}$ (Gelman & Rubin, 1992). It was calculated in this study as implemented in the software BOA (Smith, 2005).

When m parallel sequences are simulated, for each scalar estimand ψ between (B) and within-sequence (W) variances were calculated. Marginal posterior variance of the estimand var($\psi|y$) is calculated as a weighted average of B and W:

$$var^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{3.1}$$

The scale of the current distribution for ψ might be reduced by $\hat{R}$ if the simulation were continued in the limit $n \to \infty$. This potential scale reduction declines to 1 as $n \to \infty$ and calculated as

$$\hat{R} = \sqrt{\frac{\hat{var}^{+}(\psi|y)}{W}}. \tag{3.2}$$

If the potential scale reduction is high, then a longer chain of iterations is required to improve inference. The recommended value for $\hat{R}$ is near 1 (or, as a rule of thumb, the 0.975 quantile is less than 1.2).

To check the convergence, two parallel chains with over-dispersed initial values were run. Initial values for attribute discrimination parameters were set to .75 for one of the chains and the initial values for these attribute discrimination parameters were set to 1 for the second chain. Initial values for attribute difficulty, guessing, and slipping parameters were set to .25 for one chain and .10 for the other chain. Scale reduction factor $\hat{R}$ values were all close to 1. $\hat{R}$ values ranged from .9999 to 1.0212 for attribute discrimination parameters, from 1.0003 to 1.0260 for attribute difficulty parameters, from .9999 to 1.005 for slipping parameters, and from .9999 to 1.0096 for guessing parameters.

The examples of trace plots as well as the plots for the Gelman and Rubin statistics are given in Appendix B for the first condition which fit the true model and true Q–matrix. Trace plots did not show irregular patterns and they were stabilized after a few iterations which indicated convergence. Visual inspection of trace plots and also $\hat{R}$ indicated that convergence was obtained after 1,000 iterations for all structural parameters. Thus, a burn-in of 1,000 iterations and 4,000 post burn-in iterations were used in all conditions.

MCMC chains with 5000 iterations took approximately similar time across conditions. It took 9-10 hours to run each MCMC chain on a computer with 2.00 GHz Intel Xeon processor and 5GB RAM running a Windows 2003 operating system.

3.5 Recovery Analysis for the DINA Model

In this section, recovery of item parameters is evaluated under the conditions of correct and incorrect Q–matrix and model specifications. Recovery of the simulated item parameters, i.e., the slip and guessing parameters, recovery of the simulated attribute difficulty parameters for the higher-order DINA model were evaluated. Root mean squared errors (RMSEs) were calculated to examine how well the generating item parameters are estimated with the model. The RMSEs are expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{b_i} - b_i)^2}, \tag{3.3}$$

where b_i is the generating parameter for an item or an attribute parameter, $\hat{b_i}$ is the parameter estimate, and n is the number of items or attributes. Recovery of attribute mastery classification is evaluated by calculating the proportion of examinees who are correctly classified by the model as masters or non-masters on each attribute. Also, the means of the true and estimated parameters were calculated and compared for each condition.

3.6 Item Fit and Model Fit Analysis Procedures

After the data were simulated, the following analyses were done to evaluate item fit and model fit using both the raw score-based and latent class-based χ^2 proposed in this dissertation. At the same time, the item fit evaluation and model fit evaluation based on raw score-based and latent class based χ^2 was compared.

1. Run the higher-order DINA model for simulated response data in WinBUGS to obtain guessing, slipping, calculate χ^2 based on raw score groups and χ^2 calculated based on latent classes.

2. Detect item fit based on χ^2 calculated based on raw score groups for each item
3. Detect item fit based on posterior predictive p–values which used a χ^2 calculated based on raw score groups as discrepancy measures for each item
4. Detect model fit based on posterior predictive p–values which used an overall $\chi^2_{rawscore}$ as a discrepancy measure
5. Detect model fit based on $\chi^2_{latentclasses}$ calculated based on latent classes for each item
6. Detect item fit based on posterior predictive p–values which used $\chi^2_{latentclasses}$ as a discrepancy measure for each item
7. Detect model fit based on posterior predictive p–values which used overall $\chi^2_{latentclasses}$ as discrepancy measures
8. Calculate Type I error and power for all above item fit and overall fit analysis procedures

Power and Type I error rates were calculated by averaging the Type I error rate and power statistics from all replications under each condition. These results are presented and compared across all conditions. Therefore, the following questions were addressed based on the results:

1. Does χ^2 based on latent classes successfully detect problematic items?
2. Does χ^2 based on raw score groups successfully detect problematic items?
3. What is the error rate for incorrectly rejecting the items with χ^2 based on latent classes and with χ^2 based on raw scores?
4. What is the Type I error rate for χ^2 based on latent classes and for χ^2 based on raw score groups?
5. What is the power for χ^2 based on latent classes and for χ^2 based on raw score groups?

6. What is the Type I error rate for PPP-values calculated based on χ^2_{lt} and for PPP-values calculated based on χ^2_{raw}?

7. What is the power for χ^2 based on latent classes and χ^2 based on raw score groups?

8. Does the χ^2 based on latent classes perform better than the χ^2 based on raw scores for item fit evaluation and overall model fit evaluation?

9. Does the PPP-values based on χ^2_{lt} perform better than the PPP-values based on χ^2_{raw} for item fit evaluation and overall model fit evaluation?

Chapter 4

Results

First, the recovery evaluation is presented for each condition. Then the Type I error rates and power properties of item fit indices are presented. Next, Type I error rates and power are presented for overall fit indices in the form of PPP-values calculated based on χ^2 discrepancy measures. Finally the performances of fit indices that were calculated based on raw scores and based on latent classes are compared.

4.1 The recovery of model parameters

Estimation of item parameters is expected to be biased when either the Q–matrix or the model is misspecified. The amount of bias that misspecification introduces to the item parameter estimates is expected to have an impact on the model fit indices. The items that were misspecified with the wrong Q–matrix are naturally expected to be estimated with an error. It should also be remembered that the Q–matrix reflects a dynamic relationship between the items and attributes. Thus, the items that are correctly specified while some other items on the test are misspecified with the current Q–matrix, might also be estimated with an error. This error would be reflected in the item fit indices. Item parameter recovery was investigated for correct and incorrect Q–matrix and model specification conditions to shed light on the model and item fit procedures and to see the relationship between the model fit indices and the precision of parameter estimation.

4.1.1 The recovery of attribute parameters

Attribute parameters are included in the higher-order DINA model to define the relationship among the attributes. These are attribute discrimination, a, and attribute difficulty, β, parameters. The recovery of these parameters was evaluated by RMSEs. RMSEs are presented for all seven conditions that modeled the attribute relations. The first condition assumed the correct model and the correct Q–matrix specification. The recovery results from this condition reflect the potential of the algorithm to correctly estimate the parameters. Next, RMSEs are presented across the remaining six conditions where the model specification was correct but the Q–matrix was misspecified. These results convey the potential of recovering true parameters when the Q–matrix is misspecified. Comparison of these results provides information on the impact of Q–matrix misspecification on the recovery of true attribute parameters.

Table 4.1
Mean Estimates of the Attribute Parameters of the Higher Order DINA Model with True Q–Matrix

Item	a	$\hat{a}$	$SD(\hat{a})$	β	$\hat{\beta}$	$SD(\hat{\beta})$
1	1	1.04	0.15	0	0.01	0.09
2	1	0.83	0.17	0	0.06	0.08
3	1	1.05	0.17	0	-0.01	0.07
4	1	0.93	0.17	0	0.11	0.08
5	1	1.24	0.21	0	0.03	0.09

Table 4.2
Mean Estimates of the Attribute Parameters of the Higher Order DINA Model over Six Conditions with Misspecified Q–Matrix

Item	a	$\hat{a}$	$SD(\hat{a})$	β	$\hat{\beta}$	$SD(\hat{\beta})$
1	1	1.09	0.18	0	0.00	0.13
2	1	0.97	0.30	0	0.10	0.19
3	1	1.16	0.33	0	-0.02	0.14
4	1	1.00	0.18	0	0.11	0.12
5	1	1.39	0.32	0	0.09	0.12

Data were simulated by setting attribute discrimination parameters to 1 for each one of the five attributes that were measured by the simulated test in this study. True attribute difficulties for these five attributes were set to 0. The estimation procedure is satisfactory when the amount of the variations from 1 for the attribute discrimination and the amount of variation from 0 for the attribute difficulty parameters are small. Table 4.1 presents the average estimated attribute difficulty and discrimination values over 50 replications. This condition used the true Q–matrix. The results indicate that attribute parameters of the higher order DINA model are accurately estimated using the WinBUGS software and MCMC algorithm.

For the five attributes, the mean estimates deviated from the true (i.e., simulated) values by 0.04, 0.17, 0.05, 0.07, and 0.24, respectively, for discrimination parameters for attributes 1 to 5. These attribute discrimination parameter estimates across 50 replications had relatively small variabilities. Standard deviations for the attribute discrimination parameter estimates ranged from 0.15 to 0.21. Attribute difficulty parameter estimates have averages of 0.01, 0.06, -0.01, 0.11, and 0.03 across 50 replications for the first condition. These estimates were close to the true parameters, which were 0s for all attribute difficulty parameters. Standard deviations for the attribute difficulty parameters ranged from .07 to .09. Table 4.2 presents the average attribute discrimination and difficulty parameters over six conditions and 50 replications within each condition. These six conditions all used one type of misspecified Q–matrix. The mean estimates of the attribute parameters were larger in the presence of the Q–matrix misspecification than the mean estimates obtained with the use of the true Q–matrix. Mean estimates of the first, third, and fifth attribute discrimination parameters under the first condition were closer to the true value of 1. Mean estimates of the second and fourth parameters were closer to 1 for misspecification conditions. It should be noted that the averages over misspecification conditions had larger variances.

The recovery results for attribute discrimination and attribute difficulty parameters under the seven conditions are presented in Table 4.3 and Table 4.4, respectively. RMSE values for

Table 4.3
RMSE for the Attribute Discrimination Parameter of the Higher Order DINA Model for Seven Conditions

Fitting model	a_1	a_2	a_3	a_4	a_5
HODINA+Q_0	0.16	0.24	0.18	0.18	0.32
HODINA+Q_{O1}	0.16	0.23	0.18	0.18	0.33
HODINA+Q_{O5}	0.19	0.22	0.19	0.20	0.31
HODINA+Q_{U1}	0.15	0.25	0.26	0.18	0.29
HODINA+Q_{U5}	0.16	0.50	0.79	0.18	0.37
HODINA+Q_{B2}	0.31	0.27	0.15	0.19	0.58
HODINA+Q_{B6}	0.18	0.21	0.16	0.16	0.87

Table 4.4
RMSE for the Attribute Difficulty Parameter of the Higher Order DINA Model for Seven Conditions

Fitting model	β_1	β_2	β_3	β_4	β_5
HODINA+Q_0	0.09	0.09	0.07	0.14	0.09
HODINA+Q_{O1}	0.08	0.10	0.08	0.12	0.09
HODINA+Q_{O5}	0.13	0.10	0.15	0.10	0.10
HODINA+Q_{U1}	0.08	0.09	0.09	0.20	0.09
HODINA+Q_{U5}	0.12	0.40	0.24	0.19	0.19
HODINA+Q_{B2}	0.11	0.11	0.07	0.22	0.16
HODINA+Q_{B6}	0.20	0.31	0.17	0.09	0.23

attribute discrimination parameters were similar across the seven conditions, which included both true and false Q–matrix specifications. Relatively higher RMSE values were observed when QU5, QB2, and QB6 were used. QB2 and QB6 matrices were produced to present a balanced misspecification by incorrectly specifying 2% and 6% of indices of the Q–matrix. QU5 is an underfitting Q–matrix which was created by specifying 5% of indices as 0s instead of 1s. RMSEs for attribute difficulty parameters presented a similar pattern to that for the RMSEs for attribute discrimination parameters. RMSEs for QB2, QB6, and QU5 were relatively larger than other Q–matrices utilized in this study. From Tables 4.3 and 4.4 it can

be seen that RMSEs were smaller for difficulty parameters than discrimination parameters across all conditions, except when QB2 and QB6 were used.

4.1.2 The recovery of item parameters

Item parameters include guessing, g, and slipping, s, parameters. The recovery of these parameters was evaluated by investigating the estimated values and also RMSE. Table 4.5 presents the mean estimated guessing and slipping parameters of the higher order DINA model over 50 replications from the first condition, in which the correct Q–matrix was specified. The mean of the estimated guessing parameters over 50 replications varied from the true values, that is, the generating valued, by .002 to .032 for 20 items. The standard deviation was 0.01 for all 20 items over 50 replications. Thus, it can be claimed that the guessing parameter of the higher order DINA model could be accurately estimated using the current algorithm.

For all the items, the means of the estimated slip parameters did not deviate from the true value by more than .032. This bias ranged from 0 to .032 for all 20 items. The standard deviation ranged between .01 and .02. Thus, it could be claimed that the estimation procedure recovered the true values well.

Table 4.6 presents the average item parameter estimates over 13 Q–matrix misspecification conditions. The difference between the average estimated guessing values and true values ranged from 0 to .065 over 13 Q–matrix misspecification conditions. The range of this difference was .002 to .032 for estimation with the true Q–matrix. Thus, it is observed that the Q–matrix misspecification increased the difference between the average guessing estimates and the true values.

The difference between the average estimated slipping values and the true slipping values over 13 Q–matrix misspecification conditions ranged from .002 to .073. This difference was between 0 and .032 when the true Q–matrix specification was utilized. Q–matrix misspecification produced larger errors of the slipping parameter estimates as well.

Table 4.5
Mean Estimates of the Item Parameters of the Higher Order DINA Model over 50 Replications

Item	g	$\hat{g}$	$SD(\hat{g})$	s	$\hat{s}$	$SD(\hat{s})$
1	.128	.160	.01	.141	.134	.01
2	.201	.192	.01	.267	.247	.01
3	.111	.128	.01	.140	.151	.01
4	.145	.165	.01	.260	.261	.02
5	.278	.283	.01	.191	.191	.01
6	.254	.252	.01	.214	.233	.01
7	.227	.216	.01	.226	.239	.01
8	.252	.276	.01	.275	.243	.01
9	.240	.250	.01	.184	.170	.02
10	.161	.172	.01	.181	.157	.01
11	.265	.260	.01	.236	.242	.02
12	.256	.263	.01	.206	.178	.02
13	.247	.269	.01	.125	.095	.02
14	.151	.141	.01	.194	.183	.02
15	.112	.107	.01	.111	.103	.02
16	.280	.290	.01	.131	.109	.02
17	.268	.263	.01	.139	.138	.02
18	.211	.182	.01	.238	.209	.02
19	.106	.098	.01	.278	.281	.02
20	.243	.248	.01	.208	.217	.02

RMSEs for guessing parameters are presented in Table 4.7 for 14 conditions. Table entries are boldfaced to indicate the items for which one or more Q–matrix indices have been changed from the true specification of 1 to the false specification of 0. Italized indices indicate the Q–matrix misspecifications where the true value of 0s are changed to the false specification of 1s. RMSEs for guessing parameter estimates ranged from .01 to .06 for the items that were specified correctly with the Q–matrix. RMSEs ranged between .01 and .03 for the first condition where both the model and the Q–matrix were both correctly specified. Whenever there is a misspecification of one or more indices, RMSEs increased.

Table 4.6
Mean Estimates of the Item Parameters Across 13 Misspecification Conditions

Item	g	$\hat{g}$	$SD(\hat{g})$	s	$\hat{s}$	$SD(\hat{s})$
1	.128	.188	.08	.141	.139	.02
2	.201	.189	.02	.267	.252	.02
3	.111	.120	.03	.140	.153	.02
4	.145	.210	.08	.260	.273	.03
5	.278	.283	.02	.191	.195	.01
6	.254	.269	.05	.214	.233	.02
7	.227	.214	.02	.226	.245	.02
8	.252	.297	.05	.275	.245	.02
9	.240	.240	.02	.184	.177	.02
10	.161	.175	.02	.181	.165	.02
11	.265	.256	.01	.236	.274	.08
12	.256	.274	.03	.206	.213	.08
13	.247	.267	.01	.125	.093	.02
14	.151	.139	.01	.194	.213	.09
15	.112	.118	.03	.111	.149	.11
16	.280	.289	.01	.131	.152	.09
17	.268	.265	.01	.139	.137	.02
18	.211	.177	.01	.238	.289	.11
19	.106	.098	.01	.278	.351	.12
20	.243	.273	.03	.208	.242	.24

Italized and boldfaced RMSE values are consistently large with values ranging from .06 to .23. The guessing parameter of an item is estimated with consistently large errors when the Q–matrix indicated one or more required attributes for that item when they were not required.

Boldfaced RMSEs range from .01 to .05. These errors of guessing parameter estimates are smaller than italized RMSEs. Thus, RMSEs for guessing parameter are smaller for under-specified items than overspecified items. Besides, RMSEs for underspecified items are similar to RMSEs for the correctly specified items for guessing parameters.

RMSEs are similar for the models that specified the structure among the attributes as 2 parameter logistic model and the models that did not specify a relationship among attributes.

Thus, specification of or failure to specify the higher order relationship among the attributes did not contribute to the lower level parameter estimate errors.

When there is one or more misspecification of required attributes for one or more specific items in the test, the RMSEs for correctly specified items are not larger than the RMSEs from the true Q–matrix specification. Thus, the misspesification for one or more items are not expected to impact the guessing parameter estimates of other correctly specified items.

Table 4.7
RMSEs for the Guessing Parameters Across 14 Conditions

Fitting model		g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
1	Ho-DINA+Q0	.03	.02	.02	.02	.01	.01	.02	.03	.02	.02
2	Ho-DINA+QO1	.03	.02	.02	.02	.01	.01	.02	.02	.02	.02
3	Ho-DINA+QO5	***.23***	.03	.02	***.17***	.02	.02	.02	.02	.02	.02
4	Ho-DINA+QU1	.03	.02	.03	.04	.01	.01	.02	.03	.02	.02
5	Ho-DINA+QU5	.05	.04	.06	.03	.03	.01	.03	.04	.02	.04
6	Ho-DINA+QB2	.04	.03	.02	***.17***	.03	.01	.03	.03	.02	.03
7	Ho-DINA+QB6	.02	.02	.04	.02	.03	***.13***	.02	***.15***	.02	.05
8	DINA+Q0	.03	.02	.02	.01	.01	.01	.02	.02	.01	.01
9	DINA+QO1	.03	.02	.02	.01	.01	.01	.02	.02	.01	.01
10	DINA+QO5	***.23***	.03	.02	***.18***	.01	.03	.02	.02	.02	.01
11	DINA+QU1	.03	.02	.02	.02	.01	.01	.02	.03	.02	.01
12	DINA+QU5	.03	.02	.02	.02	.01	.02	.02	.05	.02	.02
13	DINA+QB2	.03	.03	.02	***.18***	.01	.02	.03	.02	.01	.01
14	DINA+QB6	.02	.03	.02	.01	.01	***.14***	.02	***.15***	.02	.02
Fitting model		g_{11}	g_{12}	g_{13}	g_{14}	g_{15}	g_{16}	g_{17}	g_{18}	g_{19}	g_{20}
1	Ho-DINA+Q0	.01	.01	.02	.01	.01	.01	.01	.03	.01	.01
2	Ho-DINA+QO1	.01	.01	.02	.01	.01	.01	.01	.03	.01	***.06***
3	Ho-DINA+QO5	.01	***.08***	.02	.01	.02	.01	.01	.04	.01	***.09***
4	Ho-DINA+QU1	.01	.01	.03	.01	.01	.01	.01	**.04**	.01	.01
5	Ho-DINA+QU5	**02**	**.01**	.03	.01	**.02**	.02	.01	**.05**	**.01**	.01
6	Ho-DINA+QB2	.01	.01	.02	**.03**	.01	.01	.01	.02	.01	.01
7	Ho-DINA+QB6	.01	.01	.01	.01	***.09***	**.02**	.01	.03	**.01**	**.02**
8	DINA+Q0	.01	.01	.02	.01	.01	.01	.01	.03	.01	.01
9	DINA+QO1	.01	.01	.02	.01	.01	.01	.01	.03	.01	***.07***
10	DINA+QO5	.01	***.09***	.02	.01	.01	.01	.01	.03	.01	***.10***
11	DINA+QU1	.01	.01	.02	.01	.01	.01	.01	**.04**	.01	.01
12	DINA+QU5	**.03**	**.01**	.02	.01	**.02**	.01	.01	**.03**	**.01**	.01
13	DINA+QB2	.01	.01	.02	**.03**	.01	.01	.01	.03	.01	.01
14	DINA+QB6	.01	.02	.02	.01	***.10***	**.02**	.01	.03	**.01**	**.02**

Boldfaced indices indicate underspecification, and italized and boldfaced indices indicate overspecification of the Q–matrix.

RMSEs for slipping parameters are presented in Table 4.8 for the 14 conditions. As in Table 4.7, Table 4.8 entries are boldfaced to indicate the underspecified items where 1s are changed to 0s, and italized indices indicate the overspecified items where the true value of 0s are changed to the false specification of 1s. RMSEs for slipping parameter estimates ranged from .01 to .05 for the items that were specified correctly with the Q–matrix. RMSEs ranged between .01 and .03 for the first condition, where both the model and the Q–matrix were both correctly specified. Whenever there is a misspecification of one or more indices, RMSEs increased.

Boldfaced RMSE values are consistently large with values ranging from .02 to .26. The slipping parameter of an item is estimated with consistently large errors when the Q–matrix fails to indicate one or more required attributes for that item.

Italized RMSEs range from .02 to .06. These RMSEs of slipping parameter estimates are smaller than boldfaced RMSEs. Thus, RMSEs for slipping parameters are smaller for overspecified items than underspecified items. RMSEs for overspecified items are similar to RMSEs for the correctly specified items for slipping parameters.

RMSEs are slightly different for some items between the models that specified the structure among the attributes with the two parameter logistic model and the models that did not specify a relationship among the attributes. Specification of or failure to specify the higher order relationship among the attributes didn't contribute to the lower level parameter estimate errors.

Misspecification of specific items did not increase RMSEs for the correctly specified items, a pattern which was also observed for the guessing parameter. Thus, the misspesification for one or more items is not expected to impact the item parameter estimates of other correctly specified items.

Table 4.8
RMSE for the Slipping Parameters Across 14 Conditions

Fitting model		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
1	HODINA+Q0	.02	.02	.02	.02	.01	.02	.02	.03	.02	.03
2	HODINA+QO1	.01	.02	.02	.02	.01	.02	.02	.03	.02	.03
3	HODINA+QO5	***.03***	.02	.03	***.05***	.01	.02	.03	.02	.02	.03
4	HODINA+QU1	.02	.03	.01	.02	.01	.02	.02	.04	.03	.03
5	HODINA+QU5	.02	.03	.02	.02	.01	.02	.02	.05	.03	.03
6	HODINA+QB2	.02	.02	.02	***.03***	.01	.02	.03	.04	.03	.03
7	HODINA+QB6	.02	.05	.02	.02	.01	***.03***	.02	***.04***	.02	.02
8	DINA+Q0	.01	.02	.02	.02	.01	.03	.03	.03	.01	.01
9	DINA+QO1	.01	.02	.03	.02	.01	.03	.03	.03	.02	.01
10	DINA+QO5	***.03***	.01	.03	***.04***	.02	.02	.03	.02	.02	.02
11	DINA+QU1	.01	.02	.02	.03	.01	.03	.03	.03	.02	.01
12	DINA+QU5	.02	.02	.03	.03	.01	.03	.02	.02	.02	.02
13	DINA+QB2	.01	.01	.02	***.03***	.02	.02	.03	.03	.02	.02
14	DINA+QB6	.02	.03	.02	.02	.02	***.02***	.02	***.06***	.02	.01
Fitting model		s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
1	HODINA+Q0	.02	.03	.03	.02	.02	.03	.02	.03	.02	.02
2	HODINA+QO1	.03	.03	.03	.02	.02	.02	.02	.03	.02	***.02***
3	HODINA+QO5	.03	***.02***	.03	.03	.04	.02	.02	.02	.03	***.03***
4	HODINA+QU1	.03	.03	.04	.02	.02	.04	.02	**.21**	.02	.02
5	HODINA+QU5	**.02**	**.18**	.04	.03	**.24**	.04	.02	**.19**	**.23**	.02
6	HODINA+QB2	.04	.03	.04	**.22**	.02	.02	.02	.04	.02	.02
7	HODINA+QB6	.03	.03	.02	.03	***.03***	**.21**	.02	.02	**.22**	**.23**
8	DINA+Q0	.02	.05	.04	.03	.03	.03	.02	.04	.02	.02
9	DINA+QO1	.02	.04	.04	.04	.03	.02	.02	.03	.02	***.03***
10	DINA+QO5	.03	***.04***	.04	.05	.02	.02	.02	.03	.02	***.05***
11	DINA+QU1	.02	.04	.05	.03	.03	.03	.02	**.22**	.02	.02
12	DINA+QU5	**.21**	**.19**	.05	.04	**.32**	.03	.02	**.23**	**.26**	.02
13	DINA+QB2	.03	.05	.05	**.24**	.02	.02	.02	.04	.02	.02
14	DINA+QB6	.03	.03	.03	.03	***.04***	**.26**	.02	.04	**.25**	**.26**

Boldfaced indices indicate underspecification, and italized and boldfaced indices indicate overspecification of the Q–matrix.

4.2 Item Fit Evaluation with χ^2 Calculated Based on Latent Classes

χ^2 is calculated for each item based on latent class classifications for the higher order DINA model, one true specified Q–matrix, and six misspecified Q–matrices. First, the descriptive statistics of these χ^2_{lt}s are investigated to provide an idea about their overall distribution. Table 4.9 presents the average χ^2 values that are calculated based on latent classes over 50 replications for the seven conditions that utilized the higher order DINA model. Standard deviations are given in parentheses next to the mean values of χ^2s in Table 4.9.

If the latent class classifications of examinees were known, then the χ^2_{lt} would follow a χ^2 distribution with 28 *df*. The degrees of freedom is four less than the number of latent classes when the statistical model is the higher order DINA because the number of parameters estimated with the model is four. These four parameters are attribute difficulty, attribute discrimination, item guessing, and item slipping. Thus, χ^2_{28} is used as a reference distribution to check the item fit. Values larger than 41.34 are in the 95th or greater percentile of the χ^2_{28} distribution.

Items that are misspecified by the Q–matrix are boldfaced in Table 4.9. The mean values of χ^2s are all smaller than the reference value of 41.34 for the correctly specified items, and they are larger than the reference value for all the misspecified items. Standard deviations are smaller for correctly specified items, whereas these are larger for misspecified items. The mean values of χ^2s are similar for correctly specified items across true and misspecified Q–matrices. Consistent with the estimation errors, having one or more misspecified items did not impact the fit of the correctly specified items.

Table 4.9
Mean Estimates of the χ^2_{lt} for 20 Items Utilizing the Higher Order DINA Model

	$\overline{\chi^2_1}$ (SD)	$\overline{\chi^2_2}$ (SD)	$\overline{\chi^2_3}$ (SD)	$\overline{\chi^2_4}$ (SD)	$\overline{\chi^2_5}$ (SD)	$\overline{\chi^2_6}$ (SD)	$\overline{\chi^2_7}$ (SD)	$\overline{\chi^2_8}$ (SD)	$\overline{\chi^2_9}$ (SD)	$\overline{\chi^2_{10}}$ (SD)
HoDINA+Q_0	34.2	33.8	31.9	33	32.4	30.4	36.4	32.4	31.4	33.8
	(3.5)	(3.8)	(2.8)	(3.3)	(3.7)	(2.7)	(4.0)	(3.1)	(2.9)	(2.8)
HoDINA+Q_{O1}	34.5	33.4	31.7	32.2	32.6	30.5	36.0	32.2	32.4	33.6
	(3.5)	(3.7)	(2.7)	(3.0)	(3.6)	(2.7)	(3.9)	(3.1)	(3.0)	(2.9)
HoDINA+Q_{O5}	**95.9**	32.9	31.8	**89.5**	31.9	31.9	33.8	31.9	34.3	34.1
	(9.0)	(2.7)	(2.0)	**(7.2)**	(3.0)	(2.2)	(3.1)	(2.7)	(2.6)	(2.3)
HoDINA+Q_{U1}	33.3	33.2	31.2	32.5	31.8	30.3	34.9	32.1	31.3	33.8
	(3.0)	(3.7)	(2.4)	(2.8)	(3.2)	(2.4)	(3.9)	(3.0)	(2.5)	(2.9)
HoDINA+Q_{U5}	34.8	36.8	32.1	31.9	32.6	30.4	35.7	32.6	31.1	33.0
	(2.6)	(3.0)	(1.4)	(2.1)	(2.7)	(1.9)	(2.8)	(2.5)	(1.6)	(2.2)
HoDINA+Q_{B2}	33.6	34.5	32	**88.0**	33.3	30.4	36.8	31.3	31.9	32.6
	(2.9)	(3.8)	(2.2)	**(8.6)**	(3.6)	(2.3)	(4.0)	(2.6)	(2.4)	(2.2)
HoDINA+Q_{B6}	34.0	34.6	33.5	33.9	33.1	**81.5**	34.4	**75.2**	31.9	34.5
	(2.6)	(3.0)	(1.7)	(2.4)	(2.8)	**(6.4)**	(2.6)	**(5.5)**	(2.1)	(2.3)

	$\overline{\chi^2_{11}}$ (SD)	$\overline{\chi^2_{12}}$ (SD)	$\overline{\chi^2_{13}}$ (SD)	$\overline{\chi^2_{14}}$ (SD)	$\overline{\chi^2_{15}}$ (SD)	$\overline{\chi^2_{16}}$ (SD)	$\overline{\chi^2_{17}}$ (SD)	$\overline{\chi^2_{18}}$ (SD)	$\overline{\chi^2_{19}}$ (SD)	$\overline{\chi^2_{20}}$ (SD)
HoDINA+Q_0	32.2	32.7	32.6	32.1	32.2	31.8	33.4	30.7	31.3	32.1
	(2.9)	(3.3)	(3.1)	(2.3)	(2.6)	(2.9)	(3.4)	(2.0)	(2.7)	(2.4)
HoDINA+Q_{O1}	32.2	32.6	32.4	31.7	32.1	32.1	33.6	30.8	31.5	**66.3**
	(2.9)	(3.2)	(3.0)	(2.3)	(2.5)	(3.0)	(3.3)	(2.2)	(2.7)	**(7.8)**
HoDINA+Q_{O5}	31.7	**73.7**	31.9	32.0	31.8	32.1	32.8	32.6	31.9	**77.6**
	(2.2)	**(6.4)**	(2.2)	(2.0)	(2.1)	(2.6)	(2.8)	(2.5)	(2.6)	**(8.2)**
HoDINA+Q_{U1}	31.8	33.2	32.6	32.1	31.7	31.4	33.6	**94.1**	31.6	32.1
	(2.5)	(3.2)	(3.1)	(2.1)	(2.4)	(2.7)	(3.2)	**(8.6)**	(2.6)	(2.5)
HoDINA+Q_{U5}	**68.9**	**71.3**	31.5	33.7	**84.9**	33.1	36.1	**79.5**	**84.5**	31.3
	(6.9)	**(6.8)**	(2.0)	(1.6)	**(8.5)**	(2.6)	(2.9)	**(7.2)**	**(7.4)**	(2.0)
HoDINA+Q_{B2}	31.8	32.9	33.2	**99.0**	31.3	31.3	32.4	31.6	33.6	32.1
	(2.8)	(3.2)	(2.8)	**(8.9)**	(2.3)	(2.6)	(2.7)	(1.9)	(3.0)	(2.2)
HoDINA+Q_{B6}	35.6	32.4	32.2	33.9	**118.0**	**86.5**	32.2	30.9	**100.5**	**78.8**
	(2.8)	(2.4)	(2.3)	(2.3)	**(12.8)**	**(8.6)**	(2.1)	(1.8)	**(8.3)**	**(6.0)**

Table entries are boldfaced to indicate the misspecified items.

Next, χ^2 is calculated for each item based on latent class classifications utilizing the independence DINA model, one true specified Q–matrix, and six misspecified Q–matrices. Table 4.10 presents the average χ^2 values that are calculated based on latent classes over 50 replications for the seven conditions that utilized the DINA model. Standard deviations are given in parentheses next to the mean values of χ^2s in Table 4.10.

If the latent class classifications of examinees were known, then the χ^2 would follow a χ^2 distribution with 30 *df*. The degrees of freedom is two less than the number of latent classes when the statistical model is the independence DINA because the number of parameters estimated with the model is two. These two parameters are item guessing and item slipping. Thus, χ^2_{30} is used as a reference distribution to check the item fit. Values larger than 43.77 are in the 95th or greater percentile of the χ^2_{30} distribution.

Items that are misspecified by the Q–matrix are boldfaced in Table 4.10. The mean values of χ^2s are all smaller than the reference value of 43.77 for the correctly specified items, and they are larger than the reference value for all the misspecified items. Standard deviations are smaller for correctly specified items, whereas these are larger for misspecified items. The mean values of χ^2s are similar for correctly specified items across true and misspecified Q–matrices. Consistent with the estimation errors, having one or more misspecified items did not impact the fit of the correctly specified items.

The higher order DINA model estimated two more parameters. Thus, χ^2 values are expected to be larger for the independence DINA model. 85% of the χ^2 values from the independence DINA model are larger than those from the higher order DINA model. Even though they were larger in value, the differences were in general small. Failing to include a higher order relationship among the attributes did not seem to be reflected in the χ^2 values that were calculated based on the latent classes.

Table 4.10
Mean Estimates of the χ^2_{lt} for 20 Items for the DINA Model

	$\overline{\chi^2_1}$ (SD)	$\overline{\chi^2_2}$ (SD)	$\overline{\chi^2_3}$ (SD)	$\overline{\chi^2_4}$ (SD)	$\overline{\chi^2_5}$ (SD)	$\overline{\chi^2_6}$ (SD)	$\overline{\chi^2_7}$ (SD)	$\overline{\chi^2_8}$ (SD)	$\overline{\chi^2_9}$ (SD)	$\overline{\chi^2_{10}}$ (SD)
DINA+Q_0	40.7	34.0	35.6	33.5	33.7	31.9	38.9	33.3	35.2	41.2
	(4.6)	(3.9)	(3.6)	(3.0)	(4.2)	(3.0)	(4.6)	(3.4)	(3.6)	(4.6)
DINA+Q_{O1}	40.7	33.9	35.0	32.8	34.2	32.2	38.6	33.0	36.5	41.6
	(4.5)	(3.9)	(3.6)	(2.8)	(4.1)	(3.1)	(4.5)	(3.3)	(3.8)	(4.8)
DINA+Q_{O5}	**103.0**	33.5	34.3	**93.3**	33.3	35.7	35.9	32.6	38.1	40.5
	(8.8)	(2.9)	(2.8)	**(7.5)**	(3.8)	(3.3)	(4.0)	(3.0)	(3.8)	(4.1)
DINA+Q_{U1}	39.6	33.5	33.7	33.6	33.0	31.6	37.5	32.8	35.8	41.5
	(4.2)	(3.8)	(2.9)	(2.6)	(3.8)	(2.7)	(4.4)	(3.3)	(3.6)	(4.7)
DINA+Q_{U5}	39.6	32.8	39.5	34.2	32.7	31.9	35.2	34.8	34.2	39.5
	(3.7)	(2.8)	(3.1)	(2.7)	(3.4)	(2.2)	(3.7)	(3.3)	(2.9)	(4.1)
DINA+Q_{B2}	43.5	34.7	35.7	**92.9**	33.6	32.0	38.4	32.5	37.1	36.8
	(5.0)	(3.8)	(3.1)	**(9.0)**	(3.7)	(3.0)	(4.2)	(3.1)	(3.7)	(3.6)
DINA+Q_{B6}	39.9	34.4	41.5	35.5	32.5	**84.2**	37.7	**73.4**	36.7	39.2
	(3.7)	(3.3)	(3.4)	(2.7)	(2.8)	**(6.4)**	(3.8)	**(5.8)**	(3.3)	(3.3)

	$\overline{\chi^2_{11}}$ (SD)	$\overline{\chi^2_{12}}$ (SD)	$\overline{\chi^2_{13}}$ (SD)	$\overline{\chi^2_{14}}$ (SD)	$\overline{\chi^2_{15}}$ (SD)	$\overline{\chi^2_{16}}$ (SD)	$\overline{\chi^2_{17}}$ (SD)	$\overline{\chi^2_{18}}$ (SD)	$\overline{\chi^2_{19}}$ (SD)	$\overline{\chi^2_{20}}$ (SD)
DINA+Q_0	31.7	32.3	34.0	32.4	34.3	34.9	35.9	31.0	30.8	33.4
	(2.6)	(3.4)	(3.6)	(2.5)	(3.1)	(3.6)	(4.2)	(2.2)	(2.8)	(2.6)
DINA+Q_{O1}	31.5	32.2	33.8	32.1	34.0	35.2	36.3	31.2	30.6	**67.0**
	(2.6)	(3.3)	(3.4)	(2.5)	(3.0)	(3.8)	(4.2)	(2.4)	(2.6)	**(7.9)**
DINA+Q_{O5}	31.1	**76.3**	34.0	33.8	33.2	34.9	35.7	32.5	31.4	**79.5**
	(1.9)	**(6.7)**	(3.0)	(3.0)	(2.3)	(3.4)	(3.8)	(2.4)	(2.7)	**(8.4)**
DINA+Q_{U1}	31.5	32.8	33.7	32.3	33.5	34.6	36.3	**99.5**	30.9	33.9
	(2.4)	(3.5)	(3.2)	(2.3)	(2.9)	(3.7)	(4.1)	**(9.0)**	(2.7)	(2.6)
DINA+Q_{U5}	**73.1**	**81.6**	32.7	32.5	**122.4**	35.3	37.2	**94.0**	**103.3**	32.8
	(7.3)	**(7.8)**	(2.4)	(2.0)	**(10.5)**	(3.6)	(3.7)	**(9.2)**	**(8.6)**	(2.3)
DINA+Q_{B2}	31.9	33.2	35.0	**111.7**	33.3	34.1	35.0	32.6	32.0	33.7
	(2.6)	(3.4)	(3.3)	**(10.0)**	(2.9)	(3.5)	(3.7)	(2.3)	(2.8)	(2.4)
DINA+Q_{B6}	34.3	32.1	34.9	32.3	**127.4**	**97.2**	34.3	32.8	**110.2**	**87.9**
	(2.5)	(2.5)	(2.8)	(2.0)	**(11.6)**	**(7.6)**	(3.0)	(2.1)	**(9.3)**	**(6.7)**

Table entries are italized to indicate the misspecified items.

4.2.1 The Type I Error Rate

Type I error rates of the indices are investigated at the .05 significance level. Table 4.11 presents the proportion of χ^2_{lt} indices greater than $p = .05$. Type I errors occur when an item is identified as misfitting but was simulated according to the true model and true Q–matrix, that is the higher order DINA model and Q_0. Thus, the overall empirical Type I error rate is calculated as the percent of items flagged as misfitting out of 121 correctly specified items across seven conditions which fit the higher order DINA model to the simulated data. 20, 19, 16, 19, 15, 18, and 14 items were correctly specified by Q_0, Q_{O1}, Q_{O5}, Q_{U1}, Q_{U5}, Q_{B2}, and Q_{B6} matrices, respectively. 50 replications for each resulted in an overall count of 6050. Only 51 out of 6050 of correctly specified items were detected as misfitting. This resulted in an empirical Type I error rate of .008 for χ^2_{lt}.

To further investigate the impact of having a misspecified item or items in a test, this study investigated Type I error rate for all seven conditions separately. The Type I error rate is small across the seven conditions. Having misspecified items in the test did not increase the Type I error rate calculated from the correctly specified items. In other words, correctly specified items were not detected as not fitting when there were misspecified items in the test.

The Type I error rate will have an impact on the power of the fit indices across the seven conditions. Inflated Type I error rates result in overestimated power and deflated Type I error rates result in underestimated power. The empirical Type I error rate should be close to .05 because the significance level is set at .05. Bradley's (1978) criteria suggest a range of .025 to .075 for a nominal level of .05. Type I error rates presented in Table 4.11 are smaller than the minimum value of this range, which is .025. All conditions yielded deflated Type I error rates for χ^2_{lt}.

Table 4.11
Proportion of χ^2_{lt} Indices Greater Than $p = .05$

	Type I Error Rates
	.05
Higher order DINA + Q_0	.011
Higher order DINA + Q_{O1}	.012
Higher order DINA + Q_{O5}	.003
Higher order DINA + Q_{U1}	.008
Higher order DINA + Q_{U5}	.011
Higher order DINA + Q_{B2}	.009
Higher order DINA + Q_{B6}	.004
Overall across seven conditions	.008

4.2.2 Power

Two types of problems with modeling are expected to create misfit; one is the failure to model higher-order structure among the attributes, and the other is the misspecification of the Q–matrix. These two conditions were used to examine the power of the fit indices. Power is calculated as the proportion of misspecified items that are detected as problematic by χ^2_{lt}. These percentages of correct detection of misfit are presented in Table 4.12. In Table 4.12 only the independence DINA conditions are listed. Power for the higher order DINA conditions was equal to 1 across all misspecification conditions. For the higher order DINA model, only Q–matrix misspecifications were expected to cause misfit, not the modeling of the higher order structure. Thus, the power is calculated for two overspecification, two underspecification, and two balanced misfit conditions and also for 1, 4, 1, 5, 2, and 6 misspecified items, respectively. χ^2_{lt} successfully flagged all these items as problematic over 50 simulations for each condition with the higher order DINA model, which resulted in the perfect power of 1.

It was expected that the independence DINA and Q–matrix misspecification combination would yield greater detection of misfit for all items since the independence DINA model

does not reflect the higher order structure among the attributes, which was used to generate the simulated data. However, that was not the case. Among the items correctly specified by the Q–matrix, only 3.3% of the items across all seven conditions were flagged by the χ^2_{lt} as problematic. Across the seven conditions that fit the independence DINA model to the simulated data, the proportion of misfit detection ranged from 2.6% to 4.3% for the items with correct Q–matrix specification. When all items are considered as not fitting, this proportion ranged from 3.2% to 37.7%. The highest misfit detection percentage is observed for the 5% underspecified Q–matrix, Q_{U5}, and the lowest misfit detection percentage is observed for the true Q–matrix. Similar to the misfit detection when utilizing the higher order DINA model, misspecified items were always detected by χ^2_{lt} with the independence DINA model. χ^2_{lt} detected the misspecified items by the Q–matrix either with the higher order DINA model or with the independence DINA model. Other than the Q–matrix misspecification, only 3% of the items were detected as misfitting for the false higher order structure. Thus, it can be claimed that χ^2_{lt} can not detect higher order structure problems.

Table 4.12
Proportion of χ^2_{lt} Indices with $p < .05$ for Misspecified Items

	Power		
	Misspecified items	Correctly specified items	All items
Independence DINA + Q_0		.032	.032
Independence DINA + Q_{O1}	1	.039	.094
Independence DINA + Q_{O5}	1	.026	.154
Independence DINA + Q_{U1}	1	.029	.083
Independence DINA + Q_{U5}	1	.033	.377
Independence DINA + Q_{B2}	1	.043	.159
Independence DINA + Q_{B6}	1	.032	.327
Overall across 7 conditions	1	.033	.164

4.3 Item Fit Evaluation with PPP-values Calculated Based on χ^2_{lt}

Posterior predictive p values are calculated utilizing χ^2_{lt} as the discrepancy measure. PPP values do not require any χ^2 approximation and provide Bayesian p-values. First, the distribution of PPP values which are based on $\chi^2 lt$ are investigated. Their means and ranges over 50 replications are presented in Table 4.13 for 20 items across the conditions that fit the higher order DINA model to the simulated data with one true and six misspecified Q–matrices. Extreme PPP values indicate item misfit. Values less than .05 and greater than .95 are considered extreme values and indicate item misfit.

Average PPP values over 50 replications for each condition ranged from .44 to .66 for the items whose attributes were specified correctly by the Q–matrix. For these items .25 is the minimum PPP-value, and the maximum value is .92. Neither minimum or maximum values indicated item misfit. Thus, none of the correctly specified items was flagged as not fitting by the PPP-values calculated based on χ^2_{lt}. PPP-values were calculated as the proportion of cases in which χ^2_{lt} from observed data and posterior distribution of variables in the model exceeded χ^2_{lt} from the replicated data and the posterior predictive distribution of y^{rep}. The range of the average PPP values for correctly specified items indicate that χ^2_{lt} calculated from the posterior distribution of variables in the model was close to χ^2_{lt} calculated from the posterior predictive distribution of replicated data for these items.

Average PPP-values for the items which are misspecified by the Q–matrix are all equal to 1. The minimum values ranged from .97 to 1, and the maximum value was 1. All misspecified items are flagged as not fitting by the PPP-values calculated based on χ^2_{lt} from all six conditions. For all the misspecified items, χ^2_{lt} from posterior distribution exceeded χ^2_{lt} from the posterior predictive distribution of replicated data the majority of the time, which resulted in the PPP-values in the upper end of the [0,1] interval. Therefore, the PPMC method demonstrates that the higher order DINA model with false Q–matrices did not explain the distribution of correct response proportions across latent classes for the items that are specified with a false Q–matrix.

Table 4.13

Means and Ranges of PPP Values Based on χ^2_{lt} for 20 Items Across Seven Higher Order DINA Conditions

	PPP_1 [range]	PPP_2 [range]	PPP_3 [range]	PPP_4 [range]	PPP_5 [range]	PPP_6 [range]	PPP_7 [range]	PPP_8 [range]	PPP_9 [range]	PPP_{10} [range]
Q_0	.57	.56	.50	.54	.52	.44	.64	.52	.48	.56
	[.35 - .84]	[.35 - .83]	[.31 - .72]	[.32 - .78]	[.31 - .77]	[.26 - .72]	[.37 - .92]	[.28 - .77]	[.26 - .68]	[.35 - .81]
Q_{O1}	.58	.55	.49	.51	.52	.45	.63	.51	.51	.55
	[.36-.85]	[.35-.82]	[.32-.69]	[.29-.72]	[.31-.76]	[.29-.73]	[.35-.9]	[.29-.79]	[.28-.75]	[.38-.83]
Q_{O5}	**1**	.53	.49	**1**	.50	.50	.56	.50	.58	.57
	[1-1]	[.36-.74]	[.37-.66]	**[1-1]**	[.33-.71]	[.37-.68]	[.32-.87]	[.27-.74]	[.41-.83]	[.42-.73]
Q_{U1}	.54	.54	.47	.52	.49	.44	.60	.50	.47	.56
	[.33-.77]	[.35-.82]	[.30-.67]	[.31-.72]	[.30-.76]	[.28-.66]	[.31-.91]	[.26-.74]	[.31-.66]	[.35-.84]
Q_{U5}	.59	.66	.5	.49	.52	.44	.63	.52	.47	.54
	[.43-.81]	[.43-.86]	[.39-.61]	[.32-.71]	[.34-.77]	[.34-.7]	[.48-.81]	[.36-.74]	[.34-.6]	[.43-.72]
Q_{B2}	.55	.58	.5	1	.55	.44	.65	.47	.5	.52
	[.36-.77]	[.37-.86]	[.35-.67]	[1-1]	[.33-.78]	[.28-.68]	[.42-.92]	[.27-.71]	[.29-.69]	[.37-.68]
Q_{B6}	.57	.59	.55	.57	.54	**1**	.58	**1**	.5	.59
	[.43-.75]	[.43-.8]	[.45-.68]	[.4-.74]	[.34-.78]	**[.99-1]**	[.39-.76]	**[.99-1]**	[.31-.7]	[.42-.72]

	PPP_{11} [Range]	PPP_{12} [Range]	PPP_{13} [Range]	PPP_{14} [Range]	PPP_{15} [Range]	PPP_{16} [Range]	PPP_{17} [Range]	PPP_{18} [Range]	PPP_{19} [Range]	PPP_{20} [Range]
Q_0	.51	.53	.52	.5	.5	.49	.55	.45	.48	.5
	[.29-.73]	[.3-.88]	[.29-.77]	[.36-.74]	[.34-.68]	[.32-.73]	[.31-.77]	[.25-.66]	[.31-.73]	[.27-.74]
Q_{O1}	.51	.52	.51	.49	.5	.5	.56	.46	.48	**.99**
	[.32-.73]	[.28-.85]	[.28-.74]	[.3-.72]	[.35-.69]	[.32-.74]	[.33-.76]	[.26-.66]	[.31-.72]	**[.89-1]**
Q_{O5}	.49	**1**	.5	.5	.49	.5	.53	.52	.5	**1**
	[.33-.67]	**[.98-1]**	[.35-.71]	[.38-.71]	[.35-.66]	[.31-.71]	[.32-.7]	[.37-.75]	[.34-.77]	**[.99-1]**
Q_{U1}	.49	.54	.52	.5	.49	.48	.56	**1**	.48	.5
	[.33-.68]	[.33-.85]	[.28-.83]	[.38-.72]	[.36-.69]	[.28-.76]	[.34-.79]	**[1-1]**	[.29-.71]	[.26-.75]
Q_{U5}	**1**	**1**	.48	.56	**1**	.54	.64	**1**	**1**	.47
	[.99-1]	**[.97-1]**	[.33-.65]	[.43-.7]	**[.99-1]**	[.39-.78]	[.44-.84]	**[.99-1]**	**[.99-1]**	[.32-.69]
Q_{B2}	.49	.53	.54	**1**	.48	.47	.52	.49	.55	.51
	[.3-.74]	[.31-.83]	[.36-.77]	**[1-1]**	[.3-.65]	[.3-.67]	[.35-.79]	[.33-.66]	[.33-.81]	[.28-.76]
Q_{B6}	.62	.51	.5	.57	**1**	**1**	.51	.46	**1**	**1**
	[.38-.84]	[.38-.79]	[.38-.69]	[.44-.74]	**[1-1]**	**[1-1]**	[.37-.7]	[.33-.65]	**[1-1]**	**[.99-1]**

Next, the distribution of the PPP values calculated based on the χ^2_{lt} are presented. Means and ranges of these PPP values are presented in the Table 4.14 for 20 items across the conditions that fit the independence DINA model to the simulated data with one true and six misspecified Q–matrices. Average PPP-values ranged from .45 to .81 over 50 replications for the items that utilized the correct attribute relation specifications by the true Q–matrix. For these items .27 is the minimum PPP-value and the maximum value is .98. Thus, for some items and replications, maximum value indicated item misfit. Therefore, when the independence DINA model is fit to the simulated data, some of the items that utilized correct attribute relation specifications are flagged as not fitting by the PPP-values calculated based on χ^2_{lt}. Range of the average PPP values indicate that the χ^2_{lt} calculated from the posterior distribution of variables in the model was close to the χ^2_{lt} calculated from the posterior predictive distribution of replicated data for the correctly specified items for the correctly specified items. These averages are greater for the independence DINA model than the higher order DINA model for majority of the items but didn't exceed the extreme value of .95.

Similar to the results from the higher order DINA model, average PPP-values for the misspecified items by the Q–matrix are all equal to 1. The minimum and maximum values are .97 and 1 respectively. All misspecified items are detected by the PPP-values calculated based on the χ^2_{lt} when the independence DINA model is fit to the data. Again similat to the results from higher order DINA model for all the misspecified items χ^2_{lt} from posterior distribution exceeded the χ^2_{lt} from the posterior predictive distribution of replicated data for majority of the time and produced the PPP-values in the upper end of the [0,1] interval. Therefore, the PPMC method demonstrates that independence DINA model with false Q–matrices didn't explain the distribution of correct response proportions across latent classes for the items that are misspecified with a false Q–matrix. Furthermore, even though the PPP-values from the independence DINA model are higher than the higher order DINA model, they are not significant and so model misspecification could not be detected by the PPP-values calculated for each item.

Table 4.14

Means and Ranges of PPP Values Based on χ^2_{lt} for 20 Items Across Seven Independence DINA Conditions

	PPP_1 [range]	PPP_2 [range]	PPP_3 [range]	PPP_4 [range]	PPP_5 [range]	PPP_6 [range]	PPP_7 [range]	PPP_8 [range]	PPP_9 [range]	PPP_{10} [range]
Q_O	.75	.57	.62	.56	.56	.5	.71	.55	.61	.76
	[.53 - .93]	[.36 - .85]	[.39 - .81]	[.35 - .81]	[.34 - .87]	[.3 - .72]	[.37 - .97]	[.29 - .82]	[.41 - .83]	[.5 - .97]
Q_{O1}	.75	.57	.6	.53	.57	.51	.71	.54	.65	.77
	[.55 - .93]	[.35 - .84]	[.37 - .82]	[.36 - .76]	[.34 - .89]	[.31 - .75]	[.38 - .96]	[.33 - .82]	[.44 - .87]	[.53 - .98]
Q_{O5}	**1**	.55	.57	**1**	.54	.62	.63	.52	.69	.75
	[1-1]	[.37 - .75]	[.4 - .76]	**[1-1]**	[.33 - .86]	[.43 - .92]	[.35 - .94]	[.33 - .77]	[.44 - .89]	[.52 - .94]
Q_{U1}	.73	.55	.56	.56	.53	.49	.68	.53	.63	.77
	[.51 - .91]	[.38 - .84]	[.36 - .74]	[.37 - .76]	[.31 - .83]	[.29 - .7]	[.36 - .96]	[.27 - .76]	[.41 - .88]	[.51 - .97]
Q_{U5}	.73	.53	.71	.58	.52	.5	.6	.6	.58	.72
	[.55 - .9]	[.35 - .82]	[.55 - .87]	[.38 - .77]	[.31 - .8]	[.33 - .72]	[.33 - .88]	[.37 - .88]	[.36 - .82]	[.5 - .94]
Q_{B2}	.81	.59	.62	**1**	.55	.5	.7	.52	.67	.65
	[.57 - .95]	[.38 - .86]	[.41 - .79]	**[1-1]**	[.35 - .83]	[.31 - .7]	[.44 - .96]	[.28 - .79]	[.45 - .87]	[.43 - .89]
Q_{B6}	.73	.58	.77	.62	.51	**1**	.68	**1**	.65	.72
	[.51 - .94]	[.39 - .78]	[.55 - .91]	[.4 - .76]	[.32 - .71]	**[.99 - 1]**	[.45 - .93]	**[.99 - 1]**	[.42 - .88]	[.55 - .91]

	PPP_{11} [Range]	PPP_{12} [Range]	PPP_{13} [Range]	PPP_{14} [Range]	PPP_{15} [Range]	PPP_{16} [Range]	PPP_{17} [Range]	PPP_{18} [Range]	PPP_{19} [Range]	PPP_{20} [Range]
Q_O	.49	.51	.57	.52	.57	.6	.63	.46	.46	.55
	[.29 - .71]	[.28 - .84]	[.31 - .78]	[.38 - .79]	[.36 - .76]	[.32 - .8]	[.38 - .91]	[.31 - .66]	[.27 - .69]	[.32 - .79]
Q_{O1}	.48	.51	.56	.5	.56	.61	.64	.47	.45	**.99**
	[.28 - .72]	[.28 - .81]	[.31 - .78]	[.35 - .78]	[.33 - .78]	[.34 - .81]	[.34 - .91]	[.33 - .69]	[.29 - .71]	**[.89 - 1]**
Q_{O5}	.47	**1**	.57	.56	.54	.6	.62	.52	.48	**1**
	[.34 - .65]	**[.99 - 1]**	[.42 - .77]	[.42 - .83]	[.37 - .68]	[.39 - .81]	[.4 - .86]	[.37 - .74]	[.34 - .73]	**[.99 - 1]**
Q_{U1}	.48	.53	.56	.51	.55	.59	.64	**1**	.46	.57
	[.31 - .68]	[.29 - .83]	[.33 - .8]	[.37 - .77]	[.35 - .75]	[.34 - .8]	[.38 - .9]	**[1-1]**	[.29 - .67]	[.33 - .78]
Q_{U5}	**1**	**1**	.52	.52	**1**	.61	.67	**1**	**1**	.53
	[.99 - 1]	**[1 - 1]**	[.36 - .7]	[.41 - .74]	**[1 - 1]**	[.4 - .82]	[.38 - .88]	**[1 - 1]**	**[1 - 1]**	[.35 - .73]
Q_{B2}	.5	.54	.6	**1**	.54	.57	.6	.52	.5	.56
	[.28 - .67]	[.3 - .85]	[.41 - .83]	**[1 - 1]**	[.35 - .74]	[.38 - .79]	[.39 - .9]	[.39 - .73]	[.34 - .74]	[.3 - .82]
Q_{B6}	.58	.5	.59	.51	**1**	**1**	.58	.53	**1**	**1**
	[.31 - .71]	[.35 - .75]	[.41 - .81]	[.39 - .77]	**[1 - 1]**	**[1 - 1]**	[.32 - .8]	[.37 - .69]	**[1 - 1]**	**[1 - 1]**

4.3.1 The Type I Error Rate

Type I error rates of the indices are investigated at .05 significance level. Conditions in which the higher order DINA model is utilized are evaluated for Type I error investigation. Type I error rate is calculated as the proportion of correctly specified items which are rejected by PPP-values based on χ^2_{lt}. The extreme values of the PPP reject the items and these values are the ones that are smaller than $p = .05$ or greater than $p = .95$.

Total of 121 items across these seven conditions are included in the Type I error rate investigation. 50 replications for each resulted in an overall count of 6050. None of these correctly specified items are flagged as being problematic by the PPP-values calculated based on χ^2_{lt} across the conditions that utilized the generating model and generating attribute specifications for items. This resulted in empirical Type I error rate of 0 for PPP-values based on χ^2_{lt}.

Having misspecified items in the test didn't increase the Type I error rate calculated from the correctly specified items. In other words, correctly specified items were not detected as problematic when there are misspecified items in the test. Type I error rate of 0 is smaller than the minimum value of suggested range of .025 to .075. Therefore, all conditions yielded deflated Type I error rates for PPP-values based on χ^2_{lt}. Thus deflated Type I error rates is expected to result in underestimated power.

Even though they are both small and considerably close, Type I error is committed less with the PPP-values based on χ^2_{lt} than χ^2_{lt} itself as a fit index. Type I error rate for χ^2_{lt} was .08% and it was 0 for PPP-values. There has been a small number of correctly specified items that are rejected by χ^2_{lt} which are not rejected by PPP-values based on χ^2_{lt}. Thus, both the χ^2_{lt} and PPP-values based on χ^2_{lt} are conservative tests because their Type I error rates are both smaller than the suggested range of .025 to .075.

Table 4.15
Proportion of PPP-values based on χ^2_{lt} Smaller Than $p = .05$ or Greater Than $p = .95$

	Type I Error Rates
Higher order DINA + Q_0	.00
Higher order DINA + Q_O1	.00
Higher order DINA + Q_O5	.00
Higher order DINA + Q_U1	.00
Higher order DINA + Q_U5	.00
Higher order DINA + Q_B2	.00
Higher order DINA + Q_B6	.00
Overall across 7 conditions	.00

4.3.2 Power

Failing to model the higher order structure and misspecification of items were used to examine the power of the PPP-values calculated based on χ^2_{lt}. For the conditions in which the higher order DINA model is utilized, power is calculated over the items misspecified by the Q–matrix. Power is calculated as the proportion of misspecified items that PPP-values based on χ^2_{lt} flagged as problematic. Almost all the misspecified items are successfully rejected by the PPP-values based on χ^2_{lt}. Only 2 out of 50 replications with the higher order DINA model and Q_{O1} matrix failed to reject the misspecified item.

For the conditions that utilized the independence DINA model, power is investigated separately for misspecified items only, for correctly specified items only, and for all items. All items are expected to be problematic because the independence DINA model failed to include the higher order structure. Proportions of misfit detection are presented in Table 4.16. 99.8% of the misspecified items are rejected by the PPP-value calculated based on χ^2_{lt}. Only .01 % of the correctly specified items are rejected by the PPP-value. The proportion of all the items that are rejected is 13.7%. 13.7% was obtained by mainly the contribution of

the misspecified items. Because the contribution of correctly specified items to the power was .01%. Thus, it was observed that the PPP-value based on χ^2_{lt} failed to reject the correctly specified items even though there was a lack of modeling of the higher order structure.

Power of the PPP-value based on χ^2_{lt} was similar to the power of the χ^2_{lt}. χ^2_{lt} detected all and PPP-values based on χ^2_{lt} detected 99.8% of the misspecified items. On the other hand, χ^2_{lt} detected only 3.3 % and PPP-value based on χ^2_{lt} detected only .01 % of the correctly specified items. On average, proportion of misfit detection over all the items was 16.4% for χ^2_{lt} and it was 13.7% for PPP-values based on χ^2_{lt}.

Table 4.16
Proportion of PPP-values Smaller than .05 or Greater than .95 for Misfitting Items

	Power		
	Misspecified items	Correctly specified items	All items
Independence DINA + Q_0		.002	.002
Independence DINA + Q_{O1}	.96	.002	.053
Independence DINA + Q_{O5}	1	0	.125
Independence DINA + Q_{U1}	1	.002	.055
Independence DINA + Q_{U5}	1	0	.333
Independence DINA + Q_{B2}	1	.002	.114
Independence DINA + Q_{B6}	1	0	.286
Overall across 7 conditions	.998	.001	.137

4.4 Item Fit Evaluation with χ^2 Calculated Based on Raw Score Groups

χ^2 is calculated for each item based on raw score groups utilizing the higher order DINA model and one true specified Q–matrix and six misspecified Q–matrices. Table 4.17 presents the average χ^2 values that are calculated based on raw score groups over 50 replications for the seven conditions that utilized the higher order DINA model.

The χ^2 is expected to follow a χ^2 distribution with 17 *df*. The degrees of freedom is four less than the number of raw score groups when the statistical model is higher order DINA because the number of parameters estimated with the model is four. These four parameters

are attribute difficulty, attribute discrimination, item guessing, and item slipping parameters. Thus, χ^2_{17} is used as a reference distribution to check the item fit. Values larger than 27.59 are in the 95th or greater percentile of the χ^2_{17} distribution.

Items that are misspecified by the Q–matrix are boldfaced in the Table 4.17. The mean values of χ^2s are not always smaller than the reference value of 27.59 for the correctly specified items. Specifically, for items 6, 7, 8, 13, and 16 χ^2_{raw} values are 30.4, 28.2, 33.1, 28.8, and 30.1, respectively. These items have average χ^2_{raw} values larger than the reference value of 27.6. Average χ^2_{raw} values are larger than the reference value for all the misspecified items. These average values give an overall idea that χ^2s calculated based on raw score groups detect the misspecified items correctly. However, these values flag correctly specified items as misfitting. Standard deviations are given in parentheses below the mean values of χ^2s in the Table 4.17. These are smaller for correctly specified items and larger for misspecified items. The mean values of χ^2s are similar for correctly specified items across true and misspecified Q–matrices. Even though χ^2_{raw} flagged several correctly specified items as misfit, it still can be claimed that having one or more misspecified items didn't impact the fit of the correctly specified items. Because these correctly specified items that were flagged as misfit are flagged as misfit even when the true Q–matrix was utilized in the estimation.

Table 4.17
Mean Estimates of the χ^2_{raw} for 20 Items Utilizing the Higher Order DINA Model

	$\overline{\chi^2_1}$ (SD)	$\overline{\chi^2_2}$ (SD)	$\overline{\chi^2_3}$ (SD)	$\overline{\chi^2_4}$ (SD)	$\overline{\chi^2_5}$ (SD)	$\overline{\chi^2_6}$ (SD)	$\overline{\chi^2_7}$ (SD)	$\overline{\chi^2_8}$ (SD)	$\overline{\chi^2_9}$ (SD)	$\overline{\chi^2_{10}}$ (SD)
HoDINA+Q_0	19.4	23.2	19.7	20.1	22.9	30.4	28.2	33.1	23.5	22.1
	(3.0)	(4.0)	(3.7)	(4.5)	(4.6)	(5.0)	(4.9)	(5.1)	(3.9)	(3.7)
HoDINA+Q_{O1}	19.4	23.1	19.8	20.4	23.4	30.4	28.2	33.0	24.6	22.9
	(3.0)	(4.0)	(3.7)	(4.6)	(4.5)	(5.0)	(4.9)	(5.2)	(4.0)	(3.8)
HoDINA+Q_{O5}	**54.5**	23.4	20.0	**42.5**	24.0	34.7	28.2	33.5	26.8	22.5
	(7.8)	(4.1)	(3.7)	**(6.8)**	(4.6)	(5.4)	(4.8)	(5.0)	(3.9)	(3.8)
HoDINA+Q_{U1}	19.3	22.9	17.9	22.0	23.0	30.5	28.6	31.9	23.0	22.1
	(3.0)	(3.9)	(3.3)	(4.8)	(4.6)	(5.0)	(5.0)	(5.1)	(3.6)	(3.7)
HoDINA+Q_{U5}	18.9	22.2	21.7	21.7	22.8	27.8	25.4	33.1	23.0	21.2
	(3.3)	(4.0)	(3.3)	(4.7)	(4.5)	(4.7)	(5.0)	(4.7)	(3.4)	(4.0)
HoDINA+Q_{B2}	18.8	23.6	20.0	**40.5**	22.5	30.1	27.5	32.8	26.1	19.8
	(3.2)	(4.2)	(3.6)	**(7.1)**	(4.5)	(4.9)	(4.7)	(5.1)	(4.0)	(3.5)
HoDINA+Q_{B6}	20.2	21.2	18.7	20.1	22.1	**56.0**	25.0	**55.1**	24.8	20.7
	(3.3)	(3.7)	(4.0)	(4.2)	(4.5)	**(7.8)**	(4.4)	**(7.6)**	(3.5)	(4.2)

	$\overline{\chi^2_{11}}$ (SD)	$\overline{\chi^2_{12}}$ (SD)	$\overline{\chi^2_{13}}$ (SD)	$\overline{\chi^2_{14}}$ (SD)	$\overline{\chi^2_{15}}$ (SD)	$\overline{\chi^2_{16}}$ (SD)	$\overline{\chi^2_{17}}$ (SD)	$\overline{\chi^2_{18}}$ (SD)	$\overline{\chi^2_{19}}$ (SD)	$\overline{\chi^2_{20}}$ (SD)
HoDINA+Q_0	25.3	24.0	28.8	18.3	16.4	30.1	23.3	21.5	18.4	23.1
	(4.8)	(5.6)	(5.2)	(3.0)	(3.6)	(6.0)	(4.4)	(3.5)	(3.4)	(4.4)
HoDINA+Q_{O1}	25.3	24.2	28.2	18.2	16.0	30.7	23.8	21.6	18.1	**38.1**
	(4.8)	(5.6)	(5.1)	(3.0)	(3.5)	(6.0)	(4.3)	(3.6)	(3.5)	**(6.4)**
HoDINA+Q_{O5}	25.7	**47.1**	27.8	19.6	15.8	29.3	23.5	22.0	18.3	**55.0**
	(4.9)	**(7.7)**	(4.9)	(2.9)	(3.4)	(5.8)	(4.2)	(3.8)	(3.5)	**(9.2)**
HoDINA+Q_{U1}	25.4	24.4	29.4	18.5	16.4	29.6	23.4	**52.8**	18.1	24.1
	(4.8)	(5.7)	(5.2)	(3.0)	(3.6)	(5.8)	(4.4)	**(6.4)**	(3.4)	(4.6)
HoDINA+Q_{U5}	**46.3**	**52.7**	29.2	19.6	**55.1**	31.2	23.0	**49.5**	**59.8**	24.2
	(7.7)	**(7.3)**	(5.1)	(3.2)	**(7.3)**	(6.3)	(4.3)	**(6.8)**	**(8.4)**	(4.6)
HoDINA+Q_{B2}	25.5	24.7	29.5	**46.7**	16.3	29.3	22.3	22.5	18.3	23.5
	(4.6)	(5.5)	(5.1)	**(6.6)**	(3.6)	(5.8)	(4.3)	(3.5)	(3.4)	(4.6)
HoDINA+Q_{B6}	25.9	26.8	28.8	19.4	**50.8**	**71.0**	22.6	23.7	**53.8**	**56.3**
	(4.6)	(5.6)	(4.8)	(3.1)	**(7.6)**	**(10.9)**	(4.2)	(3.8)	**(7.7)**	**(5.3)**

Table entries in boldface indicate the misspecified items. df=17 for χ^2

Similar to what is observed with the higher order DINA model, the average χ^2 values flagged correctly specified items with the independence DINA model. Table 4.18 presents the average χ^2 values that are calculated based on raw score groups over 50 replications for the seven conditions that utilized the DINA model. Standard deviations are given in parentheses below the mean values of χ^2s in the Table 4.18.

The χ^2_{raw} follows a χ^2 distribution with 21 df. The degrees of freedom is two less than the number of raw score groups. Thus, χ^2_{19} is used as a reference distribution to check the item fit. Values larger than 30.14 are in the 95th or greater percentile of the χ^2_{19} distribution.

Items that are misspecified by the Q–matrix are boldfaced in the Table 4.18. 6th, 7th, 8th, 9th, 10th, 13th, and 16th items are flagged as misfitting by the average χ^2_{raw} statistic when true Q–matrix is utilized. The mean values of χ^2s are all larger than the reference value for all the misspecified items. Standard deviations are smaller for correctly specified items whereas these are larger for misspecified items. The mean values of χ^2s are similar for correctly specified items across true and misspecified Q–matrices. Consistent with the estimation errors, having one or more misspecified items didn't impact the fit of the correctly specified items.

Average χ^2 values are larger for the independence DINA model than the higher order DINA model which was expected since higher order DINA model estimates two more parameters. Even if the average χ^2 values are different items that are flagged as misfitting are similar across the independent and higher order DINA models. Thus failing to include a higher order relationship among the attributes was not reflected in the χ^2 values that are calculated based on raw score groups.

Table 4.18
Mean Estimates of the χ^2_{raw} for 20 Items Utilizing the DINA Model

	$\overline{\chi^2_1}$ (SD)	$\overline{\chi^2_2}$ (SD)	$\overline{\chi^2_3}$ (SD)	$\overline{\chi^2_4}$ (SD)	$\overline{\chi^2_5}$ (SD)	$\overline{\chi^2_6}$ (SD)	$\overline{\chi^2_7}$ (SD)	$\overline{\chi^2_8}$ (SD)	$\overline{\chi^2_9}$ (SD)	$\overline{\chi^2_{10}}$ (SD)
D+Q_0	27.9	28.3	26.3	26.1	28.0	35.7	36.2	37.7	31.0	33.8
	(4.3)	(4.7)	(4.2)	(5.6)	(5.2)	(5.6)	(6.2)	(5.5)	(5.2)	(5.3)
D+Q_{O1}	27.9	28.2	25.7	27.1	29.2	35.6	36.2	37.3	32.9	35.5
	(4.4)	(4.6)	(4.4)	(5.7)	(5.2)	(5.6)	(6.1)	(5.5)	(5.4)	(5.5)
D+Q_{O5}	**64.3**	27.2	25.4	**47.6**	29.8	45.4	34.6	38.0	37.3	33.9
	(9.1)	(4.5)	(4.2)	**(7.9)**	(5.3)	(6.8)	(6.0)	(5.4)	(5.5)	(5.4)
D+Q_{U1}	27.9	27.9	23.0	30.5	28.2	35.6	36.6	36.3	31.7	34.0
	(4.3)	(4.7)	(3.6)	(6.3)	(5.2)	(5.6)	(6.2)	(5.4)	(5.3)	(5.4)
D+Q_{U5}	25.6	25.2	25.2	31.4	26.7	32.1	30.8	48.2	32.1	30.6
	(4.3)	(4.2)	(5.6)	(6.3)	(4.9)	(5.2)	(5.2)	(6.1)	(5.2)	(4.9)
D+Q_{B2}	32.2	27.2	26.4	**44.3**	26.9	36.4	33.4	37.3	37.1	28.0
	(5.3)	(4.5)	(4.1)	**(7.7)**	(4.9)	(5.8)	(5.8)	(5.5)	(6.0)	(4.3)
D+Q_{B6}	31.5	26.6	31.3	29.3	25.4	**60.1**	35.1	**57.3**	36.5	28.4
	(4.7)	(4.7)	(6.2)	(6.0)	(4.8)	**(8.3)**	(5.9)	**(7.7)**	(5.5)	(4.6)

	$\overline{\chi^2_{11}}$ (SD)	$\overline{\chi^2_{12}}$ (SD)	$\overline{\chi^2_{13}}$ (SD)	$\overline{\chi^2_{14}}$ (SD)	$\overline{\chi^2_{15}}$ (SD)	$\overline{\chi^2_{16}}$ (SD)	$\overline{\chi^2_{17}}$ (SD)	$\overline{\chi^2_{18}}$ (SD)	$\overline{\chi^2_{19}}$ (SD)	$\overline{\chi^2_{20}}$ (SD)
D+Q_0	26.4	25.0	31.9	20.0	18.7	34.1	26.9	22.3	20.3	24.8
	(5.0)	(5.5)	(5.3)	(3.0)	(3.7)	(6.4)	(4.8)	(3.4)	(3.7)	(4.7)
D+Q_{O1}	26.4	25.0	31.4	20.0	18.1	35.2	27.7	22.5	19.6	**40.6**
	(5.0)	(5.5)	(5.2)	(3.1)	(3.6)	(6.5)	(4.8)	(3.4)	(3.7)	**(6.8)**
D+Q_{O5}	27.5	**52.4**	32.7	23.5	17.3	34.3	27.0	22.5	19.8	**60.7**
	(5.1)	**(8.1)**	(5.2)	(3.8)	(3.6)	(6.2)	(4.7)	(3.3)	(3.8)	**(10.1)**
D+Q_{U1}	26.5	25.1	32.6	20.2	18.5	34.0	27.1	**62.5**	19.8	26.3
	(5.0)	(5.6)	(5.2)	(3.1)	(3.7)	(6.3)	(4.8)	**(7.5)**	(3.6)	(5.0)
D+Q_{U5}	**53.6**	**63.2**	31.4	19.4	**100.4**	33.4	25.6	**76.6**	**79.9**	26.0
	(8.7)	**(7.7)**	(5.0)	(3.0)	**(10.0)**	(6.0)	(4.4)	**(9.3)**	**(9.2)**	(4.8)
D+Q_{B2}	27.0	26.3	33.8	**65.2**	18.3	33.6	25.3	24.4	19.5	25.1
	(4.8)	(5.4)	(5.3)	**(8.5)**	(3.8)	(6.4)	(4.5)	(3.5)	(3.6)	(4.8)
D+Q_{B6}	26.8	28.9	35.4	19.2	**63.7**	**96.0**	26.3	27.2	**75.2**	**71.5**
	(4.8)	(5.6)	(5.3)	(3.0)	**(7.7)**	**(10.4)**	(4.5)	(3.9)	**(8.9)**	**(6.8)**

Boldfaced table entries indicate the misspecified items. df=21 for χ^2

4.4.1 The Type I Error Rate

In this section Type I error rates for the χ^2_{raw} indices are investigated at .05 significance level. Table 4.19 presents the proportion of χ^2_{raw} indices greater than $p = .05$. The Type I error investigation is based on the items that are simulated according to the true model and true Q–matrix, that is the higher order DINA model and Q_0. Thus, the values in the Table 4.19 is calculated as the percent of items flagged as misfitting out of 121 correctly specified items across 7 conditions and 50 replications for each condition which fit the higher order DINA model to the simulated data. 20, 19, 16, 19, 15, 18, and 14 are the number of items that were correctly specified by Q_0, Q_{O1}, Q_{O5}, Q_{U1}, Q_{U5}, Q_{B2}, and Q_{B6} matrices respectively. 50 replications for each resulted in an overall count of 6050. 1474 out of 6050 correctly specified items were detected as not fitting. This resulted in empirical Type I error rate of 0.244 for χ^2_{raw}.

To further investigate the impact of having misspecified item or items in a test, Type I error rate is investigated for all seven conditions separately. Type I error rate is high across the seven conditions and ranged from .161 to .303. Thus, χ^2_{raw} did not produce an acceptable Type I error rate by flagging too many correctly specified items as not fitting. The overall proportion of correctly specified items classified as not fitting is 24.4%. Similar to the results from χ^2_{lt}, χ^2_{raw} didn't produce higher Type I error rates for the correctly specified items when there were misspecified items in the simulated 20 item test. Type I error rate didn't seem to be impacted by the type of the Q–matrix or the percentage of the misspecified items. It had its smallest value when 6% of the items were misspecified. Thus it can be claimed that the percentage of misspecification didnot increase the Type I error rate of χ^2_{raw}. Items 6, 7, 8, 13, and 16 are consistently flagged as not fitting more than half of the time across seven conditions. 1, 3, 14, 15, and 19th items are consistently detected as fitting. Thus, there is an inconsistency across the items in terms of producing the Type I error. Table C.1 presents the proportions of correctly specified items that are rejected.

Type I error rate will have impact on the power of the fit indices across the seven conditions. Inflated Type I error rates result in overestimated power and deflated Type I error rates result in underestimated power. The empirical Type I error rate should be close to .05 because the significance level is set .05. Bradley's (1978) criterion suggests a range of .025 to .075 for a nominal level of .05. Type I error rates presented in the Table 4.19 are all greater than the maximum value of this range which is .075. All conditions yielded inflated Type I error rates for χ^2_{raw}.

Table 4.19
Proportion of χ^2_{raw} Indices Greater Than $p = .05$

	Type I Error Rates
	.05
Higher order DINA + Q_0	.230
Higher order DINA + Q_{O1}	.242
Higher order DINA + Q_{O5}	.303
Higher order DINA + Q_{U1}	.244
Higher order DINA + Q_{U5}	.261
Higher order DINA + Q_{B2}	.257
Higher order DINA + Q_{B6}	.161
Overall across 7 conditions	.244

4.5 Item Fit Evaluation with PPP-values Calculated Based on χ^2_{raw}

In this section, results from the posterior predictive p values which were calculated utilizing the χ^2_{raw} as the discrepancy measure are presented. The distribution of PPP values which are based on χ^2_{raw} are summarized in terms of their means and ranges over 50 replications. These are presented in the Table 4.20 for 20 items across seven conditions that utilized the higher order DINA model. In the Table 4.20 misspecified items are boldfaced.

Average PPP values based on χ^2_{raw} over 50 replications across seven conditions ranged from .58 to .96 for the items that their attributes are specified correctly by the Q–matrix. Minimum and maximum values of these PPP values are observed as .19 and 1. PPP values

larger than .95 are extreme values to indicate misfit. Thus, some of the correctly specified items are flagged as not fitting by these PPP-values. Average PPP-values based on χ^2_{raw} are larger than PPP-values based on χ^2_{lt}. While the average PPP values based on χ^2_{lt} were closer to the middle points of the [0,1] interval, average PPP-values based on χ^2_{raw} are in the upper half of the [0,1] interval. The χ^2_{raw} calculated from the posterior distribution of variables in the model was larger than the χ^2_{raw} calculated from the posterior predictive distribution of replicated data for the majority of the MCMC iterations. Thus the discrepancy between the observed and model predicted proportion of correct response is greater than the discrepancy between the replicated and model predicted proportion of correct response more frequently.

Some items were identified as misfitting more often than the rest. Average PPP-values for the 6th, 7th, 8th, 13th, and 16th items are greater than .90. PPP-values based on χ^2_{raw} didnot function consistently across items by producing higher values of average PPP-values for these items. 14th and 19th items were not flagged as misfitting items which is consistent with reality.

Average PPP-values for the items which are misspecified by the Q–matrix ranged from .98 to 1. The minimum PPP-value based on χ^2_{raw} was .86 and the maximum value was 1. Thus, some misspecified items were not flagged as not fitting by the PPP-values calculated based on the χ^2_{raw}.

Table 4.20

Averages and Ranges of PPP Values Based on χ^2_{raw} for 20 Items Across Seven Higher Order DINA Conditions

	PPP_1 [range]	PPP_2 [range]	PPP_3 [range]	PPP_4 [range]	PPP_5 [range]	PPP_6 [range]	PPP_7 [range]	PPP_8 [range]	PPP_9 [range]	PPP_{10} [range]
Q_0	.78	.83	.79	.69	.77	.94	.91	.96	.84	.83
	[.46 - .96]	[.49 - 1]	[.28 - .98]	[.23 - .98]	[.34 - .99]	[.69 - 1]	[.66 - 1]	[.79 - 1]	[.52 - .99]	[.61 - .97]
Q_{O1}	.78	.82	.79	.71	.79	.94	.91	.96	.86	.85
	[.48 - .96]	[.48 - 1]	[.27 - .98]	[.26 - .99]	[.37 - .98]	[.72 - 1]	[.63 - 1]	[.74 - 1]	[.58 - .99]	[.61 - .98]
Q_{O5}	**1**	.83	.78	**.99**	.81	.97	.91	.96	.9	.85
	[.99 - 1]	[.51 - 1]	[.31 - .98]	**[.92 - 1]**	[.38 - .99]	[.84 - 1]	[.66 - 1]	[.76 - 1]	[.71 - .99]	[.57 - .99]
Q_{U1}	.78	.82	.71	.76	.78	.94	.92	.95	.83	.83
	[.44 - .96]	[.48 - 1]	[.21 - .95]	[.32 - .99]	[.35 - .99]	[.7 - 1]	[.68 - 1]	[.71 - 1]	[.49 - .98]	[.62 - .98]
Q_{U5}	.72	.75	.82	.75	.76	.9	.83	.96	.83	.78
	[.35 - .95]	[.35 - .99]	[.47 - .96]	[.34 - .99]	[.33 - .99]	[.63 - 1]	[.31 - .99]	[.79 - 1]	[.51 - .98]	[.46 - .97]
Q_{B2}	.75	.83	.79	**.99**	.74	.94	.9	.96	.9	.73
	[.47 - .94]	[.54 - 1]	[.28 - .98]	**[.89 - 1]**	[.33 - .99]	[.74 - 1]	[.64 - .99]	[.79 - 1]	[.75 - .99]	[.44 - .96]
Q_{B6}	.8	.74	.81	.7	.72	**1**	.84	**1**	.87	.72
	[.53 - .97]	[.38 - .98]	[.56 - .98]	[.36 - .98]	[.33 - .99]	**[1 - 1]**	[.52 - .99]	**[.99 - 1]**	[.57 - .98]	[.41 - .96]

	PPP_{11} [Range]	PPP_{12} [Range]	PPP_{13} [Range]	PPP_{14} [Range]	PPP_{15} [Range]	PPP_{16} [Range]	PPP_{17} [Range]	PPP_{18} [Range]	PPP_{19} [Range]	PPP_{20} [Range]
Q_0	.82	.79	.93	.66	.65	.92	.79	.76	.63	.76
	[.45 - 1]	[.26 - 1]	[.55 - 1]	[.19 - .92]	[.27 - .98]	[.59 - 1]	[.41 - .99]	[.35 - .97]	[.26 - .93]	[.38 - .99]
Q_{O1}	.82	.79	.92	.66	.63	.93	.81	.75	.61	**.98**
	[.46 - .99]	[.26 - 1]	[.54 - 1]	[.21 - .92]	[.27 - .98]	[.63 - 1]	[.45 - .99]	[.38 - .97]	[.22 - .93]	**[.86 - 1]**
Q_{O5}	.83	**1**	.92	.73	.58	.91	.79	.75	.62	**1**
	[.44 - .99]	**[.96 - 1]**	[.49 - 1]	[.3 - .94]	[.25 - .98]	[.62 - 1]	[.39 - .99]	[.44 - .97]	[.16 - .9]	**[1 - 1]**
Q_{U1}	.83	.8	.94	.67	.64	.92	.8	**1**	.61	.79
	[.45 - .99]	[.29 - 1]	[.56 - 1]	[.2 - .92]	[.25 - .98]	[.56 - 1]	[.41 - .99]	**[.99 - 1]**	[.21 - .93]	[.37 - .99]
Q_{U5}	**.99**	**1**	.93	.7	**1**	.93	.77	**1**	**1**	.79
	[.96 - 1]	**[.91 - 1]**	[.57 - 1]	[.28 - .92]	**[.98 - 1]**	[.58 - 1]	[.37 - .99]	**[.99 - 1]**	**[.99 - 1]**	[.35 - .99]
Q_{B2}	.83	.81	.94	**1**	.63	.91	.75	.79	.61	.77
	[.53 - .99]	[.32 - 1]	[.6 - 1]	**[.97 - 1]**	[.24 - .98]	[.57 - 1]	[.29 - .99]	[.39 - .98]	[.21 - .92]	[.38 - .99]
Q_{B6}	.83	.86	.92	.67	**1**	**1**	.76	.82	**1**	**1**
	[.43 - .99]	[.3 - 1]	[.57 - 1]	[.27 - .94]	**[.97 - 1]**	**[1 - 1]**	[.37 - .99]	[.61 - .99]	**[.96 - 1]**	**[.99 - 1]**

Next, the distribution of posterior predictive p values calculated utilizing the χ^2_{raw} as the discrepancy measure are presented for the conditions that failed to include the higher order relationship among the attributes. Their means and the ranges over 50 replications are presented in the Table 4.21 for 20 items across the seven conditions that utilized one true and six false Q–matrices. The boldfaced entries in the table are for the misspecified items.

Average PPP values over 50 replications for each condition ranged from .69 to 1 for the items that their attributes are specified correctly by the Q–matrix. Minimum and maximum values of these PPP values are .3 and 1. Similar to the results from first seven conditions that fit higher order DINA model to the data, PPP values based on χ^2_{raw} discrepancy measure flagged some of the correctly specified items as not fitting because the ranges of the average PPP-values include .95 and higher values.

More items were identified as misfitting by the χ^2_{raw} with the independence DINA model than the higher order DINA model. Average PPP-values for first 10 items, as well as 13th, 16th, and 17th items are greater than .90. It is desired to detect more items as misfitting when the attribute relationship is not correctly modeled. However, keeping in mind that these average PPP-values based on χ^2_{raw} were high as well when the higher order DINA model was utilized, leads to the conclusion that these high average PPP-values might not be due to the false attribute relationship specifications.

Average PPP-values for the items which are misspecified by the Q–matrix ranged from .99 to 1. The minimum PPP-value based on χ^2_{raw} was .95 and the maximum value was 1. Thus, all the misspecified items were detected as not fitting by the PPP-values calculated based on the χ^2_{raw}.

For either grouping based on raw scores or based on latent classes, the discrepancy between the observed and model predicted proportion of correct responses were more frequently greater than the discrepancy between the replicated and predicted proportion of correct responses across all MCMC iterations because majority of the PPP-values were close to the upper end of the [0,1] interval.

Table 4.21

ppp values and ranges for 20 Items of Across 7 DINA Conditions for Raw Scores

	PPP_1 [range]	PPP_2 [range]	PPP_3 [range]	PPP_4 [range]	PPP_5 [range]	PPP_6 [range]	PPP_7 [range]	PPP_8 [range]	PPP_9 [range]	PPP_{10} [range]
Q_0	.96	.94	.94	.89	.91	.98	.98	.99	.96	.98
	[.85 - 1]	[.79 - 1]	[.63 - 1]	[.51 - 1]	[.6 - 1]	[.87 - 1]	[.86 - 1]	[.89 - 1]	[.8 - 1]	[.94 - 1]
Q_{O1}	.96	.93	.93	.91	.93	.98	.98	.98	.97	.99
	[.86 - 1]	[.77 - 1]	[.6 - 1]	[.56 - 1]	[.67 - 1]	[.86 - 1]	[.84 - 1]	[.88 - 1]	[.85 - 1]	[.94 - 1]
Q_{O5}	**1**	.92	.92	**1**	.93	1	.97	.99	.99	.98
	[1 - 1]	[.72 - 1]	[.53 - .99]	**[.96 - 1]**	[.67 - 1]	[.98 - 1]	[.8 - 1]	[.89 - 1]	[.91 - 1]	[.91 - 1]
Q_{U1}	.96	.93	.89	.95	.91	.98	.98	.98	.97	.99
	[.85 - 1]	[.78 - 1]	[.54 - .99]	[.76 - 1]	[.6 - 1]	[.86 - 1]	[.86 - 1]	[.86 - 1]	[.8 - 1]	[.94 - 1]
Q_{U5}	.92	.88	.91	.95	.89	.96	.95	1	.97	.97
	[.74 - .99]	[.6 - .99]	[.82 - .98]	[.78 - 1]	[.53 - .99]	[.8 - 1]	[.75 - 1]	[.98 - 1]	[.82 - 1]	[.87 - 1]
Q_{B2}	.98	.92	.94	**.99**	.88	.99	.97	.99	.99	.95
	[.89 - 1]	[.74 - 1]	[.64 - 1]	**[.95 - 1]**	[.53 - 1]	[.89 - 1]	[.8 - 1]	[.88 - 1]	[.94 - 1]	[.82 - 1]
Q_{B6}	.97	.91	.97	.94	.85	**1**	.98	**1**	.99	.94
	[.91 - 1]	[.75 - 1]	[.86 - 1]	[.75 - 1]	[.46 - .99]	**[1 - 1]**	[.9 - 1]	**[.99 - 1]**	[.91 - 1]	[.78 - 1]

	PPP_{11} [Range]	PPP_{12} [Range]	PPP_{13} [Range]	PPP_{14} [Range]	PPP_{15} [Range]	PPP_{16} [Range]	PPP_{17} [Range]	PPP_{18} [Range]	PPP_{19} [Range]	PPP_{20} [Range]
Q_0	.86	.84	.97	.76	.78	.96	.9	.81	.73	.82
	[.52 - .99]	[.33 - 1]	[.78 - 1]	[.35 - .97]	[.51 - .99]	[.74 - 1]	[.6 - 1]	[.4 - .97]	[.43 - .97]	[.49 - .99]
Q_{O1}	.86	.84	.96	.76	.76	.97	.91	.81	.7	**.99**
	[.5 - .99]	[.3 - 1]	[.78 - 1]	[.39 - .97]	[.46 - .98]	[.77 - 1]	[.66 - 1]	[.39 - .97]	[.38 - .96]	**[.9 - 1]**
Q_{O5}	.88	**1**	.97	.87	.7	.97	.89	.8	.7	**1**
	[.58 - 1]	**[.99 - 1]**	[.8 - 1]	[.56 - .99]	[.38 - .98]	[.82 - 1]	[.6 - 1]	[.41 - .97]	[.32 - .95]	**[1 - 1]**
Q_{U1}	.86	.84	.97	.77	.76	.96	.9	**1**	.71	.86
	[.53 - .99]	[.33 - 1]	[.78 - 1]	[.38 - .97]	[.44 - .99]	[.73 - 1]	[.6 - 1]	**[1 - 1]**	[.39 - .95]	[.53 - .99]
Q_{U5}	**1**	**1**	.96	.73	**1**	.96	.86	**1**	**1**	.85
	[.99 - 1]	**[.99 - 1]**	[.75 - 1]	[.32 - .96]	**[1 - 1]**	[.73 - 1]	[.54 - 1]	**[1 - 1]**	**[1 - 1]**	[.47 - .99]
Q_{B2}	.87	.87	.97	**1**	.75	.96	.85	.86	.69	.83
	[.61 - .99]	[.39 - 1]	[.83 - 1]	**[1 - 1]**	[.47 - .99]	[.73 - 1]	[.46 - 1]	[.47 - .99]	[.35 - .95]	[.51 - .99]
Q_{B6}	.87	.91	.98	.71	**1**	**1**	.88	.92	**1**	**1**
	[.48 - 1]	[.45 - 1]	[.88 - 1]	[.37 - .96]	**[.99 - 1]**	**[1 - 1]**	[.56 - 1]	[.76 - 1]	**[1 - 1]**	**[1 - 1]**

4.5.1 The Type I Error Rate

Type I error rates of the indices are investigated at .05 and .01 significance level for the PPP-values based on χ^2_{raw} discrepancy measures. Table 4.22 presents the proportion of PPP-values based on χ^2_{raw} indices smaller than $p = .05$ or greater than $p = .95$ for the investigation at .05 level and the proportion of PPP-values based on χ^2_{raw} indices smaller than $p = .01$ or greater than $p = .99$ for the investigation at .01 level.

Type I error rates are calculated for each condition that utilized Q_0, Q_{O1}, Q_{O5}, Q_{U1}, Q_{U5}, Q_{B2}, and Q_{B6} matrices and fit the higher order DINA model seperately. In the Table 4.22 Type I error rates are calculated from 20, 19, 16, 19, 15, 18, and 14 number of items for seven conditions. Furthermore, the overall empirical Type I error rate is calculated as the percent of items flagged as problematic out of 121 correctly specified items across seven conditions which fit the higher order DINA model to the simulated data. 50 replications for each condition resulted in an overall count of 6050. 1216 out of 6050 correctly specified items were detected as not fitting at .05 level. This resulted in empirical Type I error rate of .20 for PPP-values based on χ^2_{raw}.

Similar to Type I error rate from previous fit indices of χ^2_{lt}, χ^2_{raw}, and PPP based on χ^2_{lt}, having misspecified item or items in a test didn't impact the Type I error rate calculated from correctly specified items. Type I error rate is larger than .05 level across all the seven conditions. PPP-values based on χ^2_{raw} did not produce an acceptable Type I error rate by flagging too many items as problematic. The proportion of items classified as not fitting falsely is 20%. Similar to the results from χ^2_{raw} statistic, items 6, 7, 8, 13, and 16 are consistently flagged as not fitting almost half of the time across seven conditions. Only a few correctly specified items are consistently detected as fitting. Thus, there is an inconsistency across items in terms of contributing to the Type I error rate. Table C.2 presents the proportions of items that are falsely detected as not fitting by the PPP values based on χ^2_{raw} discrepancy measure.

It is expected that these Type I error rates will have impact on the power of the fit indices. Inflated Type I error rates result in overestimated power. Type I error rates presented in the Table 4.22 are all greater than the maximum value of the suggested range which is .025 to .075. Thus, these inflated Type I error rates are expected to result in overestimated power for PPP-values based on χ^2_{lt}.

Table 4.22
Proportion of PPP-Values Based on χ^2_{raw} Smaller Than $p = .05$ or Greater Than $p = .95$

	Type I Error Rates	
	.05	.01
Higher order DINA + Q_0	.20	.05
Higher order DINA + Q_{O1}	.21	.05
Higher order DINA + Q_{O5}	.25	.07
Higher order DINA + Q_{U1}	.20	.05
Higher order DINA + Q_{U5}	.21	.04
Higher order DINA + Q_{B2}	.21	.21
Higher order DINA + Q_{B6}	.12	.12
Overall across 7 conditions	.20	.08

Table entries are italized to indicate the misspecified items.

4.6 Overall Fit Evaluation with PPP-values based on χ^2_{lt}

PPP-values are calculated utilizing the χ^2_{lt} statistic as the discrepancy measure across all of the items to test the model fit. Extreme values are expected to indicate model misfit and reject the model with values smaller than .05 or greater than .95. Table 4.23 presents the averages and the ranges of these PPP-values across 14 conditions. Average values were high across all the conditions that utilized one type of the misspecified Q–matrix and ranged from .80 to 1.

PPP-values based on χ^2_{lt} did not reject the model when the true Q–matrix and the higher order DINA model is fit to the simulated data. PPP-values based on χ^2_{lt} ranged from .29 to

.81 for this condition over 50 replications and the average was .58. None of the PPP-values based on χ^2_{lt} exceeded .95 or was smaller than .05.

The range of PPP-values based on χ^2_{lt} didnot include the extreme values when only 1% of the Q–matrix indices have been changed from 0 to 1 and the higher order DINA model was fit to the simulated data. The average PPP-values based on χ^2_{lt} was .80 with the higher order DINA and Q_{O1} matrix and it was .94 with independence DINA model and Q_{O1}. Average values were higher when the independence DINA model was utilized. However, it doesn't seem to reject the independence DINA model when the items are specified correctly with Q_0 because the range of PPP-values are between .55 and .95 and the average is .82.

The rejection rates are going to be investigated next and will provide further information about the properties of PPP-values calculated based on χ^2_{lt} .

Table 4.23
Average PPP-Values Calculated Based on χ^2_{lt} for Overall Model Fit

		Q_0	Q_{O1}	Q_{O5}	Q_{U1}	Q_{U5}	Q_{B2}	Q_{B6}
Higher order DINA conditions	mean	.58	.80	1	.90	1	.99	1
	range	[.29-.81]	[.57-.93]	[1-1]	[.70-.98]	[1-1]	[.96-1]	[1,1]
Independence DINA conditions	mean	.82	.94	1	.98	1	1	1
	range	[.55-.95]	[.78-.99]	[1-1]	[.91-1]	[1-1]	[1-1]	[1,1]

4.6.1 The Type I Error Rate

Type I error rate is calculated for the condition that fit the higher order DINA model and the true Q–matrix to the simulated data based on 50 replications. This condition utilized the generating model and the Q–matrix to fit to the simulated data. The model was not rejected across all 50 replications and this resulted in Type I error rate of 0. Thus, Type I error is not committed with the PPP-values calculated based on χ^2_{lt} in this study.

4.6.2 Power

Type II error is committed when the test statistic fails to reject the false model. Power is the proportion of times that Type II error is not committed which is when the model is successfully rejected when it was false. Power is calculated for the conditions that fit the higher order DINA model and one type of the misspecified Q–matrix or the independence DINA model to the simulated data. Proportions of the false models that are rejected by the PPP-values based on χ^2_{lt} are presented in the Table 4.24 across all false models. Only the first condition fit the true model to the generating data and power is not calculated for that condition. Model is rejected most of the time when more than 1% of the items are misspecified. When only 1% of the items are overspecified by the Q–matrix, PPP-values based on χ^2_{lt} did not reject the model and resulted in the power value of 0 over 50 replications. 10 out of 50 replications rejected the higher order DINA model when 1% of the items are underspecified by the Q–matrix. Higher proportions of models are rejected by PPP-values based on χ^2_{lt} when the items are misspecified together with the attribute relationships. It should be noted that PPP-values based on χ^2_{lt} failed to reject independence DINA model when the true Q–matrix is used 98% of the time. Similar to the PPP-values based on χ^2_{lt} discrepancy measure which are calculated for checking item fit, PPP-values based on χ^2_{lt} indices for checking overall fit rejected the model due to the item misspecification successfully whereas it failed to reject the model due to the lack of higher order structure model.

4.7 Overall Fit Evaluation with PPP-values based on χ^2_{raw}

PPP-values are calculated utilizing the χ^2_{raw} statistic as the discrepancy measure. Table 4.25 presents the averages and the ranges of the PPP-values across 14 conditions. Average values were equal to 1 across all the conditions. Thus the PPP-values based on the χ^2_{raw} discrepancy measure rejected the model for all the conditions over 50 replications for each condition. Type I error rate and power is calculated for PPP-values based on χ^2_{raw} and presented next. From

Table 4.24
Rejection Rates of the PPP-Values Based on χ^2_{lt} for Overall Model Fit

Q–Matrix	Higher order DINA	Independence DINA	Overall
Q_0		.02	
Q_{O1}	0	.38	
Q_{O5}	1	1	
Q_{U1}	.20	.92	
Q_{U5}	1	1	
Q_{B2}	1	1	
Q_{B6}	1	1	
overall	.70	.76	.73

the first evaluation of this statistic based on these descriptives, it can be claimed that it is not a useful test to check the model fit.

Table 4.25
Average PPP-Values Calculated Based on χ^2_{raw} for Overall Model Fit

		Q_0	Q_{O1}	Q_{O5}	Q_{U1}	Q_{U5}	Q_{B2}	Q_{B6}
Higher order DINA conditions	mean	1	1	1	1	1	1	1
	range	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]
Independence DINA conditions	mean	1	1	1	1	1	1	1
	range	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]	[1-1]

4.7.1 The Type I Error Rate

Type I error rate for PPP-value based on χ^2_{raw} is calculated for the first condition that fit both the generating model and the generating Q–matrix to the simulated data. Calculation was over 50 replications. The model was rejected for all 50 replications when they were not

supposed to be rejected and this resulted in Type I error rate of 1. Thus PPP-values is not useful in detecting model misfit with the highest possible Type I error rate. It is highly likely to reject a useful model with the PPP-values based on χ^2_{raw}. As noted earlier, PPP-values based on χ^2_{raw} calculated for items acted inconsistently. Thus it is not unexpected to observe inconsistency from PPP-values calculated based on χ^2_{raw} for checking overall model fit.

4.7.2 Power

Power is calculated as the proportion of times that Type II error is not committed by the PPP-values based on χ^2_{raw} which is when the model is successfully rejected when it was false. Power is calculated for 13 conditions that fit the higher order DINA model and one type of misspecified Q–matrix or the independence DINA model to the simulated data.

Table 4.26
Rejection Rates of the PPP-Values Based on χ^2_{raw} for Overall Model Fit

Q-matrix	Higher order DINA	Independence DINA	Overall
Q_0		1	
Q_{O1}	1	1	
Q_{O5}	1	1	
Q_{U1}	1	1	
Q_{U5}	1	1	
Q_{B2}	1	1	
Q_{B6}	1	1	
overall	1	1	1

Proportions of the false models that are rejected by the PPP-values based on χ^2_{raw} are presented in Table 4.26 across all false models. Power is not calculated for the first condition. The model is rejected by the PPP-values based on χ^2_{raw} all the time for these 13 conditions and produced the power value of 1. However, considering the Type I error rate of PPP-values based on χ^2_{raw} it cannot be claimed that the PPP value based on χ^2_{raw} is successfully rejecting the false models. It simply rejected all the possible models including the true models.

4.8 Comparison of the Item Fit Indices Studied: χ^2_{lt}, χ^2_{raw}, and Associated PPP-values

Item fit indices are compared in terms of the probability of committing Type I error and in terms of their power. Overall Type I error rates for χ^2_{lt} and χ^2_{raw} and for the PPP-values based on these discrepancy measures are presented in the Table 4.27. Smallest Type I error rate was obtained by the PPP-values calculated based on the χ^2_{lt} discrepancy measures. Largest Type I error was observed with χ^2_{raw}. Type I error rates was similar for χ^2_{lt} and PPP-values calculated based on χ^2_{lt} which were .008 and .00 respectively. Type I error rates were high both from χ^2_{lt} as a test statistic and PPP-values calculated based on χ^2_{lt} as a discrepancy measure which were .244 and .20 respectively.

Table 4.27
Comparison of Type I Error Rates for Item Fit Indices

	χ^2	PPP value
Based on Latent Classes	.008	.00
Based on Raw Scores	.244	.20

Power for χ^2_{lt} and χ^2_{raw} and for the PPP-values based on these discrepancy measures are presented in Table 4.28. For all of the indices power was high and close to 1 calculated over misspecified items. For correctly specified items power was unacceptably low due to the inability to reject the model when the higher order structure was not specified by the model across all indices and it ranged from .001 to .470. However PPP-values rejected correctly specified items when the model failed to include the higher order structure among the attributes more frequently than the χ^2 statistics. When all the misspecified items and all the independence DINA model conditions were included in the calculations power was .164, .477, .137, and .538 for χ^2_{lt}, PPP-values based on χ^2_{lt}, χ^2_{raw}, and PPP-values based on χ^2_{raw} respectively.

Table 4.28
Comparison of Power for Item Fit Indices

	Based on Latent Classes		Based on Raw Scores	
	χ^2	PPP	χ^2	PPP
Misspecified Items	1	.999	.998	.998
Independence DINA Model	.033	.395	.001	.470
Overall	164	.477	.137	.538

4.9 Comparison of the Model Fit Indices Studied: PPP-values Calculated Based on χ^2_{lt} and χ^2_{raw}

Overall model fit indices are compared in terms of their empirical Type I error rates and in terms of their power. Overall Type I error rates for PPP-values based on χ^2_{lt} and χ^2_{raw} are presented in Table 4.29. PPP-values calculated based on χ^2_{lt} did not reject any model that fit the generating the higher order DINA model with generating Q–matrix over 50 replications. On the other hand PPP-values calculated based on χ^2_{raw} rejected the higher order DINA models with generating Q–matrix over all 50 replications.

Table 4.29
Comparison of Type I Error Rate for Model Fit Indices

	Based on Latent Classes	Based on Raw Scores
Overall	0	1

Power for PPP-values based on χ^2_{lt} and χ^2_{raw} are presented in Table 4.30. As noted earlier, PPP-values based on χ^2_{raw} rejected all the models and produced a high power value. It can not be really attributed to the power since the statistic tended to reject even the true models. 70% of the models that included misspecified items are rejected by PPP-values based on χ^2_{lt}. Even though it seems like a low percentage, it should be noted that these calculations included the models that changed only 1% of the items as false models. This might not be realistic. Furthermore, item fit index of chi^2_{lt} or associated PPP-values were able to detect

the problematic items all the time. It could be a better solution to discard the problematic items.

Table 4.30
Comparison of Power for Model Fit Indices

	Based on Latent Classes	Based on Raw Scores
Misspecified Items	.70	1
Independence DINA Model	.76	1
Overall	.73	1

Chapter 5

Illustrative analysis: Tatsuoka's fraction subtraction data

This chapter presents an illustrative data analysis and model fit evaluation using a Bayesian approach. Model fit is assessed through a posterior predictive model checking method which utilized the χ^2 discrepancy measures calculated based on latent classes and based on raw score groups. Further, PPP–values based on these discrepancy measures are calculated. In addition to the fit statistics, posterior predictive model checking utilized several graphics to investigate the model fit in this study. Example data set is analyzed with the DINA model. First, the independence DINA model is fit to the data which assumed no relationship among the attributes. Second, the fraction subtraction data is analyzed with the higher order DINA model which assumed a two parameter logistic model for the relationship among the attributes. Sample data is Tatsuoka's fraction subtraction data (Tatsuoka, 1990). Tasks measured by this test are, as its name reveals, fraction subtraction problems. The DINA model has been used to analyze the fraction subtraction data in many articles (e.g., de la Torre, 2009; de la Torre & Douglas, 2004, 2008; Henson, Templin, & Willse, 2009). The data which is analyzed in this study are for 536 middle school children. It is available at the Journal of the Royal Statistical Society's website:

http://www.blackwellpublishing.com/rss/Volumes/Cv51p3.htm

20 items from the fraction subtraction data are presented in Table 5.1. Table 5.1 also presents, for each item, the attributes that are required to produce a correct response to that item. For example, the combination of the skills 4, 6, and 7 are required to produce a correct response to the first item. Thus, a student could give a correct response to the first item only

Table 5.1
Fraction Subtraction Data, 20 Items, 8 Hypothesized Skills

Item Number	Item	Skills	Item Number	Item	Skills
1	$\frac{5}{3} - \frac{3}{4}$	$\alpha_4, \alpha_6, \alpha_7$	11	$4\frac{1}{3} - 2\frac{4}{3}$	$\alpha_2, \alpha_5, \alpha_7$
2	$\frac{3}{4} - \frac{3}{8}$	α_4, α_7	12	$1\frac{1}{8} - \frac{1}{8}$	α_7, α_8
3	$\frac{5}{6} - \frac{1}{9}$	α_4, α_7	13	$3\frac{3}{8} - 2\frac{5}{6}$	$\alpha_2, \alpha_4, \alpha_5, \alpha_7$
4	$3\frac{1}{2} - 2\frac{3}{2}$	$\alpha_2, \alpha_3, \alpha_5, \alpha_7$	14	$3\frac{4}{5} - 3\frac{2}{5}$	α_2, α_7
5	$4\frac{3}{5} - 3\frac{4}{10}$	$\alpha_2, \alpha_4, \alpha_7, \alpha_8$	15	$2 - \frac{1}{3}$	α_1, α_7
6	$\frac{6}{7} - \frac{4}{7}$	α_7	16	$4\frac{5}{7} - 1\frac{4}{7}$	α_2, α_7
7	$3 - 2\frac{1}{5}$	$\alpha_1, \alpha_2, \alpha_7$	17	$7\frac{3}{5} - \frac{4}{5}$	$\alpha_2, \alpha_5, \alpha_7$
8	$\frac{2}{3} - \frac{2}{3}$	α_7	18	$4\frac{1}{10} - 2\frac{8}{10}$	$\alpha_2, \alpha_5, \alpha_6, \alpha_7$
9	$3\frac{7}{8} - 2$	α_2	19	$4 - 1\frac{5}{3}$	$\alpha_1, \alpha_2, \alpha_3, \alpha_5, \alpha_7$
10	$4\frac{4}{2} - 2\frac{7}{12}$	$\alpha_2, \alpha_5, \alpha_7, \alpha_8$	20	$4\frac{1}{3} - 1\frac{5}{3}$	$\alpha_2, \alpha_3, \alpha_5, \alpha_7$

if he/she knows how to find a common denominator, column borrow to subtract the second numerator from the first, and subtract numerators. For this fraction subtraction data, de la Torre and Douglas (2004) defined the eight attributes required to produce a correct response as:

(1) Convert a whole number to a fraction,

(2) Separate a whole number from a fraction,

(3) Simplify before subtracting,

(4) Find a common denominator,

(5) Borrow from whole number part,

(6) Column borrow to subtract the second numerator from the first,

(7) Subtract numerators, and

(8) Reduce answers to simplest form. (de la Torre, 2004)

The Q–matrix presents the items and required attributes to produce a correct response for an item. The Q–matrix which is defined in de la Torre and Douglas (2004) for the fraction subtraction data is presented in Table 5.2 and this Q–matrix is adopted in this study.

Table 5.2
Transposed Q Matrix for Fraction Subtraction Data

	Item																			
Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0
2	0	0	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1
3	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1
4	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	1	0	0	0	0
5	0	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	1	1	1	1
6	1	1	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	1	0	0
7	1	1	1	1	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	1
8	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0

5.1 Estimation of the DINA Model

Fraction subtraction test is analyzed with the independence DINA and the higher order DINA models. The DINA model is a conjunctive model, that is an examinee needs to have mastered all the skills required to produce a correct response for an item. Two versions of the DINA model were used to analyze the data: one hypothesizing a two parameter logistic model among the attributes, the higher–order DINA model, and the other suggesting no relationship among the attributes, the independence DINA model. General latent ability could be modeled either as unidimensional or multidimensional when modeling with the higher order DINA model. For the fraction-subtraction data it is reasonable to assume a unidimensional general latent ability as the test is comprised of items requiring very specific pieces of information but from a somewhat narrow domain.

Both the independence DINA and the higher order DINA models are analyzed with Markov chain Monte Carlo (MCMC) algorithm employing Gibbs sampling in this study. WINBUGS software (Spiegelhalter, Thomas, Best, & Lunn, 2003) is used to estimate the model with MCMC algorithm in which values representing parameters of the model are repeatedly sampled from their full conditional posterior distributions over a large number

of iterations. This study adopted the algorithm used in Li (2008) with WINBUGS software. Model parameters were sampled from their full posterior distributions conditional upon the already sampled ability and examinee attribute mastery parameters for the DINA model.

The following priors were used to estimate the parameters of the higher-order DINA model:

$$\alpha_{jk}|\theta_j \sim \text{Bernoulli } (\pi_{jk}),\ \pi_{jk} = a(\theta_j) - (\beta_k),$$
$$\theta_j \sim \text{N}(0,\ 1),$$
$$a \sim \text{N}(0,\ 1)\ a > 0,$$
$$\beta_k \sim \text{N}(0,\ 1),$$
$$g \sim \text{Beta}(U_g,\ S_g),$$
$$s \sim \text{Beta}(U_s,\ S_s),$$
$$U_g \sim \text{Uniform}(.1,.9),$$
$$S_g \sim \text{Uniform}(0.5,10),$$
$$U_s \sim \text{Uniform}(.1,.9),$$
$$S_s \sim \text{Uniform}(0.5,10).$$

Most of these priors are employed for the independence DINA model except for the α_{jk}. For the independence DINA model, $\alpha_{jk}|\theta_j \sim$ Bernoulli (π_{jk}) and $pi_{jk} \sim$ Uniform (0,1). In addition to that, the independence DINA model did not estimate a or β parameters either.

The beta distribution as a prior for the g and s parameters produces values between 0 and 1 for g and s. The prior distributions and their parameter values are chosen to generate reasonable values for the item parameters. Li (2008) noted that when the value of U_g (or U_s) was larger than the value of S_g (or S_s), the priors for g and s were more likely to be drawn from .5 to 1 interval, otherwise, the priors for g and s are likely to be drawn from 0 to .5 interval. In light of this information, uniform distribution (.1, .9) for U_g (or U_s) and uniform distribution (0.5, 10) for S_g (or S_s) as the hyperprior distribution are selected to ensure that g and s are going to be drawn from 0 to .5 interval as these values are reasonable for both guessing and slipping parameters.

5.2 Convergence

First, the convergence is evaluated to check if a representative sample of the target distribution is obtained. The first iterations that do not represent the target distribution are discarded to diminish the effect of early iterations. This process is called as burn-in. First, convergence is investigated by visually inspecting the trace plots. Irregularities or different patterns indicate problems with the convergence. Then, Geweke's (1992) method is used to evaluate if there is evidence against the convergence. It is calculated based on the analysis of individual chain and implemented in the software BOA (Smith, 2005). None of the parameters p-values for Z statistic was smaller than .05 which concludes that there is not evidence against convergence. Heidelberger and Welch (1983) convergence diagnostic method is implemented again in the software BOA (Smith, 2005). It recommended to discard 1000 iterations only for one guessing parameter with the independence DINA model which was for the eighth item. Thus 1000 iterations were discarded and the rest of the iterations are kept for the analysis.

5.3 Results

First, the fraction subtraction data is analyzed with the independence DINA model. The estimated posterior means and the posterior standard deviations (PSD) for the independence DINA model are presented in Table 5.3. Guessing parameters ranged from .01 to .48. eight out of 20 items had guessing parameter larger than .20. Slip parameters ranged from .02 to .31. Only two of the slip parameters were larger than .20. Large values of the slip and guessing parameters are indication of incorrectly specified Q–matrix or problem with the hypothesized model. Slip and guessing parameter estimates are similar to the results reported in Tatsuoka (2002) and de la Torre and Douglas (2004).

The model and the item fit are evaluated and χ^2 values based on raw score and based on latent class groups are calculated. Table 5.3 presents the item fit statistics and corre-

Table 5.3
Parameter Estimation Using the Independence DINA Model

Item Number	$\hat{g}$	$PSD(\hat{g})$	$\hat{s}$	$PSD(\hat{s})$	χ^2_{raw}	PPP_{raw}	χ^2_{ltc}	PPP_{ltc}
1	.09	.02	.10	.02	20.8	.85	27.4	.64
2	.09	.02	.02	.01	17.8	.87	21.3	.54
3	.02	.01	.08	.02	16.3	.73	19.9	.43
4	.28	.03	.10	.02	50.5	1.0	54.2	.97
5	.37	.03	.09	.02	43.1	1.0	37.6	.79
6	.30	.05	.03	.01	59.7	1.0	20.4	.58
7	.05	.02	.12	.03	25.8	.95	39.5	.76
8	.41	.04	.13	.02	65.7	1.0	48.3	.98
9	.29	.06	.24	.03	82.9	1.0	68.4	1.0
10	.06	.02	.02	.02	22.1	.92	21.2	.35
11	.07	.02	.07	.02	21.4	.80	33.7	.68
12	.48	.03	.03	.01	90.6	1.0	42.1	.96
13	.01	.01	.31	.03	28.2	.87	28.5	.43
14	.29	.04	.03	.01	66.5	1.0	24.6	.50
15	.08	.02	.09	.02	31.6	.98	61.1	.99
16	.28	.04	.07	.01	58.1	1.0	28.9	.63
17	.05	.01	.13	.02	41.0	1.0	42.5	.84
18	.14	.02	.12	.02	45.5	1.0	37.4	.79
19	.04	.01	.12	.03	28.0	.93	33.5	.55
20	.04	.01	.07	.02	13.5	.52	44.4	.77

sponding posterior predictive p values for the analysis of fraction subtraction data with the independence DINA model. For the item fit evaluation purposes, χ^2_{raw} which is the discrepancy measure calculated based on the total raw score of examinees are also calculated and reported in Table 5.3. Degrees of freedom for χ^2_{raw} equals to 19 when number of items is 20 for the independence DINA model. Table value of χ^2 with degrees of freedom of 19 and .05 tail area probability is 30.1. Items 4, 5, 6, 8, 9, 12, 14, 15, 16, 17, and 18 indicate misfit with large χ^2 values based on raw score groups. PPP–values based on χ^2_{raw} are also presented in Table 5.3. Extreme PPP–values, larger than .95 and smaller than .05, indicate misfit. PPP–values calculated based on the discrepancy measure of χ^2_{raw} have extreme values for the same items that have significant χ^2_{raw} values, which were larger than 30.1.

Table 5.4
Parameter Estimation Using the higher–order DINA Model

Item Number	$\hat{g}$	$PSD(\hat{g})$	$\hat{s}$	$PSD(\hat{s})$	χ^2_{raw}	PPP_{raw}	χ^2_{ltc}	PPP_{ltc}
1	.04	.02	.11	.02	15.9	.66	22.3	.58
2	.04	.02	.04	.01	10.7	.60	37.7	.78
3	.01	.01	.12	.02	24.7	.86	21.7	.64
4	.23	.03	.11	.02	42.2	1.0	44.6	.97
5	.28	.03	.17	.02	48.8	1.0	33.9	.88
6	.05	.04	.04	.01	20.5	.72	20.1	.73
7	.03	.01	.19	.03	33.1	.98	68.4	.93
8	.41	.05	.18	.02	105.4	1.0	72.1	1.0
9	.12	.05	.25	.02	82.3	1.0	76.0	1.0
10	.03	.01	.21	.03	30.0	.95	50.0	.88
11	.07	.02	.08	.02	21.1	.79	31.4	.72
12	.13	.04	.05	.01	20.4	.87	15.0	.54
13	.01	.01	.32	.03	32.3	.91	30.2	.62
14	.07	.03	.06	.01	18.9	.69	23.4	.65
15	.04	.02	.11	.02	19.0	.83	30.8	.69
16	.11	.03	.11	.02	30.5	.95	31.6	.85
17	.04	.01	.14	.02	47.3	1.0	44.3	.87
18	.12	.02	.14	.02	40.2	1.0	49.1	.98
19	.02	.01	.22	.03	36.2	.97	44.4	.80
20	.02	.01	.15	.03	30.8	.95	33.9	.62

χ^2 values based on latent classes are calculated by combining the latent class groups because of the small number of examinees in each latent class. Final latent class groups are created by combining eight classes into one and resulted in 32 latent classes instead of the original 256 latent classes. Latent classes are combined based on the first five attributes. Latent classes which have the same attribute combinations for the first five attributes are combined. χ^2_{lt} is calculated based on 32 latent class groups and presented in Table 5.3 for the analysis of the fraction subtraction data with the independence DINA model. Table value of χ^2_{lt} with degrees of freedom of 30 and .05 tail area probability is 43.77. Items 4, 8, 9, 15, and 20 has χ^2_{lt} values larger than 43.77. PPP-values corresponding to χ^2_{lt} discrepancy measure produced larger values than .95 for the items 4, 8, 9, 12, and 15. χ^2_{lt} and corresponding PPP-

values both flagged the items 4, 8, 9, and 15. Item 20 is flagged by χ^2_{lt} but not by PPP-value. Item 12 is flagged by PPP-value but not by χ^2_{lt} in spite of the fact that it had considerably large χ^2_{lt} value.

Next, the higher order DINA model is fit to the fraction subtraction data. The estimated posterior means and the posterior standard deviations for the fraction subtraction data with the higher order DINA model are presented in Table 5.4. Guessing parameters ranged from .01 to .41. Three out of 20 items had guessing parameter larger than .20. Slip parameters ranged from .04 to .32. Four items had slipping parameters larger than .20.

Table 5.4 presents the item fit statistics and corresponding posterior predictive p values for the analysis of fraction subtraction data with the higher order DINA model. χ^2_{raw} discrepancy measure calculated based on the total raw scores of examinees are calculated and reported in Table 5.4 for each item in the fraction subtraction data and used for item fit evaluation. Degrees of freedom for χ^2_{raw} equals to 17 when number of items is 20 for the higher order DINA model. Table value of χ^2 with degrees of freedom of 17 and .05 tail area probability is 27.59. 12 items indicated misfit with large χ^2 values based on raw score groups. PPP–values based on χ^2_{raw} are also presented in Table 5.4. Nine items produced extreme PPP–values, which were larger than .95, and indicated item misfit. PPP–values calculated based on the discrepancy measure of χ^2_{raw} flagged nine of the 12 items that were flagged by χ^2_{raw} values.

χ^2_{lt} is calculated based on 32 latent class groups and presented in Table 5.4 for the analysis of the fraction subtraction data with the higher order DINA model. Table value of χ^2 with degrees of freedom of 28 and .05 tail area probability is 41.34. Items 4, 7, 8, 9, 10, 17, 18, and 19 has significant χ^2_{lt} values and the model is rejected for these items. PPP-values corresponding to χ^2_{lt} discrepancy measure produced larger values than .95 for the items 4, 8, 9, and 18. χ^2_{lt} and corresponding PPP-values both flagged the items 4, 8, 9, and 18. Four more items are flagged by χ^2_{lt} that were not flagged by PPP-value.

Based on the simulation results, conclusions based on χ^2_{raw} values are not considered meaningful. They are calculated to investigate how they acted with the real data. Items 4, 8 and 9 consistently produced large and significant χ^2 values based on latent classes with both the independence DINA and the higher-order DINA models. There are several ways to solve each of these items. Each of these ways produces a correct response with different combinations of specified skills. It might even be possible to solve items 8 and 9 without having any of these specified 8 attributes. Item 8 is $\frac{2}{3} - \frac{2}{3}$. Required attribute for item 8 is specified as "subtract numerators". Students might have responded this item correctly even when they did not know how to subtract numerators. A student can as well take the fraction 2/3 as a whole mathematical object. Then, resorting this object into its numerator was assumed to be the critical attribute, the α_7, that should become manifest in solving the Item 8. Student can take this fractional object as is without breaking into its parts, and observe that the subtraction of two identical fractions would yield to zero as well. Item 9 is $3\frac{7}{8} - 2$. For the item 9, it is specified that the item requires to know how to separate a whole number from a fraction. A student might simply subtract whole numbers without knowing how to separate a whole number from a fraction. Item 4 is $3\frac{1}{2} - 2\frac{3}{2}$. Item 4 might be solved first by converting the whole number to a fraction and subtracting numerators. These are attributes 1 and 7 which are different from current attribute specifications of 2,3,5, and 7 for this item. These approaches to problematic items in this study are only a few possible approaches that students might have taken. Observations from the item fit evaluations should be a part of the analysis of the diagnostic classification models and guide to the further directions such as improving the item specifications within the Q-matrix. However, this improvement should be done with the help of the content specialists. When the students are from diverse educational settings, it is also possible to observe a variety of problem solving strategies. For a small sample of students that come from a similar educational setting, it is likely that the students' approaches to solve these items are similar. Cases where more than one Q-matrices

are available should be analyzed accordingly as there are models to accommodate that kind of structure.

For the overall model fit evaluation PPP value based on χ^2_{lt} is considered. This PPP value rejected both the independence and the higher order DINA models. As observed from the simulation studies, when 5% of the Q–matrix indices were incorrectly specified χ^2_{lt} rejects the model majority of the time. It did not reject the model when only 1% of the Q-matrix indices are incorrectly specified. For the fraction subtraction data, at least four items indicated misfit. Modification to the Q-matrix for these four items could improve the model fit.

5.4 Graphical Summaries for Model and Item Fit Evaluation

First, Bayesian residual plot is obtained. Figure 5.1 shows the posterior mean of the expected number correct score versus the posterior mean of residuals. The residuals are mostly distributed below the zero-line towards the right of the plot where high scores are located. This is an indication that the model overpredicts the scores of individuals who have high proficiency.

Figure 5.2 is the direct data display of the fraction subtraction data with the independence DINA model. Prior to producing the graph, examinees are sorted by their total raw score from high to low and the items are sorted by their difficulty to get a clear picture of the data. The first image on the left is the image of the fraction subtraction data. The ten images represent ten replicated data sets from the analysis of the fraction subtraction data with the independence DINA model. Real data image is compared to the replicated data image to check if the model reproduce the data well.

From the analysis of direct data display graphs, the independence DINA model explains the data pretty good except for the students with lowest scores. The model prediction for these low achieving students to respond correctly to some questions were higher than what was observed in the real data.

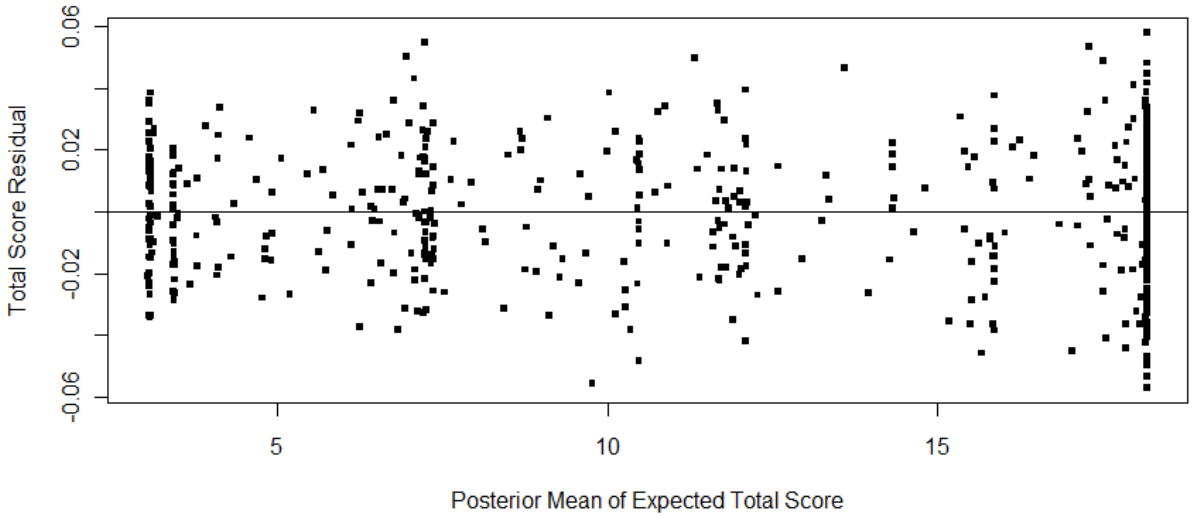

Figure 5.1: Number correct score residual versus predicted number correct for higher–order DINA

Figure 5.3 is the direct data display of the fraction subtraction data with the higher order DINA model. Again, the first image on the left is for the real data. Ten images on the right are randomly selected ten replicated data sets from the analysis of the fraction subtraction data with the higher order DINA model. Images of the replicated data sets are similar to the real data for most part. The student response data at the end of the image is somewhat different for the real data and the replicated data sets. The higher order DINA model estimated more correct responses from the low achieving students than what is observed in the real data. From the real data display graphs it can be concluded that for both the independence and higher order DINA models recovered the data for most of the sample except for the students with low scores. From the two models, the higher order DINA model recovered the data a little better for the students with low scores as the replicated data images show a little less dense expected correct responses for the low achieving students.

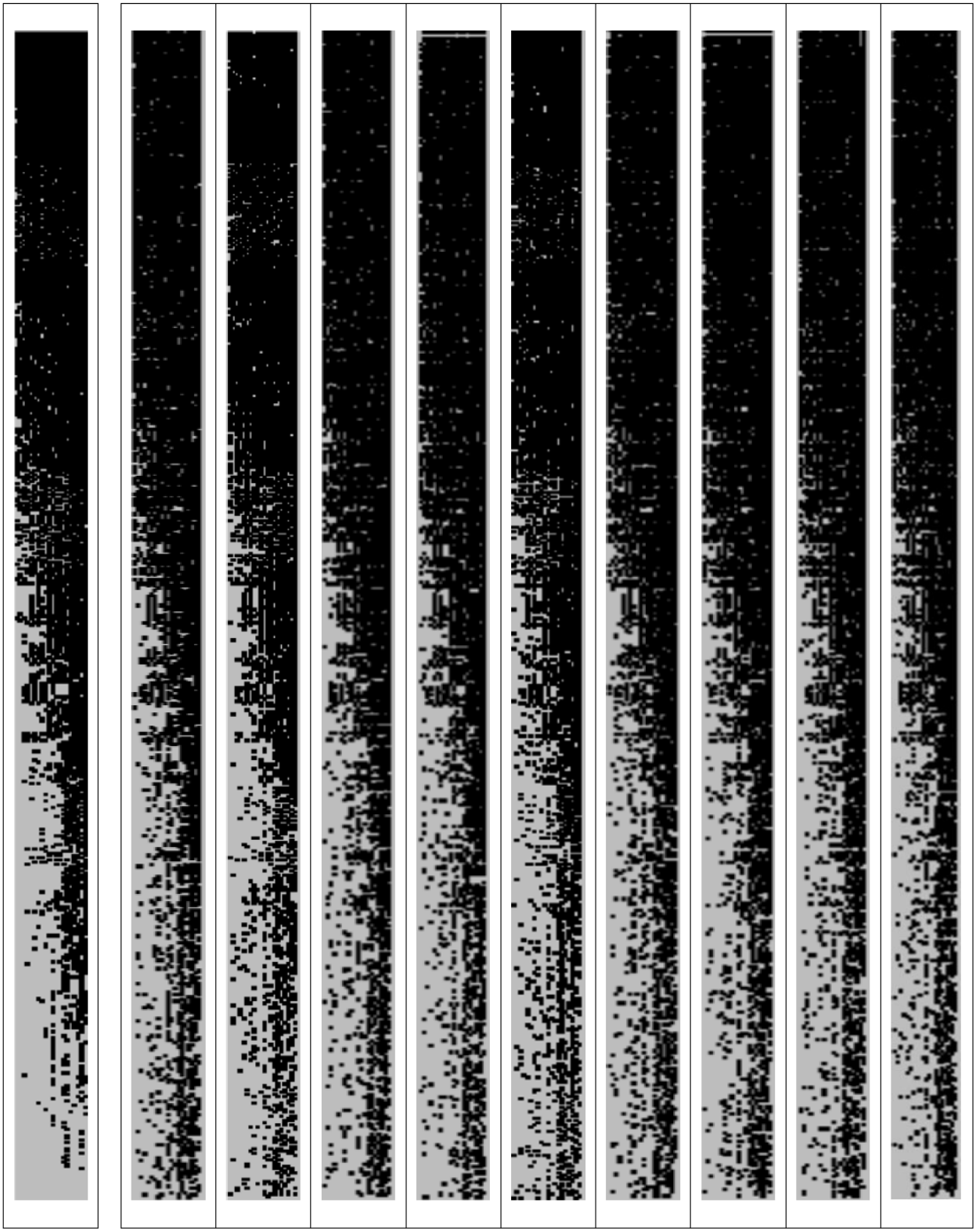

Figure 5.2: Fraction subtraction data and ten replicated data sets with the independence DINA model

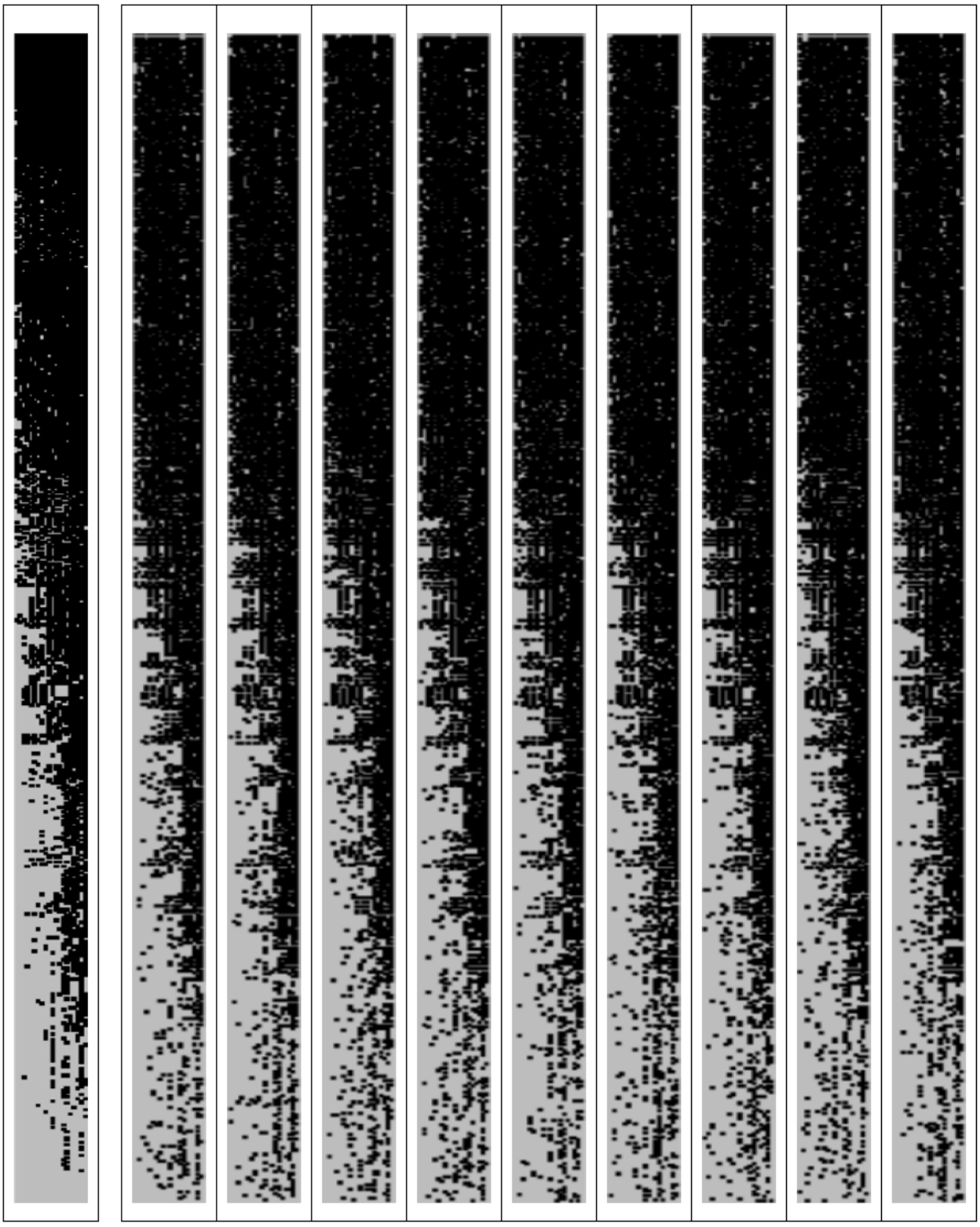

Figure 5.3: Fraction subtraction data and ten replicated data sets with the higher order DINA model

Item fit plots based on the analysis of fraction subtraction data with the independence DINA model are given in Figure 5.4. Fit plots in Figure 5.4 shows the proportion of correct response of the raw score groups for each item on the test. Horizontal axis of these fit plots represents the raw score groups. The replicated proportion of correct responses are compared to observed proportion of correct responses. Box plots represent the distribution of proportion of correct responses calculated from the replicated data for the raw score groups for each item. The whiskers of the box plot stretch from the 2.5 to 97.5 percentiles of the posterior distribution of the proportion of correct response of the raw score groups based on the replicated data. The line in the middle of the box plot represents the median of the posterior distribution of the replicated proportion of correct response. The line running over the box plots combines the observed proportion of correct responses of the raw score groups to facilitate the comparison of replicated and observed proportion of correct response values. Too many large discrepancies between the observed and replicated proportion of correct response values indicate problems with the item. When too many item fit plots indicate problems with fit, the model is suspect for the given data.

The largest discrepancies were observed for the items 3, 4, 5, and 9. For items 1, and 2, discrepancies between the observed and the replicated proportion of correct responses could be seen for the raw score groups that are in the middle of the raw score range. Item 8, 12, and 14 showed large discrepancies for the raw score groups with low scores. For the other items, fit plots did not present too many large discrepancies between the observed proportions and the posterior distribution of replicated proportion of correct response values.

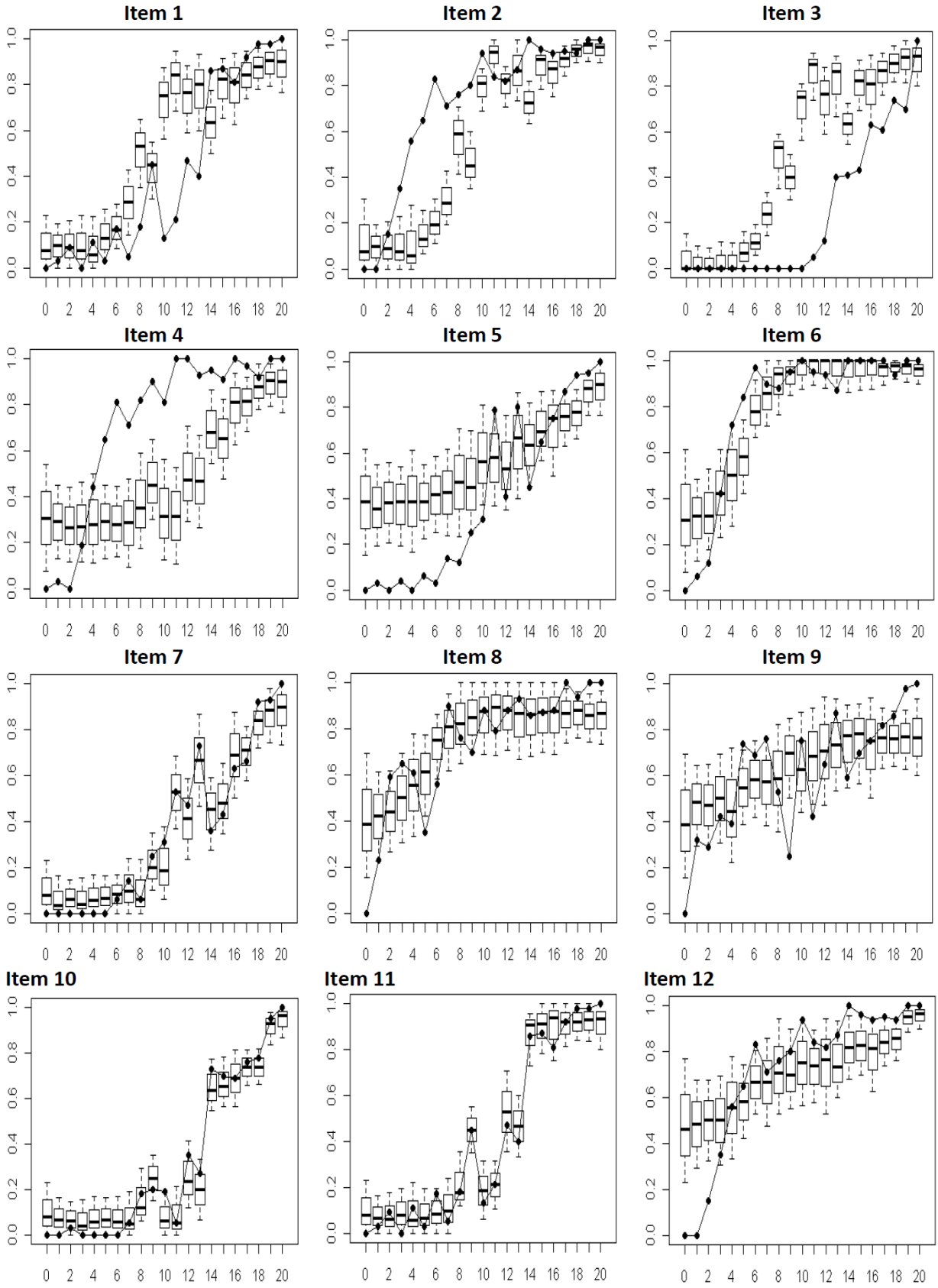

Figure 5.4: Item fit plots when the independence DINA model is fit to the fraction subtraction data.

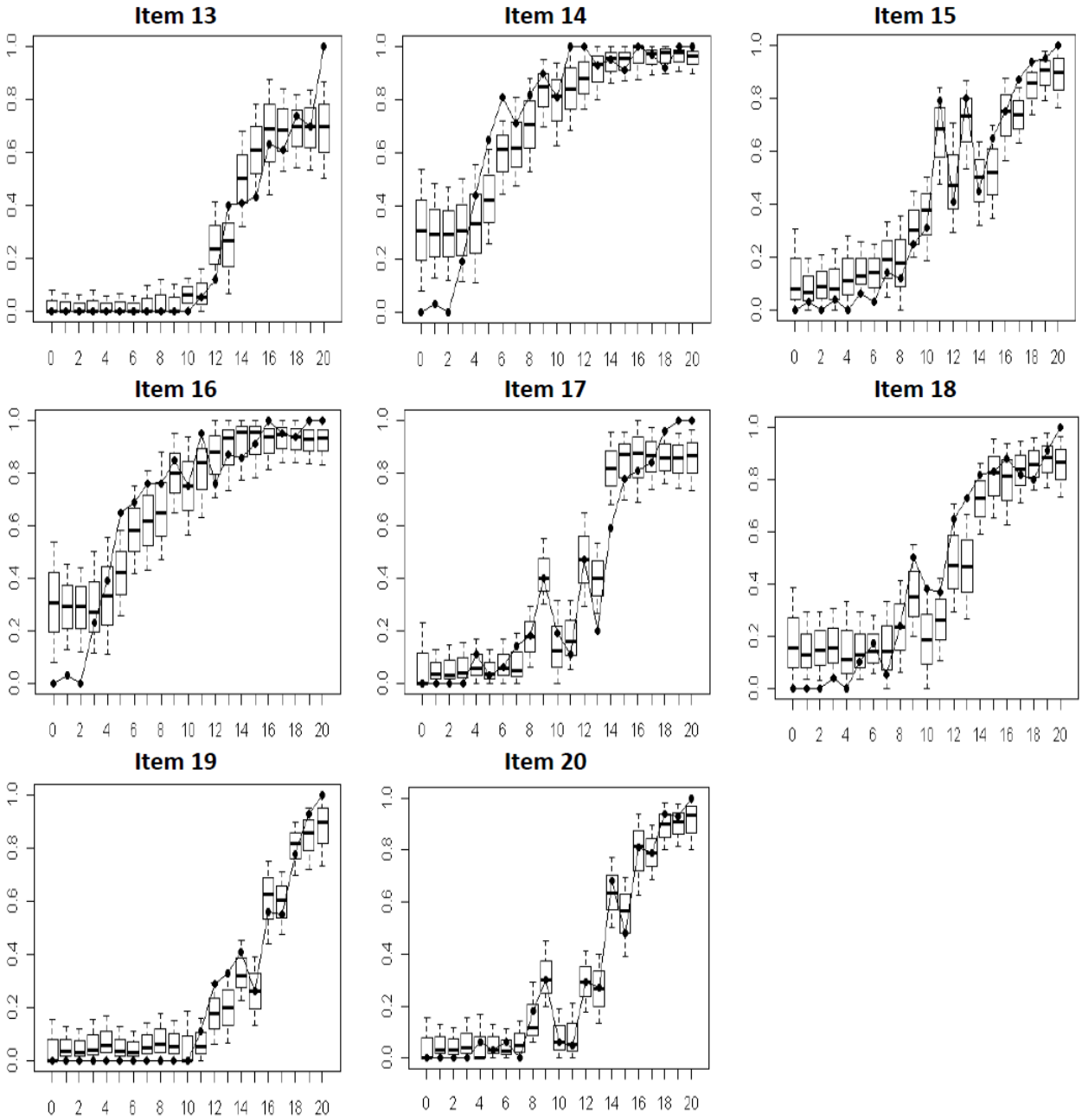

Figure 5.5: Item fit plots when the independence DINA model is fit to the fraction subtraction data.

Chapter 6

Discussion

This study presented the posterior predictive model checking method for the DINA model. Several indices for checking item fit and overall model fit were presented in this context. Two potential problems for modeling with DINA are Q–matrix misspecification and ignoring the higher order structure among the attributes. These problems created the context to investigate item fit and model fit for the fit indices for this study. Fit indices were χ^2_{lt}, χ^2_{raw}, PPP-values based on χ^2_{lt}, and PPP-values based on χ^2_{raw} for item fit evaluation. PPP-values based on χ^2_{lt} and PPP-values based on χ^2_{raw} discrepancy measures were calculated for overall model fit evaluation.

PPMC is a Bayesian tool. It is available when the Bayesian methods are utilized for statistical data analysis. In this study, Bayesian approach is undertaken. An MCMC algorithm employing Gibbs sampling was used to estimate the DINA model. A simulation study was done to examine Type I error rates, and power under possible modeling conditions. First, model recovery was evaluated to determine the correspondence between generating and estimated parameters resulting from the procedure undertaken. Q-matrix misspecification and attribute relationships were manipulated in the simulation study. Six types of Q–matrix misspecifications were created by underspecifying 1% and 5% of the indices, overspecifying 1% and 5% of the indices, and balanced misspecification of 2% and 6% of the indices. Two types of attribute relationships were created by modeling higher order structure with 2PL and failing to model the higher order structure. Finally, Tatsuoka's fraction subtraction data were used to illustrate the implementation of the PPMC method.

6.1 Summary of the Simulation Study

Recovery of the attribute and item parameters were investigated for the true Q–matrix and the misspecified Q–matrices. Attribute parameters from both true and false Q–matrices were close to generating parameters and each other. The true Q–matrix produced slightly better recovery than the misspecified Q–matrices. However, standard deviations of the estimates with the true Q–matrix were smaller for attribute difficulty and discrimination parameters. The true Q–matrix and the true model produced the smallest RMSE for attribute discrimination and difficulty parameters. The 5% underspecified Q–matrix with the higher order DINA model produced the largest RMSE for both of the attribute parameters for most of the attributes.

Recovery of item parameters which are guessing and slipping were investigated. Estimated item parameters were close to the generating values. The difference of average estimated parameter values from the generating values did not vary much across 14 conditions. However the true Q–matrix and the true model combination produced the smaller standard deviation than the rest. RMSEs for guessing parameters highlighted the overspecified items with higher values. Similarly, RMSEs produced larger values for underspecified items. The guessing parameter for an item was estimated with larger bias when the Q–matrix index or indices for that item were changed from 0 to 1. On the other hand, when indices were changed from 1 to 0, slipping parameters were estimated with larger bias. Item parameter estimates were not estimated with larger bias with the independence DINA model than the higher order DINA model. Thus, the attribute relationships did not seem to impact the item parameters.

χ^2_{lt} was calculated for each item based on latent classes. The average values were all smaller than the reference value of χ^2_{28} for the higher order DINA model and smaller than the reference value of χ^2_{30} for the independence DINA model for all the correctly specified items. Average χ^2_{lt} values were all higher than the reference values for the items whose attribute specifications were manipulated by a false Q–matrix. Changing the item specification in

the Q–matrix did not alter the overall dynamics and stayed local to the items misspecified. However, this observation was made by changing up to 6% of the indices. It would be useful to investigate whether increasing the percentage of misspecification by the Q–matrix and affected changes in the item parameters.

Type I error rate for χ^2_{lt} was calculated based on all the correctly specified items and the higher order DINA model. It was smaller than the suggested range for Type I error rate. Power for χ^2_{lt} was calculated over items that were misspecified. χ^2_{lt} rejected all the misspecified items as desired for both the higher order and the independence DINA models.

PPP-values based on χ^2_{lt} were calculated for each item. Their average values were not at the extreme values of the [0, 1] interval for correctly specified items when higher order DINA model was fit to the data. Thus PPP-values based on χ^2_{lt} did not flag correctly specified items. Similarly, PPP-values based on χ^2_{lt} did not flag any one of the correctly specified items as problematic, when the independence DINA model was fit to the data. Average PPP-values based on χ^2_{lt} indicated misfit for all of the items whose attributes were misspecified.

Type I error rate for χ^2_{lt} was calculated based on all the correctly specified items and higher order DINA model. It was smaller than the suggested range for Type I error rate. Power for χ^2_{lt} was calculated over items that were misspecified. χ^2_{lt} rejected all the misspecified items as desired for both higher order and independence DINA models.

χ^2_{raw} was calculated for each item based on raw score groups. Average values of χ^2_{raw}s were not always smaller than the reference value of χ^2_{17} for higher order DINA model. Similarly, average χ^2_{raw} values were not always smaller than the reference value of χ^2_{19} for the independence DINA model for all the correctly specified items. Thus, some of the correctly specified items are flagged as problematic by the χ^2_{raw} values. Average χ^2_{raw} values are larger than the reference value for all the misspecified items. Changing some of the item specifications by the Q–matrix did not impact all the items and stayed local to the items misspecified. However, this observation was made by changing up to 6% of the indices.

Type I error rate for χ^2_{raw} was calculated based on all the correctly specified items and higher order DINA model. It was larger than the suggested range for Type I error rate. Power for χ^2_{raw} was calculated over items that were misspecified. χ^2_{raw} rejected the misspecified items majority of the time for both higher order and independence DINA model and produced almost perfect power value of .999.

PPP-values were calculated based on χ^2_{raw} discrepancy measure for each item. Some of the correctly specified items were flagged as not fitting by these PPP-values. Average PPP-values based on χ^2_{raw} were in the upper half of the [0,1] interval. Some of the correctly specified items were flagged as problematic by the PPP-values calculated based on χ^2_{raw} values. Some items were identified as problematic more often than the others. Some misspecified items were not flagged as not fitting by the PPP-values calculated based on the χ^2_{raw} with the higher-order DINA model. On the other hand, all of the misspecified items were flagged as not fitting by the PPP-values calculated based on the χ^2_{raw} with the independence DINA model. More items were identified as misfitting by the PPP-values based on χ^2_{raw} with the independence DINA model than the higher order DINA model.

Type I error rate for PPP-values based on χ^2_{raw} was calculated over the correctly specified items with higher order DINA model. It was larger than the suggested range for Type I error rate. Power for PPP-values based on χ^2_{raw} was calculated over items that were misspecified. PPP-values based on χ^2_{raw} rejected the misspecified items a majority of the time for both higher order and independence DINA model and produced almost perfect power value of .99.

Overall fit was evaluated with PPP-values calculated based on χ^2_{lt} and χ^2_{raw}. PPP-values based on χ^2_{lt} did not reject the model when true Q–matrix and higher order DINA model was fit to the simulated data. Type I error rate was not committed when the overall fit was evaluated with the PPP-values calculated based on χ^2_{lt} in this study. Model was rejected most of the time, however, by the PPP-values calculated based on χ^2_{lt}, when more than 1% of the items are misspecified. When only 1% of the items are overspecified by the Q–matrix,

PPP-values based on χ^2_{lt} did not reject the model. When 1% of the items were underspecified by the Q–matrix, PPP-values based on χ^2_{lt} rejected the model only 20% of the time.

PPP-values based on χ^2_{raw} rejected the model when true Q–matrix and higher order DINA model is fit to the simulated data. Thus Type I error rate was equal to 1 which is unacceptable. Model is rejected all the time by the PPP-values calculated based on χ^2_{raw} for all the misspecification and higher-order structure conditions. As PPP-values based on χ^2_{raw} rejected all the models, Type II error is not committed by these. However, with the unacceptable Type I error rates, power does not hold any meaning.

6.2 Limitations and Suggestions

Type I error rate and power for χ^2 calculated based on latent classes showed a clear pattern from the simulation study results, with almost zero Type I error and almost perfect power. Even though assuming a χ^2 distribution for this discrepancy measure produced low Type I error rate and almost perfect power, its distribution was not evaluated in this study. Neither was the distribution of χ^2 based on latent classes evaluated in previous studies. Investigation of the distribution of χ^2_{lt} would be useful.

χ^2 calculated based on raw score groups did not produce acceptable Type I error rate and power. It is also the case that raw score is not sufficient for estimating an examinee's mastery status on attributes or his\her general latent ability with the independence or the higher order DINA model. The observed and expected proportions of correct responses were calculated using the sum of the expected correct responses for each examinee in each raw score group. The sums of expected correct responses were not the same for the examinees who had the same raw score. The probability of correct response given the total raw score could be calculated and used for the calculation of χ^2. Calculating χ^2 based on raw scores using the conditional probability of correct response given the total raw score could improve the Type I error rate and power of this statistic.

PPP-values based on χ^2_{lt} or χ^2_{raw} do not require distributional assumptions. PPP-values based on χ^2_{raw} did not produce acceptable Type I error rate or power. This was consistent with the use of χ^2_{raw} statistic. Calculating χ^2_{raw} by using the conditional probability of correct response given the total raw score and using this χ^2_{raw} to calculate the PPP-value could improve the PPP-values calculated based on χ^2_{raw}.

PPP-values need to reflect some features of the data not directly addressed by the probability. Other discrepancy measures might be used to calculate PPP-values instead of the χ^2s evaluated in this study. Association among the items were used to calculate PPP-values in previous studies. These were not investigated further here. They need to be investigated in terms of their Type I error rate and power values. One of these measures for the item associations was the odds ratios. Another possible measure is the point biserial correlations. Both of these reflect the associations among the items; the discrepancy would measure whether the model recovers this association.

One limitation of this study is the small number of the replications for each condition for the simulation study. Due to the long calculation time with the available software, simulation study used 50 replications for each condition. This number could be improved in the future to provide better evaluation of the statistics presented in this study. Another limitation is the small number of attributes. This number of attributes could be increased in the future as well with the availability of a faster software.

The current study fixed some properties of the data such as the sample size, item number within a test, and number of attributes measured by the test. To better understand how the model and item fit indices presented in this study perform, varying sample size, number of items, and number of attributes would be useful to study. Another aspect of the design was the use of only one Q–matrix which was moderate in terms of its density. It would be interesting to vary the density of Q–matrix for the model fit evaluation as it has consequences for the bias of the estimated model parameters due to Q–matrix misspecification. It is also possible to have different models for the relationship among the attributes with DINA

model. This study investigated only independence model which assumed the attributes are independently distributed and the two parameter logistic model for the relationship among the attributes.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification.*IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix.*Applied Psychological Measurement, 17*, 201–210.

Bayarri, M. J., & Berger, J. P. (1997). Measures of surprise in Bayesian analysis. *ISDS Discussion Paper 97-46*. Durham, NC: Duke University.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

DiBello,L. V., Stout, W. F., & Roussos, L. A. (1995). Unified congnitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, D. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis.*Psychometrika, 69*, 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624.

Gelman, A., Carlin, J. B., Stem, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. New York: Chapman & Hall.

Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica, 6*, 733–807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–72.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B, 29*, 83–100.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items, *Journal of Educational Measurement, 26*, 333–352.

Habenicht-Kunina, O. (2010). Theoretical and practical considerations for implementing diagnostic classification models (Doctoral dissertation, Humboldt–Universität of Berlin). Retrieved from http://edoc.hu-berlin.de/dissertationen/

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

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Henson, R. A., 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.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved November 20, 2011, from http://www.stat.cmu.edu/˜brian/nrc/cfa/

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 23*, 258–272.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Pschological Measurement, 33*, 519–537.

Li, F. (2008). *A modified higher–order DINA model for detecting differential item functioning and differential attribute functioning.* (Doctoral dissertation). Retrieved from http://athenaeum.libs.uga.edu/

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true–score and equipercentile observed–score "equatings". *Applied Pschological Measurement, 8*, 453–461.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Meng, X-L. (1994). Posterior predictive p-values. *Annals of Statistics, 22*, 1142–1160.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley & Sons.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6*, 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statisticia. *Annals of Statistics, 12*, 1151–1172.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement*. New York, NY: Guilford.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Sinharay, S. (2003). *Bayesian item fit analysis for dichotomous item response theory models (ETS RR-03-34)*. Princeton, NJ: ETS.

Sinharay, S. (2004). *Model diagnostics for Bayesian networks* (ETS RR-04-17). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–33.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375–394.

Sinharay, S., & Almond, R. J. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement, 67*, 239–257.

Sinharay, S., & Johnson, M. (2003, October). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (ETS RR-03-28). Princeton, NJ: Educational Testing Service.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.

Smith, B. (2005). BOA: Bayesian output analysis program user manual (Version 1.1). [Computer Software]. The University of Iowa, *http://www.public-health.uiowa.edu/boa*.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn D. J. (2003). *WinBUGS Version – 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/

Tatsuoka, K. K. (1984, January). *Analysis of errors in fraction addition and subtraction problems.* (NIE Final report for Grant no. NIE-G-81-0002). Urbana: University of Illinois, Computer-based Education Research Laboratory. (ERIC Document Reproduction Service No. ED 257665)

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 51*, 337–350.

Templin, J. (2004) *Generalized linear mixed proficiency models for cognitive diagnosis.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Tiao, G. C., & Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions. *Biometrica, 80*, 623–641.

Appendix A

WINBUGS Code for the Analysis of Tatsuoka's fraction subtraction data with Higher-order DINA Model

χ^2 is calculated based on latent classes.

```
model {
for (j in 1:NE){
for (k in 1:NS){
    logit(pi[j,k])< - a[k]*theta[j]-beta[k]
    alpha[j,k] ~ dbern(pi[j,k])
}
theta[j] ~ dnorm(0,1) }
for (j in 1:NE) {
    for ( i in 1:NI) {
        for (k in 1:NS) {
            x[i,j,k]<-pow(alpha[j,k],q[i,k]) }
            eta[i,j]<-x[i,j,1]*x[i,j,2]*x[i,j,3]*x[i,j,4]*x[i,j,5]*x[i,j,6]*x[i,j,7]*x[i,j,8]
            p[i,j]<-pow(1-s[i],eta[i,j])*pow(g[i],1-eta[i,j])
            r[j,i] ~ dbern(p[i,j]);
            r.rep[j,i] ~ dbern(p[i,j]);
        }
    }
for(k in 1:NS){
    a[k] ~ dnorm(0,1)I(0,)
```

```
    beta[k] ~ dnorm(0,1) }
for(i in 1:NI){
    g[i] ~ dbeta(Ug,Sg)
    s[i] ~ dbeta(Us,Ss)
}
Ug ~ dunif(.1,.9)
Sg ~ dunif(.5,10)
Us ~ dunif(.1,.9)
Ss ~ dunif(.5,10)
for (j in 1:NE){
    right[j]<-sum(p[,j]+1)
    rawtot[j]<-sum(r[j,]+1)
    rawtot.rep[j]<-sum(r.rep[j,]+1)
    resid[j]<-rawtot.rep[j]-right[j] }
for (i in 1:NE){
    eqc[i] <- (128*alpha[i,1])+(64*alpha[i,2])+...+8*alpha[i,5])/8+1
}
for (j in 1:NI){
    for (k in 1:32){
        for (i in 1:NE){
            resplt[k,j,i]<-(equals(eqc[i],k))*r[i,j]
            resplt.rep[k,j,i]<-(equals(eqc[i],k))*r.rep[i,j]
            expresplt[k,j,i]<-(equals(eqc[i],k))*p[j,i]
        }
        sumcorrectlt[k,j]<-sum(resplt[k,j,])
        sumcorrectlt.rep[k,j]<-sum(resplt.rep[k,j,])
        sumrightlt[k,j]<-sum(expresplt[k,j,])
```

```
        }
    }
    for (k in 1:32){
        for (i in 1:NE){
            Nlt[k,i]<-(equals(eqc[i],k))
            }
            SumNlt[k]<-sum(Nlt[k,])
        }
    for (k in 1:32){
        for (i in 1:NE){
obsprocor[k,j]< (1-equals(SumNlt[k],0))*sumcorrectlt[k,j]/(SumNlt[k]+equals(SumNlt[k],0))
obsprocor.rep[k,j]<-(1-equals(SumNlt[k],0))*(sumcorrectlt.rep[k,j]/(SumNlt[k]+equals(SumNlt[k],0)))
expprocor.rep[k,j]<-(1-equals(SumNlt[k],0))*(sumrightlt[k,j]/(SumNlt[k]+equals(SumNlt[k],0)))
        }
    }
    for (j in 1:NI){
        for (k in 1:32){
        chisqtop[k,j]<-SumNlt[k]*(pow(obsprocor[k,j]-expprocor[k,j],2))
        chisqbot[k,j]<-(expprocor[k,j]*(1-expprocor[k,j]))
        chisqtop.rep[k,j]<-SumNlt[k]*(pow(obsprocor.rep[k,j]-expprocor[k,j],2))
        chisqbot.rep[k,j]<-(expprocor[k,j]*(1-expprocor[k,j]))
        chisq[k,j]<-(1-equals(chisqbot[k,j],0))*(chisqtop[k,j]/(chisqbot[k,j]+equals(chisqbot[k,j],0)))
        chisq.rep[k,j]<-(1-equals(chisqbot.rep[k,j],0))*(chisqtop.rep[k,j]/(chisqbot.rep[k,j]+
    equals(chisqbot.rep[k,j],0)))
            }
            chi2[j]<-sum(chisq[,j])
            chi2.rep[j]<-sum(chisq.rep[,j])
```

```
        itemfitp[j]<-step(chi2[j]-chi2.rep[j])
        }          }
list(NE=536, NI=20, NS=6,
q=structure(.Data=c(
1,1,1,0,0,0,
1,1,1,0,0,0,
....
0,1,1,1,0,1),.Dim=c(20,6)),
r = structure(.Data = c(0,0,0,
```

Appendix B

Convergence diagnostics

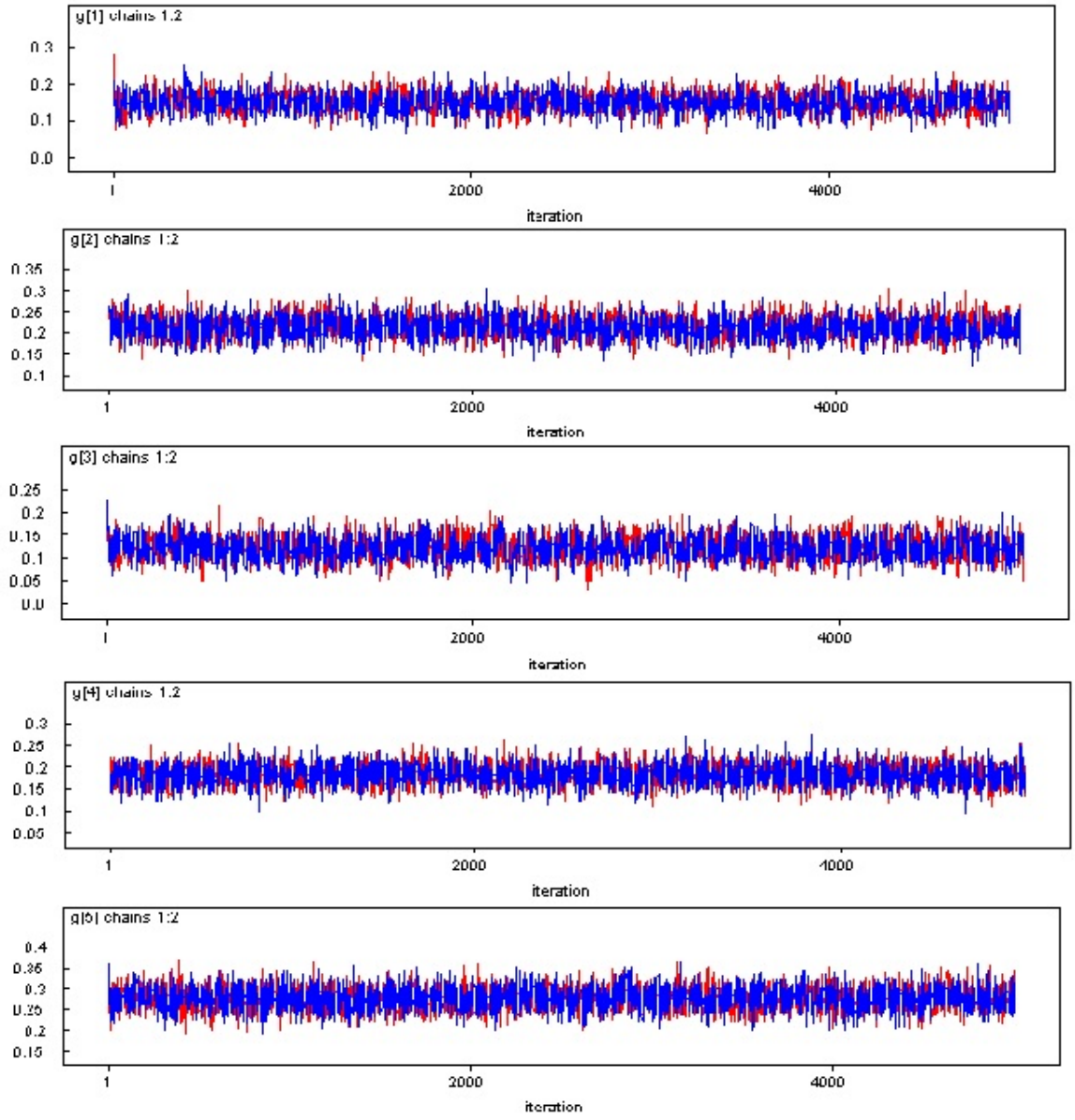

Figure B.1: The trace plots for g guessing parameter for the first 5 items for the higher–order DINA model with true Q–matrix

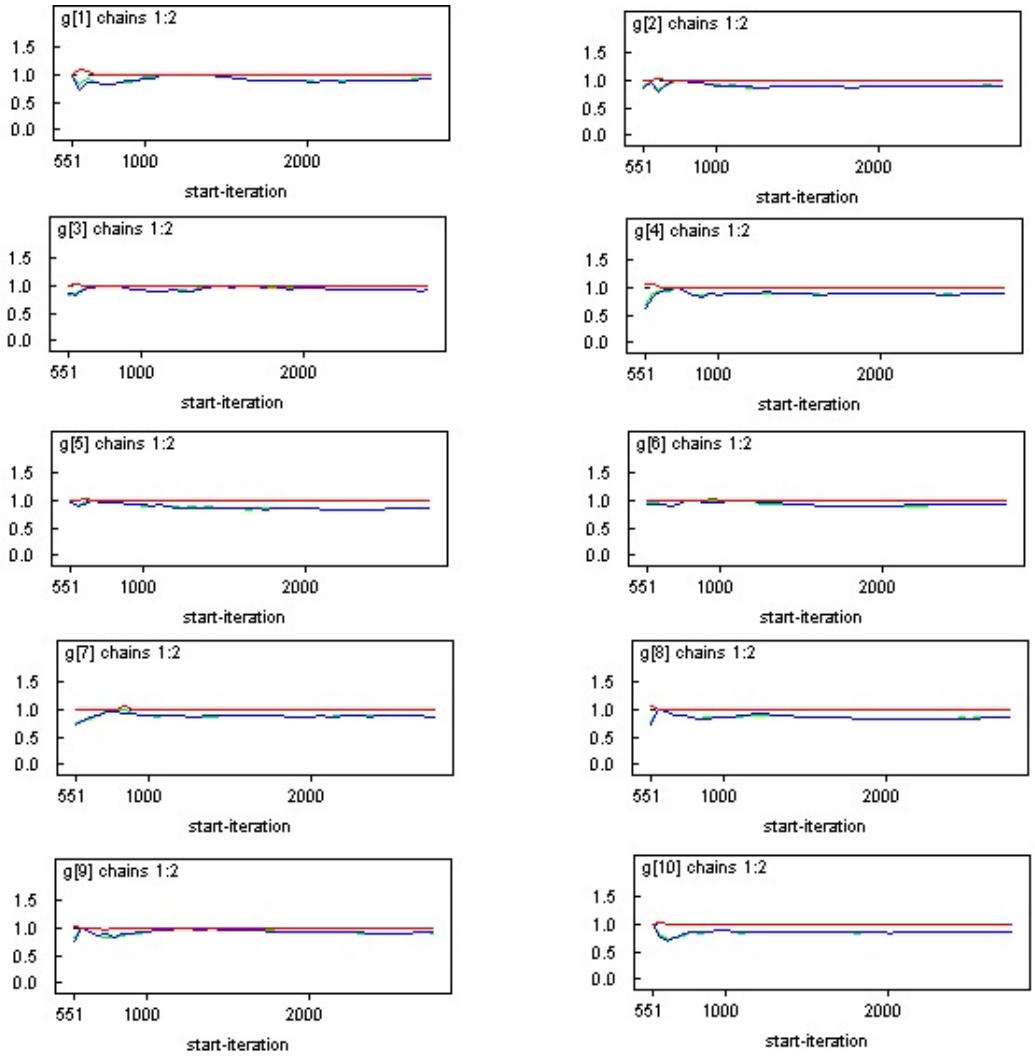

Figure B.2: The line plots for Gelman and Rubin statistic for g, guessing parameter for the first 10 items for the higher–order DINA model with true Q–matrix

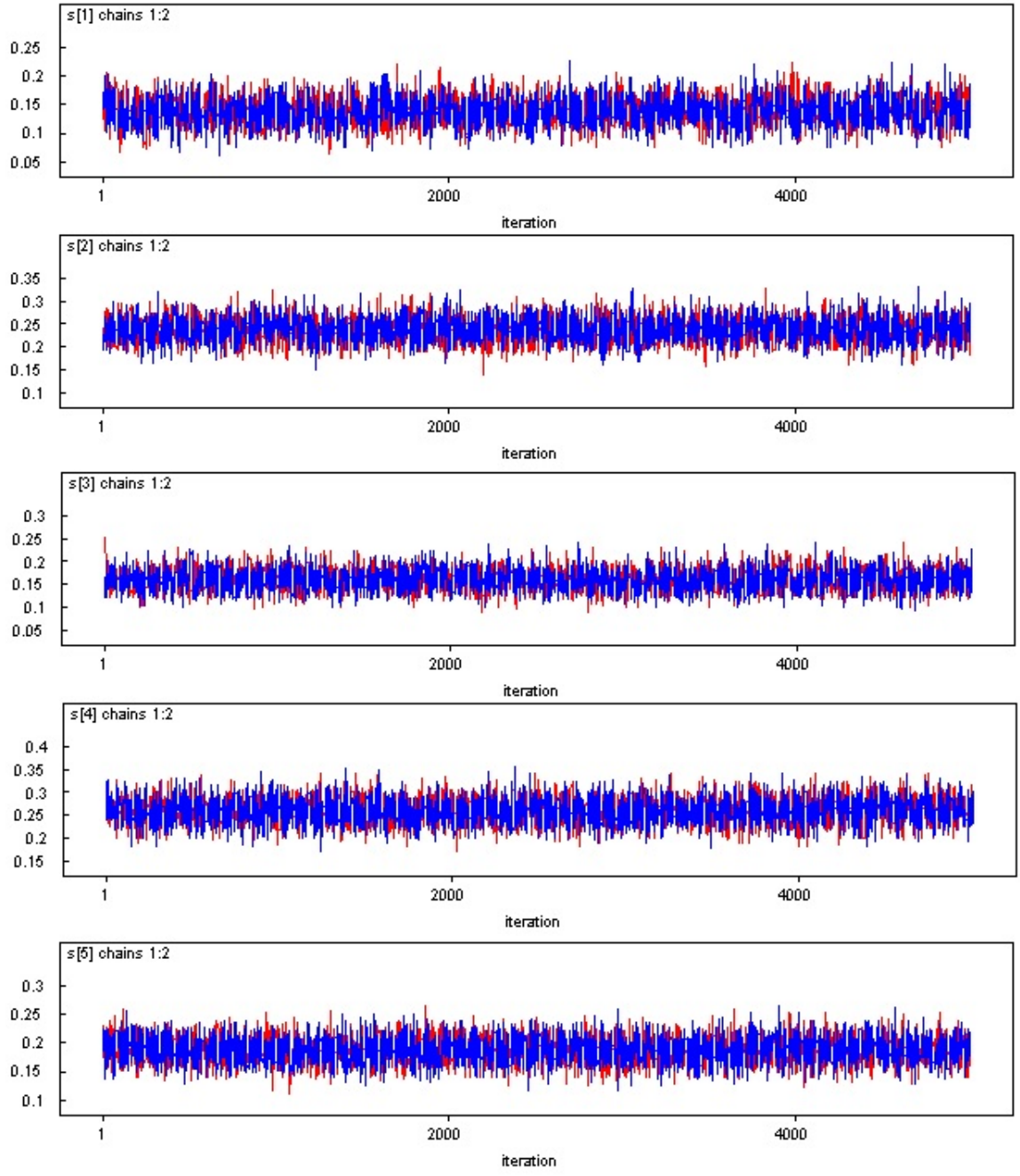

Figure B.3: The trace plots for s slip parameter for the first 5 items for the higher–order DINA model with true Q–matrix

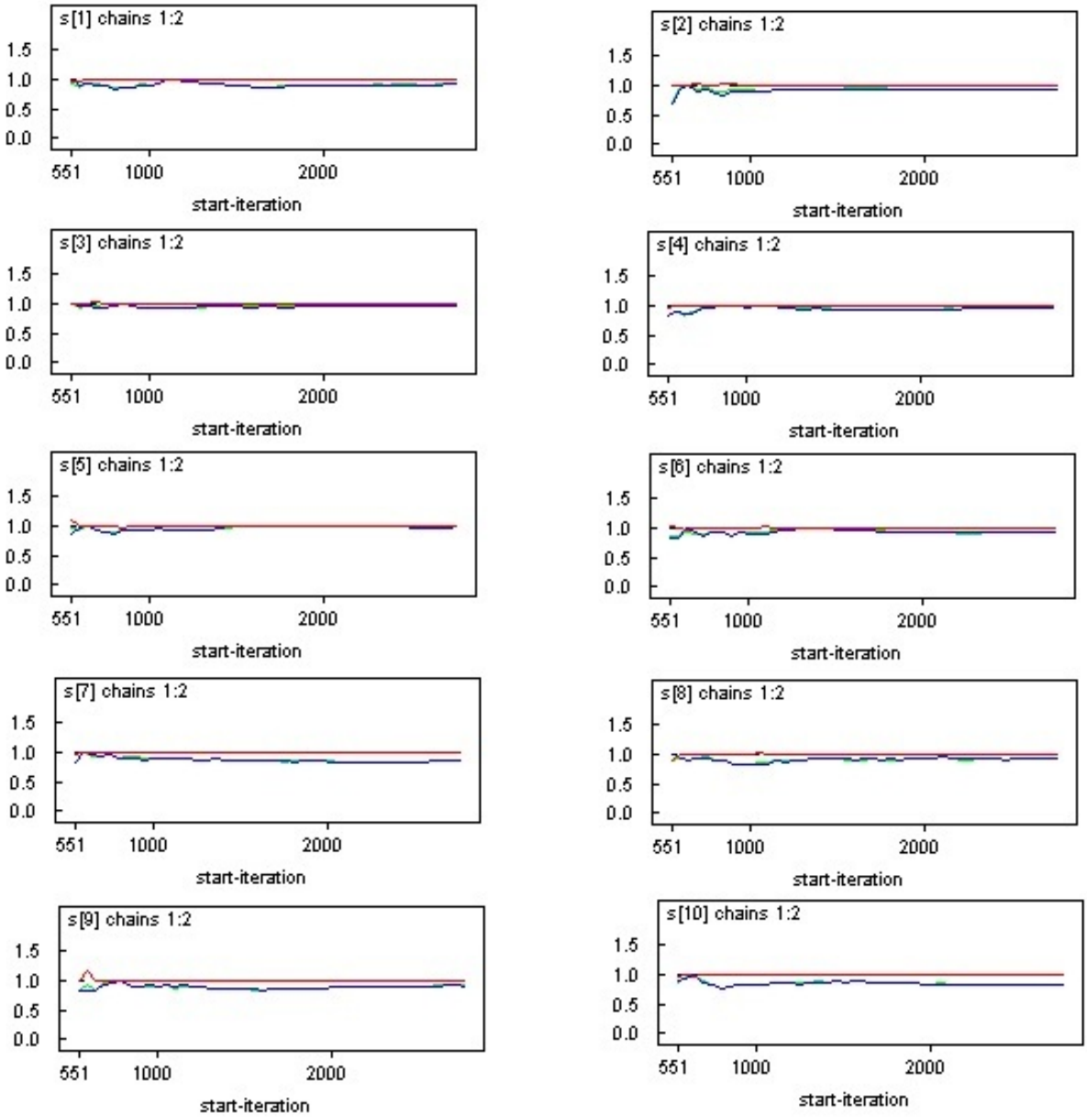

Figure B.4: The line plots for Gelman and Rubin statistic for s, slip parameter for the first 10 items for the higher–order DINA model with true Q–matrix

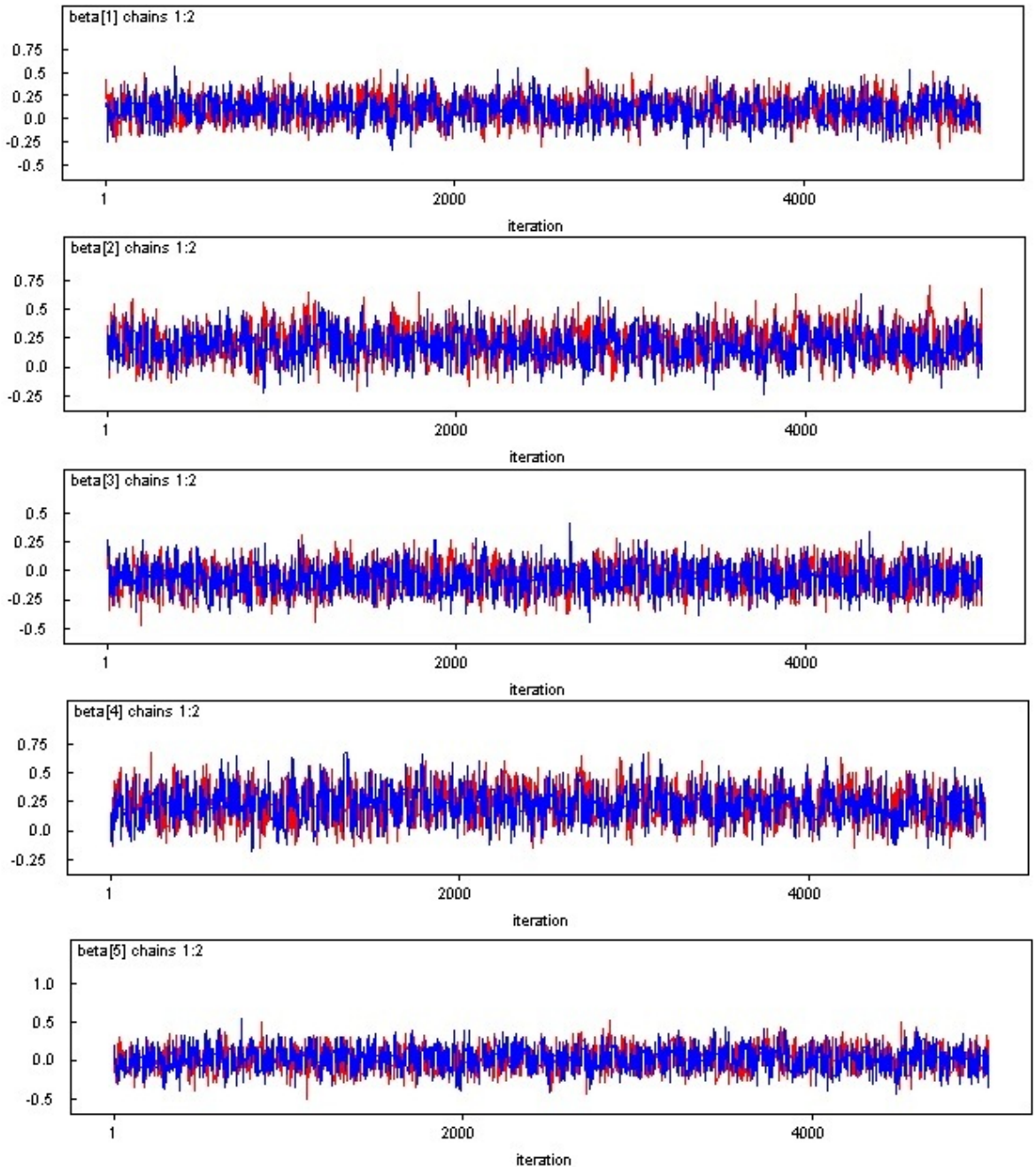

Figure B.5: The trace plots for beta attribute difficulty parameters for five attributes for the higher–order DINA model with true Q–matrix

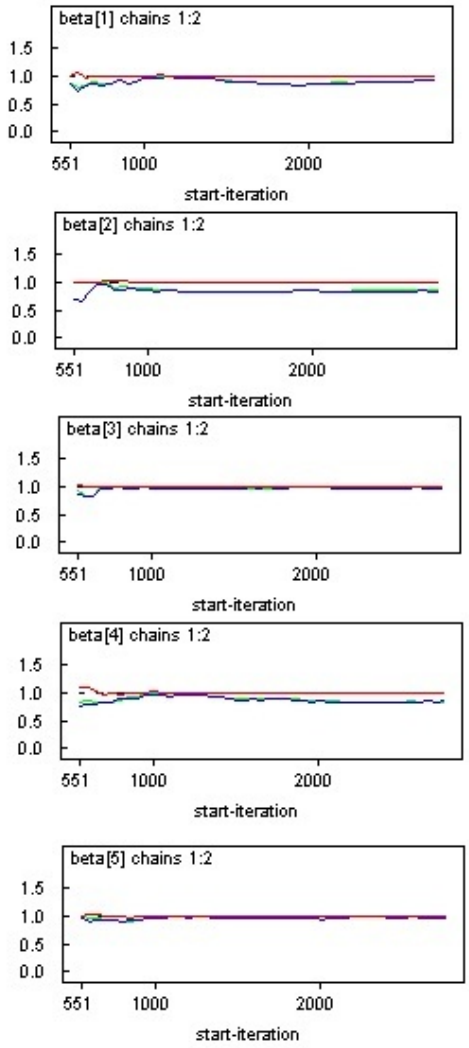

Figure B.6: The line plots for Gelman and Rubin statistic for beta attribute difficulty parameters for five attributes for the higher–order DINA model with true Q–matrix

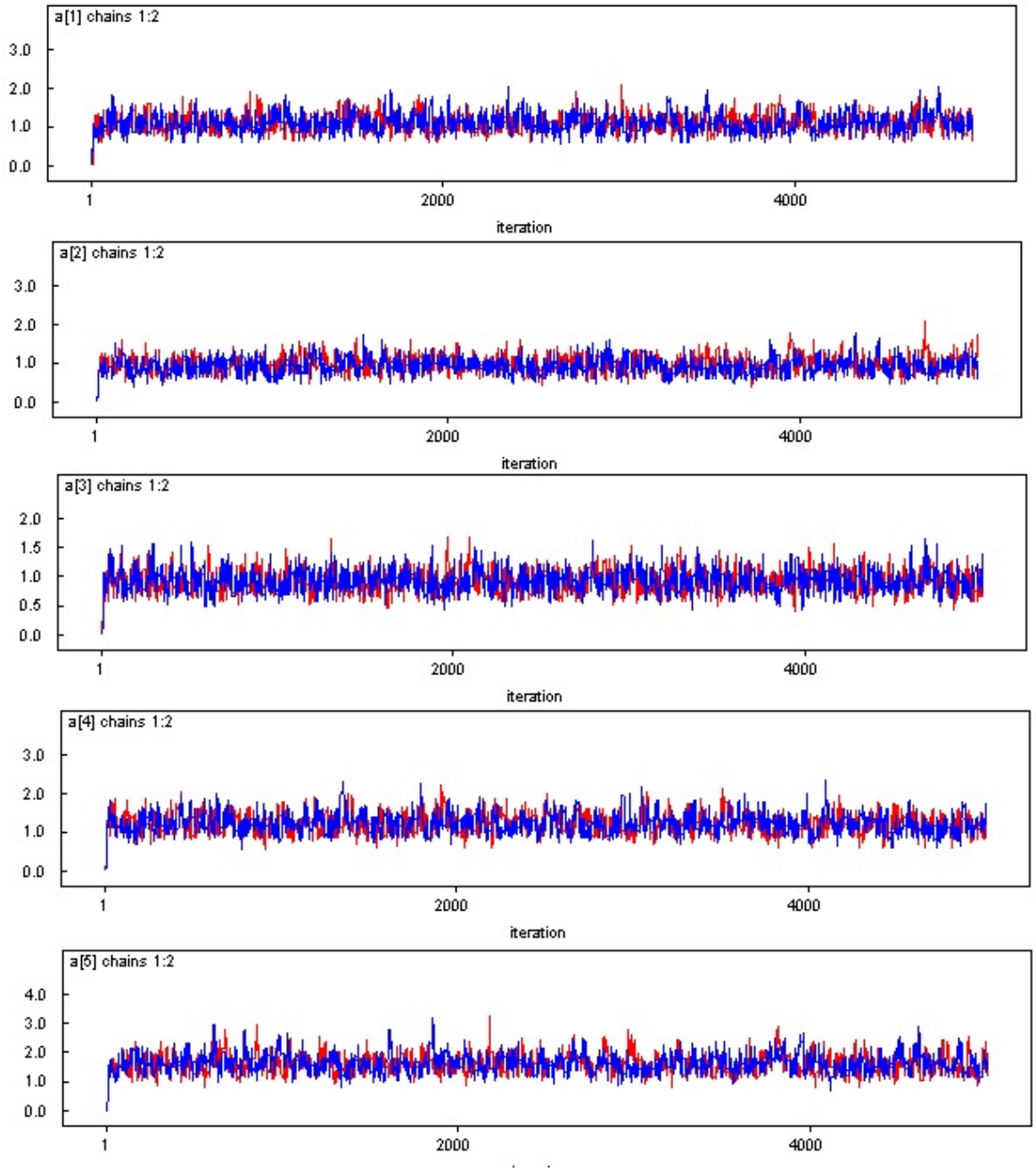

Figure B.7: The trace plots for a attribute discrimination parameters for five attributes for the higher–order DINA model with true Q–matrix

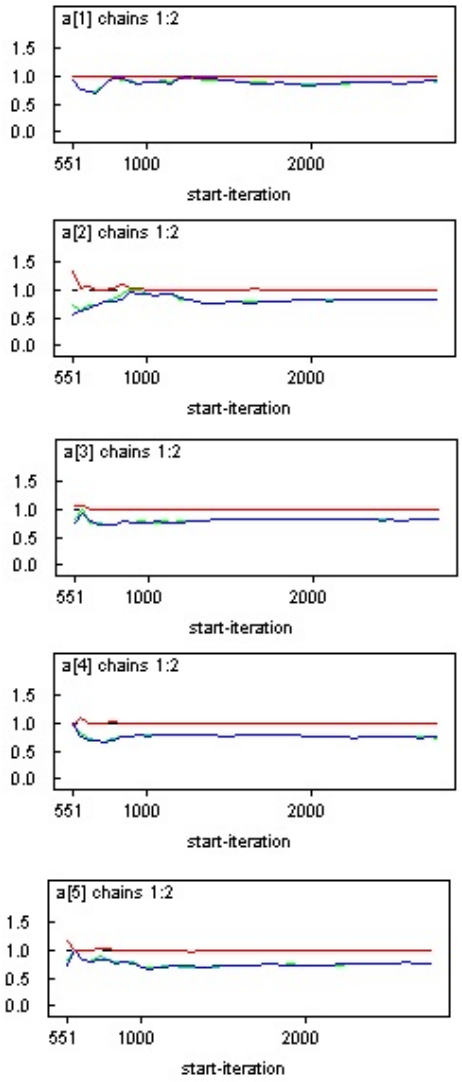

Figure B.8: The line plots for Gelman and Rubin statistic for a attribute discrimination parameters for five attributes for the higher–order DINA model with true Q–matrix

Appendix C

Type I Error Rates Across Items

Table C.1
Proportion of χ^2_{raw} Indices Greater Than $p = .05$ Across Items for Higher Order DINA

Type I Error Rates for Seven Conditions Across Items

	Overall	Q_0	Q_{O1}	Q_{O5}	Q_{U1}	Q_{U5}	Q_{B2}	Q_{B6}
1	0	0	0	*	0	0	0	0
2	.09	.08	.08	.12	.06	.08	.14	.04
3	.01	0	0	0	0	.04	0	0
4	.1	.06	.08	*	.14	.12	*	.08
5	.17	.16	.18	.22	.18	.16	.16	.14
6	.67	.7	.66	.92	.66	.48	.62	*
7	.46	.52	.52	.54	.52	.38	.46	.30
8	.84	.84	.84	.88	.80	.86	.82	*
9	.19	.14	.20	.38	.10	.06	.32	.14
10	.08	.08	.12	.10	.10	.10	.02	.06
11	.31	.3	.30	.36	.30	*	.32	.30
12	.28	.24	.24	*	.24	*	.26	.40
13	.56	.56	.50	.52	.58	.60	.60	.56
14	0	0	0	0	0	0	*	0
15	.02	.02	.02	.02	.02	*	.02	*
16	.61	.58	.64	.56	.58	.72	.58	*
17	.12	.14	.14	.14	.12	.12	.08	.10
18	.09	.06	.08	.08	*	*	.08	.14
19	0	0	0	0	0	*	0	*
20	.18	.12	*	*	.24	.20	.14	*

* indicates the items for which Type I error is not calculated

Table C.2
Proportion of PPP-values based on χ^2_{raw} Indices Smaller Than $p = .05$ or Greater Than $p = .95$ or Across Items for Higher Order DINA

Type I Error Rates for Seven Conditions Across Items								
	Overall	Q_0	Q_{O1}	Q_{O5}	Q_{U1}	Q_{U5}	Q_{B2}	Q_{B6}
1	.01	.02	.02	*	.02	0	0	.02
2	.05	.04	.04	.08	.04	.02	.08	.04
3	.07	.08	.08	.08	.02	.06	.06	.01
4	.08	.06	.08	*	.12	.10	*	.06
5	.11	.12	.12	.14	.14	.12	.10	.04
6	.56	.54	.56	.86	.54	.38	.50	*
7	.37	.42	.40	.42	.44	.22	.42	.24
8	.73	.72	.72	.76	.66	.74	.78	*
9	.15	.12	.18	.34	.08	.04	.20	.10
10	.11	.14	.14	.12	.14	.12	.02	.06
11	.20	.22	.20	.20	.20	*	.14	.22
12	.20	.18	.18	*	.18	*	.22	.24
13	.49	.50	.46	.38	.58	.52	.54	.44
14	0	0	0	0	0	0	*	0
15	.02	.02	.02	.02	.02	*	.02	*
16	.52	.54	.54	.46	.48	.62	.48	*
17	.07	.08	.10	.08	.08	.06	.04	.06
18	.08	.08	.08	.08	*	*	.06	.08
19	0	0	0	0	0	*	0	*
20	.10	.10	*	*	.10	.10	.10	*

* indicates the items for which Type I error is not calculated