Title
Summed score likelihood based indices for testing latent variable distribution fit in item response theory

Author
Zhen, Li

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Summed Score Likelihood Based Indices for Testing Latent Variable Distribution Fit in Item Response Theory

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Zhen Li

2015

Abstract of the Dissertation

Summed Score Likelihood Based Indices for Testing Latent Variable Distribution Fit in Item Response Theory

by

Zhen Li

Doctor of Philosophy in Education

University of California, Los Angeles, 2015

Professor Li Cai, Chair

In item response theory (IRT), the underlying latent variables are typically assumed to be normally distributed. If the assumption of normality is violated, the item and person parameter estimates can become biased. Therefore, it is necessary in practical data analysis situations to examine the adequacy of this assumption in an effective manner. There is a recent surge of interest in limited-information overall goodness-of-fit test statistics for IRT models (see e.g., Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Joe & Maydeu-Olivares, 2010; Cai & Hansen, 2013), but their appropriateness for diagnosing latent variable distributional fit has not been studied.

The approach undertaken in this research is to use summed score likelihood based indices. The idea itself is not new (see e.g., Ferrando & Lorenzo-Seva, 2001; Hambleton & Traub, 1973; Lord, 1953; Ross, 1966; Sinharay, Johnson, & Stern, 2006; Thissen & Wainer, 2001), but this study recasts the problem using the framework of limited-information goodness of fit testing. The summed score based indices can be viewed as a particular form of reduction of the full underlying multinomial that are potentially sensitive to the latent variable distributional

misspecifications.

Results from a pilot study (Li & Cai, 2012) show that summed score likelihood based indices enjoy high statistical power for detecting latent variable distributional assumption violations, and are not sensitive (correctly) to other forms of model misspecification such as unmodeled multidimensionality. Meanwhile, the limited-information overall fit statistic M_2 (Maydeu-Olivares & Joe, 2005) has relatively low power against latent variable non-normality. However, technically the statistical indices proposed by Li and Cai (2012) don't follow an exactly chi-squared distribution. They proposed a heuristic degrees of freedom adjustment, but more rigorous justifications could be developed along the lines of Sattora-Bentler type moment adjustment popular in structural equation modeling (Satorra & Bentler, 1994). In IRT, the moment adjustment approaches have been used by Cai et al. (2006) and Maydeu-Olivares (2001).

The major methodological contributions of my dissertation come from simulation studies that examine the calibration and power of the moment adjusted test statistics across various conditions: number of items, sample size, item type, generating latent variable distribution, and the values of generating item parameters. The performance of these fit statistics are also compared with the limited-information overall fit statistic M_2 (Maydeu-Olivares & Joe, 2005). Simulation study results show that the proposed moment-adjusted statistics improve upon the unadjusted statistics in the null and alternative conditions, especially when generating item parameters are dispersed. Finally, performance of the indices is illustrated with empirical data from educational and psychological assessment development projects.

The dissertation of Zhen Li is approved.

Steven Reise

Mark Patrick Hansen

Michael Seltzer

Li Cai, Committee Chair

University of California, Los Angeles

2015

to my mom & sister

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisor and committee chair, Professor Li Cai, for providing me academic guidance, scholarship, and support. I could not have accomplished this dissertation without him. Many times, he motivates me to learn new things and corrects my mistakes with great patience.

I am sincerely thankful to my other committee members for their advice and feedback: Professor Mike Seltzer, Professor Mark Hansen, and Professor Steven Reise.

I thank my seniors from the psychometrics lab at UCLA: Professors Mark Hansen, Ji Seung Yang, Scott Monroe, Carl Falk; Dr. Moonsoo Lee; and Larry Thomas, for their generous help in the past four years. I thank my best cohort in SRM: Megan Kuhfeld, Jenn Ho, Liz Perez-LoPresti, Danny Dockterman, Kevin Schaaf, and Jason Tsui. I thank my other colleagues Seungwon Chung and Julie Liao. And I thank my friends I met in graduate schools: Professors Wei Tian, Sihan Xiao and Xiaochen Chen, for their encouragement.

I thank my family for their love and support.

Vita

EDUCATION

2008	Bachelors of Science, Beijing Normal University, Beijing, China. Psychology.
2011	Masters of Education, Beijing Normal University, Beijing, China. Psychometrics.

WORK

Summer 2009	Research Intern, National Education Examinations Authority, Beijing, China
2007-2011	Research Assistant, National Assessment of Educational Quality Center, Beijing, China
Summer 2014	Research Intern, CTB, Monterey, USA.
2011-Present	Graduate Student Researcher, University of California, Los Angeles, USA.

Publications

Cai, L. & Li, Z. (in preparation). Summed score likelihood based indices for examining latent variable distribution fit in Item Response Theory.

Edelen, M.O., Tucker, J. S., Shadel, W. G., Stucky, B. D., Cerully, J., Li, Z.,

Hansen, M., & Cai, L. (2014). *Development of the PROMIS health expectancies of smoking item banks.* Nicotine & Tobacco Research, 16(3), S223-S231.

Hansen M., Cai. L., Monroe, S., & Li, Z. (2014). *Limited-information goodness-of-fit testing of diagnostic classification item response theory models.* CRESST Report No. 840. Los Angeles, CA: UCLA.

Li, Z. & Cai, L. (in preparation). Item response growth modeling: A new application of multilevel factor model to analyze longitudinal item response data.

Li, Z. & Smith, J. (in preparation). Detecting aberrant behaviors with a hierarchical lognormal response time model.

Tucker, J. S., Shadel, W. G., Edelen, M. O., Stucky, B. D., Li, Z., Hansen, M., & Cai, L. (2014). *Development of the PROMIS positive emotional and sensory expectancies of smoking item banks.* Nicotine & Tobacco Research, 16(3), S212-S222.

Li, Z., Xin, T., & Chen, P. (2010). *Standard setting: Steps, methods and evaluating criteria.* Examinations Research (in Chinese), 2, 83-95.

CHAPTER 1

Introduction

In the area of educational and psychological measurement, item response theory (IRT) models are widely utilized for test development and scoring (Embretson & Reise, 2000; Thissen & Wainer, 2001; Brennan, 2006). In standard IRT models, the distribution of the latent variable is often assumed to be normal (Lord, 1980). Then, the expected probability of a particular item response is modeled as a function of item parameters and the latent variables. This function, known as an item response function (IRF), can be considered the building block for item parameter estimation.

As with any statistical model, many assumptions are made in the application of IRT models, and verification of these assumptions is desirable. For instance, when an IRT model is fitted with maximum marginal likelihood estimation (MMLE, Bock & Lieberman, 1970), the latent variable is often assumed to follow a normal distribution. While many latent variables may arguably follow a normal distribution, this assumption might be unrealistic in other cases (Woods & Thissen, 2006). For example, in large-scale educational assessment, two subpopulations with different means and variances might be grouped together (e.g., English Language Learners and the general population). As a consequence, the population distribution of the proficiency latent variable might be nonnormal.

The purpose of the current research is to develop a set of statistical indices to test the appropriateness of the latent variable normality assumption for realistic data analysis situations. It is valuable because if the assumption of latent vari-

able normality is violated, item parameter estimates might be biased (Woods & Thissen, 2006). For example, in Computer Adaptive Testing (CAT) applications, item parameter estimates are ultilized in item selection and later on for test scoring, thus the bias of parameter estimates could lead to undesirable consequences for both. Furthermore, many quantities of interest, such as score estimates and test information, are functions of the item parameter estimates. Consequently, bias in the parameter estimates will lead to bias in these other quantities.

Certainly, there are alternative approaches to specifying the latent variable distribution in IRT modeling. Researchers have developed several methods to estimate the shape of the latent variable distribution. These include the empirical histogram (EH) method (Bock & Aitkin, 1981), Ramsay-Curve IRT (Woods & Thissen, 2006), and Davidian Curve IRT (Woods & Lin, 2009) as well as its multidimensional extension (Monroe, 2014). The various approaches can provide reasonable item and structural parameter estimates in the case of a non-normal latent variable distribution. However, all of these methods are more computationally demanding than standard IRT estimation (assuming normality) and require special software. Clearly, it would be convenient and useful for practitioners to consult a statistical test of the assumption of normality before more complex models are applied.

In the application of IRT models, overall goodness-of-fit statistics, based on discrepancies between model-implied and observed response pattern probabilities (Reiser, 1996) are calculated to assess model fit. These include Pearsons chi-square test, the likelihood ratio test, and Maydeu-Olivares and Joe's (2006) M_2 index. But recent studies (Li & Cai, 2012; Hansen, Cai, Monroe, & Li, 2015) show that the overall goodness-of-fit test statistics are not sensitive to the violation of the latent variable distribution assumption. Bayesian methods, such as the Posterior Predictive Model Checking (PPMC, Sinharay et al., 2006) approach, could also be applied to check the latent variable distribution assumption. However, predictive

model checking has not been applied to IRT model fit test in the context when item parameters are estimated by MMLE, which remains by far the more popular approach to item parameter estimation in practical settings. An ideal test statistic should be easily computed along with MMLE, and be specifically sensitive to the violation of the assumption of latent variable normality.

Summed score likelihood based indices (Li & Cai, 2012), a special case of limited-information goodness-of-fit tests, were demonstrated to meet these requirements. These indices are derived from a power divergence family of goodness-of-fit (GOF) statistics (Cressie & Read, 1984). Compared with the limited-information goodness-of-fit test statistics, summed score likelihood based statistics are more powerful in detecting latent variable non-normality (Li & Cai, 2012). However, in Li and Cai's (2012) study, a heuristic formula for the degrees of freedom (df) was applied for the statistics, and the empirical distribution of these statistics did not always seem to follow a chi-squared distribution. In this dissertation, I propose two moment-matching approaches (Satorra & Bentler, 1994; Cai et al., 2006) to adjust the summed score likelihood based indice $\bar{X}^2$. An adapted Lord-Wingersky algorithm (Lord & Wingersky, 1984) was developed to calculate the Jacobian matrix useful for finding the first and second moments of $\bar{X}^2$. To distinguish the unadjusted and adjusted statistics, the chi-square statistic with a heuristic df (Li & Cai, 2012) is renamed as $\bar{X}^2_H$. The proposed statistic with first-moment adjustment is named as $\bar{X}^2_{C1}$, while the one with both first- and second-moment adjustments is named as $\bar{X}^2_{C2}$. A simulation study and an empirical study are proposed to examine the performance of $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ under various conditions, in comparison with overall GOF statistics, such as M_2.

My dissertation consists of nine chapters in total. Chapter 2 introduces some typical unidimensional IRT models. Chapter 3 demonstrates the importance of normal latent variable distribution assumption in IRT modeling. Chapter 4 presents popular multinomial goodness-of-fit statistical tests (full-information and

limited-information) for IRT. Chapter 5 introduces the proposed summed score likelihood based indices as a member of the limited-information statistical test family, the moment adjustment procedure, and the adapted Lord-Wingersky algorithm for the Jacobian matrix. Chapter 6 describes the simulation study design and the data generation process. Chapter 7 illustrates the simulation study results. Chapter 8 presents an application of the proposed statistics to two sets of empirical data. Chapter 9 provides a conclusion for the current study and some directions for future exploration.

CHAPTER 2

Item Response Theory Model

The statistical aspects of item response theory can be understood from the generalized linear model formulation (Skrondal & Rabe-Hesketh, 2004) for categorical item level data. In unidimensional IRT models, the examinee's responses to items are assumed to be related to a general underlying dimension, representing proficiency, ability, achievement, preference, etc. Item response probabilities are assumed to be independent conditional on the latent variable (Thissen & Steinberg, 2009). Some commonly used item response theory models are introduced in this chapter.

In standard IRT models, the conditional item response probabilities (also referred to as item tracelines) are represented as a function of item parameters and the latent variable θ. For example, the 3-parameter logistic (3-PL) model can be written as:

$$T_i(1|\theta) = g_i + \frac{1 - g_i}{1 + \exp[-(c_i + a_i\theta)]}, \tag{2.1}$$

where $T_i(1|\theta)$ represents item i's traceline for the 1 category (indicating correct/ endorsement response in most contexts) as a function of θ. The item parameters include: g_i, which is the pseudo-guessing probability for the item (lower asymptote parameter); a_i, which is the slope (discrimination) parameter, and c_i, which is the item intercept parameter. The slope parameter indicates to what extent the item can discriminate individuals on the latent variable continuum. The classical difficulty (threshold) parameter is obtained as $b_i = -c_i/a_i$, which is defined as the value of θ when correct/ endorsement response probability equals to $(1 + g_i) * 0.5$.

If g_i is zero, the model reduces to a 2-parameter logistic (2-PL) model, and if all the item slopes are constrained to be equal to a common slope ($a_i \equiv a$), the 1-parameter logistic (1-PL) model results. For an item with two categories, the incorrect/non-endorsement response probability is equal to $T_i(0|\theta) = 1 - T_i(1|\theta)$.

For an item with K_i ordered polytomous responses, the graded response model (Samejima, 1969) is often utilized. Let the response categories be coded as $k = 0, ..., K_i - 1$. The cumulative response probability for item i in categories k and above is

$$T_i^+(k|\theta) = \frac{1}{1 + \exp[-(c_{ik} + a_i\theta)]}, \tag{2.2}$$

for $k = 1, ..., K_i - 1$. Having defined the boundary cases $T_i^+(0|\theta) = 1$ and $T_i^+(K|\theta) = 0$, the category response probability can be written as

$$T_i(k|\theta) = T_i^+(k|\theta) - T_i^+(k+1|\theta), \tag{2.3}$$

for $k = 0, ..., K_i - 1$. Let U_i be a random variable whose realization u_i is a response to item i. Regardless of the number of categories or the form of the model, the probability mass function of U_i, conditional on θ, is that of a multinomial with trial size 1:

$$P(U_i = u_i|\theta) = \prod_{k=0}^{K_i-1} [T_i(k|\theta)]^{1_k(u_i)}, \tag{2.4}$$

where $1_k(u_i)$ is an indicator function such that

$$1_k(u_i) = \begin{cases} 1: & \text{if } k = u_i, \\ 0: & \text{otherwise.} \end{cases}$$

A 2-PL model is actually a special case of the graded response model when the number of response categories equals to two: correct or incorrect. In addition, there are alternative unidimensional IRT models for polynomial data, such as Muraki's (1992) generalized partial credit (GPC) model and Bock's (1972)

nominal categories model. Details about these models can be found elsewhere (e.g., Thissen & Wainer, 2001). In my dissertation, the graded response model (Samejima, 1969) is applied for polytomous item response data.

CHAPTER 3

The Latent Variable Distribution in IRT

As previously mentioned, in the application of IRT, the population distribution of latent variable $g(\theta)$ is often conveniently assumed to be normal for the purposes of item parameter calibration and latent trait estimation (Thissen & Wainer, 2001). In reality, however, the distribution of the latent variables can be nonnormal. Woods and Thissen (2006) described several potential situations where θ might be nonnormal. For example, as severe symptoms of psychological disorders rarely exist in the general population and most people have low levels of psychopathological symptoms, the population distribution of latent variables reflecting these symptoms may be positively skewed. Another possible cause arises in the situation when the population is heterogeneous. For instance, when two or more subpopulations with different means and variances are grouped together, potentially multimodal population distributions may be the result. Calibrating the items with respect to the combined population renders the normality assumption suspect.

When the assumption of normal latent variable distribution is challenged, nonparametric or semi-parametric estimation methods can be applied to estimate $g(\theta)$ along with item parameters. The EH method is now an established strategy for detecting and correcting latent variable nonnormality in IRT (Bock & Aitkin, 1981; Mislevy, 1984; Woods, 2006). Newer semi-parametric density estimation procedures (e.g., Ramsay Curve IRT, Woods & Thissen, 2006; Davidian Curve IRT, Woods & Lin, 2009) offer more efficient alternatives. Monroe and Cai (2014)

and Monroe (2014) extended these estimation methods for multidimensional IRT models, where semi-parametric densities are estimated for more than one latent variable.

In practice, however, estimating latent variable densities often requires additional computation and specialized software. Although the EH method has been implemented in several IRT software, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and flexMIRT (Cai, 2013), more complex latent variable distributions involve more parameters to be estimated from the data, increasing the need for larger calibration sample sizes to achieve stable estimation. Moreover, even as nonnormal latent densities may be modeled, e.g., using a Ramsay Curve IRT model (Woods & Thissen, 2006), and the relative model fit may be evaluated against a baseline using likelihood ratio tests, it does not circumvent the need for absolute goodness of fit indices to establish the adequacy of the least restrictive model in the class of models being compared (see Maydeu-Olivares & Cai, 2006 for further explanation). It would be highly desirable to use statistical tests to examine the extent to which a normal latent variable distribution may in fact be a reasonable characterization before more computational demanding methods and software programs for semi-parametric density estimation are employed.

In developing such a family of test statistics for latent variable distribution fit, several requirements should be considered. First, the statistics should be easily computed, preferably using only standard byproducts of the MMLE procedure. Second, the statistics should have well-grounded heuristic motivation and theoretical justification. Third, the asymptotic distribution of the statistics under the null hypothesis should be sufficiently accurate. Finally, the statistics should have adequate power that is focused on latent variable distribution assumption violation and sufficient diagnostic specificity, comparing with the overall GOF fit test statistics.

The guiding insight has been provided elsewhere in the literature. For uni-

dimensional IRT modeling, the observed and model-implied summed score distribution can be a basis for inferring the adequacy of the latent variable distribution specification in the IRT model (Thissen & Wainer, 2001). After model fitting, residual summed score probabilities may be used to construct chi-square test statistics. While the idea has been discussed from different standpoints (see Ferrando & Lorenzo-Seva, 2001; Hambleton & Traub, 1973; Lord, 1953; Ross, 1966; Sinharay et al., 2006, among others), the recently developed theory of limited-information goodness-of-fit testing is utilized to formally demonstrate that the summed score likelihood based fit indices proposed here belong to the general family of multinomial limited-information tests.

CHAPTER 4

Multinomial Goodness-of-fit Tests

With no loss of generality, consider a test with N examinees and n dichotomous items. The observed responses can be summarized in an n-dimensional contingency table, with 2^n cells. Statistically, the adequency of an IRT model can always be tested using goodness-of-fit indices by comparing the observed probabilities in each of the 2^n cells with expected probabilities based on the model, with test statistics such as likelihood ratio and Pearson's statistics (Reiser, 1996). Moreover, there are different ways to construct an asymptotically chi-square distributed test statistic with a quardrature form (Cressie & Read, 1984). In this chapter, several quardratic-form statistics for model misfit detection in IRT are described, as well as the relationship among these statistics. Summed score likelihood based indices turn out to belong to the general family of multinomial limited-information tests.

4.1 Multinomial IRT model

Based on n-dimensional contingency table, and under the conditional independence assumption, the IRT model specifies the conditional response pattern probability as follows:

$$P\left(\bigcap_{i=1}^{n} U_i = u_i|\theta\right) = \prod_{i=1}^{n} P(U_i = u_i|\theta). \tag{4.1}$$

Assuming that $g(\theta)$ is the distribution of the latent variable (also known as the prior distribution), the marginal response pattern probability is the following integral:

$$P\left(\bigcap_{i=1}^{n} U_i = u_i\right) = \int \prod_{i=1}^{n} P(U_i = u_i|\theta) g(\theta) d\theta = \pi_{\mathbf{u}}(\boldsymbol{\gamma}), \tag{4.2}$$

where $\mathbf{u} = (u_1, ..., u_n)'$ is the response pattern, and $\boldsymbol{\gamma}$ is a $d \times 1$ vector that collects together the free item parameters from all n items. The parenthetical notation $\pi_{\mathbf{u}}(\boldsymbol{\gamma})$ in Equation 4.2 is used to emphasize the fact that it *is* the model. The marginal response probability depends on the item parameters, the item-level response models, and the assumed latent variable distribution. Recall that the number of categories for item i is two. For n items, the IRT model generates a total of $C = 2^n$ cross-classifications or possible item response patterns in the form of a contingency table. Based on a sample of N respondents, let the observed proportion associated with pattern $\mathbf{u}$ be denoted as $p_{\mathbf{u}}$. The sampling model for this contingency table is a multinomial distribution with C cells and N trials. The multinomial log-likelihood for the item parameters $\boldsymbol{\gamma}$ is proportional to

$$\log L(\boldsymbol{\gamma}) \propto N \sum_{\mathbf{u}} p_{\mathbf{u}} \log \pi_{\mathbf{u}}(\boldsymbol{\gamma}), \tag{4.3}$$

where the summation is over all C response patterns. Maximization of the log-likelihood (e.g., with the EM algorithm; Bock & Aitkin, 1981) leads to the maximum marginal likelihood estimator $\hat{\boldsymbol{\gamma}}$. Upon finding $\hat{\boldsymbol{\gamma}}$, the IRT model generates model-implied probabilities for each response pattern $\pi_{\mathbf{u}}(\hat{\boldsymbol{\gamma}}) = \hat{\pi}_{\mathbf{u}}$. Suppose the model-implied response pattern probabilities $\hat{\pi}_{\mathbf{u}}$ are collected into a $C \times 1$ vector $\hat{\boldsymbol{\pi}}$ of all model-implied response pattern probabilities. By analogy, let a $C \times 1$ vector $\boldsymbol{\pi}$ contain the true (population) response pattern probabilities. Similarly, the observed proportions $p_{\mathbf{u}}$ can be collected into a $C \times 1$ vector $\mathbf{p}$. For example, for 3 dichotomously scored items there are $2^3 = 8$ item response patterns, and the

response pattern probabilities and observed proportions are:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \pi_{111} \end{pmatrix}, \qquad \hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix} = \begin{pmatrix} \pi_{000}(\hat{\boldsymbol{\gamma}}) \\ \pi_{001}(\hat{\boldsymbol{\gamma}}) \\ \pi_{010}(\hat{\boldsymbol{\gamma}}) \\ \pi_{011}(\hat{\boldsymbol{\gamma}}) \\ \pi_{100}(\hat{\boldsymbol{\gamma}}) \\ \pi_{101}(\hat{\boldsymbol{\gamma}}) \\ \pi_{110}(\hat{\boldsymbol{\gamma}}) \\ \pi_{111}(\hat{\boldsymbol{\gamma}}) \end{pmatrix}, \qquad \mathbf{p} = \begin{pmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{pmatrix}. \tag{4.4}$$

From results in discrete multivariate analysis (e.g., Bishop, Fienberg, & Holland, 1975), $\hat{\boldsymbol{\gamma}}$ is consistent, asymptotically normal, and asymptotically efficient, which can be summarized as follows:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{D} N_d(0, \mathbf{F}^{-1}), \tag{4.5}$$

where $\mathbf{F} = \boldsymbol{\Delta}'[diag(\boldsymbol{\pi})]^{-1}\boldsymbol{\Delta}$ is the $d \times d$ Fisher information matrix, with the Jacobian matrix $\boldsymbol{\Delta}$ defined as a $C \times d$ matrix of all first-order partial derivatives of the response pattern probabilities with respect to the item parameters:

$$\boldsymbol{\Delta} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}. \tag{4.6}$$

4.2 Distribution of Multinomial Cell Residuals

Based on Equation 4.6, it can be shown that the asymptotic distribution of the multinomial cell residual vector $(\mathbf{p} - \boldsymbol{\pi})$ is C-variate normal:

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{D} N_C(\mathbf{0}, \boldsymbol{\Xi}), \tag{4.7}$$

where $\boldsymbol{\Xi} = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ is the covariance matrix associated with the multinomial. The cell residual vector $(\mathbf{p} - \hat{\boldsymbol{\pi}})$ under MMLE of item parameters is asymptotically C-variate normal:

$$\sqrt{N}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} N_C(0, \boldsymbol{\Gamma}), \tag{4.8}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Xi} - \boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}'$, and the second term $(\boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}')$ reflects variability due to estimation of item parameters.

4.2.1 Lower-order Marginal Probabilities

The IRT model implies marginal probabilities. Consider the 3-item example from above. There are 3 mathematically independent first-order marginal probabilities $\dot{\pi}_i \, (i = 1, ..., 3)$, one per item. There are also 3 mathematically independent second-order marginal probabilities $\ddot{\pi}_{ij}$ for the unique item pairs $(1 \leq j < i \leq 3)$. In general, these probabilities correspond to the n univariate and $n(n-1)/2$ bivariate margins that can be obtained from the full C-dimensional contingency table using a reduction operator matrix (see e.g., Maydeu-Olivares & Joe, 2005). An example is given below:

$$\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \ddot{\pi}_{21} \\ \ddot{\pi}_{31} \\ \ddot{\pi}_{32} \end{pmatrix} = \mathbf{L}\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix}, \tag{4.9}$$

where $\mathbf{L}$ is a fixed operator matrix of 0s and 1s that reduces the response pattern probabilities and proportions into marginal probabilities and proportions up to or-

der 2. $\hat{\boldsymbol{\pi}}_2$ is the vector of first and second order marginal probabilities. Obviously $\mathbf{p}_2 = \mathbf{L}\mathbf{p}$ is the vector of first and second order observed marginal proportions.

More general versions of the reduction operator matrices for multiple categorical IRT models can be derived using similar logic (see e.g., Maydeu-Olivares & Joe, 2006; Cai & Hansen, 2013). Note that $\mathbf{L}$ has full row rank. It implies that the marginal residual vector $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2) = \mathbf{L}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is a full rank linear transformation of the multinomial cell residual vector $(\mathbf{p} - \hat{\boldsymbol{\pi}})$. Therefore, the marginal residual vector $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is asymptotically normal:

$$\sqrt{N}(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2) = \sqrt{N}\mathbf{L}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} N_Z(0, \boldsymbol{\Gamma}_2), \tag{4.10}$$

and $\boldsymbol{\Gamma}_2 = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}' = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}' - \mathbf{L}\boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}'\mathbf{L}' = \boldsymbol{\Xi}_2 - \boldsymbol{\Delta}_2\mathbf{F}^{-1}\boldsymbol{\Delta}_2'$, where $\boldsymbol{\Xi}_2 = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}'$, and $\boldsymbol{\Delta}_2 = \mathbf{L}\boldsymbol{\Delta}$ is the Jacobian for the marginal probabilities. Z is the number of first and second order marginal residuals. For example, in the case of dichotomous items, the dimensionality of the marginal residual vector is $Z = n + n(n-1)/2 = n(n+1)/2$.

4.2.2 Summed Score Probabilities

In addition to the response pattern and marginal probabilities, the IRT model also generates model-implied summed score probabilities. For a test with n dichotomous items, there are a total of $S = 1 + n$ summed scores ranging from 0 to n. Suppose the observed summed probabilities based on a sample of size N are equal to $\bar{p}_s$ for $s = 0, ..., n$. Under maximum likelihood estimation of item parameters, the corresponding IRT model-implied summed score probabilities are formally defined as

$$\bar{\pi}_s = \sum_{\mathbf{u}} 1_s(\|\mathbf{u}\|)\hat{\pi}_{\mathbf{u}}, \tag{4.11}$$

where $\|\mathbf{u}\| = \sum_{i=1}^{n} u_i$ is a notational shorthand for the summed score associated with response pattern $\mathbf{u}$, and the indicator function takes a value of 1 if and only if $s = \|\mathbf{u}\|$:

$$1_s(\|\mathbf{u}\|) = \begin{cases} 1: \text{ if } s = \|\mathbf{u}\|, \\ 0: \text{ otherwise.} \end{cases}$$

Equation 4.11 shows that the IRT model-implied probability for summed score s is a sum over all such response pattern probabilities leading to summed score s, in other words, it may also be obtained by a reduction operator matrix.

Let $\mathbf{S}$ be a matrix of fixed 0s and 1s such that the pre-multiplication of $\boldsymbol{\pi}$ by $\mathbf{S}$ yields the summed score probabilities. Each row of $\mathbf{S}$ can be understood as a set of binary logical relations for a particular summed score. An entry in row m of $\mathbf{S}$ is equal to 1 if and only if the corresponding response pattern in $\boldsymbol{\pi}$ leads to summed score $m-1$. In general, for n items, there are S rows and C columns in $\mathbf{S}$. In particular, $\mathbf{S}$ has full row rank and the rows of $\mathbf{S}$ are mutually orthogonal.

Returning to the 3-item example, there are 4 summed scores in this case: 0, 1, 2, and 3. The 4×8 matrix $\mathbf{S}$ (below) relates the summed score probabilities to the original multinomial probabilities:

$$\bar{\boldsymbol{\pi}} = \begin{pmatrix} \bar{\pi}_0 \\ \bar{\pi}_1 \\ \bar{\pi}_2 \\ \bar{\pi}_3 \end{pmatrix} = \mathbf{S}\boldsymbol{\pi} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \pi_{111} \end{pmatrix}, \qquad (4.12)$$

and

$$\hat{\bar{\boldsymbol{\pi}}} = \begin{pmatrix} \hat{\bar{\pi}}_0 \\ \hat{\bar{\pi}}_1 \\ \hat{\bar{\pi}}_2 \\ \hat{\bar{\pi}}_3 \end{pmatrix} = \mathbf{S}\hat{\boldsymbol{\pi}}. \tag{4.13}$$

The observed summed score proportions can be obtained in a similar way:

$$\bar{\mathbf{p}} = \begin{pmatrix} \bar{p}_0 \\ \bar{p}_1 \\ \bar{p}_2 \\ \bar{p}_3 \end{pmatrix} = \mathbf{S}\mathbf{p}. \tag{4.14}$$

From Equation 4.8, under MMLE, the summed score residual vector $\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}$ is asymptotically S-variate normally distributed:

$$\sqrt{N}(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}) = \sqrt{N}(\mathbf{S}\mathbf{p} - \mathbf{S}\hat{\boldsymbol{\pi}}) = \sqrt{N}\mathbf{S}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} N_S(0, \bar{\boldsymbol{\Gamma}}), \tag{4.15}$$

and $\bar{\boldsymbol{\Gamma}} = \mathbf{S}\boldsymbol{\Gamma}\mathbf{S}' = \mathbf{S}\text{diag}(\boldsymbol{\pi})\mathbf{S}' - \mathbf{S}\boldsymbol{\pi}\boldsymbol{\pi}'\mathbf{S}' - \mathbf{S}\boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}'\mathbf{S}' = diag(\bar{\boldsymbol{\pi}}) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}' - \bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'$, with $\bar{\boldsymbol{\Delta}} = \mathbf{S}\boldsymbol{\Delta}$. Recall that $\bar{\boldsymbol{\pi}}$ contains the true (population) summed score probabilities. $\bar{\boldsymbol{\Delta}}$ is the $S \times d$ Jacobian matrix for $\hat{\bar{\boldsymbol{\pi}}}$, $\mathbf{F}$ is the $d \times d$ Fisher information matrix, and the term $\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'$ reflects variability due to item parameter estimation.

The reason for introducing the reduction operator matrix $\mathbf{S}$ is primarily a theoretical one. It facilitates the subsequent derivations of summed score likelihood based indices for testing latent variable distribution fit. Pragmatically, the Lord-Wingersky algorithm (Lord & Wingersky, 1984) could be applied to compute the model-implied summed score probabilities. If summed score to scale score conversion tables are computed (for unidimensional IRT model, see Thissen & Wainer, 2001; for multidimensional IRT model, see Cai, 2014), the probabilities become

automatic byproducts.

4.3 Model-fit Test Statistics in IRT

4.3.1 Overall Goodness of Fit Statistics

Overall goodness of fit indices may be used for testing latent variable distribution fit. The full-information test statistics such as likelihood ratio G^2 and Pearson's X^2 use residuals based on the full response pattern cross-classifications to test the IRT model against the general multinomial alternative. The comparison between $\hat{\pi}_{\mathbf{u}}$ and $p_{\mathbf{u}}$ (on logarithmic or linear scales) leads to well-known goodness of fit statistics such as the likelihood ratio G^2 and Pearson's X^2:

$$G^2 = 2N \sum_{\mathbf{u}} p_{\mathbf{u}} \log \frac{p_{\mathbf{u}}}{\hat{\pi}_{\mathbf{u}}}, \tag{4.16}$$

$$X^2 = N \sum_{\mathbf{u}} \frac{(p_{\mathbf{u}} - \hat{\pi}_{\mathbf{u}})^2}{\hat{\pi}_{\mathbf{u}}}. \tag{4.17}$$

Under the null hypothesis that the IRT model fits exactly, these two statistics have the same asymptotic reference distribution, which is a central chi-square with degrees of freedom equal to $C - 1 - d$ (Bishop et al., 1975). For subsequent development, it is instructive to rewrite Pearson's statistic as a quadratic form in multinomial residuals:

$$X^2 = N(\mathbf{p} - \hat{\boldsymbol{\pi}})'[diag(\hat{\boldsymbol{\pi}})]^{-1}(\mathbf{p} - \hat{\boldsymbol{\pi}}). \tag{4.18}$$

Unfortunately as the number of items increases, the number of response patterns increases exponentially. For more than a dozen or so dichotomous items (or perhaps a handful of polytomous items), the contingency table upon which the multinomial is defined becomes sparse for any realistic N. It is well known that the asymptotic chi-square approximations for the full-information test statistics

break down under sparseness (see e.g., Bartholomew & Tzamourani, 1999) and the utility of the full-information overall goodness of fit indices for routine IRT applications is questionable at best.

Recently, limited-information overall fit statistics such as Maydeu-Olivares and Joe's (2005) M_2 have been developed. Limited-information fit statistics use residuals based on lower order (e.g., first and second order) margins of the contingency table. These lower order margins are far better filled when compared to the sparse full contingency table. There is growing awareness that limited-information tests can maintain correct size and can be more powerful than the full-information tests (Cai et al., 2006; Joe & Maydeu-Olivares, 2010).

Under the assumption that the number of first and second order margins (Z) is larger than the number of free parameters (d): $Z > d$, and that $\mathbf{\Delta}_2$ has full column rank (local identification), M_2 can be written as

$$M_2 = N(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)'\tilde{\mathbf{\Delta}}_2[\tilde{\mathbf{\Delta}}_2'\mathbf{\Xi}_2\tilde{\mathbf{\Delta}}_2]^{-1}\tilde{\mathbf{\Delta}}_2'(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2), \tag{4.19}$$

evaluated at the MMLEs. With $\hat{\boldsymbol{\gamma}}$ being the MMLEs, the statistic M_2 can be obtained with Equation 4.19, where $\hat{\boldsymbol{\pi}}_2 = \mathbf{L}\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$, $\mathbf{\Xi}_2 = \mathbf{\Xi}_2(\hat{\boldsymbol{\gamma}})$, and $\tilde{\mathbf{\Delta}}_2 = \tilde{\mathbf{\Delta}}_2(\hat{\boldsymbol{\gamma}})$. $\tilde{\mathbf{\Delta}}_2$ is a $Z \times (Z - d)$ orthogonal complement of $\mathbf{\Delta}_2$ such that $\tilde{\mathbf{\Delta}}_2'\mathbf{\Delta}_2 = 0$. It is not unique, but the choice of $\tilde{\mathbf{\Delta}}_2$ should not influence M_2. From Equation 4.10, $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is asymptotically normal with zero means and covariance matrix $\mathbf{\Xi}_2 - \mathbf{\Delta}_2\mathbf{F}^{-1}\mathbf{\Delta}_2'$, which implies that the covariance matrix of $\tilde{\mathbf{\Delta}}_2'(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is $\tilde{\mathbf{\Delta}}_2'\mathbf{\Xi}_2\tilde{\mathbf{\Delta}}_2$. Thus, M_2 is asymptotically chi-square distributed with $Z - d$ degrees of freedom. In the simulation study M_2 will be used as a benchmark due to its numerous desirable properties identified in the literature (see e.g., Cai & Hansen, 2013). Performance of the proposed latent variable distribution fit indices will be evaluated against M_2.

While an overall test may be used to detect specification errors of latent vari-

able distributions, the fact that they are also sensitive to other forms of model error (e.g., unmodeled multidimensionality) makes it difficult to pinpoint the source of model misspecification. To that end, more specific diagnostic indices have been created for IRT. For example, Chen and Thissen's (1997) local dependence indices are particularly sensitive to violations of the local independence assumption. Orlando and Thissen's (2000) item fit diagnostics is another example where the extent to which the IRT model fits the empirical operating characteristics for an item (e.g., whether monotonicity holds) can be examined. The next section develops summed score likelihood based indices that specifically target latent variable distribution fit for IRT models.

CHAPTER 5

The Proposed Indices

This chapter describes two motivations for summed score likelihood based statistical indices: $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$. As Li and Cai's (2012) statistical indice $\bar{X}^2$ with a heuristic *df* (renamed as $\bar{X}^2_H$ in this study) does not follow an asymptotically chi-squre distribution under some conditions, two moment-matching approaches to adjust $\bar{X}^2$ are developed (the statistic with first-order moment adjustment is named as $\bar{X}^2_{C1}$ and the statistic with two-moment adjustment is named as $\bar{X}^2_{C2}$). Similar to Satorra and Bentler's (1994) approach, the first-order moment adjustment involves the computation of a scaling constant to correct the mean of test statistics, and the two-moment adjustment matches both the mean and variance of the statistics with fixed degrees of freedom. One challenge for the correction is the calculation of Jacobian matrix of summed score likelihoods with respect to item parameters. In this chapter, we develop and illustrate an efficient procedure of Jacobian matrix calculation based on a modification of the Lord-Wingersky algorithm (Lord & Wingersky, 1984).

5.1 Summed Score Likelihood Based Indices

There are two important lines of reasoning for the derivation of these model fit indices. The first is a recognition based on heuristics: IRT model-implied summed score probabilities may provide useful diagnostic information about the latent variable distributional assumption (Thissen & Wainer, 2001; p. 130). The second recognition is that the summed score based indices proposed here are formally

limited-information test statistics.

5.1.1 A Heuristic Motivation

Researchers (Ferrando & Lorenzo-Seva, 2001; Hambleton & Traub, 1973; Li & Cai, 2012; Lord, 1953; Ross, 1966; Sinharay et al., 2006; Thissen & Wainer, 2001) noticed that when the latent variable distribution assumed in the IRT model does not represent the population distribution of the respondents adequately, the model-implied summed score probabilities $\hat{\bar{\pi}}_s$ will depart from the observed summed score probabilities $\bar{p}_s$. Hence all that is needed is to find appropriate test statistics that can summarize the degree to which the model-implied and observed summed score probabilities diverge. It is also preferable if the indices are approximately chi-squared distributed test statistics.

The power divergence family (Cressie & Read, 1984) meets this requirement. Recall that the total number of summed scores is $S = 1 + \sum_{i=1}^{n}(K_i - 1)$, where K_i is the number of categories for each item. The power divergence family of goodness of fit statistics yields a direct comparison between the model-implied summed score probability $\hat{\bar{\pi}}_s$ and the observed summed score probability $\bar{p}_s$:

$$\bar{D}(\lambda) = \frac{2N}{\lambda(\lambda+1)} \sum_{s=0}^{S-1} \bar{p}_s \left\{ \left(\frac{\bar{p}_s}{\hat{\bar{\pi}}_s} \right)^{\lambda} - 1 \right\}, \tag{5.1}$$

where a real-valued scalar λ parameterizes the family. Members of this family include Pearsons statistic $\bar{X}^2 = \bar{D}(1)$ and (defined by continuity) the likelihood ratio statistic $\bar{G}^2 = \bar{D}(0)$. These test statistics are different from the full-information test statistics shown in Equation 4.16 because they are based on summed score probabilities as opposed to response pattern probabilities. Cressie and Read (1984) also advocated a compromise statistic $\bar{D}^2 = \bar{D}(2/3)$ that is "between" the two classical statistics $\bar{X}^2$ and $\bar{G}^2$. As my major interest is not to compare these various statistics, I will focuse on $\bar{X}^2$ in this study, because Pear-

sons statistic $\bar{X}^2$ is algebraically the most straightforward member of the power divergence family.

It is conjectured that under a wide variety of conditions $\bar{X}^2$ statistic might have an asymptotic distribution whose tail-area probability can be approximated by a central chi-squared random variable under the null hypothesis that the latent variable distribution $g(\theta)$ is correctly specified in the IRT model. Li and Cai (2012) proposed a heuristic approach for computing the *df* of $\bar{X}^2$. The rationale behind the heuristic *df* is as follows.

The S summed scores probabilities must sum to 1. The first minus 1 is to reflect that constraint. If the item parameters were known, the *df* would be exactly $S - 1$. When the item parameters are estimated (assuming with MMLE), an additional penalty must be introduced to reflect the effect of parameter estimation. While the location and scale of the latent variable θ are typically fixed for model identification, the model-implied summed score distribution does not have an inherent location and scale. The location and scale is determined as a result of estimating the item parameters. Hence the estimation of item parameters amounts to adding at least two more constraints for the model-implied summed score probability distribution. $\bar{X}^2$ with the heuristic *df* is simple and performs well in some conditions (Li & Cai, 2012). However, a relatively more complex moment-matching approach can adjust $\bar{X}^2$ so that the goodness-of-fit statistics more closely follow an asymptotic chi-squared distribution under the null hypothesis and may lead to a more powerful test under the alternative.

5.1.2 A More Formal Derivation

While the proposed test statistics are not associated with particular marginal probabilities in the same manner as Maydeu-Olivares and Joe's (2005) M_2, they are nevertheless related to the response pattern probabilities via the reduction

operator matrix $\mathbf{S}$ defined earlier (see Equations 4.12). It is the choice of this particular reduction operator that leads to more focused tests targeting latent variable distribution fit (see Joe & Maydeu-Olivares, 2010). Using the reduction operator $\mathbf{S}$, the derivations above imply that $\bar{X}^2$ can be rewritten as

$$\bar{X}^2 = N \sum_{s=0}^{S-1} \frac{(\bar{p}_s - \hat{\bar{\pi}}_s)^2}{\hat{\bar{\pi}}_s} = N(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}})'[diag(\hat{\bar{\boldsymbol{\pi}}})]^{-1}(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}), \tag{5.2}$$

where $(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}) = \mathbf{S}(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is the summed score residual vector (see Equation 4.12). Under the null hypothesis that the IRT model is correctly specified, one can obtain the probability limit of the weight matrix as $plim\,([diag(\hat{\bar{\boldsymbol{\pi}}})]^{-1}) = [diag(\bar{\boldsymbol{\pi}})]^{-1}$ by the consistency of the maximum likelihood estimator (see Equation 4.4), the continuity of the mapping from $\boldsymbol{\gamma}$ to the summed score probabilities, and the continuity of the matrix inverse. Following results on quadratic forms in asymptotically normal random vectors (e.g., Mathai & Provost, 1992, p. 53), the asymptotic expected value of $\bar{X}^2$ is equal to

$$\begin{aligned} tr\,\{\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\} &= tr\,\{[diag(\bar{\boldsymbol{\pi}}) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'][diag(\bar{\boldsymbol{\pi}})]^{-1}\} - tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} \\ &= tr\,(I_S) - tr\,\{\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} - tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} \\ &= S - 1 - tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\}. \end{aligned} \tag{5.3}$$

In this equation, S is the number of all possible summed scores for the test. Minus one is due to the constraint that summed score probabilities sum up to 1. The third part in Equation 5.3 reflects additional uncertainty due to estimation of item parameters. "tr" indicates the "trace" of a matrix. "I_S" indicates an $(S \times S)$ identity matrix. The Jacobian matrix $\bar{\boldsymbol{\Delta}}$ $(S \times d)$ and Fisher information matrix $\mathbf{F}$ $(d \times d)$ can be calculated once the MMLEs of item parameters are available. Similarly, the asymptotic variance of $\bar{X}^2$ is equal to $2tr\,\{\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\}$,

and

$$
\begin{aligned}
&tr\,\{\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\} \\
&\quad = tr\,\big\{\big([diag(\bar{\boldsymbol{\pi}}) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'][diag(\bar{\boldsymbol{\pi}})]^{-1} - \bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\big) \\
&\qquad\quad \big(\{[diag(\bar{\boldsymbol{\pi}}) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'][diag(\bar{\boldsymbol{\pi}})]^{-1}\} - \bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\big)\big\} \\
&\quad = tr\,(I_S) - 2tr\,\{\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} - 2tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} \\
&\qquad + 2tr\,\{\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} + tr\,(\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'1_S) \\
&\qquad + tr\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} \\
&\quad = S - 2 - 2tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\} + 2sum\,(\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}') + sum\,(\bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}') \\
&\qquad + tr\,\{\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Delta}}\mathbf{F}^{-1}\bar{\boldsymbol{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\}. \qquad (5.4)
\end{aligned}
$$

In this equation, S, $\bar{\boldsymbol{\Delta}}$, $\mathbf{F}$ are the same components as in Equation 5.3. "$sum()$" means the sum of all elements of a matrix. The first-order and second-order moments of the statistics are essential elements for the adjustement of the indices. From now on, $tr\,\{\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\}$ will be indicated by U_1 for short, and $tr\,\{\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\}$ will be indicated by U_2. The next section will illustrate how to use the first and second moments (U_1 & U_2) to adjust $\bar{X}^2$.

5.2 Adjustment of Statistics

From Equation 5.2, it is clear that the statistic $\bar{X}^2$ cannot be asymptotically chi-squared. Even though it is a quadratic form in asymptotically normally distributed random vectors, a key condition for its chi-squaredness is not met. That is, the product of the probability limit of the weight matrix $[diag(\bar{\boldsymbol{\pi}})]^{-1}$ and the covariance matrix of the normal random vector ($\bar{\boldsymbol{\Gamma}}$) is not idempotent in general, i.e., $\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1} \neq \bar{\boldsymbol{\Gamma}}[diag(\bar{\boldsymbol{\pi}})]^{-1}$. On the other hand, according to Satorra and Bentler's (1994) paper, test statistics which do not follow an asymptotically chi-squared distribution can be modified, using a scaling to correct the

mean (or the mean and variance) of the test statistics or an adjustment to *df*. Cai et al. (2006) also illustrated an approach of moment adjustment to goodness-of-fit testing of IRT models when parameter estimation is taken into account.

In this study, two Satorra-Bentler type modification approaches (Satorra & Bentler, 1994) to adjust $\bar{X}^2$ are developed: a first-order moment adjustment and a two-moment adjustment. The rational for the first-order moment adjustment is as following. Equation 5.3 shows that the asymptotic expected value of $\bar{X}^2$ is equal to $S-1$ minus a constant that depends on the trace of matrix $\mathbf{F}^{-1}\bar{\mathbf{\Delta}}'[diag(\bar{\boldsymbol{\pi}})]^{-1}\bar{\mathbf{\Delta}}$, which reflects additional uncertainty due to estimation of item parameters. According to the properties of a chi-squared distribution, the *df* of $\bar{X}^2$ should be equal to the asymptotic expected value of $\bar{X}^2$: U_1. Therefore, the *df* of $\bar{X}^2$ should be better approximated by U_1 than the heuristic *df*. On the other hand, if the *df* is fixed to a constant (e.g., $S-1-2$), the statistic could be rescaled so that the asymptotic expected value of the adjusted statistic would approach the fixed *df*. For the two-moment adjustment, both the asymptotic mean and variance of $\bar{X}^2$ are considered (Asparouhov & Muthen, 2010). Not only the asymptotic expected value of the two-moment adjusted statistic would approximate a fixed *df*, but also the asymptotic variance of the two-moment adjusted statistic would approximate $2df$. Next, equations for the moment adjustment approaches are illustrated step by step.

Obviously, the index $\bar{X}^2_H$ equals to $\bar{X}^2$:

$$\bar{X}^2_H = \bar{X}^2 = N(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}})'[diag(\hat{\bar{\boldsymbol{\pi}}})]^{-1}(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}). \tag{5.5}$$

Let

$$C = \frac{U_1}{df_{fixed}}. \tag{5.6}$$

Thus the first-order moment adjusted statistic $\bar{X}^2_{C1}$ is

$$\begin{aligned} \bar{X}^2_{C1} &= C^{(-1)}\ \bar{X}^2 \\ &= \frac{\bar{X}^2\ df_{fixed}}{U_1}, \end{aligned} \tag{5.7}$$

where

$$df_{fixed} = S - 1 - 2. \tag{5.8}$$

Similiar to the first-order moment adjustment, the df of $\bar{X}^2_{C2}$ is fixed to some constant df_{fixed}, and the two-moment adjusted statistics is:

$$\bar{X}^2_{C2} = \bar{X}^2 \sqrt{\frac{df_{fixed}}{U_2}} + df_{fixed} - \sqrt{\frac{df_{fixed}(U_1)^2}{U_2}}, \tag{5.9}$$

where

$$df_{fixed} = S - 1 - 2. \tag{5.10}$$

Theoritically, the constant df_{fixed} can take on an arbitrary value. For the purpose of comparison among the summed score likelihood based indices, all the proposed statistics have the same df as $\bar{X}^2_H$ in this study. Though all the elements in the adjusted part can be calculated using the MMLEs of item parameters, the process of calculation might be computationally demanding when the number of items is large. Some commercial software for IRT (e.g., flexMIRT®; Cai, 2013) provides the Fisher information matrix $\mathbf{F}$ and model-implied summed score probabilities $\hat{\bar{\boldsymbol{\pi}}}$ in the output file, but none of them produces the Jacobian matrix of summed score likelihoods $\bar{\boldsymbol{\Delta}}$. Direct numerical calculation of the Jacobian matrix can be very computationally demanding. Take the 2-PL model as an example. It requires the computation of 2×2^n first-order derivatives to obtain $\bar{\boldsymbol{\Delta}}$. For a test with 12

items, 8,192 first-order derivatives need to be computed. When the number of items increases to 24, 33,554,432 first-order derivatives need to be computed. To solve this problem, a modified Lord-Wingerskty algorithm (Lord & Wingersky, 1984) for calculating $\bar{\mathbf{\Delta}}$ is developed in the next section.

5.2.1 Lord-Wingersky Algorithm for Calculating the Jacobian Matrix

As the calculation of the Jacobian matrix for the moment adjustment becomes a computational problem when the number of items is large, a modified Lord-Wingersky algorithm (Lord & Wingersky, 1984) for computing the Jacobian matrix is proposed in this section.

The Lord-Wingersky algorithm (Lord & Wingersky, 1984) for summed score probabilities in unidimensional and multidimensional IRT applications was thoroughly demonstrated in a recent publication (Cai, 2014), and the algorithm could be adapted to calculate the Jacobian matrix using the chain rule. The logic of reasoning is as following. When Lord-Wingersky algorithm is applied, the summed score likelihoods are accumulated by multiplying items' tracelines one at a time. When the nth item is added, the likelihood of a summed score s is a function of summed score likelihoods from the previous steps and the current item's tracelines. The Jacobian matrix can be obtained by recursively multiplying the first-order derivatives for each item with all the other items' summed score probabilities, and the latter is easily found via the Lord-Wingersky algorithm. An example of the procedure for the case of dichotomously scored items is presented as follows.

Consider a test with n dichotomous items, calibrated by a 2-PL IRT model. Recall that $T_i(1|\theta)$ is the ith item's traceline for category 1 (Equation 2.1), with $T_i(0|\theta) = 1 - T_i(1|\theta)$ for category 0. Theoretically, there should be 2^n response patterns. The variable of response patterns, indicated by $\mathbf{u} = (u_1, ..., u_n)'$. Under the assumption of items' conditional independence, the likelihood for $\mathbf{u}$ can be

expressed as

$$L(\mathbf{u}|\theta) = \prod_{i=1}^{n} T_i(u_i|\theta). \tag{5.11}$$

For n dichotomous items, the summed score s ranges from 0 to n. $S = n + 1$ is the number of all possible summed scores. Recall that $\|\mathbf{u}\| = \sum_{i=1}^{n} u_i$ is a notational shorthand for the summed score associated with response pattern $\mathbf{u}$ (see Equation 4.12). The likelihood for summed score $s = 0, ..., n$ is defined as

$$L(s|\theta) = \sum_{\|\mathbf{u}\|=s} L(\mathbf{u}|\theta) = \sum_{s=\|\mathbf{u}\|} \prod_{i=1}^{n} T_i(u_i|\theta). \tag{5.12}$$

Clearly, the likelihood of a summed score s is the sum of all response pattern likelihoods for $\|\mathbf{u}\| = s$. In Lord-Wingersky algorithm, the summed score likelihoods are built up recursively, one at a time (Lord & Wingersky, 1984; Cai, 2014). Let $L_i(s|\theta)$ indicate the likelihood for summed score s after item i has been added into the computation. In the first step, two summed score likelihoods are computed based on the tracelines of item 1: $L_1(0|\theta) = T_1(0|\theta)$ and $L_1(1|\theta) = T_1(1|\theta)$; In the second step, we have three summed score likehoods based on the summed score likelihoods from step 1 and tracelines of item 2:

$$\begin{aligned} L_2(0|\theta) =& L_1(0|\theta)T_2(0|\theta), \\ L_2(1|\theta) =& L_1(1|\theta)T_2(0|\theta) + L_1(0|\theta)T_2(1|\theta), \\ L_2(2|\theta) =& L_1(1|\theta)T_2(1|\theta). \end{aligned} \tag{5.13}$$

Suppose n items have been added in. The likelihoods for summed scores (0,

..., n) would be:

$$\begin{aligned} L_n(0|\theta) =& L_{n-1}(0|\theta)T_n(0|\theta), \\ & \cdots \\ L_n(s|\theta) =& L_{n-1}(s|\theta)T_n(0|\theta) + L_{s-1}(s-1|\theta)T_n(1|\theta), \\ & \cdots \\ L_n(n|\theta) =& L_{n-1}(n-1|\theta)T_n(1|\theta). \end{aligned} \tag{5.14}$$

To obtain the Jacobian matrix of summed scores with respective to item parameters, the Lord-Wingersky algorithm is adapted slightly. As previously mentioned, in the first step, there are only two summed score likelihoods based on item 1: $L_1(0|\theta)$ and $L_1(1|\theta)$. Suppose I have a parameter γ_1 for this item. The first-order derivatives of summed score likelihoods with respect to γ_1 are computed:

$$\begin{aligned} \frac{\partial L_1(0|\theta)}{\partial \gamma_1} &= \frac{\partial T_1(0|\theta)}{\partial \gamma_1}, \\ \frac{\partial L_1(1|\theta)}{\partial \gamma_1} &= \frac{\partial T_1(1|\theta)}{\partial \gamma_1}. \end{aligned} \tag{5.15}$$

Notice that

$$\frac{\partial T_1(0|\theta)}{\partial \gamma} = -\frac{\partial T_1(1|\theta)}{\partial \gamma}, \tag{5.16}$$

where γ is a parameter in a 2PL model.

In the second step, one more item is added. The first-order derivatives of summed score likelihoods with respect to γ_1 and γ_2 are computed based on the

first step:

$$
\begin{aligned}
\frac{\partial L_2(0|\theta)}{\partial \gamma_1} =& \frac{\partial L_1(0|\theta)}{\partial \gamma_1} T_2(0|\theta), \\
\frac{\partial L_2(1|\theta)}{\partial \gamma_1} =& \frac{\partial L_1(0|\theta)}{\partial \gamma_1} T_2(1|\theta) + \frac{\partial L_1(1|\theta)}{\partial \gamma_1} T_2(0|\theta), \\
\frac{\partial L_2(2|\theta)}{\partial \gamma_1} =& \frac{\partial L_1(1|\theta)}{\partial \gamma_1} T_2(1|\theta), \\
\frac{\partial L_2(0|\theta)}{\partial \gamma_2} =& L_1(0|\theta) \frac{\partial T_2(0|\theta)}{\partial \gamma_2}, \\
\frac{\partial L_2(1|\theta)}{\partial \gamma_2} =& L_1(0|\theta) \frac{\partial T_2(1|\theta)}{\partial \gamma_2} + L_1(1|\theta) \frac{\partial T_2(0|\theta)}{\partial \gamma_2}, \\
\frac{\partial L_2(2|\theta)}{\partial \gamma_2} =& L_1(1|\theta) \frac{\partial T_2(1|\theta)}{\partial \gamma_2}.
\end{aligned}
\tag{5.17}
$$

Then, the first-order derivatives of summed score likelihood functions with

respect to the n item's parameters $(\gamma_1, ..., \gamma_n)$ are:

$$
\begin{aligned}
\frac{\partial L(0|\theta)}{\partial \gamma_1} =& \frac{\partial L_{n-1}(0|\theta)}{\partial \gamma_1} T_n(0|\theta), \\
& \cdots \\
\frac{\partial L(s|\theta)}{\partial \gamma_1} =& \frac{\partial L_{n-1}(s-1|\theta)}{\partial \gamma_1} T_n(1|\theta) + \frac{\partial L_{n-1}(s|\theta)}{\partial \gamma_1} T_n(0|\theta), \\
& \cdots \\
\frac{\partial L(n|\theta)}{\partial \gamma_1} =& \frac{\partial L_{n-1}(1|\theta)}{\partial \gamma_1} T_n(1|\theta), \\
& \cdots \\
& \cdots \\
& \cdots \\
\frac{\partial L(0|\theta)}{\partial \gamma_s} =& \frac{\partial L_{n-1}(0|\theta)}{\partial \gamma_s} T_n(0|\theta), \\
& \cdots \\
\frac{\partial L(s|\theta)}{\partial \gamma_s} =& \frac{\partial L_{n-1}(0|\theta)}{\partial \gamma_s} T_n(1|\theta) + \frac{\partial L_{n-1}(1|\theta)}{\partial \gamma_s} T_n(0|\theta), \\
& \cdots \\
\frac{\partial L(n|\theta)}{\partial \gamma_s} =& \frac{\partial L_{n-1}(1|\theta)}{\partial \gamma_s} T_n(1|\theta), \\
& \cdots \\
& \cdots \\
& \cdots \\
\frac{\partial L(0|\theta)}{\partial \gamma_n} =& L_{n-1}(0|\theta) \frac{\partial T_n(0|\theta)}{\partial \gamma_n}, \\
& \cdots \\
\frac{\partial L(s|\theta)}{\partial \gamma_n} =& L_{n-1}(s|\theta) \frac{\partial T_n(0|\theta)}{\partial \gamma_n} + L_{n-1}(s-1|\theta) \frac{\partial T_n(1|\theta)}{\partial \gamma_n}, \\
& \cdots \\
\frac{\partial L(n|\theta)}{\partial \gamma_n} =& L_{n-1}(n-1|\theta) \frac{\partial T_n(1|\theta)}{\partial \gamma_n}.
\end{aligned}
\tag{5.18}
$$

The resulting first-order derivatives of summed score likelihoods with respect to n slope parameters and n intercept parameters is collected into a $S \times 2n$ Jacobian matrix $\bar{\mathbf{\Delta}}$. Then the first and second moments of $\bar{X}^2$ for the adjustments of statistics follow.

5.2.2 An Illustrative Example

The process of modified Lord-Wingersky algorithm for calculating the Jacobian matrix is illustrated with an example. Consider a simple test with three dichotomous items. The traceline for each item is defined using a 2-PL IRT model:

$$T_i(1|\theta) = \frac{1}{1 + \exp[-(c_i + a_i\theta)]}. \tag{5.19}$$

The values of slope parameters are $\mathbf{a} = (1.0, 0.8, 1.2)$, and the values of intercept parameters are $\mathbf{c} = (-0.2, 0.6, -1.0)$. Recall that the marginal probability for summed scores with known $g(\theta)$ (see Equation 4.2) is

$$p(s) = \int L(s|\theta)g(\theta)d\theta. \tag{5.20}$$

The integrals in Equation 5.20 must be approximated by quadrature. In addition, a good way to demonstrate the algorithm is to show the calculations over a set of quadrature points (Cai, 2014). We approximate the marginal probability using Q quadrature points and quadrature weights:

$$p(s) = \int L\left(s|\theta\right)g(\theta)d\theta = \sum_{q=1}^{Q} L\left(s|X_q\right)W(X_q). \tag{5.21}$$

where X_q is a quadrature node and $W\left(X_q\right)$ is the corresponding quadrature weights. To obtain $W\left(X_q\right)$, a set of normalized ordinates of the prior density are applied (Cai, 2014), i.e., $W\left(X_q\right) = g(X_q)/\sum_{q=1}^{Q} g(X_q)$.

Table 5.1 shows the recursive computations with numbers for item 3. This illustration is similar to Thissen and Wainer's (2001) Table 3.8, as well as Cai's (2014) Table 2. It shows the values of summed score likelihoods, 1st-order derivatives of tracelines, and the 1st-order derivatives of summed score likelihoods at five equally spaced quadrature points: -2, -1, 0, 1, and 2.

Table 5.1: Calculating 1st-order derivatives of summed score likelihoods with respect to item 3's slope parameter at five rectangular quadrature points

Quadrature points of θ	-2	-1	0	1	2
Summed score likelihoods for all the other items					
$L_2(0\|\theta)$	.658	.423	.195	.061	.014
$L_2(1\|\theta)$	.315	.473	.515	.385	.213
$L_2(2\|\theta)$	.027	.104	.291	.553	.773
Derivatives of tracelines with respect to item 3's a parameter					
$\frac{\partial T_n(1\|\theta)}{\partial a_3}$	-.063	-.090	.000	.248	.317
$\frac{\partial T_n(0\|\theta)}{\partial a_3}$	.063	.090	.000	-.248	-.317
First-order derivatives of summed score likelihoods					
$\frac{\partial L_3(0\|\theta)}{\partial a_3} = L_2(0\|\theta)\frac{\partial T_3(0\|\theta)}{\partial a_3}$	.041	.038	.000	-.015	-.004
$\frac{\partial L_3(1\|\theta)}{\partial a_3} = L_2(1\|\theta)\frac{\partial T_3(0\|\theta)}{\partial a_3} + L_2(0\|\theta)\frac{\partial T_3(1\|\theta)}{\partial a_3}$	-.021	.005	.000	-.080	-.063
$\frac{\partial L_3(2\|\theta)}{\partial a_3} = L_2(2\|\theta)\frac{\partial T_3(0\|\theta)}{\partial a_3} + L_2(1\|\theta)\frac{\partial T_3(1\|\theta)}{\partial a_3}$	-.018	-.033	.000	-.042	-.177
$\frac{\partial L_3(3\|\theta)}{\partial a_3} = L_2(2\|\theta)\frac{\partial T_3(1\|\theta)}{\partial a_3}$	-.002	-.009	.000	.137	.245

Note. The first block presents the summed score likelihoods after the 1st and 2nd items are added in. The second block presents the 1st-order derivatives of item 3's tracelines with respect to its slope parameter. The third block presents the 1st-order derivatives of summed score likelihoods with respect to item 3's slope parameter.

Table 5.2 presents the 1st-order derivatives of the marginal probabilities of the

summed scores with quardrature weights. Notice that, more quadrature points should be used for better precision in practice.

Table 5.2: Calculating 1st-order derivatives of marginal summed score likelihoods with respect to item 3's slope parameter with quadrature weights

	Quadrature points of θ					
	-2	-1	0	1	2	
$W(\theta)$	.054	.244	.403	.244	.054	
First-order derivatives of summed score likelihoods						
$\frac{\partial L_3(0\|\theta)}{\partial a_3}$	.041	.038	.000	-.015	-.004	
$\frac{\partial L_3(1\|\theta)}{\partial a_3}$	-.021	.005	.000	-.080	-.063	
$\frac{\partial L_3(2\|\theta)}{\partial a_3}$	-.018	-.033	.000	-.042	-.177	
$\frac{\partial L_3(3\|\theta)}{\partial a_3}$	-.002	-.009	.000	.137	.245	
	Weighted derivatives					Jacobian
$\frac{\partial L_3(0\|\theta)}{\partial a_3} * W(\theta)$	.002	.009	.000	-.004	.000	.008
$\frac{\partial L_3(1\|\theta)}{\partial a_3} * W(\theta)$	-.001	.001	.000	-.020	-.003	-.023
$\frac{\partial L_3(2\|\theta)}{\partial a_3} * W(\theta)$	-.001	-.008	.000	-.010	-.010	-.029
$\frac{\partial L_3(3\|\theta)}{\partial a_3} * W(\theta)$	.000	-.002	.000	.033	.013	.044

Note. $W(\theta)$ indicates quadrature weights at each θ level. "Weighted derivatives" are found by multiplying (point to point) the first-order derivatives of summed score likelihoods with $W(\theta)$. The last column "Jacobian" indicates the 1st-order derivatives of marginal summed score likelihoods with respect to item 3's slope parameter. It is the summation of the weighted derivatives over all quadrature points for each summed score likelihood.

CHAPTER 6

Simultion Study Design

The primary goal of the simulation study is to examine the performance of summed score likelihood based indices $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ for detecting latent variable nonnormality. Recall that the statistic $\bar{X}^2_H$ has the same formula as $\bar{X}^2$ (see Equation 5.5), with fixed degrees of freedom. $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are adjusted indices based on the first-order and second-order moments of $\bar{X}^2$.

This chapter illustrates the simulation conditions, process of data generation, and evaluative statistics of interest for the simulation study. There are three objectives in the simulation study. First, I will check whether the tail areas of the statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ can be well approximated by their purported chi-squared distributions under null conditions, where the latent variables are generated from a normal distribution. Second, I will examine the statistical power of $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ against latent variable nonnormality in alternative conditions, comparing with overall limited-information goodness-of-fit statistics, specifically Maydeu-Olivares and Joe's (2005) M_2. Finally, I am interested in the influence of item parameters' dispersion on the difference among $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ with respect to their performance.

6.1 Simulation Conditions

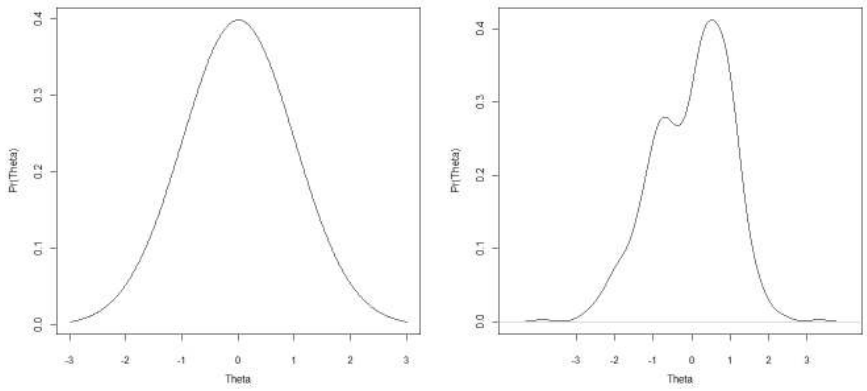
Simulations were undertaken to evaluate the summed score likelihood based indices $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$, comparing with the limited-information overall fit statis-

tics M_2. There are 96 conditions in total ($2 \times 2 \times 4 \times 3 \times 2$), with 500 replications in each condition. Studied variables are: the item types (2-PL or graded response with 4 categories), the number of items (12 or 24), the value of item parameters (e.g., the extent to which generating item slope/threshold parameters are fixed to be equal across items or dispersed), the sample size (500, 1000, or 1500), and last but not the least, the distribution of generating θ (normal or non-normal), as shown in Table 6.1.

Table 6.1: *Manipulated factors and conditions for simulation study*

Levels	Conditions
Item types (2)	2-PL, Graded response with 4 categories
Number of items (2)	12, 24
Value of item parameters (4)	Equal a and Equal b
	Random a and Random b
	Random a and Dispersed b
	Dispersed a and Dispersed b
Sample size (3)	500, 1000, 1500
Distribution of θ (2)	Normally distributed θ
	Nonnormally distributed θ

In condition "Value of item parameters", "Equal a" means all the slope (a) parameters are fixed to 1; "Equal b" means all the threshold (b) parameters are fixed to 0; "Random a" means all the a parameters are generated randomly from a log-normal distribution (M=0.5, SD=0.2); "Random b" means all the b parameters are generated randomly from a standard normal distribution; "Dispersed a" means b parameters are equally spaced from 1 to 3. "Dispersed b" means b parameters are equally spaced from -2 to 2. In condition "Distribution of θ", "Normally distributed θ" indicates that the latent variable θ was generated from a standard normal distribution; "Nonnormally distributed θ" indicates that a nonnormal θ



(a) Normal (b) Non-normal

Figure 6.1: Density plots for illustrating normal vs non-normal latent variable distribution

was generated from a distribution obtained from a combination of two normal distribution with different means and variances (group1: M=1, SD=0.4; group2: M=0, SD=1; $N_{group1} : N_{group2} = 1 : 4$). Figure 6.1 shows an example of normal vs non-normal generating latent variable distributions in the simulation study.

6.2 Data Generation

Item parameters were generated with properties mimicking item parameters found in typical educational assessments, and then manipulated to meet the requirements of research design. Table 6.2 presents the generated item parameters for 2-PL model with 12 items. For item slope parameters, there are three groups of values: "equal", "random" and "dispersed". The value of "equal" slope parameters was chosen to be 1. In the group of "dispersed", the generated item slope parameters were equally spaced from 1 to 3, which is an unrealistic condition. We only use it to illustrate a high level of item parameter dispersion in this study and does not represent any real data. Values in the group of "random" were randomly drawn from a log-normal distribution ($M = 0.5, SD = 0.2$). Item threshold

parameters also have three group of values. The value of "equal" threshold parameters was chosen to be 0. Values in the group of "random" were drawn from a standard normal distribution ($M = 0, SD = 1$). In the group of "dispersed", the generated item threshold parameters were equally spaced from -2 to 2. after completing the generation of item slopes and thresholds, combining one group of item slope parameter with one group of item threshold parameter in Table 6.2 formed a set of generating item parameters. As mentioned in Table 6.1, there would be four sets of generating item parameters for 2-PL model with 12 items: 1) Both slope and threshold parameters equal; 2) Both slope and threshold parameters random; 3) The slope parameters random, but the threshold parameters dispersed; 4) Both slope and threshold parameters dispersed.

Table 6.3 presents the generated item parameters for graded model with 12 items. Item slope parameters and the first threshold paramters took the same value as 2-PL model. The second threshold parameters were generated by adding 0.5 to their first threshold parameters correspondingly, and the third threshold parameters were generated by adding 0.5 to the second threshold parameters. For both 2-PL model and graded model with 24 items, the random item parameters were generated with the same distribution as in Table 6.2 and 6.3 and then manipulated in the same way. More details can be found in the Appendix: Table A.1 and Table A.2.

After the item parameters were obtained, response pattern data were simulated with randomly generated θ drawn from normal or nonnormal distributions. For each simulation condition, the true item parameters remained unchanged, and 500 sets of response data were generated.

6.3 Evaluative Statistics of Interest

The means, variances, minimum & maximum values of $\bar{X}^2_H$, $\bar{X}^2_{C1}$, and $\bar{X}^2_{C2}$ were calculated for each simulation condition. To compare the performance of these indices, empirical rejection rates were computed in the null conditions at three levels: 0.01, 0.05, and 0.1. The rejection rates are expected to be close to their α-levels. The confidence intervals for all the rejection rates were calculated as well. Moreover, the Kolmogorov-Smirnov test (KS test) was carried out to compare the distributions of the statistics with their purported chi-squared distributions. In the alternative conditions, statistical power against model misspecification were also evaluated at three α-levels: 0.01, 0.05, and 0.1. Additionally, a fourth model fit index, the limited-information overall goodness-of-fit statistic M_2 was employed as a comparison in this study.

Table 6.2: Generating item parameters for two-parameter IRT models, n=12

		a			b	
Item	Equal	Random	Dispersed	Equal	Random	Dispersed
1	1	1.85	1.00	0.00	-0.28	-2.00
2	1	1.87	1.18	0.00	-0.29	-1.64
3	1	1.80	1.36	0.00	1.05	-1.27
4	1	1.72	1.55	0.00	-0.52	-0.91
5	1	1.39	1.73	0.00	-1.32	-0.55
6	1	1.24	1.91	0.00	0.29	-0.18
7	1	1.19	2.09	0.00	0.62	0.18
8	1	1.66	2.27	0.00	0.50	0.55
9	1	1.71	2.45	0.00	0.55	0.91
10	1	2.11	2.64	0.00	0.81	1.27
11	1	1.68	2.82	0.00	1.49	1.64
12	1	1.98	3.00	0.00	0.52	2.00

Note. In this table, all possible generating item slope parameters (a) and item threshold parameters (b) for a 12-item multiple-choice test with 2-PL item type are presented. Both item slope and threshold parameters could be equal, random or dispersed across items. The range of a parameters is from 1 to 3, and the values are equally spaced in this range in the "dispersed parameter" condition. The range of b parameters is from -2 to 2. Their values are also equally spaced in this range in the last condition. When both item slope and threshold parameters are set to equal across items, a parameters of all the twelve items equal to 1 and b parameters all equal to 0.

Table 6.3: Generating item parameters for graded-response IRT models, n=12

	a			b 1			b 2			b 3		
Item	Equal	Random	Dispersed	Equal	Random	Dispersed	Equal	Random	Dispersed	Equal	Random	Dispersed
1	1.00	1.85	1.00	0.00	-0.28	-2.00	0.50	0.22	-1.50	1.00	0.72	-1.00
2	1.00	1.87	1.18	0.00	-0.29	-1.64	0.50	0.21	-1.14	1.00	0.71	-0.64
3	1.00	1.80	1.36	0.00	1.05	-1.27	0.50	1.55	-0.77	1.00	2.05	-0.27
4	1.00	1.72	1.55	0.00	-0.52	-0.91	0.50	-0.02	-0.41	1.00	0.48	0.09
5	1.00	1.39	1.73	0.00	-1.32	-0.55	0.50	-0.82	-0.05	1.00	-0.32	0.45
6	1.00	1.24	1.91	0.00	0.29	-0.18	0.50	0.79	0.32	1.00	1.29	0.82
7	1.00	1.19	2.09	0.00	0.62	0.18	0.50	1.12	0.68	1.00	1.62	1.18
8	1.00	1.66	2.27	0.00	0.50	0.55	0.50	1.00	1.05	1.00	1.50	1.55
9	1.00	1.71	2.45	0.00	0.55	0.91	0.50	1.05	1.41	1.00	1.55	1.91
10	1.00	2.11	2.64	0.00	0.81	1.27	0.50	1.31	1.77	1.00	1.81	2.27
11	1.00	1.68	2.82	0.00	1.49	1.64	0.50	1.99	2.14	1.00	2.49	2.64
12	1.00	1.98	3.00	0.00	0.52	2.00	0.50	1.02	2.50	1.00	1.52	3.00

Note. in this table, all possible generating item slope parameters (a) and item threshold parameters ($b1$, $b2$, $b3$) for a 12-item multiple-choice test with graded response item type are presented. Notice that the slope parameter and the first threshold parameter for each item are the same as those for 2-PL IRT models. However, two more columns of threshold parameters indicates that item difficulties are monotonically increasing from category 1 to category 4 within an item.

CHAPTER 7

Simulation Study Results

This chapter presents results from the simulation study. Findings include the means, variances, minimum & maximum values of summed score likelihood based statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$, as well as those values of the overall goodness-of-fit statistic M_2 in various simulation conditions. In the null conditions, where the latent variable follows a standard normal distribution, emperical rejection rates at three α levels and the Kolmogorov-Smirnov test (KS test) p-values are reported for all the statistics. In the alternative conditions, where the latent variable is generated from a non-normal distribution, the statistical power of the proposed statistics are reported and compared with overall fit statistic M_2. In addition, the influence of number of items, item type, sample size, and dispersion of item parameters on the performance of these statistics are examined.

7.1 Simulation Results for 2-PL IRT Models

7.1.1 Type I Error Rates

In the null condition, the generating latent variable follows a normal distribution, and simulated item response patterns were calibrated with a standard unidimensional 2-PL IRT model. Since the fitted model was the same as the generating model, the distribution of the proposed statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ should be well approximated by their purported chi-squared distribution under null hypothesis. That is, the means of the statistics should be close to df, the variances should be

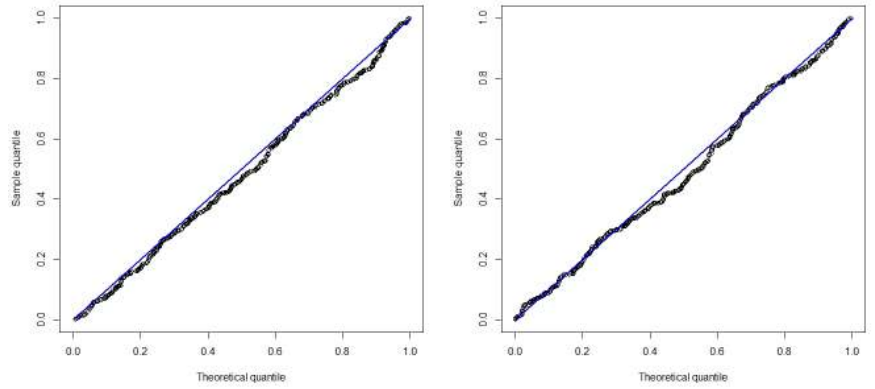
close to $2df$, and empirical rejection rates should approximate their corresponding α levels respectively.

Table 7.1, Table 7.2, Table 7.3 and Table 7.4 present simulation study results under the null hypothesis for 2-PL model. The number of items is 12, and sample size has three levels, $N = 500, 1000, 1500$. The result tables for larger number of items ($n = 24$) could be found in the Appendix (see Table A.3, Table A.4, Table A.5, and Table A.6). From Table 7.1 to Table 7.4, the extent of item parameter dispersion is increasing: Table 7.1 for the condition of "Equal a & Equal b"; Table 7.2 for the condition of "Random a & Random b"; Table 7.3 for the condition of "Random a & Dispersed b"; Table 7.4 for the condition of "Dispersed a & Dispersed b". As mentioned in Chapter 6, the values of generating item parameters could exert an influence on the adjustment of summed score likelihood based statistics. Therefore, I present these four tables side by side, and further discussion will be carried out later. These tables report the means, variances, minimum & maximum values of the statisics, and empirical rejection rates at three α levels (95% confidence inverval): $\alpha = 0.01(0.004, 0.016)$, $\alpha = 0.05$ $(0.036, 0.064)$, $\alpha = 0.1(0.081, 0.119)$, as well as the p values of KS test.

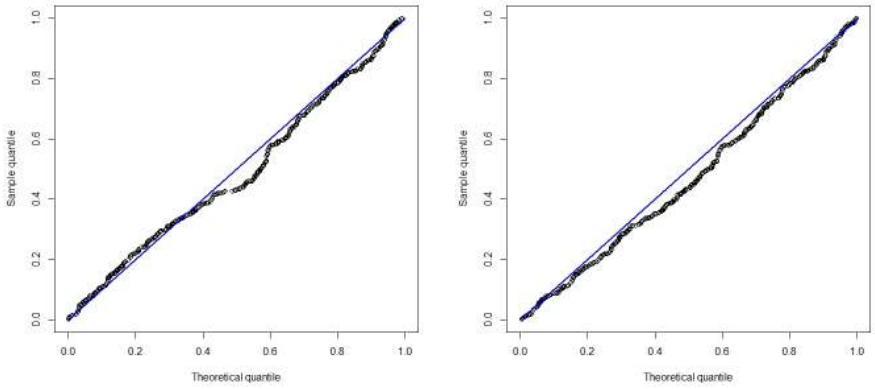
Results in Table 7.1 show that when the generating item parameters are equal across items, all the proposed indices work well under the null hypothesis. As suggested earlier, the means of these indices should be close to the expected value of their purported chi-squared distribution, specifically, the degrees of freedom, and the variances of these indices should approximate the expected variance of a chi-squared distribution, that is, twice the degrees of freedom. Data in Table 7.1 confirm this speculation: Most of the means are close to their corresponding df, and the variances are approximating $2df$. Empirical rejection rates are close to their α levels, too. Take the empirical rejection rates at the level $\alpha = 0.05$ as an example. All rejection rates fall into its confidence interval $(0.036, 0.064)$. When the sample size is small, the rejection rates for summed score likelihood based statistics are slightly smaller than 0.05, and some KS test p-values are smaller than 0.05. But their performance improves as the sample size increases. Furthermore, Q-Q plots in Figure 7.1 show that under the null conditions when the item parameters are equal and sample size is small (N=500), the distribution of statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$, and $\bar{X}^2_{C2}$ can be well approximated by a chi-squared distribution with a fixed df.

From Table 7.2, we can see that when the generating item parameters are random, the empirical rejection rates of $\bar{X}^2_H$ are lower than their corresponding α levels, indicating that the tail area of the distribution of $\bar{X}^2_H$ cannot be well approximated by a chi-squared distribution with $df = 10$. In addition, KS test p-values for this index are smaller than 0.001. Results in Table 7.3 show that when the item parameters become more dispersed, the empirical rejection rates for $\bar{X}^2_H$ are even lower. As reflected in both Table 7.2 and Table 7.3, the moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ perform well in these two conditions, with reasonable empirical rejection rates and KS test p-values.

Table 7.4 show that in the condition of "Dispersed a & Dispersed b", $\bar{X}^2_{C2}$ performs better than both $\bar{X}^2_H$ and $\bar{X}^2_{C1}$. Specifically, the empirical rejection rates



(a) $\bar{X}^2_H$ (b) $\bar{X}^2_{C1}$

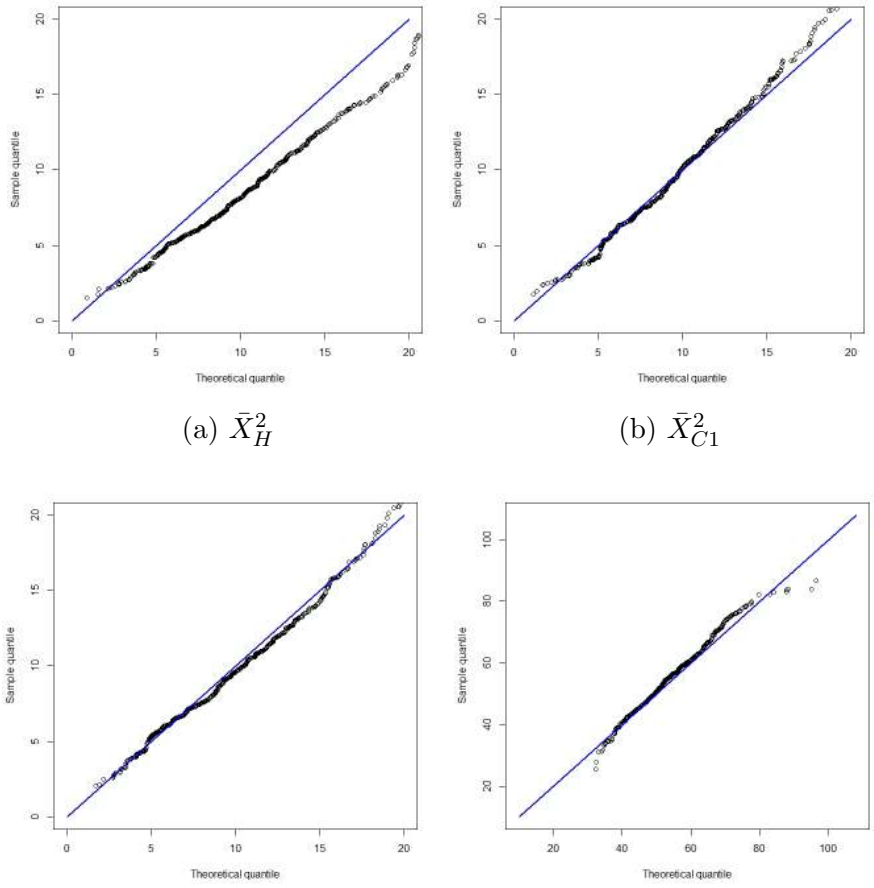


(c) $\bar{X}^2_{C2}$ (d) $\bar{M}_2$

Figure 7.1: Q-Q plots for a null condition (normal θ, n=12, N=500, Equal a & b parameters)

of $\bar{X}^2_{C2}$ reach its corresponding α levels. The first-moment adjusted index $\bar{X}^2_{C1}$ seems to be a little more sensitive than it should be in this condition, but it still improves upon $\bar{X}^2_H$. Furthermore, Q-Q plots in Figure 7.2 show that under the null condition where the item parameters are widely dispersed, the distributions of statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are better approximated by a chi-squared distribution than that of $\bar{X}^2_H$.

Limited-information overall fit statistic M_2 appears to be well calibrated in the null conditions, with relatively stable empirical rejection rates.



(a) $\bar{X}^2_H$ (b) $\bar{X}^2_{C1}$

(c) $\bar{X}^2_{C2}$ (d) $\bar{M}_2$

Figure 7.2: Q-Q plots for a null condition (normal θ, n=12, N=500, dispersed a & b parameters)

7.1.2 Statistical Power under Alternative Hypothesis: Non-normal Latent Variable Distribution

Since the proposed statistics' distribution can be well approximated by a chi-squared distribution with a fixed *df* in the null conditions, the performance of these statistics in the alternative conditions could be further examined. In the alternative conditions, item response data were generated with a non-normally distributed latent variable, but calibrated under the assumption that the latent variable follows a normal distribution. Since the model is misspecified, it is expected that the proposed statistics will not follow a central chi-squared distribution in these conditions. Statistical power indicates to what extent the proposed statistical indices are sensitive to model misspecification.

Table 7.5, Table 7.6, Table 7.7, and Table 7.8 present simulation study results under the alternative hypothesis for 2-PL model. Similiar to the null condition, the number of items is 12, and sample size has three levels, $N = 500, 1000, 1500$. From Table 7.5 to Table 7.8, the level of item parameter dispersion is inceasing. These tables together present the indices' statistical power against latent variable nonnormality across different sample sizes and levels of item parameter dispersion. Results for a longer test ($n = 24$) can be found in the Appendix (see Table A.7, Table A.8, Table A.9, and Table A.10). It is worth mentioning that the number of items exerts a non-ignorable influence on these indices' statistical power. The larger the number of items, the more powerful the statistics are against non-normal latent variable distribution.

Results in Table 7.5 show that in the condition of "Equal a & Equal b", summed score likelihood based statistics, $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$, perform identically well. Comparing with the overall model-fit statistic M_2, they are more sensitive to the nonnormality of θ distribution. With larger sample size, the statistical power of these indices is increasing. Similiar to the results in the null condition, there is not much difference among $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ when item parameters are equal across items.

Table 7.6 presents results in the condition of "Random a & Random b". All the summed score likelihood based statistics still have larger power than the overall model-fit index M_2. But the statistical power of the unadjusted index $\bar{X}^2_H$ becomes consistently lower than the moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$. The difference between the unadjusted and adjusted statistics increases in Table 7.7, where b parameters are highly dispersed.

Table 7.8 demonstrates that when the generating item parameters are most widely dispersed ("Dispersed a & Dispersed b), the $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ indices have larger power than the $\bar{X}^2_H$ statistic to detect latent variable nonnormality. Take the first block ($N = 500$) for example. $\bar{X}^2_H$ has almost no power against latent variable nonnormality at all α levels. On the contrary, the statistical power of $\bar{X}^2_{C1}$ is 0.036 at level $\alpha = 0.1$, 0.142 at $\alpha = 0.05$, and 0.246 at $\alpha = 0.1$, indicating a moderate degree of sensitivity to detection of model misspecification. In this condition, the power of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are very similiar. $\bar{X}^2_{C1}$ enjoys a slightly higher power than $\bar{X}^2_{C2}$. One explanation is that $\bar{X}^2_{C1}$ is generally more sensitive than $\bar{X}^2_{C2}$, in both null and alternative conditions.

7.2 Simulation Results for Graded Response IRT model

7.2.1 Type I Error Rates

Four-category polytomous data were generated for tests with 12 or 24 items. The sample size has three levels: 500, 1000, and 1500. As with 2-PL model, the null hypothesis is that the unidimensional latent variable follows a standard normal distribution. Graded response data were fitted using the same IRT model as data generation. Thus, the statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ should follow their purported chi-squared distribution under the null hypothesis.

Table 7.9, Table 7.10, Table 7.11, and Table 7.12 present simulation results in the null conditions for the graded model. The number of items is 12. Table 7.9 show that when the item parameters are equal, all the proposed statistics, as well as the overall model-fit index M_2, perform equivalently well. Most of the empirical rejection rates fall into their 95% confidence intervals. There is one exception: when the sample size is relatively large ($N = 1500$), M_2 seems to be slightly more sensitive than it should be. At the same time, the expected values of the statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are close to each other.

Table 7.12 presents simulation results when the item parameters are widely dispersed. In the condition where the sample size is small, KS test p-values indicate that $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ cannot be well approximated by their purported chi-squared distribution. But when the sample size increases to 1500, the performance of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ improves, while $\bar{X}^2_H$ has a KS test p-value that is smaller than 0.05. Though, all of these empirical rejection rates fall into its confidence interval at $\alpha = 0.05(0.036, 0.064)$, indicating that the tail area of the statistics' sampling distribution could be well approximated by a chi-squared distribution and these statistics could be used for detecting IRT model misfit.

Results for 24-item graded response model are presented in the Appendix (see Table A.11, Table A.12, Table A.13, and Table A.14). When the number of items

is large and the sample size is small, the empirical rejection rates of the proposed statistics tend to be higher than their α levels, especially when the item parameters are dispersed. However, when the sample size increases, the rejection rates would approach their corresponding α levels.

7.2.2 Statistical Power under Alternative Hypothesis: Non-normal Latent Variable Distribution

Table 7.13, Table 7.14, Table 7.15 and Table 7.16 present simulation results for the graded model in alternative conditions. The statistical power of the proposed statistics were calculated for 12-item tests with 4-category polytomous data. Table 7.13 shows results for the condition where all the item parameters are equal across items. Table 7.16 contains results for the condition where both item slope and threshold parameters are widely dispersed. Table 7.14 and Table 7.15 present results for conditions with moderately dispersed item parameters.

Results in Table 7.13 show that when latent variable distribution is misspecified for the graded models, the proposed statistics can detect this kind of model misspecification, and their statistical power improves when the sample size increases. For example, when the sample size is 500, the means of the statistics $\bar{X}^2_H$, $\bar{X}^2_{C1}$, and $\bar{X}^2_{C2}$ are respectively 37.59, 37.78 and 37.76, slightly greater than the degrees of freedom ($df = 34$). Their statistical power at $\alpha = 0.05$ are 0.11, 0.116, and 0.114 respectively. When the sample size is 1500, the statistical power of the three indices increase to 0.404, 0.412, and 0.408 respectively. Obviously, the power of the unadjusted and moment adjusted statistics are very close to each other.

In the condition where item parameters are widely dispersed, as shown in Table 7.16, even though all the summed score likelihood based statistics are sensitive to latent variable nonnormality, the moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ has higher statistical power than the unadjusted indice $\bar{X}^2_H$. Likewise, all the three indexes have more statistical power against model misfit with larger sample size. In this condition, M_2 always has much lower statistical power than the other three indices.

Another finding is that for the graded model, even when the item parameters

are dispersed, the first-moment adjusted indice $\bar{X}^2_{C1}$ performs as well as the two-moment adjusted indice $\bar{X}^2_{C2}$, and has a slightly larger power than $\bar{X}^2_{C2}$ in most conditions. More results for the graded model can be found in the Appendix (see Table A.14, Table A.15, Table A.16, and Table A.17)

7.3 Discussion

With simulated data, the properties of the proposed moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ for detecting latent variable nonnormality were examined across different numbers of items, sample sizes, item types and values of generating item parameters. Meanwhile, their performance were compared to unadjusted summed score likelihood based statistic $\bar{X}^2_H$ (Li & Cai, 2012) and overall GOF statistic M_2 (Maydeu-Olivares & Joe, 2005). This chapter presents all the simulation study results.

In the null conditions, when the fitted models and data generating models are the same, the expected values of statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are mostly close to the value of *df*, indicating that the distribution of moment-adjusted statistics could be well approximated by a chi-squared distribution. Both moment adjusted indices $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ perform better than the unadjusted statistic $\bar{X}^2_H$, especially when the values of generating item parameters are widely dispersed. It's worth mentioning that, values of item parameters are very likely to spread to some extent between completely equality and extreme dispersion in a real test. Thus, the use of moment adjusted statistics is recommended.

In the alternative conditions, when the generating latent variable is nonnormally distributed, the statistics $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ turned out to be sensitive to the violation of assumption of latent variable normality, while the overall limited-information GOF statistic M_2 has almost no power against the nonnormal alternative. This could be explained by the observation that M_2 is based only on first and second order margins of the underlying contingency table, but to detect latent variable distributional misfit, information from higher order margins and interactions might be necessary. Additionally, when the values of generating item parameters are equal across items, $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ perform equivalently well. Otherwise, when the item parameters are dispersed, the adjusted indices $\bar{X}^2_{C1}$ and

$\bar{X}^2_{C2}$ enjoy greater power than the unadjusted index $\bar{X}^2_H$.

7.3.1 Do the Moment Adjusted Statistics Improve upon the Unadjusted One?

Summed score likelihood based statistic $\bar{X}^2_H$ (Li & Cai, 2012) proved to be sensitive to latent variable nonnormality for IRT models. One challenge of applying this statistic is that its distribution is not asymptotically chi-squared. In some scenarios, adjustments to the statistic or degrees of freedom might increase its statistical power. By matching its first one or two moments with those of a referenced chi-squared distribution, moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ were expected to perform better than $\bar{X}^2_H$.

In most of the conditions I tested, when compared with $\bar{X}^2_H$, simulation study results show that $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ can be better approximated by a chi-squared distribution in the null conditions, and enjoy increased statistical power in the alternative conditions. Furthermore, the dispersion of generating slope and threshold parameters exerts a non-ignorable influence on the comparative performance of the statistics. When the generating item parameters are equal, $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ perform equally well. However, in the case of high dispersion of generating item parameters, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ would improve upon $\bar{X}^2_H$. In order to illustrate this finding, Table 7.17 and Table 7.18 present the empirical rejection rates and statistical power of $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ at three α levels across different values of generating item parameters: "Equal a & b", "Random a & b", and "Dispersed a & b".

According to Table 7.17, the empirical rejection rates and statistical power of $\bar{X}^2_H$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are very close in the condition of "Equal a & b". But when the generating item parameters are unequal, moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ improve upon unadjusted statistic $\bar{X}^2_H$ in general. For example, the statistical

power of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are higher than that of $\bar{X}^2_H$, even when both slope and threshold parameters were randomly generated. This difference increases when item parameters become more dispersed. However, this is not the case for the graded model (see Table 7.18). The statistical power of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are greater than that of $\bar{X}^2_H$ only when item parameters are widely dispersed.

The reason why the values of item parameters exert an influence on the performance of the test statistics is demonstrated as following. As illustrated in Chapter 5, the *df* of $\bar{X}^2_H$ is equal to $S - 1 - 2$. "-2" is a heuristic value to adjust the degrees of freedom due to item parameter estimation. When the values of a and b parameters are equal, the expected values of $\bar{X}^2_H$ are very close to its heuristic *df*. But when the values of item parameters are widely dispersed, the expected values of $\bar{X}^2_H$ are often smaller than the heuristic *df*. With moment adjustments, the expected values of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ better approach their purported degrees of freedom. Therefore, the performance of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ improve upon $\bar{X}^2_H$ when generating item parameters are dispersed.

7.3.2 The Influence of Other Factors on the Performance of Statistics

In addition to item parameter dispersion, other properties of tests also exert an influence on the performance of the proposed statistics. Results from simulation studies for both 2-PL and the graded model show that the empirical rejection rates in the null conditions and statistical power in the alternative conditions are highly influenced by the sample size and number of items.

In the null conditions, the empirical rejection rates approach their α levels with increased sample size increases. In the alternative conditions, the statistical power of the proposed statistics grows when the sample size is augmented. For instance, as shown in section 7.1.2, when the sample size is small ($N = 500$), $\bar{X}^2_H$ has a very low power against model misfit for 2-PL models, and it has a slightly larger power when the sample size becomes 1000. However, when the sample size changes from 1000 to 1500, $\bar{X}^2_H$ becomes certainly sensitive to latent variable nonnormality. Meanwhile, the moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ always have more statistical power than $\bar{X}^2_H$ across different sample sizes.

It is also evident that the statistical power of the proposed indices improves when the number of items increases. As previously discussed, when the number of items is small ($n = 12$), the unadjusted statistic $\bar{X}^2_H$ has very low power against latent variable nonnormality. But when the number of items increases to 24, the statistical power of $\bar{X}^2_H$ becomes moderately large. The adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are in general more sensitive than $\bar{X}^2_H$, but the difference between the adjusted and unadjusted indexes decreases with a growing number of items.

7.3.3 Summary

In sum, simulation study results suggest the following: 1) Compared with $\bar{X}^2_H$ (Li & Cai, 2012), the moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ can be better approximated by a chi-squared distribution in the null conditions and enjoy higher

statistical power in the alternative conditions, especially in the situation when the generating item parameters are dispersed; 2) Compared with the overall GOF statistic M_2, all of the summed score likelihood based statistics are more sensitive to the violation of the assumption of normal latent variable distribution in alternative conditions. Therefore, the proposed statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are prefered for detecting latent variable nonnormality. The two-moment adjusted statistic $\bar{X}^2_{C2}$ and the first-moment adjusted statistic $\bar{X}^2_{C1}$ perform equally well in most conditions, and $\bar{X}^2_{C1}$ seems to have a slightly larger power.

Results from simulation studies also show that the performance of the proposed statistics could be influenced by several factors. First of all, sample sizes and numbers of items are highly influential. For both 2-PL and graded models, when the sample size increases, the empirical rejection rates of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ approach their corresponding α levels, and their statistical power against latent variable nonnormality become larger in alternative conditions. Likewise, the statistical power of $\bar{X}^2_{H}$, $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ grow with increased number of items.

Nonetheless, the proposed statistics should be used with caution in some situations, especially when the number of items is large and the sample size is small. Results for graded models show that in the condition of $n = 24$ and $N = 500$, all of the summed score likelihood statistics have empirical rejection rates larger than their corresponding α levels, indicating that the tail area of their distribution cannot be well approximated by a chi-squared distribution.

Another important finding is the influence of values of generating item parameters on the comparison between unadjusted and adjusted statistics. It is evident that the more dispersed the generating item parameters are, the more effective the proposed moment adjustment approaches turn out to be. For 2-PL models, when item slope and threshold parameters are both randomly generated, the statistical power of $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ are mostly greater than $\bar{X}^2_{H}$. However, it is worth noting that this study only considers four levels of item parameter dispersion, and the

differences between two adjacent levels are not equivalent.

Table 7.1: Results of simulation study for the null conditions (*2-PL*, Equal a & Equal b, $n = 12$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	10	10.20	18.57	1.31	28.12	.008	.038	.104	.100
	$\bar{X}^2_{C1}$	100%	10	10.26	18.82	1.32	28.03	.010	.040	.106	.039∗
	$\bar{X}^2_{C2}$	100%	10	10.26	18.60	1.38	28.11	.010	.038	.106	.038∗
	M_2	100%	54	54.59	104.49	26.90	91.74	.008	.060	.104	.562
1000	$\bar{X}^2_H$	100%	10	9.99	19.62	.94	27.61	.009	.044	.098	.621
	$\bar{X}^2_{C1}$	100%	10	10.02	19.71	.95	27.70	.009	.044	.100	.475
	$\bar{X}^2_{C2}$	100%	10	10.02	19.62	.96	27.65	.009	.044	.100	.485
	M_2	100%	54	54.09	111.21	28.64	96.27	.010	.051	.110	.626
1500	$\bar{X}^2_H$	100%	10	10.07	21.75	1.89	31.43	.012	.060	.108	.988
	$\bar{X}^2_{C1}$	100%	10	10.09	21.79	1.90	31.48	.012	.058	.110	.989
	$\bar{X}^2_{C2}$	100%	10	10.09	21.74	1.91	31.42	.012	.058	.110	.982
	M_2	100%	54	53.44	103.53	27.33	85.71	.008	.048	.090	.183

Note. Valid% indicates the percentage of valid calculations in 500 replications; "Valid calculation" means that the generating response data were calibrated normally, and that values of the statistics with quadratic forms are not negative or infinite. $*$ indicates that p values of KS test are smaller than .05.

Table 7.2: Results of simulation study for the null conditions (*2-PL*, $n = 12$, Random a & Random b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	10	9.27	18.92	1.21	27.19	.012	.042	.076	.000∗
	$\bar{X}^2_{C1}$	100%	10	9.94	21.83	1.28	28.88	.018	.056	.112	.642
	$\bar{X}^2_{C2}$	100%	10	9.94	21.18	1.34	28.72	.018	.052	.106	.641
	M_2	100%	54	53.36	123.33	27.25	91.54	.014	.062	.112	.012
1000	$\bar{X}^2_H$	100%	10	9.17	17.96	1.03	26.80	.010	.036	.066	.002∗
	$\bar{X}^2_{C1}$	100%	10	9.79	20.51	1.10	28.90	.012	.044	.102	.495
	$\bar{X}^2_{C2}$	100%	10	9.79	20.07	1.20	28.65	.012	.044	.102	.605
	M_2	100%	54	53.64	116.48	31.67	94.64	.016	.046	.098	.554
1500	$\bar{X}^2_H$	100%	10	9.43	16.31	1.23	24.10	.004	.024	.068	.127
	$\bar{X}^2_{C1}$	100%	10	10.05	18.54	1.31	25.93	.006	.038	.100	.472
	$\bar{X}^2_{C2}$	100%	10	10.05	18.19	1.41	25.68	.006	.036	.100	.418
	M_2	100%	54	53.75	112.83	25.28	90.29	.012	.050	.092	.895

Note. $*$ indicates that p values of KS test are smaller than .05.

Table 7.3: Results of simulation study for the null conditions (*2-PL*, $n = 12$, Random a & Random b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	10	9.34	17.99	1.56	31.97	.004	.024	.078	.002∗
	$\bar{X}^2_{C1}$	100%	10	10.20	21.71	1.71	35.81	.016	.062	.120	.704
	$\bar{X}^2_{C2}$	100%	10	10.19	20.77	1.86	34.71	.014	.062	.120	.824
	M_2	100%	54	54.64	101.92	23.56	85.54	.006	.058	.098	.047∗
1000	$\bar{X}^2_H$	100%	10	9.16	16.85	1.18	27.08	.003	.033	.063	.000∗
	$\bar{X}^2_{C1}$	100%	10	9.94	19.94	1.28	29.62	.010	.047	.096	.802
	$\bar{X}^2_{C2}$	100%	10	9.94	19.28	1.40	29.25	.007	.045	.096	.942
	M_2	100%	54	54.69	103.49	28.20	97.97	.011	.058	.104	.019∗
1500	$\bar{X}^2_H$	100%	10	9.52	15.64	1.80	25.92	.004	.026	.064	.064
	$\bar{X}^2_{C1}$	100%	10	10.30	18.23	1.96	27.72	.010	.046	.094	.009∗
	$\bar{X}^2_{C2}$	100%	10	10.29	17.76	2.11	27.60	.010	.046	.092	.005∗
	M_2	100%	54	54.45	112.34	31.91	87.32	.012	.062	.112	.644

Note. $*$ indicates that p values of KS test are smaller than .05.

Table 7.4: Results of simulation study for null conditions (*2-PL*, Dispersed a & Dispersed b, $n = 12$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	10	8.61	16.10	1.56	25.44	.004	.032	.054	.000∗
	$\bar{X}^2_{C1}$	100%	10	9.81	20.97	1.78	29.37	.008	.056	.102	.105
	$\bar{X}^2_{C2}$	100%	10	9.82	19.71	2.06	28.48	.008	.050	.090	.227
	M_2	100%	54	54.80	122.02	25.49	86.79	.016	.078	.132	.239
1000	$\bar{X}^2_H$	100%	10	9.01	15.81	1.15	24.01	.003	.024	.059	.000∗
	$\bar{X}^2_{C1}$	100%	10	10.17	20.19	1.29	27.33	.008	.056	.112	.248
	$\bar{X}^2_{C2}$	100%	10	10.16	19.19	1.46	26.67	.006	.052	.108	.218
	M_2	100%	54	54.06	102.99	27.48	98.20	.013	.048	.098	.644
1500	$\bar{X}^2_H$	100%	10	8.56	15.35	.48	22.45	.000	.018	.044	.000∗
	$\bar{X}^2_{C1}$	100%	10	9.62	19.34	.54	25.25	.012	.042	.082	.022∗
	$\bar{X}^2_{C2}$	100%	10	9.63	18.51	.75	24.99	.012	.040	.078	.051
	M_2	100%	54	53.94	103.98	28.26	93.98	.014	.048	.092	.339

Note. ∗ indicates that p values of KS test are smaller than .05.

Table 7.5: Results of simulation study for the alternative conditions (*2-PL*, $n = 12$, Equal a & Equal b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	10	12.95	30.76	3.16	36.05	.044	.160	.282
	$\bar{X}^2_{C1}$	100%	10	13.10	31.87	3.17	36.20	.052	.164	.294
	$\bar{X}^2_{C2}$	100%	10	13.04	30.91	3.20	35.98	.050	.162	.284
	M_2	100%	54	54.14	101.26	31.82	85.89	.004	.050	.104
1000	$\bar{X}^2_H$	100%	10	15.51	38.69	1.43	43.11	.114	.296	.421
	$\bar{X}^2_{C1}$	100%	10	15.63	39.67	1.43	43.46	.121	.302	.431
	$\bar{X}^2_{C2}$	100%	10	15.58	38.78	1.47	43.06	.115	.298	.426
	M_2	100%	54	53.78	111.28	25.45	99.35	.011	.046	.098
1500	$\bar{X}^2_H$	100%	10	18.98	48.81	4.33	46.96	.260	.504	.626
	$\bar{X}^2_{C1}$	100%	10	19.11	49.90	4.35	47.30	.268	.516	.632
	$\bar{X}^2_{C2}$	100%	10	19.03	48.82	4.36	46.93	.262	.510	.630
	M_2	100%	54	53.82	111.84	25.71	96.63	.008	.060	.112

Note. Valid% indicates the percentage of valid calculations in 500 replications; "Valid calculation" means that the generating response data were calibrated normally, and that values of the statistics with quadratic forms are not negative or infinite.

Table 7.6: Results of simulation study for the alternative conditions (*2-PL*, $n = 12$, Random a & Random b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	10	13.44	28.44	2.83	36.46	.046	.170	.290
	$\bar{X}^2_{C1}$	100%	10	14.53	33.88	3.06	40.04	.082	.234	.368
	$\bar{X}^2_{C2}$	100%	10	14.40	31.97	3.18	38.91	.076	.226	.366
	M_2	100%	54	54.39	119.57	28.55	100.65	.014	.068	.104
1000	$\bar{X}^2_H$	100%	10	16.83	45.29	4.07	42.02	.174	.352	.478
	$\bar{X}^2_{C1}$	100%	10	18.11	53.04	4.36	44.83	.220	.432	.556
	$\bar{X}^2_{C2}$	100%	10	17.93	50.59	4.43	44.32	.214	.422	.552
	M_2	100%	54	54.17	119.67	24.33	95.83	.018	.052	.120
1500	$\bar{X}^2_H$	100%	10	20.86	58.85	4.97	49.26	.332	.612	.708
	$\bar{X}^2_{C1}$	100%	10	22.41	68.75	5.31	52.56	.414	.668	.746
	$\bar{X}^2_{C2}$	100%	10	22.16	65.60	5.36	51.64	.404	.664	.742
	M_2	100%	54	54.35	116.05	30.41	88.66	.014	.066	.112

Note. Valid% indicates the percentage of valid calculations in 500 replications.

Table 7.7: Results of simulation study for the alternative conditions (*2-PL*, $n = 12$, Random a & Dispersed b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	10	11.23	20.76	2.53	28.62	.014	.080	.156
	$\bar{X}^2_{C1}$	100%	10	12.39	25.84	2.74	31.72	.038	.122	.236
	$\bar{X}^2_{C2}$	100%	10	12.29	24.10	2.87	30.83	.030	.112	.226
	M_2	100%	54	54.85	106.30	27.24	89.34	.004	.062	.128
1000	$\bar{X}^2_H$	100%	10	13.53	30.10	2.54	38.17	.053	.181	.305
	$\bar{X}^2_{C1}$	100%	10	14.82	36.63	2.77	42.54	.095	.259	.375
	$\bar{X}^2_{C2}$	100%	10	14.68	34.56	2.88	41.16	.085	.247	.369
	M_2	100%	54	54.01	109.72	27.80	93.05	.012	.056	.099
1500	$\bar{X}^2_H$	100%	10	15.62	28.81	3.39	38.77	.076	.306	.448
	$\bar{X}^2_{C1}$	100%	10	17.04	34.70	3.70	42.74	.142	.390	.528
	$\bar{X}^2_{C2}$	100%	10	16.87	32.92	3.83	41.64	.138	.378	.522
	M_2	100%	54	54.73	102.22	31.78	89.09	.008	.064	.106

Note. Valid% indicates the percentage of valid calculations in 500 replications.

Table 7.8: Results of simulation study for the alternative conditions (*2-PL*, $n = 12$, Dispersed a & Dispersed b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	10	10.94	20.50	2.03	27.00	.010	.064	.134
	$\bar{X}^2_{C1}$	100%	10	12.69	28.07	2.30	32.46	.036	.142	.246
	$\bar{X}^2_{C2}$	100%	10	12.55	25.43	2.47	30.49	.028	.132	.230
	M_2	100%	54	56.52	146.36	29.21	119.05	.024	.092	.148
1000	$\bar{X}^2_H$	100%	10	12.49	26.34	2.29	41.87	.030	.116	.214
	$\bar{X}^2_{C1}$	100%	10	14.29	34.81	2.57	48.43	.067	.209	.337
	$\bar{X}^2_{C2}$	100%	10	14.13	32.30	2.74	47.01	.062	.199	.319
	M_2	100%	54	55.62	117.32	29.28	110.31	.016	.073	.141
1500	$\bar{X}^2_H$	100%	10	14.34	26.98	3.17	35.51	.074	.202	.338
	$\bar{X}^2_{C1}$	100%	10	16.33	35.11	3.58	40.71	.128	.330	.490
	$\bar{X}^2_{C2}$	100%	10	16.12	32.86	3.76	39.53	.114	.312	.480
	M_2	100%	54	56.25	136.41	26.43	98.99	.024	.092	.146

Note. Valid% indicates the percentage of valid calculations in 500 replications.

Table 7.9: Simulation study results for graded model in the null conditions ($n = 12$, Equal a & Equal b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	34	34.589	70.734	16.195	71.457	.012	.058	.108	.174
	$\bar{X}^2_{C1}$	100%	34	34.661	70.770	16.211	71.265	.012	.060	.114	.121
	$\bar{X}^2_{C2}$	100%	34	34.66	70.59	16.22	71.36	.012	.060	.114	.120
	M_2	100%	30	30.20	58.91	12.72	57.95	.012	.052	.122	.627
1000	$\bar{X}^2_H$	100%	34	34.27	74.49	15.72	77.22	.018	.064	.104	.711
	$\bar{X}^2_{C1}$	100%	34	34.30	74.52	15.74	77.03	.018	.064	.106	.703
	$\bar{X}^2_{C2}$	100%	34	34.30	74.49	15.75	77.13	.018	.064	.106	.682
	M_2	99%	30	30.09	67.53	2.27	77.40	.014	.040	.097	.733
1500	$\bar{X}^2_H$	100%	34	33.96	65.97	15.10	65.05	.012	.048	.100	.753
	$\bar{X}^2_{C1}$	100%	34	33.98	66.01	15.12	65.01	.012	.050	.100	.751
	$\bar{X}^2_{C2}$	100%	34	33.99	65.99	15.13	65.04	.012	.050	.100	.759
	M_2	98.8%	30	31.18	80.92	2.85	81.86	.020	.075	.142	.003*

Note. Valid% indicates the percentage of valid calculations in 500 replications; "Valid calculation" means that the generating response data were calibrated normally, and that values of the statistics with quadratic forms are not negative or infinite. $*$ indicates that KS test $p < 0.05$

Table 7.10: Simulation study results for graded model in the null conditions ($n = 12$, Random a & Random b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	34	34.21	65.40	15.21	62.76	.014	.050	.102	.530
	$\bar{X}^2_{C1}$	100%	34	34.60	66.83	15.43	63.59	.014	.052	.110	.369
	$\bar{X}^2_{C2}$	100%	34	34.60	66.50	15.51	63.48	.014	.052	.110	.372
	M_2	100%	30	30.16	68.64	11.88	55.79	.016	.074	.118	.452
1000	$\bar{X}^2_H$	100%	34	34.19	72.62	12.65	67.11	.010	.054	.118	.570
	$\bar{X}^2_{C1}$	100%	34	34.52	74.04	12.79	67.71	.013	.060	.127	.061
	$\bar{X}^2_{C2}$	100%	34	34.52	73.89	12.84	67.71	.013	.058	.127	.060
	M_2	100%	30	30.20	57.55	11.35	59.99	.009	.037	.090	.078
1500	$\bar{X}^2_H$	100%	34	33.88	72.37	12.65	62.17	.014	.054	.106	.745
	$\bar{X}^2_{C1}$	100%	34	34.20	73.78	12.77	62.73	.016	.058	.112	.673
	$\bar{X}^2_{C2}$	100%	34	34.19	73.66	12.77	62.73	.016	.058	.112	.681
	M_2	100%	30	29.93	57.87	13.38	54.74	.006	.052	.114	.982

Note. $*$ indicates that KS test $p < 0.05$

Table 7.11: Simulation study results for graded model in the null conditions ($n = 12$, Random a & Dispersed b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	34	33.73	66.57	11.88	63.69	.008	.032	.098	.910
	$\bar{X}^2_{C1}$	99.6%	34	34.42	69.16	12.21	64.83	.010	.042	.118	.107
	$\bar{X}^2_{C2}$	99.6%	34	34.42	68.52	12.50	64.72	.008	.042	.116	.095
	M_2	100%	30	30.13	68.00	10.63	71.96	.012	.064	.122	.878
1000	$\bar{X}^2_H$	100%	34	34.25	59.48	15.78	63.45	.008	.046	.086	.405
	$\bar{X}^2_{C1}$	100%	34	34.87	61.58	16.04	64.51	.010	.052	.100	.127
	$\bar{X}^2_{C2}$	100%	34	34.87	61.19	16.07	64.42	.010	.052	.098	.123
	M_2	100%	30	29.64	52.70	13.46	57.34	.006	.034	.084	.397
1500	$\bar{X}^2_H$	100%	34	33.57	68.68	15.10	60.14	.012	.054	.098	.092
	$\bar{X}^2_{C1}$	100%	34	34.14	71.10	15.34	60.93	.014	.064	.112	.750
	$\bar{X}^2_{C2}$	100%	34	34.14	70.72	15.37	60.94	.014	.064	.112	.762
	M_2	100%	30	30.12	64.31	14.84	66.75	.012	.074	.114	.758

Note. $*$ indicates that KS test $p < 0.05$

Table 7.12: Simulation study results for graded model in the null conditions ($n = 12$, Dispersed a & Dispersed b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	99.2%	34	34.52	71.10	15.68	68.31	.014	.054	.107	.097
	$\bar{X}^2_{C1}$	93.2%	34	35.51	74.65	16.41	69.80	.017	.064	.135	.000*
	$\bar{X}^2_{C2}$	93.2%	34	35.50	73.83	16.86	69.70	.017	.062	.133	.000*
	M_2	100%	30	29.45	64.55	11.88	71.55	.018	.040	.092	.077
1000	$\bar{X}^2_H$	99.8%	34	34.02	61.43	16.06	60.34	.008	.044	.094	.893
	$\bar{X}^2_{C1}$	99.8%	34	34.82	64.36	16.50	61.57	.010	.054	.110	.280
	$\bar{X}^2_{C2}$	99.8%	34	34.81	63.85	16.61	61.47	.010	.052	.110	.247
	M_2	100%	30	29.31	53.16	13.65	57.30	.008	.036	.070	.185
1500	$\bar{X}^2_H$	100%	34	33.18	61.25	15.70	62.10	.006	.046	.082	.018*
	$\bar{X}^2_{C1}$	100%	34	33.91	63.93	15.98	63.26	.008	.048	.090	.747
	$\bar{X}^2_{C2}$	100%	34	33.91	63.56	16.01	63.24	.008	.048	.090	.731
	M_2	100%	30	29.99	60.39	10.29	57.73	.004	.050	.100	.840

Note. $*$ indicates that KS test $p < 0.05$

Table 7.13: Simulation study results for graded model in the alternative conditions ($n = 12$, Equal a & Equal b)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	34	37.59	78.22	17.36	70.58	.026	.110	.196
	$\bar{X}^2_{C1}$	100%	34	37.78	79.43	17.39	71.22	.030	.116	.208
	$\bar{X}^2_{C2}$	100%	34	37.76	78.47	17.42	70.80	.028	.114	.202
	M_2	100%	30	30.50	61.18	11.05	57.38	.008	.054	.122
1000	$\bar{X}^2_H$	100%	34	42.10	111.62	13.96	84.17	.086	.266	.372
	$\bar{X}^2_{C1}$	100%	34	42.26	113.06	14.00	84.34	.090	.268	.380
	$\bar{X}^2_{C2}$	100%	34	42.22	111.98	14.05	84.21	.090	.268	.380
	M_2	99.4%	30	30.75	70.72	2.47	82.35	.006	.066	.127
1500	$\bar{X}^2_H$	100%	34	46.94	119.20	19.44	82.98	.198	.404	.562
	$\bar{X}^2_{C1}$	100%	34	47.11	120.47	19.48	83.26	.204	.412	.564
	$\bar{X}^2_{C2}$	100%	34	47.05	119.42	19.50	83.07	.204	.408	.564
	M_2	99.2%	30	31.83	93.30	12.66	82.90	.034	.101	.157

Note. Valid% indicates the percentage of valid calculations in 500 replications; "Valid calculation" means that the generating response data were calibrated normally, and that values of the statistics with quadratic forms are not negative or infinite.

Table 7.14: Simulation study results for graded model in the alternative conditions ($n = 12$, Random a & Random b)

N	statistics	valid %	df	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	34	39.30	82.50	17.62	80.13	.048	.128	.246
	$\bar{X}^2_{C1}$	100%	34	39.94	85.71	17.79	81.22	.054	.140	.276
	$\bar{X}^2_{C2}$	100%	34	39.88	83.90	17.79	81.02	.052	.138	.274
	M_2	100%	30	31.86	67.01	14.42	67.21	.020	.076	.154
1000	$\bar{X}^2_H$	100%	34	45.25	114.63	21.01	83.08	.139	.344	.468
	$\bar{X}^2_{C1}$	100%	34	45.89	118.86	21.22	84.82	.156	.365	.485
	$\bar{X}^2_{C2}$	100%	34	45.80	116.86	21.25	84.05	.154	.362	.483
	M_2	100%	30	31.59	64.84	11.29	64.42	.019	.072	.144
1500	$\bar{X}^2_H$	100%	34	52.44	110.32	25.64	83.26	.342	.616	.744
	$\bar{X}^2_{C1}$	100%	34	53.17	114.32	25.91	84.74	.368	.646	.756
	$\bar{X}^2_{C2}$	100%	34	53.03	112.33	25.93	84.12	.366	.642	.754
	M_2	100%	30	31.85	70.13	13.97	59.60	.030	.096	.162

Note. Valid% indicates the percentage of valid calculations in 500 replications

Table 7.15: Simulation study results for graded model in the alternative conditions ($n = 12$, Random a & Dispersed b)

N	statistics	valid %	df	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	99.6%	34	33.91	75.51	15.53	65.12	.024	.092	.155
	$\bar{X}^2_{C1}$	99.6%	34	36.83	79.24	16.01	66.58	.032	.104	.183
	$\bar{X}^2_{C2}$	99.6%	34	36.80	77.64	16.26	66.35	.030	.102	.181
	M_2	100%	30	30.68	61.05	12.00	57.25	.010	.064	.124
1000	$\bar{X}^2_H$	100%	34	39.31	80.15	16.26	72.34	.038	.146	.250
	$\bar{X}^2_{C1}$	100%	34	40.16	83.80	16.61	73.59	.046	.160	.292
	$\bar{X}^2_{C2}$	100%	34	40.12	82.62	16.71	73.44	.044	.156	.292
	M_2	100%	30	30.21	62.63	13.11	53.62	.010	.058	.120
1500	$\bar{X}^2_H$	100%	34	43.46	99.25	19.45	78.38	.106	.286	.408
	$\bar{X}^2_{C1}$	100%	34	44.36	104.10	19.78	79.71	.128	.324	.442
	$\bar{X}^2_{C2}$	100%	34	44.29	102.52	19.83	79.58	.126	.324	.442
	M_2	100%	30	30.68	64.47	13.65	54.78	.012	.060	.132

Note. Valid% indicates the percentage of valid calculations in 500 replications

Table 7.16: Simulation study results for graded model in the alternative conditions ($n = 12$, Dispersed a & Dispersed b)

N	statistics	valid %	df	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	95.8%	34	37.36	72.74	16.59	71.30	.013	.094	.190
	$\bar{X}^2_{C1}$	80.6%	34	38.84	79.28	17.13	73.14	.040	.129	.236
	$\bar{X}^2_{C2}$	80.6%	34	38.75	76.95	17.30	72.80	.032	.127	.233
	M_2	100%	30	30.71	69.63	10.02	71.64	.014	.058	.112
1000	$\bar{X}^2_H$	99%	34	40.91	83.56	18.05	78.23	.057	.188	.311
	$\bar{X}^2_{C1}$	96.2%	34	42.14	88.21	18.49	80.07	.079	.227	.360
	$\bar{X}^2_{C2}$	96.2%	34	42.07	86.70	18.60	79.87	.077	.225	.356
	M_2	100%	30	30.85	66.94	9.05	53.83	.012	.064	.136
1500	$\bar{X}^2_H$	99.8%	34	46.19	107.04	22.44	87.23	.172	.389	.515
	$\bar{X}^2_{C1}$	99.4%	34	47.43	113.70	22.99	89.86	.203	.427	.559
	$\bar{X}^2_{C2}$	99.4%	34	47.32	111.73	23.04	89.28	.201	.425	.557
	M_2	100%	30	31.31	72.75	9.45	64.02	.020	.076	.146

Note. Valid% indicates the percentage of valid calculations in 500 replications

Table 7.17: Comparison of $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ across values of generating item parameters for 2-PL models ($n = 12$, $N = 500$)

	Levels	Equal a & b			Random a & b			Dispersed a & b		
		$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$	$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$	$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$
Rejection rates	$\alpha = .01$	.008	.010	.010	.012	.018	.018	.004	.008	.008
	$\alpha = .05$	.038	.040	.038	.042	.056	.052	.032	.056	.050
	$\alpha = .10$	.104	.106	.106	.076	.112	.106	.054	.102	.090
Power	$\alpha = .01$	.044	.052	.050	.046	.082	.076	.010	.036	.028
	$\alpha = .05$	.160	.164	.162	.170	.234	.226	.064	.142	.132
	$\alpha = .10$	.282	.294	.284	.290	.368	.366	.134	.246	.230

Table 7.18: Comparison of $\bar{X}^2_{C1}$, $\bar{X}^2_{C2}$ and $\bar{X}^2_H$ across values of generating item parameters for graded models ($n = 12$, $N = 500$)

	Levels	Equal a & b			Random a & b			Dispersed a & b		
		$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$	$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$	$\bar{X}^2_H$	$\bar{X}^2_{C1}$	$\bar{X}^2_{C2}$
Rejection rates	$\alpha = .01$	.012	.012	.012	.014	.014	.014	.014	.017	.017
	$\alpha = .05$	.058	.060	.060	.050	.052	.052	.054	.064	.062
	$\alpha = .10$	.108	.114	.114	.102	.110	.110	.106	.135	.133
Power	$\alpha = .01$	.026	.030	.028	.048	.054	.052	.012	.040	.032
	$\alpha = .05$	.110	.116	.114	.128	.140	.138	.090	.129	.127
	$\alpha = .10$	.196	.208	.202	.246	.276	.274	.182	.236	.233

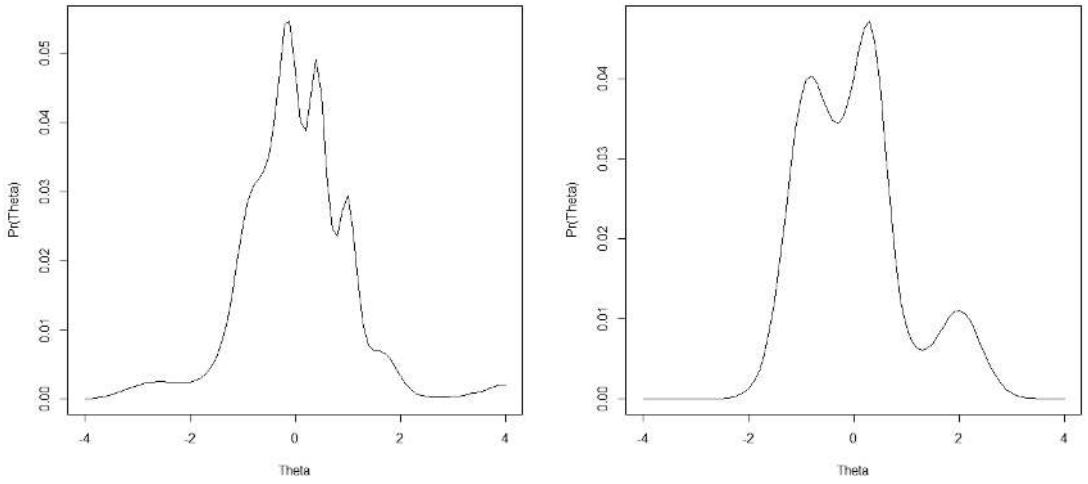
CHAPTER 8

Empirical Applications

In this chapter, I present results from applications of the proposed summed score likelihood based statistics to psychological and educational datasets. The first dataset consists of 2,717 cigarette smokers' responses to 12 items regarding positive consequences of nicotine (PCN), as part of a questionnaire dealing with various attitudes, beliefs and behaviors related to smoking (see Shadel, Edelen, & Tucker, 2011). The data were collected as part of the National Institute of Health's Patient Reported Outcomes Measurement Information System (PROMIS) Smoking Initiative (Edelen, 2014). It is plausible that these 12 items measure a common latent dimension. The density plot (Figure 8.1a) of the latent variable distribution shows its deviation from a standard normal distribution as there are two maximum points in the middle instead of a "bell curve" shape.

The second data set contains 10 items from the PISA 2012 mathematics assessment (Kastberg, Roey, Lemanski, Chan, & Murray, 2014). The sample consists of 1,648 students from three different countries, and the latent variable is very likely to be non-normal when subgroups with different means and variances are combined together. Figure 8.1b shows the density plot of the latent variable distribution for the 10-item test data. The distribution appears to have three modes at different values of θ, enforcing the conjecture that there are three different subgroups.

Figure 8.1: Latent variable distribution for empirical data sets



(a) PROMIS Smoking Initiative (b) PISA Mathematical Achievement Test

8.1 PROMIS Smoking Initiative

Table 8.1 presents the contents of the 12 items from PROMIS smoking assessment. All 12 items assess positive consequences of nicotine (Tucker et al., 2014), and each item is rated on a 5-point ordinal scale: "Not at all", "A little bit", "Somewhat", "Quite a bit", or "Very much".

The fitted IRT model contains a normal latent variable. Item parameter estimates and standard errors of parameter estimation are presented in the Appendix. The unadjusted and moment adjusted nonnormality detecting indices were calculated. Results show that $\bar{X}^2_H$ equals to 208.46; $\bar{X}^2_{C1}$ equals to 179.28, and $\bar{X}^2_{C2}$ equals to 195.17. All three statistics indicate significant model misfit (df=46, p=0.000). However, when the empirical histogram latent density estimation was used in the parameter estimation, $\bar{X}^2_H$ equals to 51.21 (df=46; p=0.277), $\bar{X}^2_{C1}$ equals to 49.38 (df=46; p=0.340), and $\bar{X}^2_{C2}$ equals to 49.45 (df=46; p=0.340). Thus, the latent variable distribution for the 12 items is non-normal, and the proposed statistics were sensitive to latent variable nonnormality.

Table 8.1: Items from PROMIS Smoking Initiative

	Item wordings
Item 1	Smoking helps me concentrate.
Item 2	Smoking helps me think more clearly.
Item 3	Smoking helps me stay focused.
Item 4	Smoking makes me feel better in social situations.
Item 5	Smoking makes me feel more self-confident with others.
Item 6	Smoking helps me feel more relaxed when I'm with other people.
Item 7	Smoking helps me deal with anxiety.
Item 8	Smoking calms me down.
Item 9	If I'm feeling irritable, a cigarette will help me relax.
Item 10	Smoking a cigarette energizes me.
Item 11	Smoking makes me feel less tired.
Item 12	Smoking perks me up.

8.2 PISA Mathematical Assessment

The proposed statistics were applied to an educational data set, 1,648 students' responses to 10 items from the paper-based mathematical test in PISA 2012 (Kastberg et al., 2014). As in other large-scale assessments, test items in PISA were compiled into clusters, and administerd in the form of booklets. The items used in this study were located in the cluster "PM2" of Booklet 12 in PISA 2012. This cluster was also administered in PISA 2003, 2006, and 2009. The selected items were scored as either "no credit" or "full credit". The name and labels for these items are presented in Table 8.2.

To fit the data with an IRT model, I recoded "no credit" into "0", and "full credit" into "1". All students have complete responses. The 10 items were modeled with a normal unidimensional 2-PL model using flexMIRT® (Cai, 2013). The

Table 8.2: Items from PISA Mathematics Assessment

	Name	Item labels
Item 1	PM305Q01	MATH - P2000 Map Q1
Item 2	PM406Q01	MATH - P2003 Running Tracks Q1
Item 3	PM406Q02	MATH - P2003 Running Tracks Q2
Item 4	PM423Q01	MATH - P2003 Tossing Coins Q1
Item 5	PM496Q02	MATH - P2003 Cash Withdrawal Q2
Item 6	PM496Q01T	MATH - P2003 Cash Withdrawal Q1
Item 7	PM564Q01	MATH - P2003 Chair Lift Q1
Item 8	PM564Q02	MATH - P2003 Chair Lift Q2
Item 9	PM571Q01	MATH - P2003 Stop the Car Q1
Item 10	PM603Q01T	MATH - P2003 Number Check Q1

proposed summed score likelihood based statistics were calculated. Results show that $\bar{X}^2_H$ is equal to 31.46; $\bar{X}^2_{C1}$ is equal to 32.97, and $\bar{X}^2_{C2}$ is equal to 32.03. All the summed score likelihood based statistics are significant ($df = 8$; $p < 0.000$) and close to one another. After implementing the empirical histogram latent density estimation, $\bar{X}^2_H$ is equal to 5.66 ($df = 8$; $p = 0.686$), $\bar{X}^2_{C1}$ is equal to 5.00 ($df = 8$; $p = 0.758$), and $\bar{X}^2_{C2}$ is equal to 4.70 ($df = 8$; $p = 0.758$). All the proposed statistics are not significant anymore. Thus, the latent variable is very likely to follow a nonnormal distribution in this dataset, and summed score likelihood based statistics are able to detect the violation of assumption of latent variable normality.

CHAPTER 9

Conclusion

Normality of latent variable distribution is a critical assumption in standard maximum marginal likelihood estimation for IRT models. However, this assumption could be violated in real data analysis (Woods, 2006; Woods & Lin, 2009). Consequently, item parameter estimation and test scoring might suffer from bias. Even though alternative approaches to estimate the latent variable distribution have been proposed, most of them are computationally demanding and not available in commercial software programs. Therefore, it is necessary to develop statistical indices to detect latent varibale nonnormality before more "expensive" approaches are ultilized.

A family of summed score likelihood based indices has been proposed for detecting the violation of latent variable normal ditribution assumption (Li & Cai, 2012). However, as stated in the introduction, those statistics do not asymptotically follow a chi-squared distribution. In this study, two Satorra-Bentler type moment adjustment approaches (Satorra & Bentler, 1994) are proposed to correct the summed score likelihood based indice $\bar{X}^2$ (Li & Cai, 2012). These moment adjustment methods have been utilized widely for the modification of goodness-of-fit statistics (Satorra & Bentler, 1994; Cai et al., 2006; Asparouhov & Muthen, 2010). Nonetheless, the calculation of first-order and second-order moments of $\bar{X}^2$ is quite computationally challenging, which involoves the computation of a Jacobian matrix. To solve this problem, an adapted Lord-Wingersky algorithm (Lord & Wingersky, 1984) was developed to calculate the Jacobian matrix re-

cursively. The adapted Lord-Wingersky algorithm and an illustration example could be found in Chapter 5. The algorithm provides an efficient way to obtain the Jacobian matrix, upon which the first- and second-order moments of $\bar{X}^2$ are worked out. Once the first- and second-order moments of the statistics are known, the process of correction is relatively straightforward (see Chapter 5; Satorra & Bentler, 1994; Asparouhov & Muthen, 2010).

The simulation study findings reinforced the conjecture that moment adjusted statistics $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ have improved properties in the null and alternative conditions. Their sampling distribution could be well approximated by a chi-squared distribution under the null hypothesis, and the indices are sensitive to the violation of latent variable normality in the alternative conditions. Simulation results also provided evidence that $\bar{X}^2_{C1}$ and $\bar{X}^2_{C2}$ improve the unadjusted statistic index $\bar{X}^2_H$ the most when the generating item parameters are highly dispersed. In this study, the two-moment adjusted statistic $\bar{X}^2_{C2}$ and the first-moment adjusted statistic $\bar{X}^2_{C1}$ perform equivalently well in most conditions.

In addition, This study is not without its limitations. Firstly, among the three summed score likelihood based indices proposed by Li and Cai (2012), only the Pearson's $\bar{X}^2$ indice was considered in my dissertation. While the simulation study results have demonstrated the feasibility and effectivity of the moment adjustment approaches for $\bar{X}^2$, future efforts could focuse on the adjustments of other summed score likelihood based indices. Secondly, limited simulation conditions were examined. More conditions could be tested in future simulation studies, for example, other types of IRT models, different kinds of latent variable nonnormality (e.g., skewed, bi-modal, or multi-modal), or various levels of model misspecification.

Finally, this study only considered the conditions when item response data are unidimensional. Multidimensional IRT models (MIRT, Reckase, 2009) should be considered in the subsequent work. One particularly popular model in educational

and psychological research is the full-information item bifactor model (Gibbons & Hedeker, 1992; Cai, Yang, & Hansen, 2011; Reise, 2012). In this model, all items load on a general dimension, and an item is permitted to load on at most one specific dimension that influences non-overlapping subsets of items. This feature of bifactor models implies that there exits valuable relation between an observed summed score and the distribution of the latent general dimension (Cai, 2014). This relation implies an opportunity to test the underlying assumption about the distribution of general latent dimension with summed score likelihood based statistics.

APPENDIX A

Appendix

	Slope			Threshold		
Item	Equal	Random	Dispersed	Equal	Random	Dispersed
1	1.00	1.85	1.00	0.00	-0.28	-2.00
2	1.00	1.87	1.09	0.00	-0.29	-1.83
3	1.00	1.80	1.17	0.00	1.05	-1.65
4	1.00	1.72	1.26	0.00	-0.52	-1.48
5	1.00	1.39	1.35	0.00	-1.32	-1.30
6	1.00	1.24	1.43	0.00	0.29	-1.13
7	1.00	1.19	1.52	0.00	0.62	-0.96
8	1.00	1.66	1.61	0.00	0.50	-0.78
9	1.00	1.71	1.70	0.00	0.55	-0.61
10	1.00	2.11	1.78	0.00	0.81	-0.43
11	1.00	1.68	1.87	0.00	1.49	-0.26
12	1.00	1.98	1.96	0.00	0.52	-0.09
13	1.00	1.73	2.04	0.00	-0.27	0.09
14	1.00	1.81	2.13	0.00	0.01	0.26
15	1.00	1.98	2.22	0.00	-0.39	0.43
16	1.00	1.59	2.30	0.00	-0.15	0.61
17	1.00	1.59	2.39	0.00	-0.69	0.78
18	1.00	2.05	2.48	0.00	0.36	0.96
19	1.00	1.47	2.57	0.00	1.09	1.13
20	1.00	1.89	2.65	0.00	0.52	1.30
21	1.00	2.08	2.74	0.00	0.41	1.48
22	1.00	1.89	2.83	0.00	2.10	1.65
23	1.00	1.70	2.91	0.00	1.12	1.83
24	1.00	1.81	3.00	0.00	0.65	2.00

Table A.1: Generating item parameters for two-parameter IRT model, n=24

	Slope			Threshold 1			Threshold 2			Threshold 3		
Item	Equal	Random	Dispersed	Equal	Random	Dispersed	Equal	Random	Dispersed	Equal	Random	Dispersed
1	1.00	1.85	1.00	0.00	-0.28	-2.00	0.50	0.22	-1.50	1.00	0.72	-1.00
2	1.00	1.87	1.09	0.00	-0.29	-1.83	0.50	0.21	-1.33	1.00	0.71	-0.83
3	1.00	1.80	1.17	0.00	1.05	-1.65	0.50	1.55	-1.15	1.00	2.05	-0.65
4	1.00	1.72	1.26	0.00	-0.52	-1.48	0.50	-0.02	-0.98	1.00	0.48	-0.48
5	1.00	1.39	1.35	0.00	-1.32	-1.30	0.50	-0.82	-0.80	1.00	-0.32	-0.30
6	1.00	1.24	1.43	0.00	0.29	-1.13	0.50	0.79	-0.63	1.00	1.29	-0.13
7	1.00	1.19	1.52	0.00	0.62	-0.96	0.50	1.12	-0.46	1.00	1.62	0.04
8	1.00	1.66	1.61	0.00	0.50	-0.78	0.50	1.00	-0.28	1.00	1.50	0.22
9	1.00	1.71	1.70	0.00	0.55	-0.61	0.50	1.05	-0.11	1.00	1.55	0.39
10	1.00	2.11	1.78	0.00	0.81	-0.43	0.50	1.31	0.07	1.00	1.81	0.57
11	1.00	1.68	1.87	0.00	1.49	-0.26	0.50	1.99	0.24	1.00	2.49	0.74
12	1.00	1.98	1.96	0.00	0.52	-0.09	0.50	1.02	0.41	1.00	1.52	0.91
13	1.00	1.73	2.04	0.00	-0.27	0.09	0.50	0.23	0.59	1.00	0.73	1.09
14	1.00	1.81	2.13	0.00	0.01	0.26	0.50	0.51	0.76	1.00	1.01	1.26
15	1.00	1.98	2.22	0.00	-0.39	0.43	0.50	0.11	0.93	1.00	0.61	1.43
16	1.00	1.59	2.30	0.00	-0.15	0.61	0.50	0.35	1.11	1.00	0.85	1.61
17	1.00	1.59	2.39	0.00	-0.69	0.78	0.50	-0.19	1.28	1.00	0.31	1.78
18	1.00	2.05	2.48	0.00	0.36	0.96	0.50	0.86	1.46	1.00	1.36	1.96
19	1.00	1.47	2.57	0.00	1.09	1.13	0.50	1.59	1.63	1.00	2.09	2.13
20	1.00	1.89	2.65	0.00	0.52	1.30	0.50	1.02	1.80	1.00	1.52	2.30
21	1.00	2.08	2.74	0.00	0.41	1.48	0.50	0.91	1.98	1.00	1.41	2.48
22	1.00	1.89	2.83	0.00	2.10	1.65	0.50	2.60	2.15	1.00	3.10	2.65
23	1.00	1.70	2.91	0.00	1.12	1.83	0.50	1.62	2.33	1.00	2.12	2.83
24	1.00	1.81	3.00	0.00	0.65	2.00	0.50	1.15	2.50	1.00	1.65	3.00

Table A.2: Generating item parameters for graded response IRT model, n=24

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	22	22.06	43.98	4.69	62.78	.010	.064	.102	.574
	$\bar{X}^2_{C1}$	100%	22	22.20	44.47	4.72	62.93	.010	.066	.106	.415
	$\bar{X}^2_{C2}$	100%	22	22.20	43.91	4.80	62.83	.010	.066	.106	.345
	M_2	100%	252	251.63	454.41	194.50	321.04	.004	.044	.102	.324
1000	$\bar{X}^2_H$	100%	22	21.85	44.71	7.13	56.85	.008	.056	.100	.128
	$\bar{X}^2_{C1}$	100%	22	21.91	44.85	7.16	56.70	.008	.056	.104	.202
	$\bar{X}^2_{C2}$	100%	22	21.91	44.63	7.21	56.79	.008	.056	.103	.229
	M_2	100%	252	251.52	516.47	167.81	340.83	.012	.047	.096	.713
1500	$\bar{X}^2_H$	100%	22	22.01	49.22	7.40	55.22	.014	.052	.108	.791
	$\bar{X}^2_{C1}$	100%	22	22.05	49.37	7.43	55.40	.014	.052	.108	.872
	$\bar{X}^2_{C2}$	100%	22	22.05	49.20	7.47	55.28	.014	.052	.108	.858
	M_2	100%	252	250.94	490.41	183.78	318.69	.012	.048	.086	.421

Table A.3: Results of simulation study for the null conditions (*2-PL*, Equal a & Equal b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	22	21.97	46.19	6.35	56.14	.014	.048	.096	.961
	$\bar{X}^2_{C1}$	100%	22	22.42	48.15	6.46	57.26	.020	.056	.108	.261
	$\bar{X}^2_{C2}$	100%	22	22.42	47.58	6.53	56.99	.018	.056	.108	.264
	M_2	100%	252	253.36	512.59	193.96	317.29	.010	.060	.114	.222
1000	$\bar{X}^2_H$	100%	22	22.15	45.18	6.67	46.90	.012	.064	.108	.975
	$\bar{X}^2_{C1}$	100%	22	22.53	46.65	6.76	47.70	.018	.070	.116	.397
	$\bar{X}^2_{C2}$	100%	22	22.52	46.38	6.76	47.59	.018	.070	.116	.378
	M_2	100%	252	251.97	551.94	179.69	351.34	.014	.062	.116	.405
1500	$\bar{X}^2_H$	100%	22	21.59	46.81	6.52	46.29	.014	.054	.104	.175
	$\bar{X}^2_{C1}$	100%	22	21.94	48.33	6.62	47.01	.014	.058	.106	.566
	$\bar{X}^2_{C2}$	100%	22	21.94	48.09	6.65	46.97	.014	.058	.106	.584
	M_2	100%	252	253.71	478.89	191.35	324.40	.006	.054	.114	.209

Table A.4: Results of simulation study for the null conditions (*2-PL*, Random a & Random b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	22	21.65	40.28	9.21	47.79	.010	.050	.092	.196
	$\bar{X}^2_{C1}$	100%	22	22.36	43.22	9.59	50.04	.012	.058	.108	.468
	$\bar{X}^2_{C2}$	100%	22	22.35	42.25	9.76	49.25	.012	.058	.106	.366
	M_2	100%	252	252.35	579.52	178.25	323.23	.018	.068	.112	.483
1000	$\bar{X}^2_H$	100%	22	21.63	44.33	6.58	54.02	.011	.045	.090	.060
	$\bar{X}^2_{C1}$	100%	22	22.22	46.79	6.77	55.55	.016	.057	.108	.501
	$\bar{X}^2_{C2}$	100%	22	22.22	46.23	6.86	55.29	.016	.054	.105	.511
	M_2	100%	252	252.43	474.75	192.68	323.58	.008	.041	.095	.208
1500	$\bar{X}^2_H$	100%	22	21.91	46.12	7.86	43.25	.010	.058	.118	.901
	$\bar{X}^2_{C1}$	100%	22	22.47	48.52	8.05	44.19	.016	.072	.126	.540
	$\bar{X}^2_{C2}$	100%	22	22.47	48.06	8.11	44.16	.016	.072	.126	.558
	M_2	100%	252	251.16	492.08	188.56	313.65	.006	.048	.096	.513

Table A.5: Results of simulation study for the null conditions (*2-PL*, Random a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	22	21.61	36.81	6.11	44.66	.002	.026	.066	.138
	$\bar{X}^2_{C1}$	100%	22	22.55	40.13	6.38	46.60	.008	.040	.084	.007
	$\bar{X}^2_{C2}$	100%	22	22.54	39.20	6.54	46.34	.008	.040	.084	.007
	M_2	100%	252	250.84	479.00	188.57	313.68	.010	.042	.098	.352
1000	$\bar{X}^2_H$	100%	22	21.43	41.89	5.47	50.79	.009	.041	.082	.011
	$\bar{X}^2_{C1}$	100%	22	22.22	45.06	5.71	53.27	.014	.052	.101	.360
	$\bar{X}^2_{C2}$	100%	22	22.22	44.37	5.89	52.64	.014	.051	.099	.389
	M_2	100%	252	252.88	507.93	178.47	346.01	.014	.050	.097	.237
1500	$\bar{X}^2_H$	100%	22	21.47	42.98	8.12	42.65	.004	.040	.080	.006
	$\bar{X}^2_{C1}$	100%	22	22.22	46.08	8.41	44.12	.006	.062	.110	.133
	$\bar{X}^2_{C2}$	100%	22	22.21	45.50	8.50	44.00	.006	.062	.108	.139
	M_2	100%	252	251.48	489.88	196.77	329.62	.018	.050	.090	.239

Table A.6: Results of simulation study for the null conditions (*2-PL*, Dispersed a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	22	28.71	71.60	9.12	61.36	.092	.224	.360
	$\bar{X}^2_{C1}$	100%	22	29.03	73.80	9.18	62.27	.104	.242	.380
	$\bar{X}^2_{C2}$	100%	22	28.92	71.60	9.26	61.54	.094	.230	.376
	M_2	100%	252	250.46	581.79	188.42	339.43	.010	.054	.106
1000	$\bar{X}^2_H$	100%	22	34.74	89.34	11.34	80.41	.245	.501	.632
	$\bar{X}^2_{C1}$	100%	22	34.99	91.26	11.39	81.26	.262	.512	.639
	$\bar{X}^2_{C2}$	100%	22	34.87	89.26	11.44	80.36	.252	.509	.638
	M_2	100%	252	250.93	503.30	179.18	352.68	.014	.042	.081
1500	$\bar{X}^2_H$	100%	22	41.62	121.70	13.58	85.76	.506	.750	.842
	$\bar{X}^2_{C1}$	100%	22	41.88	124.14	13.61	86.34	.510	.756	.844
	$\bar{X}^2_{C2}$	100%	22	41.72	121.63	13.63	85.71	.510	.754	.844
	M_2	100%	252	249.36	551.25	175.22	306.95	.000	.038	.088

Table A.7: Results of simulation study for the alternative conditions (*2-PL*, Equal a & Equal b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	22	28.70	63.46	11.72	62.91	.084	.246	.350
	$\bar{X}^2_{C1}$	100%	22	29.45	67.62	11.99	64.60	.116	.264	.378
	$\bar{X}^2_{C2}$	100%	22	29.33	65.46	12.07	63.74	.106	.258	.374
	M_2	100%	252	249.63	470.51	190.59	310.46	.002	.022	.090
1000	$\bar{X}^2_H$	100%	22	37.07	86.88	17.15	67.07	.350	.608	.748
	$\bar{X}^2_{C1}$	100%	22	37.89	91.40	17.43	68.69	.392	.640	.778
	$\bar{X}^2_{C2}$	100%	22	37.73	89.32	17.46	68.08	.384	.632	.778
	M_2	100%	252	249.94	532.80	177.26	308.78	.004	.054	.098
1500	$\bar{X}^2_H$	100%	22	45.03	116.03	22.77	87.22	.650	.862	.922
	$\bar{X}^2_{C1}$	100%	22	45.98	122.05	23.21	89.22	.684	.870	.934
	$\bar{X}^2_{C2}$	100%	22	45.76	119.05	23.21	88.41	.674	.870	.932
	M_2	100%	252	251.07	470.99	185.86	322.46	.012	.036	.084

Table A.8: Results of simulation study for the alternative conditions (*2-PL*, Random a & Random b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	22	26.60	57.78	8.49	55.84	.058	.156	.256
	$\bar{X}^2_{C1}$	100%	22	27.65	63.19	8.73	58.53	.072	.186	.312
	$\bar{X}^2_{C2}$	100%	22	27.54	60.61	8.80	57.53	.070	.184	.302
	M_2	100%	252	252.96	494.10	188.57	366.77	.010	.046	.106
1000	$\bar{X}^2_H$	100%	22	31.86	63.24	11.67	59.63	.141	.388	.519
	$\bar{X}^2_{C1}$	100%	22	32.92	68.02	12.05	61.38	.183	.428	.555
	$\bar{X}^2_{C2}$	100%	22	32.78	66.07	12.14	60.95	.177	.424	.552
	M_2	100%	252	252.99	551.36	181.83	345.63	.014	.062	.110
1500	$\bar{X}^2_H$	100%	22	37.59	86.13	14.54	80.77	.340	.648	.758
	$\bar{X}^2_{C1}$	100%	22	38.77	92.43	14.95	83.83	.416	.688	.782
	$\bar{X}^2_{C2}$	100%	22	38.58	90.01	15.01	82.81	.410	.684	.780
	M_2	100%	252	252.59	485.21	194.09	318.26	.004	.036	.094

Table A.9: Results of simulation study for the alternative conditions (*2-PL*, Random a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	22	26.48	50.67	10.47	55.91	.032	.160	.244
	$\bar{X}^2_{C1}$	100%	22	27.91	56.95	10.95	58.75	.056	.202	.320
	$\bar{X}^2_{C2}$	100%	22	27.77	54.37	11.09	58.09	.050	.202	.314
	M_2	100%	252	257.94	593.51	178.77	333.55	.022	.106	.174
1000	$\bar{X}^2_H$	100%	22	31.91	61.23	10.72	65.63	.144	.355	.519
	$\bar{X}^2_{C1}$	100%	22	33.34	67.25	11.14	68.17	.189	.430	.581
	$\bar{X}^2_{C2}$	100%	22	33.16	65.14	11.25	67.47	.181	.424	.576
	M_2	100%	252	259.51	576.21	188.36	328.93	.038	.101	.183
1500	$\bar{X}^2_H$	100%	22	37.62	79.51	14.27	70.31	.356	.658	.774
	$\bar{X}^2_{C1}$	100%	22	39.20	86.95	14.82	73.83	.414	.712	.812
	$\bar{X}^2_{C2}$	100%	22	38.98	84.56	14.89	72.78	.404	.706	.808
	M_2	100%	252	259.76	493.64	200.87	332.91	.020	.096	.170

Table A.10: Results of simulation study for the alternative conditions (*2-PL*, Dispersed a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	70	71.96	156.20	33.82	117.91	.018	.074	.134	.018
	$\bar{X}^2_{C1}$	100%	70	72.19	156.92	33.96	118.20	.018	.080	.142	.008
	$\bar{X}^2_{C2}$	100%	70	72.18	155.88	34.13	118.04	.018	.080	.142	.009
	M_2	100%	204	202.45	402.28	142.17	260.48	.004	.036	.078	.407
1000	$\bar{X}^2_H$	100%	70	70.12	138.33	37.71	109.24	.014	.048	.090	.497
	$\bar{X}^2_{C1}$	100%	70	70.21	138.55	37.74	109.27	.014	.048	.090	.352
	$\bar{X}^2_{C2}$	100%	70	70.21	138.21	37.76	109.26	.014	.048	.090	.349
	M_2	100%	204	205.72	1079.56	147.11	700.00	.010	.048	.098	.763
1500	$\bar{X}^2_H$	100%	70	70.54	140.11	35.27	111.69	.014	.060	.098	.058
	$\bar{X}^2_{C1}$	100%	70	70.59	140.20	35.27	111.71	.014	.060	.100	.047
	$\bar{X}^2_{C2}$	100%	70	70.59	140.03	35.27	111.70	.014	.060	.100	.047
	M_2	100%	204	203.69	405.62	147.67	285.06	.012	.056	.096	.631

Table A.11: Results of simulation study for the null conditions (*graded model*, Equal a & Equal b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	70	72.04	137.82	43.94	111.54	.018	.058	.120	.000
	$\bar{X}^2_{C1}$	100%	70	72.45	138.84	44.33	112.40	.018	.070	.128	.000
	$\bar{X}^2_{C2}$	100%	70	72.45	137.98	44.42	112.18	.018	.068	.126	.000
	M_2	100%	204	202.77	471.19	149.91	277.31	.008	.054	.118	.068
1000	$\bar{X}^2_H$	100%	70	70.63	151.71	39.21	118.27	.016	.058	.121	.267
	$\bar{X}^2_{C1}$	100%	70	70.89	152.74	39.38	118.75	.017	.060	.126	.123
	$\bar{X}^2_{C2}$	100%	70	70.89	152.40	39.43	118.67	.017	.060	.126	.126
	M_2	100%	204	203.09	422.56	139.26	288.26	.011	.045	.089	.480
1500	$\bar{X}^2_H$	100%	70	70.27	137.86	43.40	110.67	.014	.048	.098	.427
	$\bar{X}^2_{C1}$	100%	70	70.49	138.55	43.53	111.02	.016	.050	.100	.221
	$\bar{X}^2_{C2}$	100%	70	70.49	138.41	43.55	110.99	.016	.050	.100	.221
	M_2	100%	204	203.55	418.46	155.23	274.53	.012	.052	.098	.818

Table A.12: Results of simulation study for the null conditions (*graded model*, Random a & Random b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	100%	70	71.98	173.41	37.55	175.13	.020	.064	.132	.001
	$\bar{X}^2_{C1}$	100%	70	72.60	175.66	37.92	176.02	.020	.078	.138	.000
	$\bar{X}^2_{C2}$	100%	70	72.59	174.48	38.08	175.93	.020	.078	.138	.000
	M_2	100%	204	202.36	377.83	148.23	253.60	.000	.042	.086	.118
1000	$\bar{X}^2_H$	100%	70	70.71	136.79	41.01	123.16	.010	.052	.100	.326
	$\bar{X}^2_{C1}$	100%	70	71.14	138.17	41.24	123.54	.010	.052	.110	.089
	$\bar{X}^2_{C2}$	100%	70	71.14	137.74	41.28	123.58	.010	.052	.108	.084
	M_2	100%	204	204.33	416.84	130.73	287.58	.010	.048	.116	.862
1500	$\bar{X}^2_H$	100%	70	70.82	133.01	38.19	103.48	.004	.056	.114	.159
	$\bar{X}^2_{C1}$	100%	70	71.19	134.32	38.39	103.96	.004	.058	.118	.045
	$\bar{X}^2_{C2}$	100%	70	71.19	134.05	38.41	103.95	.004	.058	.118	.045
	M_2	100%	204	204.30	418.29	149.80	270.81	.018	.056	.114	.847

Table A.13: Results of simulation study for the null conditions (*graded model*, Random a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Empirical rejection rates			KS test
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	p-value
500	$\bar{X}^2_H$	98%	70	72.84	160.66	43.68	148.97	.014	.080	.153	.000*
	$\bar{X}^2_{C1}$	92.2%	70	73.51	164.45	44.22	150.52	.022	.095	.165	.000*
	$\bar{X}^2_{C2}$	92.2%	70	73.50	163.01	44.38	150.23	.020	.095	.163	.000*
	M_2	100%	204	202.28	428.91	130.33	278.06	.012	.046	.096	.082
1000	$\bar{X}^2_H$	100%	70	71.52	148.50	40.84	111.23	.014	.076	.130	.035*
	$\bar{X}^2_{C1}$	99.4%	70	72.09	150.87	41.07	112.14	.016	.080	.143	.004*
	$\bar{X}^2_{C2}$	99.4%	70	72.09	150.34	41.07	112.03	.016	.080	.143	.004*
	M_2	100%	204	205.21	408.54	150.44	267.11	.012	.044	.116	.482
1500	$\bar{X}^2_H$	100%	70	70.44	138.12	36.77	106.81	.012	.060	.104	.565
	$\bar{X}^2_{C1}$	100%	70	70.89	139.87	37.01	107.70	.014	.062	.106	.128
	$\bar{X}^2_{C2}$	100%	70	70.89	139.52	37.07	107.57	.014	.062	.106	.127
	M_2	100%	204	204.61	443.71	147.69	290.67	.012	.052	.120	.652

Table A.14: Results of simulation study for the null conditions (*graded model*, Dispersed a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	70	76.28	174.79	44.03	126.81	.040	.136	.232
	$\bar{X}^2_{C1}$	100%	70	76.69	176.64	44.14	126.99	.046	.142	.240
	$\bar{X}^2_{C2}$	100%	70	76.65	174.32	44.21	126.90	.044	.138	.238
	M_2	100%	204	206.22	444.88	156.19	268.20	.010	.064	.118
1000	$\bar{X}^2_H$	100%	70	84.19	169.96	51.69	128.02	.116	.280	.432
	$\bar{X}^2_{C1}$	100%	70	84.47	171.44	51.81	128.61	.122	.306	.442
	$\bar{X}^2_{C2}$	100%	70	84.40	169.86	51.86	128.26	.120	.300	.440
	M_2	100%	204	207.31	416.69	148.64	265.67	.012	.050	.140
1500	$\bar{X}^2_H$	100%	70	92.52	196.68	53.29	146.23	.294	.552	.670
	$\bar{X}^2_{C1}$	100%	70	92.77	198.34	53.36	146.68	.304	.558	.672
	$\bar{X}^2_{C2}$	100%	70	92.69	196.75	53.39	146.37	.300	.554	.672
	M_2	99.8%	204	206.15	482.91	80.67	273.72	.012	.072	.142

Table A.15: Results of simulation study for the alternative conditions (*graded model*, Equal a & Equal b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	99.8%	70	79.72	168.44	36.22	136.46	.058	.200	.331
	$\bar{X}^2_{C1}$	99.8%	70	80.50	172.00	36.47	136.96	.060	.224	.355
	$\bar{X}^2_{C2}$	99.8%	70	80.40	168.90	36.62	136.89	.060	.220	.353
	M_2	100%	204	215.48	486.46	151.61	280.23	.050	.146	.240
1000	$\bar{X}^2_H$	100%	70	88.83	197.51	52.80	135.78	.216	.430	.567
	$\bar{X}^2_{C1}$	100%	70	89.45	201.23	53.13	136.84	.226	.453	.580
	$\bar{X}^2_{C2}$	100%	70	89.34	198.65	53.21	136.38	.223	.452	.579
	M_2	100%	204	213.89	469.67	155.81	288.50	.034	.134	.239
1500	$\bar{X}^2_H$	100%	70	98.58	225.08	56.43	150.98	.424	.692	.804
	$\bar{X}^2_{C1}$	100%	70	99.19	228.88	56.71	151.83	.446	.702	.818
	$\bar{X}^2_{C2}$	100%	70	99.05	226.33	56.74	151.50	.440	.700	.818
	M_2	100%	204	215.82	472.29	157.65	295.84	.040	.160	.254

Table A.16: Results of simulation study for the alternative conditions (*graded model*, Random a & Random b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	100%	70	77.559	159.210	46.960	129.745	.046	.138	.240
	$\bar{X}^2_{C1}$	100%	70	78.494	162.970	47.475	130.673	.050	.154	.264
	$\bar{X}^2_{C2}$	100%	70	78.416	160.094	47.620	130.551	.050	.152	.262
	M_2	100%	204	207.734	412.355	157.580	270.120	.020	.068	.138
1000	$\bar{X}^2_H$	100%	70	84.736	188.988	54.808	127.664	.148	.308	.434
	$\bar{X}^2_{C1}$	100%	70	85.480	192.435	55.261	128.437	.152	.324	.456
	$\bar{X}^2_{C2}$	100%	70	85.397	190.363	55.330	128.347	.150	.322	.452
	M_2	100%	204	207.417	415.702	146.270	263.820	.012	.076	.128
1500	$\bar{X}^2_H$	100%	70	93.699	209.900	56.329	147.661	.290	.586	.702
	$\bar{X}^2_{C1}$	100%	70	94.434	213.985	56.695	148.924	.312	.594	.722
	$\bar{X}^2_{C2}$	100%	70	94.326	211.867	56.729	148.516	.312	.594	.722
	M_2	100%	204	208.300	404.623	146.110	277.430	.016	.078	.142

Table A.17: Results of simulation study for the alternative conditions (*graded model*, Random a & Dispersed b, $n = 24$)

N	statistics	valid %	*df*	Mean	Var	Min	Max	Power		
								$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
500	$\bar{X}^2_H$	92.4%	70	78.13	172.25	46.60	117.83	.048	.190	.281
	$\bar{X}^2_{C1}$	75.8%	70	79.35	180.96	47.27	122.15	.061	.219	.301
	$\bar{X}^2_{C2}$	75.8%	70	79.23	176.11	47.44	118.02	.058	.208	.298
	M_2	100%	204	209.07	466.60	146.71	274.02	.028	.094	.160
1000	$\bar{X}^2_H$	99.8%	70	85.39	183.43	48.34	147.25	.138	.322	.451
	$\bar{X}^2_{C1}$	96.6%	70	86.23	189.70	48.83	148.64	.151	.342	.480
	$\bar{X}^2_{C2}$	96.6%	70	86.14	187.63	48.93	148.35	.149	.340	.476
	M_2	100%	204	210.40	448.98	158.14	282.10	.026	.086	.162
1500	$\bar{X}^2_H$	100%	70	93.41	206.37	58.52	149.61	.298	.558	.698
	$\bar{X}^2_{C1}$	99.2%	70	94.28	212.59	58.99	152.12	.310	.589	.716
	$\bar{X}^2_{C2}$	99.2%	70	94.17	210.39	59.02	151.12	.308	.587	.716
	M_2	100%	204	210.33	443.85	152.32	293.92	.030	.086	.180

Table A.18: Results of simulation study for the alternative conditions (*graded model*, Dispersed a & Dispersed b, $n = 24$)

Item	a		$b1$		$b2$		$b3$		$b4$	
	est.	*s.e.*	*est.*	*s.e.*	*est.*	*s.e.*	*est.*	*s.e.*	*est.*	*s.e.*
1	3.37	0.08	-0.53	0.02	0.12	0.02	0.86	0.02	1.53	0.03
2	3.12	0.08	-0.47	0.02	0.15	0.02	0.96	0.02	1.65	0.04
3	3.56	0.09	-0.54	0.02	0.08	0.02	0.87	0.02	1.56	0.03
4	1.68	0.05	-0.44	0.03	0.32	0.03	1.29	0.04	2.11	0.06
5	1.86	0.06	0.21	0.03	0.8	0.03	1.67	0.05	2.4	0.07
6	1.86	0.05	-0.82	0.03	0.01	0.02	0.94	0.03	1.82	0.05
7	1.94	0.05	-1.59	0.04	-0.63	0.03	0.13	0.02	0.95	0.03
8	2.11	0.06	-1.96	0.05	-0.74	0.03	0.15	0.02	1.02	0.03
9	2.14	0.06	-1.88	0.05	-0.77	0.03	0.08	0.02	0.93	0.03
10	2.52	0.07	-0.47	0.02	0.32	0.02	1.23	0.03	1.96	0.05
11	2.06	0.06	-0.13	0.02	0.56	0.03	1.56	0.04	2.31	0.06
12	2.59	0.07	-0.82	0.02	0.03	0.02	0.94	0.03	1.69	0.04

Table A.19: Item parameter estimates for PROMIS smoking initiative

Item	a	$s.e.$	b	$s.e.$
1	0.66	0.07	-0.03	0.12
2	2.83	0.2	0.95	0.06
3	3.07	0.21	1.18	0.06
4	0.81	0.1	-2.22	0.25
5	1.42	0.12	-0.48	0.05
6	1.68	0.14	0.09	0.04
7	0.87	0.08	0.45	0.07
8	1.18	0.1	0.36	0.05
9	1.18	0.1	0.43	0.05
10	0.85	0.08	0.92	0.09

Table A.20: Item parameter estimates for PISA 2012 mathematical test

Bibliography

Asparouhov, T., & Muthen, B. (2010). *Simple second order chi-square correction.* Retrieved 2014-7-1, from `http://www.statmodel.com`

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research*, *27*(4), 525–546.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

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

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

Bock, R. D., & Lieberman, M. (1970, December). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Brennan, R. (Ed.). (2006). *Educational measurement* (4th Ed.). Westport, CT : Praeger Publishers.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel item factor analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L. (2014). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, 1-25.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *The British journal of mathematical and*

statistical psychology, *66*(2), 245–276.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173–194.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248.

Chen, W., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(3), 440–464.

Edelen, M. O. (2014). The promis® smoking assessment toolkit-background and introduction to supplement. *Nicotine and Tobacco Research*, *16*(Suppl 3), S170-174.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologist*. Psychology Press.

Ferrando, P. J., & Lorenzo-Seva, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program EP-fit. *Educational and Psychological Measurement*, *61*(5), 895–902.

Gibbons, R., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423436.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, *26*(2), 195–211.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2015). *Limited-information goodness-of-fit testing of diagnostic classification item response models.* (Manuscript

submitted for publication)

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419.

Kastberg, D., Roey, S., Lemanski, N., Chan, J., & Murray, G. (2014). *Technical report and user guide for the program for international student assessment (pisa). (NCES 2014-025)*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Li, Z., & Cai, L. (2012, July). Summed score based fit indices for testing latent variable distribution assumption in IRT. Paper presented at the 2012 International Meeting of the Psychometric Society, Lincoln, NE.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*(4), 517–549.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*(4), 453–461.

Mathai, A. M., & Provost, S. B. (1992). *Quadratic forms in random variables: theory and applications*. New York: Marcel Dekker.

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*(2), 209–227.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55–64.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the*

American Statistical Association, *100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Mislevy, D. R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381.

Monroe, S. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities.* University of California, Los Angeles. (Unpublished Doctoral Dissertation)

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the MetropolisHastings RobbinsMonro Algorithm. *Educational and Psychological Measurement*, *74*(2), 343369.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Reckase, M. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences).* New York: Springer.

Reise, S. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*(3), 509–528.

Ross, J. (1966). An empirical study of a logistic mental test model. *Psychometrika*, *31*(3), 325–340.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*(17).

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von & C. C. Clogg (Eds.),

Latent variables analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA, US: Sage Publications, Inc.

Shadel, W. G., Edelen, M., & Tucker, J. S. (2011). A unified framework for smoking assessment: the PROMIS smoking initiative. *Nicotine and Tobacco Research*, *13*(5), 399–400.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R.Millsap & A.Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). London: Sage Publications.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Routledge.

Tucker, J., Shadel, W. G., Stucky, B., Cerully, J., Li, Z., Hansen, M., & Cai, L. (2014). Development of the promis® positive emotional and sensory expectancies of smoking item banks. *Nicotine and Tobacco Research*, *16*(Suppl 3), S212–222.

Woods, C. M. (2006). Empirical histograms in item response theory with ordinial data. *Educational and Psychological Measurement*, *67*(1), 73–87.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102–117.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281–301.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (1996). *BILOG-MG: Multiple-*

group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International.