USE OF SCORE-BASED TESTS IN IRT MODELS

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Psychology

by
Ting Wang

Dr. Edgar C. Merkle, Dissertation Supervisor

July 2015
The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

USE OF SCORE-BASED TESTS IN IRT MODELS

presented by Ting Wang,
a candidate for the degree of Doctor of Psychology and hereby certify that, in their opinion, it is worthy of acceptance.

______________________________
Dr. Edgar C. Merkle

______________________________
Dr. Xianyang Zhang

______________________________
Dr. Clintin P. Davis-Stober

______________________________
Dr. Jeffery Rouder

______________________________
Dr. Timothy Trull
ACKNOWLEDGMENTS

First of all, I would like to deeply thank my advisor Dr. Edgar Merkle for his guidance in my studies. I can not list all the things he helped, so simply, I will summarize the result of his impact: research becomes much more interesting and graduate school life becomes happier since I worked with him.

Besides, I would like to thank my committee members. Dr. Xianyang Zhang gave very good comments on detecting multiple change points. Dr. Jeff Rouder clarified how to compare this score-based method and traditional interaction method. Besides, he also sharply pointed out the issue in my writing. It helped me a lot in even other writings. Dr. Clintin P. Davis-Stober and Dr. Tim Trull have been always helpful along my five years’ graduate school. I am thankful for their patience and good advice.

In the end, I am grateful for my families and friends who were always there to listen to my all sorts of feelings during this graduate school journey.
ACKNOWLEDGMENTS ............................................................ ii
LIST OF TABLES ................................................................. vi
LIST OF FIGURES ............................................................... vii
ABSTRACT ........................................................................ xi

CHAPTER

1 Introduction ................................................................. 1

2 Background ................................................................. 6
  2.1 IRT ................................................................. 6
    2.1.1 Model Descriptions .............................................. 6
    2.1.2 Model Estimation ............................................... 11
  2.2 Measurement Invariance ........................................... 13
    2.2.1 Mantel-Haenszel ................................................. 14
    2.2.2 Raju’s Area ....................................................... 15
    2.2.3 Lord’s Wald Test ............................................... 16
    2.2.4 Logistic Regression ............................................. 16
  2.3 Structural Change Tests .......................................... 18
    2.3.1 Theoretical History ............................................. 18
    2.3.2 Implementation ................................................ 22
    2.3.3 Estimation ....................................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.2 Multilevel Model</td>
<td>68</td>
</tr>
<tr>
<td>5.3.3 Longitudinal Studies</td>
<td>68</td>
</tr>
<tr>
<td>5.4 Summary</td>
<td>69</td>
</tr>
<tr>
<td>APPENDIX</td>
<td></td>
</tr>
<tr>
<td>A Simulation 1.1: Estimated by MML</td>
<td>71</td>
</tr>
<tr>
<td>References</td>
<td>76</td>
</tr>
<tr>
<td>VITA</td>
<td>81</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Contingency table</td>
<td>14</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>2.1 Example of Raju’s Area</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Example of one Brownian bridge</td>
<td>19</td>
</tr>
<tr>
<td>3.1 Simulation 1.1. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\alpha_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.</td>
<td>36</td>
</tr>
<tr>
<td>3.2 Simulation 1.1. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\gamma_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.</td>
<td>37</td>
</tr>
</tbody>
</table>
3.3 Simulation 1.1. Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\alpha_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.

3.4 Simulation 1.1. Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\gamma_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.

3.5 Simulation 1.2. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting extended Model). The parameter violating measurement invariance is $\alpha_3$. The number of categories is $m = 4$. Panel labels denote the parameter(s) being tested and sample size.

3.6 Simulation 1.2. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting extended Model). The parameter violating measurement invariance is $\gamma_3$. The number of categories is $m = 4$. Panel labels denote the parameter(s) being tested and sample size.
3.7 Simulation 2.1. Simulated empirical fluctuation process summary of max $WDM_o$ 25% quantile, median and 75% quantile across four changing levels for violating parameter $\alpha_3$ (left) and $\gamma_3$ (right), $n = 960$, estimated by PML. Sub-panel labels denote the changing level and violation magnitude $d = 4$. .................................................. 48

3.8 Simulation 2.2. Simulated empirical fluctuation process summary of max $WDM_o$ 25% quantile, median and 75% quantile with two changing levels for violating parameter $\alpha_3$ (left) and $\gamma_3$ (right), $n = 960$, estimated by PML. Sub-panel labels denote the changing level and violation magnitude $d = 4$. .................................................. 52

4.1 Empirical fluctuation processes of the $WDM_o$ statistic (first row) and max $LM_o$ (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using 2PL model ........................................... 57

4.2 Fluctuation processes of the $WDM_o$ statistic (first row) and max $LM_o$ (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using extended model. ................................. 59

4.3 Fluctuation processes of the $WDM_o$ statistic (first row) and max $LM_o$ (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using model only estimating group mean. ............................. 60
A.1 Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\alpha_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.

A.2 Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\gamma_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.

A.3 Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\alpha_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$.

A.4 Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\gamma_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 

x
ABSTRACT

Measurement invariance is a fundamental assumption in item response theory models, where the relationship between a latent construct (ability) and observed item responses are of interest. Violation of this assumption would render the scale misinterpreted or cause systematic bias against certain groups of people. While a number of methods have been proposed to detect measurement invariance violations, they all require definition of problematic model parameters and respondent grouping information in advance. However, these “locating” pieces of information are typically unknown in practice.

As an alternative, this dissertation focuses on a family of recently-proposed tests based on stochastic processes of casewise derivatives of the likelihood function (i.e., scores). These score-based tests only require estimation of the null model (when measurement invariance assumption holds), with problematic subgroups of respondents and model parameters being identified in a factor-analytic, continuous data context. In this dissertation, I aim to generalize these tests to item response theory models for categorical data. The tests’ theoretical background and implementation are detailed. The tests’ ability to identify problematic subgroups and model parameters is studied via simulation. An empirical example involving the tests is also provided. In the end, potential applications and future development are discussed.
Chapter 1

Introduction

To make inferences and predictions from a parametric model, we need to check if parameters stay stable with respect to (w.r.t.) various conditions, typically time. Parameter instability is often studied in longitudinal data analysis, where it indicates a change in trajectory (a "regime shift"). This is of interest in areas such as econometrics (Brown, Durbin, & Evans, 1975; Andrews, 1993; Nyblom, 1989; Hansen, 1992; Horn & McArdle, 1992; Hjort & Koning, 2002), policy analysis (Zeileis & Hornik, 2007) and drug intervention (Hothorn & Zeileis, 2008). In all of the above applications, regression or ANOVA models are generally implemented.

In the social sciences, parameter stability is also an important concern. The only difference is that the object being measured is often unobservable. Thus, instead of regression or ANOVA approach, item response theory (IRT) and structural equation modeling (SEM) are two main approaches utilized to assess unobserved constructs. The constructs could be ability, propensity of emotions, and attitude. Underlying these models is also a fundamental assumption that item/scale parameters are in-
variant across observations. This assumption is often called measurement invariance (Mellenbergh, 1989) when it is related to unobserved measurement construct. Violation of this assumption could render the scales misunderstood or cause bias. That is, if a set of scales violates measurement invariance, then individuals with the same ability (“amount” of latent variable) may systematically receive different scale scores.

For example, consider an item that is supposed to assess respondents’ general vocabulary knowledge: “What does alto mean?”. For Hispanic respondents, the word “alto” means “high”. However, for non-Hispanic respondents, it might mean “counter-tenor”. Therefore, respondents’ performance on this item not only depends on vocabulary proficiency but also on the respondents’ ethnic group (De Ayala, 2009). This will make the item unfair with respect to ethnicity. Horn and McArdle (1992) summarized the impact of this issue, stating that “Lack of evidence of measurement invariance equivocates conclusions and casts doubt on theory in the behavioral sciences”.

A formal definition of measurement invariance is given below (Mellenbergh, 1989):

\[ f(X|v, \eta) = f(X|\eta), \]  

(1.1)

where \( X \) is a vector of observed variables, \( \eta \) is the latent variable vector, \( v \) is the auxiliary variable against which we are testing measurement invariance, and \( f(\cdot) \) is an assumed parametric distribution.

To frame this as a hypothesis testing problem, we assume a potential observation-specific parameter vector \( \theta_i \). Then the null hypothesis of measurement invariance can be expressed as all observations arising from a common set of population parameters
\( \theta_0: \)

\[ H_0 : \theta_i = \theta_0 \quad (i = 1, \ldots n). \quad (1.2) \]

The alternative hypothesis is

\[ H_1 : \theta_i = \theta(v_i) \quad (i = 1, \ldots n), \quad (1.3) \]

where \( v_i \) is an auxiliary variable against which we are testing measurement invariance. It can be time, income, age, country etc. \( \theta(v_i) \) is typically an unknown function w.r.t. \( v_i \). If the function were known, the alternative hypothesis can be expressed more specifically. For example, one function of particular interest involves \( v \) dividing individuals into two subgroups with different parameter vectors:

\[ H_1 : \theta_i = \begin{cases} 
\theta^{(A)} & v_i \leq v \\
\theta^{(B)} & v_i > v.
\end{cases} \quad (1.4) \]

For this hypothesis testing problem, the most popular method is to use a likelihood ratio test (LRT). The basic idea of LRT involves comparing full to reduced models. The full model is a multiple group model with parameters free to vary across group A and group B; the reduced model constrained parameters to be equal across all observations. The LRT statistic can be expressed as

\[ LR(v) = -2[\ell(\hat{\theta}; x_1, \ldots, x_n) - \{\ell(\hat{\theta}^{(A)}; x_1, \ldots, x_m) + \ell(\hat{\theta}^{(B)}; x_{m+1}, \ldots, x_n)\}] 
\]

where \( \ell \) represents the log-likelihood function. This LRT statistic has an asymptotic \( \chi^2 \) distribution with degrees of freedom equal to the number of parameters in \( \theta \).
Further, $\hat{\theta}^{(A)}$ is the MLE of $\theta^{(A)}$ based on $\{x_1, \ldots, x_m\}$, for which $v_i \leq v$ and $\hat{\theta}^{(B)}$ is the MLE of $\theta^{(B)}$ based on $\{x_{m+1}, \ldots, x_n\}$ for which $v_i > v$.

However, when the grouping information is unknown, a natural extension of the above LRT is to compute $LR(v)$ for each possible value of $v$ in some interval $[v, \bar{v}]$, and the statistic of interest now becomes:

$$\max_{v \in [v, \bar{v}]} LR(v).$$

The asymptotic distribution of the maximum LR statistics is not $\chi^2$ any more. Andrews (1993) showed that the asymptotic distribution of Equation (1.6) under null hypothesis converges in distribution to some stochastic process (details will be illustrated in Chapter 2). Therefore, the problem is framed as hypothesis testing again, but only requires specifying the null hypothesis. These tests were widely used in econometric studies (see Perron, 2006, for further details) and have been recently proposed to detect measurement invariance (Merkle & Zeileis, 2013). This family of tests have been called score-based tests in the literature (Merkle & Zeileis, 2013; Merkle, Fan, & Zeileis, 2014; T. Wang, Merkle, & Zeileis, 2014). All this work focused on factor analysis models with continuous data assumed arising from multivariate normal distributions. For categorical data in social sciences, item response theory (IRT) models are commonly used. In this dissertation, I am going to implement these tests in IRT models.

Under the condition that the score-based test shows a measurement invariance violation, the next natural question is to locate the measurement invariance. As Millsap (2005) stated, locating the invariance violation is one of the major outstanding problems in the field. This locating problem can be divided into two aspects. One
is to locate which parameter violates the measurement invariance assumption. The other is to locate point/level of the auxiliary variable \((v)\) at which the violation occurs. Traditionally, these two questions were dealt with by the LRT. However, LRT requires the researcher to specify the violating parameter and the violating level of \(v\). In most applications, this information is not available. In this dissertation, I am going to utilize these score-based tests to solve measurement invariance locating issues in IRT models.

To use the proposed method, I will first introduce background information (Chapter 2). This background includes IRT models, traditional methods of testing measurement invariance, and the score-based tests. Then locating issue related to IRT models will be examined by simulation (Chapter 3) and empirical studies (Chapter 4). Finally, I will discuss future development in the application of score-based tests.
Chapter 2

Background

2.1 IRT

Item response theory (IRT) is a family of models describing the relationship between a latent variable and observed discrete variables. This research area has been studied and popularized since the 1960s. In the following, we will describe the model first and then introduce model estimation methods.

2.1.1 Model Descriptions

Rasch (1961) developed several psychological measurement models to separately estimate properties of the items and people. The Rasch model can predict the probability that person $i$ is correct on item $j$. After transforming this probability with the logit function, a model very similar to factor analysis model could be applied to the mean
response. The Rasch model can be described as below:

\[ y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad (2.1) \]
\[ \logit(p_{ij}) = \eta_i - \gamma_j. \quad (2.2) \]

Equation (2.1) assumes each person’s response to each item \( y_{ij} \) arises from a Bernoulli distribution with parameter \( p_{ij} \). Then Equation (2.2) transforms \( p_{ij} \) to \( \logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) \), which is described by the person’s ability \( \eta_i \) and the item’s difficulty \( \gamma_j \).

After development of the Rasch model, more general models were proposed, such as 2PL model. The 2PL model assumes the following functional form:

\[ \logit(p_{ij}) = \alpha_j \eta_i + \gamma_j, \quad (2.3) \]

where on the right hand, \( \gamma_j, \alpha_j \) are the difficulty and discrimination parameter for item \( j \), respectively; \( \eta_i \) is the latent variable (ability) for individual \( i \), typically assumed to follow \( N(0,1) \). Let \( \alpha \) be the row vector containing \( \alpha_1, \ldots, \alpha_p \) (\( p \) represents the number of items), \( \gamma \) be the row vector containing \( \gamma_1, \ldots, \gamma_p \), Equation (2.3) can be transformed into

\[ \logit(p_i) = \alpha \eta_i + \gamma, \quad (2.4) \]

where \( p_i \) represents the predicted probablity vector for individual \( i \), which includes \( p_{i1}, \ldots, p_{ip} \). Just looking at the right hand side of the equation, we can see that it is related to the one dimension factor analysis model, which is defined as

\[ y_i^* = \Lambda \eta_i + \epsilon, \quad (2.5) \]
where $\Lambda$ is $p \times 1$ factor loading vector, with components $\lambda_1, \ldots, \lambda_p$; $\eta_i \sim N(0, 1)$; $\epsilon$ is error variance, which follows distribution $N(0, \psi)$, with $\psi$ being a diagonal matrix, $\psi = I - \text{diag}(\Lambda \Lambda')$. The continuous variable vector $y_i^*$ are composed by $y_{ij}^*$ ($j = 1, \ldots, p$). They are dichotomized as

$$y_{ij} = \begin{cases} 1 & y_{ij}^* \geq \tau_j \\ 0 & y_{ij}^* < \tau_j. \end{cases}$$

(2.6)

Therefore, we can see that $\lambda_j$ corresponds to $\alpha_j$ in Equation (2.3). They both serve to describe the random variable $\eta_i$. Error variance $\epsilon$ corresponds to the Bernoulli distribution assumed by Equation (2.1), with both serving to describe the uncertainty/randomness in the model. Threshold $\tau_j$ corresponds to $\gamma_j$ which describes item $j$’s difficulty.

Takane and de Leeuw (1987) formally proved the equivalence of parameters in 2PL and factor analysis model. By using 2PL model approach, the marginal log likelihood of observed data $y$ (composed by $y_i$), given the parameter vector $\theta$ (including item parameter $\alpha_j$, $\gamma_j$ and person parameter $\eta_i$) is

$$\ell(\theta; y) = \sum_{i=1}^{n} \log \int f(y_i|\eta_i)g(\eta_i)d\eta_i,$$

(2.7)

where $g(\eta_i)$ is density function of $N(0, 1)$ and

$$f(y_i|\eta_i) = \prod_{j=1}^{p} p_{ij}^{y_{ij}}(1 - p_{ij})^{1-y_{ij}}$$

(2.8)

where $p_{ij} = \text{logit}^{-1}(\alpha_j \eta_i + \gamma_j)$ as shown by Equation (2.3). Instead of using logit func-
tion as the link function, we can also use cumulative density of standard distribution \( \Phi \) as link function, serving the same purpose of transforming \((-\infty, \infty)\) to \((0, 1)\) scale. So \( p_{ij} \) can be written as \( p_{ij} = \Phi(\alpha_j \eta_i + \gamma_j) \). The shape of logit link function and probit link function are very similar, with logit link function has thinner tails. The parameter estimation based on probit link function can be transformed to parameter estimation based on logit link function by multiply a constant 1.702. I will use probit link function in this section to make the comparison with factor analysis approach easier.

By using factor analysis model approach, the marginal log likelihood of observed data \( \mathbf{y} \) (composed by \( \mathbf{y}_i \)), given the parameter vector \( \mathbf{\theta} \) (including \( \lambda_j, \tau_j \)) \(^1\) is

\[
\ell(\mathbf{\theta}; \mathbf{y}) = \sum_{i=1}^n \log \int f(\mathbf{y}_i^*|\eta_i) d\mathbf{y}_i^* \\
= \sum_{i=1}^n \log \left( \int f(\mathbf{y}_i^*|\eta_i) g(\eta_i) d\eta_i \right) d\mathbf{y}_i^* \\
= \sum_{i=1}^n \log \left( \int f(\mathbf{y}_i^*|\eta_i) d\mathbf{y}_i^* \right) g(\eta_i) d\eta_i. \tag{2.11}
\]

Thus we can see Equation (2.11) is in the form as Equation (2.7), \(^2\) where

\[
\int_{\mathbf{y}_i} f(\mathbf{y}_i^*|\eta_i) d\mathbf{y}_i^* = \prod_{j=1}^p \left\{ \left( \int_{\tau_j}^{\infty} f(y_{ij}^*|\eta_i) dy_{ij}^* \right)^{y_{ij}} \left( 1 - \int_{\tau_j}^{\infty} f(y_{ij}^*|\eta_i) dy_{ij}^* \right)^{1-y_{ij}} \right\}, \tag{2.12}
\]

\(^1\psi_j\) is not included because it is determined by \(1-\lambda_j^2\). \(^2\) integration space is determined by threshold \(\tau_j\). It will be illustrated in Equation (2.12)
where

\[ y_{ij}^* | \eta_i \sim N(\lambda_j \eta_i, \psi_j), \quad (2.13) \]

\[ \sim N(\lambda_j \eta_i, 1 - \lambda_j^2). \quad (2.14) \]

Thus,

\[ \int_{\tau_j}^{\infty} f(y_{ij}^* | \eta_i) dy_{ij}^* = \Phi \left( \frac{\lambda_j \eta_i - \tau_j}{\sqrt{1 - \lambda_j^2}} \right). \quad (2.15) \]

Comparing Equation (2.15) to the IRT approach based on probit link function

\[ p_{ij} = \Phi(\alpha_j \eta_i + \gamma_j), \]

we can see they are equivalent after the following transformation:

\[ \alpha_j^* = \frac{\lambda_j}{\sqrt{(1 - \lambda_j^2)}}, \quad (2.16) \]

\[ \gamma_j = - \frac{\tau_j}{\sqrt{(1 - \lambda_j^2)}}. \quad (2.17) \]

The only difference between these two approaches is where integration is performed. In the IRT formulation, dichotomization is conducted conditional on the latent variable (ability); the integration is performed on the conditional probability. In the factor analysis model formulation, the integration is undertaken directly on the assumed continuous underlying variable. Dichotomization is dealt with as threshold. Based on the different integration places, these two approaches utilize different estimation methods.
2.1.2 Model Estimation

Marginal Maximum Likelihood (MML) is usually implemented to estimate IRT models (Thissen, 1982). The main idea is that since person parameters $\eta_i$ has a distribution associated with it, we can integrate it out first and concentrate on item parameters. The difficulty of this approach is the integration part, because the integral typically does not have a closed form. The solution is to estimate the integral by using quadrature. The basic idea of quadrature is to divide the area covered by the distribution curve into many small bars. Then the sum of these bars’ area should be close to the area under the curve. After we estimate item parameters, we can estimate the person parameters by using Bayesian reasoning. In particular, the estimates of person parameter $\eta_i$ can be viewed as arising from the posterior distribution of $\eta_i$ which is proportional to the product of the likelihood function (Bernoulli distribution) and the prior distribution of $\eta_i$ (the prior can be set as non-informative).

Generalized Least Squares estimation (GLS) is typically utilized in factor analysis models (Muthén, 1978). Three stages are involved. In the first stage, thresholds are estimated by maximizing univariate marginal likelihoods separately. Second, given these estimated thresholds, polychoric correlations are estimated by maximizing bivariate marginal likelihoods. In the third stage, given the polychoric correlation matrix, all other parameters are estimated by using a version of generalized least squares methods.

Pairwise Maximum Likelihood (PML) was recently developed for factor models with ordinal data (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). The advantage of PML over existing estimation methods is mainly computational. Instead of integrating over all latent variables for all items, PML constrained the integration
as pairwise, so the estimation only deals with bivariate integration. In the end, all item parameters are estimated simultaneously instead of in multiple stages.

The above estimation methods can handle both binary data and ordinal data. Ordinal data contain several categories as opposed to 0, 1 in binary data. By changing the Bernoulli distribution to Multinomial, the same IRT modeling ideas can be applied. The model can be generally expressed in the following way:

\[ y \sim \text{Multinomial}(n = 1, p_1, p_2, \ldots, p_J) \]  
\[ \log\left( \frac{p_{11}^*}{p_{21}^*} \right) = \alpha_1 + \beta_{11}x_1 + \ldots \]  
\[ \log\left( \frac{p_{12}^*}{p_{22}^*} \right) = \alpha_2 + \beta_{12}x_1 + \ldots \]  
\[ \vdots \]  
\[ \log\left( \frac{p_{1J-1}^*}{p_{2J-1}^*} \right) = \alpha_{(J-1)} + \beta_{1(J-1)}x_1 + \ldots \]  

Depending on how \( p_{1j}^*, p_{2j}^* \) for \( j = 1, \ldots, (J - 1) \) is defined, multiple models have been proposed. For example, when \( p_{1j}^* = P(y \leq j), p_{2j}^* = P(y > j) \) is defined for \( j = 1, \ldots, (J - 1) \), and \( \alpha_1 < \alpha_2 < \ldots < \alpha_{J-1} \) is constrained to ensure cumulative probabilities \( P(y \leq j) \) are monotonically increasing with \( j \), it is called a Graded Response Model (GRM; Samejima, 1969). Another way to define is to let \( p_{1j}^* = P(y = j), p_{2j}^* = P(y = j + 1) \) for \( j = 1, \ldots, (J - 1) \). This model is called a Partial Credit Model (PCM; Muraki, 1992). In summary, we have models and estimation methods to handle latent variables that are measured by both binomial and ordinal data.

After using IRT to develop a set of items, researchers often want to know whether the items are fair to all respondents. In the following section, I will illustrate the
“bias” issue. Although this issue is generally called differential item functioning (DIF) in IRT research literature, it is more general to define the question in the framework of measurement invariance originally developed from factor analysis studies. This more general conceptualization makes it easier to utilize methods developed in other related fields.

2.2 Measurement Invariance

Applying the general measurement invariance assumption to a parametric, IRT framework, Equation (1.1) can be interpreted as construct (ability) $\eta$ and response $X$ (binary or ordinal observed data) relationship holds regardless of the value of $v$. This is not the same as looking for correlations between $X$ and $v$ in regression models. For example, a positive correlation is found in students’ GRE scores ($X$) and parental income ($v$). As parental income increases, GRE scores will also increase. However, this correlation does not necessarily indicate measurement invariance violation. If parental income ($v$) simply increases ability ($\eta$) on the GRE, then this is not a measurement invariance violation; If parental income ($v$) leads students to interpret some items differently, then there exists a violation.

Under this definition, several procedures to detect item bias were introduced. In general, they can be divided into two categories according to whether model comparison is required. Mantel-Haenszel statistic (Holland & Thayer, 1988) and Raju’s Area approach (Raju, 1988) do not involve model comparison, while logistic regression methods (Swaminathan & Rogers, 1990), Lord’s Wald test (Lord, 1980) and LRT (D. Thissen, Steinberg, & Wainer, 1988) are based on comparing models. Below I
will introduce each approach’s main idea.

2.2.1 Mantel-Haenszel

The Mantel-Haenszel (MH) statistic is based on a series of $2 \times 2$ contingency tables where each table is made for each sum score (except full sum score and zero sum score because they make the $2 \times 2$ table collapse into single column). If $v$ does not matter, then the proportion correct should be similar for each category of $v$. Usually, two categories are compared. They are called the focal group and the reference group. The former is often the one being investigated to see if it is disadvantaged by the item, the latter is the comparison group. For example, assuming we want to know whether a 4 item scale is biased against females, then the focal group is female and the reference group is male; the possible sum score would be 0 to 4. Excluding 0 and 4 sum score, we will have three $2 \times 2$ contingency tables for sum score 1, 2 and 3. The $2 \times 2$ contingency table for one of these sum scores is shown in Table 2.1. Based on this contingency table, we can obtain the observed accuracy for each group and the expected accuracy based on two groups. We sum the results across these three contingency tables, resulting in a single statistic.

\begin{table}
\centering
\caption{Contingency table}
\begin{tabular}{lll}
\hline
\text{Gender} & \text{Response} & \text{Total by Gender} \\
\hline
Female & 43 & 9 & 52 \\
Male & 44 & 4 & 48 \\
Total by Response & 87 & 13 & 100 \\
\hline
\end{tabular}
\end{table}
2.2.2 Raju’s Area

Raju’s area approach is based on plots of $\eta$ vs predicted probability of being correct. If an item exhibits no measurement invariance violation across two groups, then the groups’ item response curves should lie on top of each other. If there exists a measurement invariance violation between groups, the differences between item response curve areas describe the magnitude of DIF. Based on the relative position of two group item response curves, the measurement invariance violation can be classified as uniform or nonuniform. For example, in Figure 2.1 left panel, throughout the $\eta$ continuum, the probability of response of 1 is higher for reference group, the shaded area represents the magnitude of measurement invariance. This is referred to as uniform measurement invariance. On the right panel, the reference curve is higher than focal group for part of $\eta$ continuum. This demonstrates nonuniform measurement invariance.
2.2.3 Lord’s Wald Test

Unlike the above two methods, Lord’s Wald test is a model-based approach. This method first fits a model individually to each group, then tests equality of item parameters across groups. Assuming we still have two groups, focal group (F) and reference group (R), we can test a specific parameter by using the $z$ statistic. For example, we can test discrimination parameter $\alpha_j$ as

$$z_j = \frac{\hat{\alpha}_{Fj} - \hat{\alpha}_{Rj}}{\sqrt{\text{Var}(\hat{\alpha}_{Fj}) + \text{Var}(\hat{\alpha}_{Rj})}}.$$ (2.23)

When we simultaneously test equality of discrimination and difficulty parameters for an item across groups, we get a similar statistic which follows a $\chi^2$ distribution with two degrees of freedom:

$$\chi_j^2 = v'_j \Sigma_j^{-1} v_j,$$ (2.24)

where $v'_j = \{(\hat{\alpha}_{Fj} - \hat{\alpha}_{Rj}), (\hat{\gamma}_{Fj} - \hat{\gamma}_{Rj})\}$, $\Sigma_j$ is the covariance matrix of differences between item $j$’s parameter estimates.

2.2.4 Logistic Regression

Logistic regression uses the sum score, group, and interaction of these two variables as predictors in a logistic regression, with item response (on a single item) as the response variable. Specifically, let $x_i$ be the sum score of individual $i$ and let $g_i$ be individual $i$’s group (0=reference group, 1=focal group). For a specific item $j$, we fit
the model:

\[
y_{ij} \sim \text{Bernoulli}(p_{ij})
\]

\[
\logit(p_{ij}) = \tau_0 + \tau_1 x_i + \tau_2 g_i + \tau_3 x_i g_i
\]

If coefficient of \(\tau_2\) is significant, that is evidence for uniform DIF. If coefficient for \(\tau_3\) is significant, that is evidence for nonuniform DIF. The significance of coefficients are often tested by comparing full (including \(\tau_2, \tau_3\) terms) and reduced models (excluding \(\tau_2\) and/or \(\tau_3\) terms).

As stated before, the most popular method for detecting measurement invariance is LRT. Particularly, D. Thissen et al. (1988) proposed to examine the discrimination and difficulty parameters for each item simultaneously by applying the LRT (with \(p\)-value adjusted by Bonferroni correction). If one item demonstrates violation of measurement invariance, that item will be allowed to vary for each group or deleted. This procedure is called “purification”.

However, there are several drawbacks of using a LRT. First, it is time consuming to conduct the purification, especially when there are many items. Second, the LRT is sensitive to large sample size (Bentler & Bonett, 1980). Thus not all significant results have substantial “real” meaning. To deal with this issue in practice, Meredith (1993) provided the notion that the measurement invariance amount or the effect size matters more than the measurement invariance detected. In practice, effect size is dependent on specific research context (Meredith, 1993). In other words, if we can explain possible reason(s) for biased items, it will provide information when effect size is small enough to ignore. Several methods were proposed, such as qualitative (i.e. word detecting), correlational, quasi-experimental and experimental (Mellenbergh,
But these methods are subjective to some extent (qualitative and correlational) and require explanation variables’ property (quasi-experimental and experimental). Besides, these explanations generally involve conducting another study.

Ideally, if a test could detect meaningful measurement invariance violations (as opposed to the sensitivity of large sample size as LRT) according to a selected auxiliary variable while conducting the same study, it could save time and resources. A solution is potentially found in score-based tests, which are related to ideas introduced in econometrics during the 1970s (Brown et al., 1975). We will describe the score-based tests’ theory and application in the following section.

2.3 Structural Change Tests

2.3.1 Theoretical History

Brown et al. (1975) proposed methods for studying the stability over time of regression parameters. CUSUM (recursive residual type), defined to be the sum of recursive residuals decorrelated by the estimate of error variance and the square root of the number of recursive residuals \((n - k)\), where \(k\) is the number of predictors in regression model, was proposed to detect parameter change in regression models. Emphasis was placed on graphical methods, because the distributional theory of this sum was not available. In the 1980s, instead of applying a completely graphical approach, the asymptotic distribution of CUSUM was studied. This distribution relies only on the null hypothesis that parameters are stable (Nyblom, 1989). If the null hypothe-
sis holds, CUSUM will wander around zero (a tied-down random walk) along time. Formally, let $T$ represent a continuous auxiliary variable such as time and let $W(t)$ represent the value of the stochastic process at the point $T = t$. The CUSUM that obeys measurement invariance will follow Brownian bridge, which is defined as:

\[
\mu(t) = 0 \forall t \quad \text{(2.27)}
\]

\[
\text{Cov}(t_1, t_2) = \min(t_1, t_2) \quad \text{(2.28)}
\]

\[
W(0) = 0 \quad \text{(2.29)}
\]

\[
W(1) = 0 \quad \text{(2.30)}
\]

where $\mu(t)$ and $\text{Cov}(t_1, t_2)$ respectively represent the mean at time $t$ and covariance between time $t_1$ and $t_2$. The beginning and ending are restricted to be 0 (“tied down”). A graph of one Brownian bridge is shown in Figure 2.2.

If the measurement invariance assumption is violated, CUSUM will not behave
like a random walk and test statistics based on CUSUM will be large. Tests based on CUSUM were proved to be the locally most powerful test for detecting regression coefficients’ ($\beta$) nonstationary movements in regression models.

Further, Hansen (1992) generalized the CUSUM test to the empirical cumulative sums of first-order conditions. The procedure to obtain the test statistic can be described as below. First, we need to get the first derivatives of log likelihood of observation $i$ for each parameter evaluated at the maximum likelihood estimates (these are called scores); see Equation (2.31) and (2.32), where $\ell$ is the log likelihood function, $x_i$ is the observed data vector for unordered $i$th observation, and $x(i)$ is the ordered vector of observations. The order is determined by auxiliary variable $v$. Then the test statistics are based on the empirical cumulative score (accumulated along ordered observations $1, 2, \ldots, n$), see Equation (2.33).

$$s(\hat{\theta}; x(i)) = \frac{\partial}{\partial \theta} \ell(x(i), \theta)|_{\theta = \hat{\theta}}, \text{ where}$$

$$\hat{\theta} = \arg\max_\theta \ell(\theta; x_1, \ldots, x_n) \quad \text{(2.32)}$$

$$\sum_{i=1}^{n} s(\hat{\theta}; x(i)) = 0, \quad \text{(2.33)}$$

where

$$s(\theta; x(i)) = \left(\frac{\partial \ell(\theta; x(i))}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta; x(i))}{\partial \theta_k}\right) \quad \text{(2.34)}$$

is the partial derivative of the case wise likelihood contributions w.r.t. the parameter $\theta$.

Similarly to the CUSUM test, the sum of the score contributions over all observations is zero for each parameter. If there are no systematic deviations, the score contributions should fluctuate around zero. Conversely, the score contributions will
shift away from zero for subgroups where the model does not fit.

The test statistic based on the empirical cumulative score has two advantages over CUSUM test. First, it has asymptotic local power against nonsationary movements in all parameters instead of only the regression coefficients. This means we can test any parameters within the model with non-trivial power. Moreover, Hansen (1992) showed that the null hypothesis asymptotic distribution is nonstandard but only related to the number of parameters being tested. Thus the critical values can be obtained by simulation. Second, the empirical cumulative score is based on full sample estimation (n scores/score vectors) instead of sub-sample estimation as CUSUM (n − k recursive residuals). Monte Carlo studies demonstrated that test statistics based on cumulative scores have better power than CUSUM tests in almost all scenario considered (Andrews, 1990).

Further, Hjort and Koning (2002) showed that after decorrelation by some consistent estimate of the covariance matrix of the scores, each parameter’s cumulative score converges in distribution to an independent Brownian bridge (in a time interval [0, 1], it starts at 0 when t = 0 and returns to 0 when t = 1). Specifically, it can be written as

$$B(t; \hat{\theta}) = \hat{I}^{-1/2}n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} s(\theta; x_{(i)}) \quad (0 \leq t \leq 1),$$  \hspace{1cm} (2.35)

$$B(\cdot; \hat{\theta}) \xrightarrow{d} B^0(\cdot),$$  \hspace{1cm} (2.36)

where $\hat{I}$ is an estimate of the covariance matrix of the scores, which serves to decorrelate the fluctuation processes associated with individual model parameters; $\lfloor nt \rfloor$ is the integer part of nt (i.e., a floor operator); $x_{(i)}$ reflects the observation vector.
with the $i$-th smallest value of the auxiliary variable $v$ (the ordered $\mathbf{x}$ as shown in Equation (2.34)); and $t$ is $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$. When $t = 1$, the score vector always equals $\mathbf{0}$, as defined in Equation (2.33). $\mathbf{B}^0(\cdot)$, is a $k$ (number of parameters) dimensional Brownian bridge where each column represents a univariate Brownian bridge and pertains to a single parameter.

So far, the score-based tests' theoretical background has been established. In the next section, we will introduce the implementation of these score-based tests.

### 2.3.2 Implementation

Zeileis and Hornik (2007) summarized use of the cumulative score-based tests. The standard procedure to conduct these tests is described as follows. First, parameter estimates are obtained from a model (typically linear regression) estimated via ML. Second, the scores are collected, as shown in Equation (2.31) for every observation (with the empirical cumulative score typically being ordered according to time). Third, the fluctuations in the empirical cumulative score need to be aggregated to a scalar, serving as the test statistic. The asymptotic distribution of this test statistic can be easily obtained by applying the same aggregation to the asymptotic Brownian bridge, so that critical values and the corresponding $p$-values can be derived. Depending on the aggregation function, various statistics have been proposed. Selection of these statistics is typically based on the plausible patterns of violation of measurement invariance.

The simplest aggregation strategy is to reject measurement invariance if the largest component of the empirical cumulative score matrix is greater than a critical value. Based on the location of the detected component, we can easily locate the violating
parameter and timing. Because this statistic is searching for the maximum over
the parameters (columns of the empirical cumulative score matrix) and time points
(rows of the empirical cumulative score matrix), this statistic is called the “double
maximum” (DM).

\[
DM = \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} |B(\hat{\theta})_{ij}|. \quad (2.37)
\]

For example, Hsu (1979) investigated the stability of weekly Dow Jones industrial
average stock returns from 1971 to 1974. There are two parameters in the model,
mean and variance. The DM statistic indicates significant instability in the variance
parameter, and the corresponding time point of the instability is March 1973.

However, taking maximums “wastes” power if many of the parameters change
and/or there exist multiple changing timepoints instead of one. In such cases, sums
across parameters and times are more suitable. The Cramér-von Mises (CvM) statis-
tic falls in this category,

\[
CvM = n^{-1} \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,k} B(\hat{\theta})_{ij}^2. \quad (2.38)
\]

If we know there is definitely only one change point, but that change point affects
multiple parameters, we can aggregate by summing over parameters, then taking the
maximum over the time interval (scaled by variance). This statistic is max LM.

\[
\max LM = \max_{i=1,\ldots,n} \left\{ \frac{i}{n} \left(1 - \frac{i}{n}\right) \right\}^{-1} \sum_{j=1,\ldots,k} B(\hat{\theta})_{ij}^2. \quad (2.39)
\]

For example, an “Operation Ceasefire” intervention was launched to prevent homicide
rates in Boston from early 1995. However, it is hard to tell precisely when this
intervention works. Zeileis and Hornik (2007) utilized a Poisson model and max LM statistic to identify the changing time point.

Note that this statistic is asymptotically equivalent to the max LR mentioned in Chapter 1. The asymptotic distributions of them are the same: the maximum of a squared, \( k \)-dimensional tied-down Bessel process, which is represented in Equation (2.39) by the Euclidean norm of a \( k \)-dimensional Brownian bridge \( \sum B(\hat{\theta})_{ij}^2 \).

Across the above applications, the auxiliary variable is always time. The models used are generally regression utilized in time series. Merkle and Zeileis (2013) first applied these score-based tests to measurement invariance issues in the framework of factor analysis. The auxiliary ordered variable could be any variable of interest (e.g. age). Further, Merkle et al. (2014) introduced two statistics that could deal with the situation when the auxiliary variable is ordinal, such as grade or income level. They are described below.

For an ordinal auxiliary variable with \( m \) levels, the modifications are based on \( t_\ell \) \((\ell = 1, \ldots, m - 1)\), which are the empirical, cumulative proportions of individuals observed at the first \( m - 1 \) levels. The modified statistics are then given by

\[
WDM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n} \left( 1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j=1,\ldots,k} |B(\hat{\theta})_{ij}|, \quad (2.40)
\]

\[
\text{max } LM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n} \left( 1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\ldots,k} B(\hat{\theta})_{ij}^2, \quad (2.41)
\]

where \( i_\ell = \lfloor n \cdot t_\ell \rfloor \) \((\ell = 1, \ldots, m - 1)\).

If the auxiliary variable \( v \) is only nominal/categorical, the empirical cumulative sums of scores can be used to obtain a Lagrange multiplier statistic by first summing scores within each of the \( m \) levels of the auxiliary variable, then “summing the sums”
(Hjort & Koning, 2002). This test statistic can be formally written as

\[
LM_{uo} = \sum_{\ell=1}^{m} \sum_{j=1}^{k} \left( B(\hat{\theta})_{i\ell j} - B(\hat{\theta})_{i\ell-1j} \right)^2, \tag{2.42}
\]

where \( B(\hat{\theta})_{i0j} = 0 \) for all \( j \). This statistic is asymptotically equivalent to the usual, likelihood ratio statistic, and it is advantageous over the LRT (shown in Equation (1.5)) because it requires the estimation of only one model (the restricted model). In general, all test statistics’ (continuous, ordinal and categorical) properties are retained when they are applied to models involving latent variables.

Aside from the factor models considered above, other popular latent variable models arise from IRT. The only difference is that the observed variables are not continuous as the factor analysis model explored by Merkle and Zeileis (2013) and Merkle et al. (2014). Strobl, Kopf, and Zeileis (2013) applied the tests to the Rasch model (the simplest IRT model, as mentioned in Chapter 1) to replace the traditional DIF test. This is the first attempt to apply score-based tests in detecting DIF in items. However, conditional maximum likelihood estimation is utilized in this study. This estimation method is difficult to extend to other IRT models.

Motivated by promising results of Merkle and Zeileis (2013) and Strobl et al. (2013), I am aiming to generalize these tests to the most popular IRT model (2PL). The key for utilizing the score-based tests in the 2PL model is to get the empirical score vector for each individual. The specific estimation method employed influences how we get the score vector, as described below.
2.3.3 Estimation

Marginal Maximum Likelihood

Assuming the 2PL model described above, the score vector for each individual can be described as:

\[ s(\theta; y_i) = \left( \frac{\partial \ell(\theta; y_{i1})}{\partial \alpha_1}, \ldots, \frac{\partial \ell(\theta; y_{ip})}{\partial \alpha_p}, \frac{\partial \ell(\theta; y_{i1})}{\partial \gamma_1}, \ldots, \frac{\partial \ell(\theta; y_{ip})}{\partial \gamma_p} \right), \]  (2.43)

where the likelihood function for each observation \( \ell(\theta; y_{ij}) \) can be expressed as

\[ \ell(\theta; y_{ij}) = \log \int_{-\infty}^{\infty} f(y_{ij}|\eta_i) g(\eta_i) d\eta_i \]  (2.44)

Then we plug in Equation (2.3) to Equation (2.44), obtaining the following functional form:

\[ \ell(\theta; y_{ij}) = \int_{-\infty}^{\infty} \left( \alpha_j \eta_i + x_{ij} \gamma_j + \log(1 - \frac{1}{1 + \exp(-\alpha_j \eta_i + \gamma_j)}) \right) g(\eta_i) d\eta_i. \]  (2.45)

Therefore, the component in the score vector (2.43) is:

\[ \frac{\partial \log(L(p|x_i))}{\partial \alpha_j} = \int_{-\infty}^{\infty} \left\{ x_{ij} \eta_i - \frac{\eta_i \exp(\gamma_j + \alpha_j \eta_i)}{1 + \exp(\gamma_j + \alpha_j \eta_i)} \right\} g(\eta_i) d\eta_i, \]  (2.46)

\[ \frac{\partial \log(L(p|x_i))}{\partial \gamma_j} = \int_{-\infty}^{\infty} \left\{ x_{ij} - \frac{\exp(\gamma_j + \alpha_j \eta_i)}{1 + \exp(\gamma_j + \alpha_j \eta_i)} \right\} g(\eta_i) d\eta_i. \]  (2.47)

The integral associated with \( \eta_i \) does not have closed form and we estimate the integral by using Gauss-Hermite quadrature, which is one form of quadrature to calculate integrals involving \( e^{-x^2} \).
As dimension of ability ($\eta$) increases, the multidimensional quadrature will become computationally difficult. To avoid this constraint, Pairwise Maximum Likelihood estimation (PML) can instead be used. This is described in the following section.

**Pairwise Maximum Likelihood**

PML utilizes the factor analysis framework. The log likelihood function for the model is shown as Equation (2.9) (before transformation).

\[
\ell(\theta; y) = \sum_{i=1}^{n} \log \int_{y_i} f(y_i^*) dy_i^* \quad (2.48)
\]

We can see the difficulty involves integration over the $p$-dimensional normal distribution. We can change the integral notation to predicted probability. This change is convenient for later comparison.

\[
\ell(\theta; y) = \sum_{i=1}^{n} \log \pi_i(\theta). \quad (2.49)
\]

Jöreskog and Moustaki (2001) proposed that the model likelihood function can be modified as:

\[
\ell(\theta; y) = \sum_{i=1}^{n} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) - ap \sum_{j} \ell(\theta; y_{ij}) \right\} \quad (2.50)
\]

where $\sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})$ is likelihood for all item pairs of individual $i$, which is similar to the multiple 2-way contingency tables; $a$ is a constant to be chosen for optimal efficiency. Katsikatsou et al. (2012) showed that the most appropriate choice of $a$ is zero and it will not affect either the accuracy or efficiency of estimation. Thus, maximizing $\ell(\theta; y)$ will be equivalent to maximizing the composite pairwise log likelihood.
$p\ell(\theta; y)$, which can be expressed in the following form:

$$
p\ell(\theta; y) = \sum_{i=1}^{n} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\},
$$

$$
= \sum_{i=1}^{n} \left\{ \sum_{j<k} \left( m_j \sum_{c_j=1}^{m_j} \sum_{c_k=1}^{m_k} \log \pi(y_{ij}y_{ik})^{c_jc_k}(\theta) \right) \right\}, \tag{2.51}
$$

where $m_j$ and $m_k$ represents the total category for each item ($m_j = 2$, $m_k = 2$). Further, $\pi^{(y_{ij}y_{ik})}(\theta)$ is the probability for individual $i$ in category $c_j$ ($c_j = 1$) and $c_k$ ($c_k = 1$) under the model. This can be expressed explicitly in the following form:

$$
\pi^{(y_{ij}y_{ik})}(\theta) = \pi(y_{ij} = c_j, y_{ik} = c_k; \theta) = \Phi_2(\tau^{(y_{ij})}, \tau^{(y_{ik})}; \rho_{y_{ij}y_{ij}}) - \Phi_2(\tau^{(y_{ik})}, \tau^{(y_{ik})}; \rho_{y_{ik}y_{ik}})
$$

$$
- \Phi_2(\tau^{(y_{ij})}, \tau^{(y_{ik})}; \rho_{y_{ik}y_{ik}}) + \Phi_2(\tau^{(y_{ij})}, \tau^{(y_{ik})}; \rho_{y_{ij}y_{ik}}), \tag{2.52}
$$

where $c_{j-1} = 0$, $c_{k-1} = 0$ and $\Phi_2(a, b, \rho)$ is the bivariate cumulative standard normal distribution with correlation $\rho$ evaluated at the point $(a, b)$,

$$
\rho_{y_{ij}y_{ik}} = \lambda_j \lambda_k, \tag{2.53}
$$

where $j = 1, \ldots, p - 1$, $k = j + 1, \ldots, p$.

Comparing Equation (2.51) with Equation (2.48), we can see that the $p$-dimensional integral is reduced to all possible pairwise ($j < k$) integrals. This significantly reduces the complexity of computation.

Maximizing the log-likelihood function in Equation (2.51) over the parameter $\theta$, we obtain the composite pairwise maximum likelihood estimator $\hat{\theta}_{PML}$. Again, this
is equivalent to solving for $\theta$ so that the sum of scores equals zero. The score vector of the pairwise likelihood for each individual can be decomposed in two blocks: the first derivative with respect to the factor loading $\Lambda$ and thresholds $\mathbf{\tau}$:

$$
\mathbf{s}(\theta; \mathbf{y}_i) = \left( \frac{\partial}{\partial \mathbf{\Lambda}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\}, \frac{\partial}{\partial \mathbf{\tau}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\} \right). \quad (2.54)
$$

The elements of the subvector $\frac{\partial}{\partial \mathbf{\tau}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\}$ are the first derivatives with respect to thresholds $\frac{\partial}{\partial \tau_{cj}}$ (same pattern for $k$) and are given as below:

$$
\frac{\partial}{\partial \tau_{cj}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\} = \sum_{c_j=1}^{m_j} \frac{1}{\pi_{c_j c_k} c_j} - \frac{1}{\pi_{(c_j-1)c_k} c_j} \frac{\partial \pi_{(y_{ij} y_{ik})}}{\partial \tau_{cj}}, \quad (2.55)
$$

where

$$
\frac{\partial \pi_{(y_{ij} y_{ik})}}{\partial \tau_{cj}} = \phi_1(\tau_{cj}) \left[ \Phi_1 \left( \frac{\tau_{cj}}{\sqrt{1 - \rho^2 y_{ij} y_{ik}}} \right) - \Phi_1 \left( \frac{\tau_{cj-1}}{\sqrt{1 - \rho^2 y_{ij} y_{ik}}} \right) \right], \quad (2.56)
$$

where $\phi_1$ and $\Phi_1$ are the standard univariate normal distribution and cumulative density, respectively.
The score with respect to $\Lambda$ are given by using chain rules.

\[
\frac{\partial}{\partial \Lambda} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\} = \frac{\partial}{\partial\rho_{y_{ij}y_{ik}}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\} \frac{\partial\rho_{y_{ij}y_{ik}}}{\partial\Lambda} \tag{2.57}
\]

\[
(2.58)
\]

The partial derivative with respect to $\rho_{y_{ij}y_{ik}}$ is

\[
\frac{\partial}{\partial \rho_{y_{ij}y_{ik}}} \left\{ \sum_{j<k} \ell(\theta; (y_{ij}, y_{ik})) \right\} = \sum_{c_j=1}^{m_j} \sum_{c_k=1}^{m_k} \frac{1}{\pi^{(y_{ij}y_{ik})}_{c_jc_k}} \frac{\partial\pi^{(y_{ij}y_{ik})}_{c_jc_k}}{\partial\rho_{y_{ij}y_{ik}}} \tag{2.59}
\]

where $\frac{\partial\pi^{(y_{ij}y_{ik})}_{c_jc_k}}{\partial\rho_{y_{ij}y_{ik}}}$ is given in Olsson (1979) as below:

\[
\frac{\partial\pi^{(y_{ij}y_{ik})}_{c_jc_k}}{\partial\rho_{y_{ij}y_{ik}}} = \phi(\tau^{(y_{ij})}_{c_j}, \tau^{(y_{ik})}_{c_k}; \rho_{y_{ij}y_{ik}}) - \phi(\tau^{(y_{ij})}_{c_j}, \tau^{(y_{ik})}_{c_k-1}; \rho_{y_{ij}y_{ik}}) - \phi(\tau^{(y_{ij})}_{c_j-1}, \tau^{(y_{ik})}_{c_k}; \rho_{y_{ij}y_{ik}}) + \phi(\tau^{(y_{ijk})}_{c_j-1}, \tau^{(y_{ik})}_{c_k-1}; \rho_{y_{ij}y_{ik}}) \tag{2.60}
\]

The partial derivative of $\rho_{y_{ij}y_{ik}}$ with respect to $\Lambda$ can be obtained from chain rule.

The final form of the derivatives is:

\[
\frac{\partial\rho_{y_{ij}y_{ik}}}{\partial\Lambda} = \lambda_k \frac{\partial\lambda_j}{\partial\Lambda} + \lambda_j \frac{\partial\lambda_k}{\partial\Lambda} \tag{2.61}
\]

where $\frac{\partial\lambda_j}{\partial\Lambda}$, $\frac{\partial\lambda_k}{\partial\Lambda}$ are matrices of zeros and ones. The dimensions are determined by $\Lambda$.

The above score matrix can be easily extended to multidimensional and multiple
categories by defining $\eta$ as a $q$-dimensional vector of latent variable and

$$y_{ij} = c_j \iff \tau^{(y_{ij})}_{c_j-1} < y^*_{ij} < \tau^{(y_{ij})}_{c_j},$$

(2.62)

where $\tau^{(y_{ij})}_{c_j}$ is the $c_j$ threshold of variable $y_{ij}$ and $-\infty = \tau^{(y_{ij})}_{0} < \tau^{(y_{ij})}_{1} < \ldots < \tau^{(y_{ij})}_{m_j-1} < \tau^{(y_{ij})}_{m_j} = +\infty$. Then the only changing place in the score matrix is the place involving $\lambda_j$ and $\lambda_k$. For further details, see Katsikatsou et al. (2012).

After we obtain the score vector for each individual, we need to get the estimate of score covariance matrix, which is shown in Equation 2.35 as $\hat{I}$. For regular Maximum Likelihood estimation, the covariance matrix is equal to the information matrix (Merkle & Zeileis, 2013). However, this identity does not hold for PML (Katsikatsou et al., 2012). Therefore, we use the most common form of covariance matrix estimate $\hat{I} = 1/n \sum_{i=1}^n s(\hat{\theta}, x_{(i)})s(\hat{\theta}, x_{(i)})^T$ instead of information matrix.

### 2.4 Summary

Comparing these two estimation methods, we can see that scores obtained from PML are less computationally intensive. We can obtain scores easily for the 2PL model and other extended models. Therefore, I will focus on the score-based tests for PML.

Specifically, after we obtain the score vector for individual, we will order these row vectors according to an auxiliary variable (e.g. age or income level). Then the empirical cumulative score matrix will then be built based on the ordered score rows. The tests can thus be carried out as the previous mentioned examples (Merkle & Zeileis, 2013; Merkle et al., 2014). Depending on the scale of the auxiliary variable (continuous, ordinal, categorical), we will choose an appropriate statistic. If the test
When the score-based test demonstrates a measurement invariance violation, researchers generally want to know where and when the measurement invariance happens (Zeileis & Hornik, 2007; Millsap, 2005) (as the “locating issue” mentioned in Chapter 1). We can see this problem is naturally embedded in score-based tests. The parameter is represented by the columns of the empirical cumulative score matrix, while the changing point/level is represented by the rows of the empirical cumulative score matrix. As argued before, the double maximum statistic (ordinal or continuous) is particularly appealing for this because the \( k \) columns of empirical cumulative score can be graphed along with boundaries for the associated critical values separately. Any column that demonstrates boundary crossing implies a violation of measurement invariance. In each column, the location of the most extreme deviation(s) (peak(s)) in the process convey the changing point. Thus in this dissertation, I am going to examine the “locating” properties of the family of score-based tests. In the following chapter, I will demonstrate the score-based tests’ locating uses via simulations.
Chapter 3

Simulation

In this chapter, I explore parameter locating and level locating properties in Simulation 1 and 2, respectively. Within Simulation 1, parameter locating property is examined in the frame of 2PL model (Simulation 1.1) and extended model (Simulation 1.2); Within Simulation 2, locating property was divided into a single parameter shift resulting in two groups to identify (Simulation 2.1) and two parameter shifts resulting in multi groups to identify (Simulation 2.2).

3.1 Simulation 1.1: Parameter Locating Property: 2PL Model

In this study, I aim to examine score-based tests’ ability to locate item parameters that violate measurement invariance. Consider a hypothetical battery of five items administered to students in eight age groups \((v = 1, 2, 3, 4, 5, 6, 7, 8)\). All five items are intended to measure verbal ability. Because the items are described by two types
of parameters (discrimination and difficulty), we may observe two types of measurement invariance violations. One possibility is that certain items can discriminate young students’ \( (v = 1, 2, 3, 4) \) verbal ability very well, but cannot discriminate older students’ \( (v = 5, 6, 7, 8) \) verbal ability so well. The other possibility is that certain items are very difficult for young students \( (v = 1, 2, 3, 4) \) but less difficult for older students \( (v = 5, 6, 7, 8) \). It is plausible that violations in an item’s discrimination parameter impacts the item’s difficulty parameter and vice versa. It is also plausible that one item’s violation impacts the other items. Thus, the goal of Simulation 1.1 is to examine the extent to which the score-based tests attribute the measurement invariance violation to the parameters that are truly in violation.

Method

To examine the above hypothetical example, data were generated from a 2PL model which consisted of the 5 items described earlier. A violation occurred in one of the two places: discrimination parameter of item 3, \( \alpha_3 \), or difficulty parameter of item 3, \( \gamma_3 \). The fitted models matched the data generating model, and parameter estimates are obtained by one of the estimation methods described in Chapter 2, MML or PML. Measurement invariance violations were tested in eight subsets of parameters: each item’s difficulty parameter (or discrimination parameter, depending on the true violation place), the true violating items’ non-violating parameter (\( \gamma_3 \) or \( \alpha_3 \)), all items’ difficulty parameters, and all items’ discrimination parameters.

Power and Type I error were examined across three sample sizes \( (n = 120, 480, 960) \), three numbers of categories \( (m = 4, 8, 12) \) and 17 magnitudes of invariance violations (described in the following sentences). The measurement invariance violations began
at age level $m/2 + 1$ and were consistent thereafter. Students below age level $m/2 + 1$ deviated from students at or above age level $m/2 + 1$ by $d$ times the parameters’ asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.25, 0.5, \ldots, 4$.

For each combination of sample size $(n) \times$ violation magnitude $(d) \times$ violating parameter $\times$ categories $(m)$, 5000 data sets were generated and tested. In all conditions, we maintained equal sample sizes in each subgroup of the categories $m$. Statistics from Equations (2.40), (2.41) (because the auxiliary variable $m$ is ordinal) and (2.42) (categorical statistic, which ignored the ordering information) were examined. As mentioned previously, (2.42) is asymptotically equivalent to the usual likelihood ratio test. Thus, this statistic provides information about the relative performance of the ordinal statistics vs. the LRT.

**Results**

Full simulation results for PML are presented in Figures 3.1 to 3.4. (Similar results for MML are shown in Appendix A). In these figures, I generally show results exhibited nonzero power curves. Because item 1, 2, 4 and 5 display similar near-zero power curves in all conditions, I only show item 2’s results as they are representative of all the non-violating items. Figure 3.1 and Figure 3.2 display power differences among statistics. Figure 3.3 and Figure 3.4 display power differences among sample sizes.

Figure 3.1 demonstrates power curves (of sample size 960) as a function of violation magnitude in item 3’s discrimination parameter $\alpha_3$, with the parameters being tested changing across rows, the number of levels $m$ of the ordinal variable $v$ across columns, and lines reflecting different test statistics. Figure 3.2 demonstrates similar power curves when the violating parameter is item 3’s difficulty parameter $\gamma_3$. 
Figure 3.1: Simulation 1.1. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\alpha_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 

Violation Magnitude($\alpha_3$)

<table>
<thead>
<tr>
<th>Power</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_3$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha_1, \ldots, \alpha_5$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\gamma_1, \ldots, \gamma_5$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha_1, \ldots, \alpha_5$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\gamma_1, \ldots, \gamma_5$</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

$\maxLM_o$ $WDM_o$ $LM_{uo}$

maxLM_o $\bullet$ $\bullet$ $\bullet$

WDM_o $+$ $+$ $+$

LM_uo $\nabla$ $\nabla$ $\nabla$

Violation Magnitude($\alpha_3$)
Figure 3.2: Simulation 1.1. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\gamma_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 
Figure 3.3: Simulation 1.1. Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\alpha_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 
Figure 3.4: Simulation 1.1. Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_o$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{m}$), estimated by PML 2PL Model. The parameter violating measurement invariance is $\gamma_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 

<table>
<thead>
<tr>
<th>$n=120$</th>
<th>$n=480$</th>
<th>$n=960$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$, $\gamma_2$</td>
<td>$\gamma_1$, $\gamma_2$</td>
<td>$\gamma_1$, $\gamma_2$</td>
</tr>
<tr>
<td>$\alpha_1$, $\alpha_2$, $\alpha_3$</td>
<td>$\alpha_1$, $\alpha_2$, $\alpha_3$</td>
<td>$\alpha_1$, $\alpha_2$, $\alpha_3$</td>
</tr>
<tr>
<td>$\eta_1$, $\eta_2$, $\eta_3$</td>
<td>$\eta_1$, $\eta_2$, $\eta_3$</td>
<td>$\eta_1$, $\eta_2$, $\eta_3$</td>
</tr>
</tbody>
</table>

Violation Magnitude($\gamma_3$) vs. Power

Violation Magnitude($\gamma_3$) vs. Power

Violation Magnitude($\gamma_3$) vs. Power

Violation Magnitude($\gamma_3$) vs. Power
Within each panel of Figure 3.1 to Figure 3.2, the three lines reflect the three test statistics. It is seen that the two ordinal statistics exhibit similar results, with max $LM_{uo}$ demonstrating lower power across all situations. This demonstrates the sensitivity of the ordinal statistics to invariance violations that are monotonic with $v$. In situations where only one parameter is tested, $WDM_o$ and max $LM_o$ exhibit equivalent power curves. This is because, these two statistics are equivalent when only one parameter is tested.

Figure 3.3, and Figure 3.4 display similar power curves (of statistics $WDM_o$), but with lines reflecting different sample sizes. Figure 3.3 demonstrates results when the violating parameter is $\alpha_3$. We can see with increasing sample sizes, the power increases. Figure 3.4 displays the results when violating parameter is $\gamma_3$. All three sample sizes demonstrate similar power curves.

From these figures, one generally observes that the tests isolate the parameter violating measurement invariance. Comparing Figure 3.1 and Figure 3.2 in general, we can see the tests have somewhat higher power to detect measurement invariance violations in the difficulty parameter as opposed to discrimination parameter. I speculate the reason is that difficulty can be seen as main effect while discrimination is more like an interaction term. Meanwhile, comparing Figure 3.3 and Figure 3.4 in general, it can be seen that we need large sample size ($n = 960$) to get decent power (with increasing violation magnitude) when the violating parameter is discrimination; whereas sample size has much less effect on detecting difficulty parameter violation. With sample size as small as $n = 120$, the power can be more than 0.8 (when the violating magnitude is larger than 3).

Finally, simultaneous tests of all discrimination parameters, all difficulty parame-
ters result in decreased power, as compared to the situation where one tests only the violating parameter. This occurs because, in testing a subset of parameters (only one of which violates measurement invariance), we are dampening the signal of a measurement invariance violation. This “dampening” effect is more apparent for the max $LM_o$ statistic, because it involves a sum across all tested parameters (see Equation (2.41)). Conversely, $WDM_o$ takes the maximum over parameters (see Equation (2.40)), so that invariant parameters have no impact on this statistic. However, this does not imply that $WDM_o$ is always better than max $LM_o$. In situations where multiple parameters change simultaneously, max $LM_o$ will have higher power.

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct 2PL parameter. This provides evidence that, in practice, one can have confidence in the tests’ abilities to locate the measurement invariance violation in item parameters.

### 3.2 Simulation 1.2: Parameter Locating Property: Extended Model

In the previous simulation, the test can locate the measurement invariance violating item parameters $\alpha_j/\gamma_j$ when the testing model is the same as data generating model (2PL Model), which does not estimate person hyperparameters. However, as stated in measurement invariance section, person hyperparameters (ability mean, variance) could be unstable, too. Although person hyperparameter instability is not counted as measurement invariance violation, ignoring person hyperparameters may cause a “false alarm” in item parameters.
Specifically, in a regular 2PL model, we assume person parameter as \( \eta_i \sim N(0, 1) \). But person parameter could follow a distribution described by other hyperparameters, such as \( \eta_i^* \sim N(\mu_l, \sigma^2_l) \). To illustrate person hyperparameters’ effect, I will do the following transformation:

\[
\logit(p_{ij}) = \gamma_j + \alpha_j \eta_i^*,
\]

\[= \gamma_j + \alpha_j (\mu_l + \sigma_l \eta_i),
\]

\[= (\gamma_j + \alpha_j \mu_l) + \sigma_l \alpha_j \eta_i.
\]

We can see changes in \( \sigma_l \) will cause evidence of measurement invariance violation in \( \alpha_j \) (for any \( j \)) and changes in \( \mu_l \) will cause evidence of measurement invariance violation in \( \gamma_j \) (for any \( j \)). Besides, since \( \mu_l \) is not 0 any more, changes in \( \alpha_j \) will also cause evidence of measurement invariance violation in \( \gamma_j \) (works through the term \( \alpha_j \mu_l \)). Therefore, the item specificity and discrimination \( \alpha_j \)/difficulty \( \gamma_j \) specificity are both lost. So we should estimate person hyperparameter \( \mu_l \) and \( \sigma^2_l \) when there is uncertainty about person ability mean and variance.

However, there is no clear evidence about the impact of estimating \( \mu_l \) and \( \sigma_l \) on the power of detecting measurement invariance violation in item parameters \( \alpha_j \) and \( \gamma_j \). In the following, I will conduct simulation on this specific topic.

### 3.2.1 Method

To examine the impact of person hyperparameter, data were generated from an extended model which consisted of person hyperparameters and the 5 items described in Simulation 1.1. Without loss of generality, I continue to use the hypothetical example
described in Simulation 1.1. The only changing place is the person’s hyperparameters. Students (younger) in level 1, 2, 3, 4 are generated from $\eta_i \sim N(0, 1)$, while students (older) in level 5, 6, 7, 8 are generated from $\eta_i \sim N(-1, 2)$. The fitted model estimates each levels’ $\mu_l$ (with level 1 fixed to be zero for identification), $\sigma_{l}^2$ (with level 1 fixed to be 1 for identification) and the 5 items’ parameters as Simulation 1.1. Parameter estimates are obtained by PML. Because the extended model has more parameters to be estimated (7 mean parameters $\mu_l$ and 7 variance parameters $\sigma_{l}^2$), to maintain a similar parameter/sample size ratio, the sample sizes are increased to be $n = 1200, 4800, 9600$ in this Simulation.

A violation still occurred in one of the two places: discrimination parameter of item 3, $\alpha_3$, or difficulty parameter of item 3, $\gamma_3$. Measurement invariance violations were tested in the same subsets of parameters as Simulation 1.1. Power and Type I error were examined across three sample sizes and 17 magnitudes of invariance violations (manipulated in the same way as Simulation 1.1). For each combination of sample size ($n$) $\times$ violation magnitude ($d$), 5000 data sets were generated and tested. In all conditions, we still maintained equal sample sizes in each level of the categories $m = 8$. Statistics from Equations (2.40), (2.41) and (2.42) were examined as Simulation 1.1.

3.2.2 Results

Full simulation results for extended model estimated by PML are presented in Figures 3.5 and 3.6. In both figures, item 2’s results are still representative as non-violating items’.

Figure 3.5 demonstrates power curves as a function of violation magnitude in item
Figure 3.5: Simulation 1.2. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting extended Model). The parameter violating measurement invariance is $\alpha_3$. The number of categories is $m = 4$. Panel labels denote the parameter(s) being tested and sample size.
Figure 3.6: Simulation 1.2. Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by PML (fitting extended Model). The parameter violating measurement invariance is $\gamma_3$. The number of categories is $m = 4$. Panel labels denote the parameter(s) being tested and sample size.
3’s discrimination parameter $\alpha_3$, with the parameters being tested changing across rows, the sample sizes $n$ across columns, and lines reflecting different test statistics. Figure 3.6 demonstrates similar power curves when the violating parameter is item 3’s difficulty parameter $\gamma_3$.

From these two figures, one generally observes that the tests isolate the parameter violating measurement invariance in the extended model (across rows); With increasing number of sample sizes, the power increases (across columns). This sample size effect is still more present in detecting difficulty parameters. We need sample size as large as 9600 to get decent power (with increasing violation magnitude); whereas there is no big difference across columns when violating parameter is difficulty.

Within each panel of Figure 3.5 to Figure 3.6, the three lines reflect the three test statistics. It is seen that the two ordinal statistics still exhibit similar results, with max $LM_{uo}$ demonstrating lower power across all situations. Therefore, the sensitivity of the ordinal statistics is preserved in the extended model.

Comparing Figure 3.5 and Figure 3.6 in general, we can see the tests still have somewhat higher power to detect measurement invariance violations in the difficulty parameter as opposed to discrimination parameter. The speculated reason remain the same as Simulation 1.1. Moreover, the “dampening” effect in simultaneous tests of all discrimination parameters, all difficulty parameters also remains.

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct extended model’s parameter. Although estimating extended model requires much more observations than 2PL model to obtain decent power curves, this is very necessary in practice when there is uncertainty about person’s hyperparameter changes in each level. Otherwise, there will be serious “false
alarm” as illustrated by the Equation 3.1 to Equation 3.3.

3.3 Simulation 2.1: Level Locating Property: Single Changing Level Locating Property

In this study, I aim to examine the proposed tests’ ability to locate the correct changing level. Assuming the same hypothetical battery of five scales administered to students in eight age groups ($v = 1, 2, 3, 4, 5, 6, 7, 8$), the changing point of differing younger and older students does not necessarily need to be near the middle level 5 (represented by $m/2 + 1$ in Simulation 1.1). The changing point is specifically defined as the level where people at or above has a different set of parameters from people below. ¹

Using the same hypothetical 2PL model described in Simulation 1.1 and assuming one changing point, I manipulated the age level ($v$) where the parameter changes. The other set up is the same as Simulation 1.1.

Method

Test statistics were examined across change point at level $m = 2, \ldots, 7$. Without loss of generality, the sample size is set to be 960. Equal sample sizes were maintained in each level. (Therefore, sample size is 120 in each level). Thus, except changing point at 5 ($m/2 + 1$) as Simulation 1.1, there will be all unequal sample sizes between non-violation group and violation group. Due to the fact that the pattern is symmetric for changing points (3, 4) and (6, 7), results only for changing points at (2, 3, 4, 5) are

¹When the changing point is the smallest or largest level, it means people at the smallest or largest level has a different set of parameters from people not. For simplicity, I did not simulate results for these two cases. Results should retain in a similar manner.
Figure 3.7: Simulation 2.1. Simulated empirical fluctuation process summary of max $WDM_0$ 25% quantile, median and 75% quantile across four changing levels for violating parameter $\alpha_3$ (left) and $\gamma_3$ (right), $n = 960$, estimated by PML. Sub-panel labels denote the changing level and violation magnitude $d = 4$.

Results

Full simulation results of the statistic $WDM_0$ (same pattern for max $LM_0$) for 2PL model estimated by PML with various changing points are presented in Figures 3.7. The left panel demonstrates empirical fluctuation process (of sample size 960) across 5000 replications when the violating parameter is $\alpha_3$, with the parameters being tested displayed.
is the violating parameter and lines reflecting statistics $WDM_o$’s 25% percentile, median and 75% percentile among 5,000 replication. The right panel of Figure 3.7 demonstrates similar empirical fluctuation process when the violating parameter is $\gamma_3$.

Within each sub-panel of Figure 3.7, we can see $WDM_o$’s 25% percentile, median and 75% percentile demonstrates the maximum across levels corresponding to the change level shown in the sub-panel title.

The difference between the left panel and right panel is that when the violating parameter is $\alpha_3$, 25% percentile line at the largest value is lower than the critical value, which indicates that the power of detecting discrimination parameter violation will be lower than 75% (1-25%); while the power of detecting difficulty parameter violation will be more than 75%.

Specifically, among 5,000 replications of each combination, when the violating parameter is $\alpha_3$, the percentage of the maximum value of $WDM_o$ statistic corresponds to the correct changing level and larger than the critical value (so this percentage is close to but smaller than power) is 0.54, 0.58, 0.6, 0.57 for changing level 2, 3, 4, 5; the percentage of max $LM_o$ statistic is the same as one parameter is tested. When the violating parameter is $\gamma_3$, the corresponding percentage for $WDM_o$ ranges from 0.93 to 0.95, much higher than the percentage of detecting $\alpha_3$’s correct changing level. These results are consistent with Simulation 1.1: Power for detecting $\gamma_3$ measurement invariance violation is higher than that for $\alpha_3$.

In summary, we found that the proposed tests’ maximum across levels can attribute measurement invariance violations to the correct level. This provides evidence that, in practice, one can have some confidence in the tests’ abilities to locate the measurement invariance violation level, assuming there is only one change point.
3.4 Simulation 2.2: Level Locating Property: Multiple Changing Level Detection

If there is more than one change point, a natural extension of the above method is to observe “peaks” (in analogy of the maximum in Simulation 2.1) in the empirical fluctuation process. “Peak” is defined as the level satisfying the following two conditions: (1) its associated statistic is larger than the pre and post level’s associated statistic, (2) its associated statistic is larger than the critical value.\(^2\) The drawback of this approach is that it is impossible to demonstrate “peaks” in some situations (i.e. when changing points are adjacent). The second approach involves using a model based recursive partitioning method (Zeileis, Hothorn, & Hornik, 2008). It proceeds in the following way: (1) fit the 2PL model to the whole sample, (2) test for parameter instability over all parameters, (3) if there is some overall parameter instability detected, split the sample w.r.t. the auxiliary variable associated with the highest statistics (highest instability), (4) repeat the procedure in each of the resulting subsamples until there is no overall parameter instability detected. The drawback of this approach is that the parameter locating is lost because the splitting decision is based on overall parameter instability. All model parameters are re-estimated for each splitting node. The third approach is adapted from the second. Instead of re-estimate all parameters for each splitting node, we can only estimate parameter(s) of interest and fix the other parameter values to the full sample estimates. Thus, we proceed by: (1) Estimate a model with full sample, (2) test for parameter instability for chosen parameter(s) while hold the other parameters to be the full sample estimates, (3) given the split points resulting from (2), estimate a model with segmented chosen parameter(s) while hold the other parameters to be the full sample estimates, (4) repeat the procedure in each of the resulting subsamples until there is no overall parameter instability detected. The drawback of this approach is that the parameter locating is lost because the splitting decision is based on overall parameter instability. All model parameters are re-estimated for each splitting node.

\(^2\)For the smallest and largest level, the condition (1) is adapted to be larger than the post and pre statistic, respectively.
ter(s) and constant other parameters, (4) iterate (2) and (3) until the segmentation for the chosen parameter(s) does not change anymore. I did not pursue this approach in this dissertation due to potential difficulty involving iteration between (2) and (3). In the following, I will examine the “peak” approach to locating multiple changing levels by simulation.

3.4.1 Method

The simulation setup is the same as Simulation 2.1, the only difference is that instead of one changing level, there are two changing levels, set to be 3 and 6. The difference between each pre and at/post changing level (until the next changing level) is manipulated to the same degree: $0.5 \times d^3$ times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 4$.

3.4.2 Results

Results of the “peaks” approach for statistic $WDM_o$ (the same pattern for statistic $max LM_o$) are demonstrated in Figure 3.8. The left panel shows the empirical fluctuation process summary among 5000 replications when the violating parameter is $\alpha_3$ and the right panel shows similar process summary when the violating parameter is $\gamma$. Lines still represent 25% percentile, median and 75% quantile as Simulation 2.1. We can see there are two “peaks” appeared on the two correct changing levels. The percentage of correctly identifying two changing levels by the previously defined “peak” among 5000 replications are 0.32 and 0.45 for violating parameter $\alpha_3$ and $\gamma_3$.

\[^{3}\text{The changing degree can not be too large, otherwise the } p_{ij} \text{ will be transformed to the boundary 0, or 1.} \]
Figure 3.8: Simulation 2.2. Simulated empirical fluctuation process summary of \( \max WDMo \) 25% quantile, median and 75% quantile with two changing levels for violating parameter \( \alpha_3 \) (left) and \( \gamma_3 \) (right), \( n = 960 \), estimated by PML. Sub-panel labels denote the changing level and violation magnitude \( d = 4 \).
respectively. So it is still relatively easier to detect difficulty parameter but the percentage of identifying two changing level is much lower than identifying one changing level in Simulation 2.1 (The percentage was above 0.93).

In summary, the tests’ ability to identify multiple changing levels is problematic. The straightforward extension of one changing level detection method is not satisfying. Traditional model based recursive method loses the parameter information and adapted recursive method has difficulty in iterations. All the above methods are related to the CUSUM function. One alternative to cumulative sums is moving sums, whereby we use sums of first derivatives within a moving window of levels (Zeileis & Hornik, 2007). The window size can be defined as two levels (assuming each level contains equal sample size) when the auxiliary variable is ordinal. Then measurement invariance will be examined within the moving window (all possible adjacent two levels). Further details will be discussed in Chapter 5.

Taking all simulation studies together, I examined the tests’ properties to locate the violating parameter and the violating level. To identify the violating parameter, the key is to eliminate the potential impact of person hyperparameters by modeling. Meanwhile, the tests have adequate power to identify single changing points. For multiple changing points, further research is needed. In the next chapter, I will illustrate the tests’ locating properties in practice.
Chapter 4

Empirical Study

In this chapter, a real world data set is examined. The main goal is similar to the simulation studies: to locate parameters/levels that violate measurement invariance (if there are any).

In this data set, 2156 grade eight students responded to 18 dichotomously scored mathematics items taken from the graduation examination developed by the Netherlands National Institute for Educational Measurement. One student-intake characteristic was measured: socioeconomic status (SES). This characteristic only has 40 unique values in this sample. So SES is treated as an ordinal variable and divided into 6 levels to maintain roughly equal sample sizes in each level: no greater than 

\[-1, (-1, -0.5], (-0.5, 0], (0, 0.5], (0.5, 1], (1, 3].

This SES level variable can be treated as the ordinal auxiliary variable.

The correlation between SES and mathematics achievement (sum of the 18 items) is as high as 0.49. This positive relationship has been widely observed in education. There are two plausible reasons for this positive relationship: person ability and item
fairness. Person ability means the observed higher math score is a true reflection of higher ability of students with higher SES. There is nothing wrong with the tests’ items. Whereas item fairness means the tests’ one or multiple items have different difficulty/discrimination parameters for students with different SES. The observed higher math score for higher SES students is due to unfair items. Thus, educators should identify and replace the unfair items (DIF items).

As described in the Background section, DIF detection research has a long history in item response theory (IRT). The most often used approach is based on LRT as described in Chapter 2. The basic procedure is (1) fit a model with all parameters equal across groups, (2) test every parameter for DIF one by one, (3) If DIF is detected in any parameter(s), refit the model with the “worst” parameter freed and retest every parameter again until no parameter demonstrates DIF. There are several drawbacks of this LRT based approach. First, it is computationally demanding to fit $2^m$ models, where $m$ is the category of the ordinal auxiliary variable. Second, the LRT ignores the ordering information contained in the ordinal auxiliary variable. Third, sample size above 1000 is typical in educational research, but LRT is overly sensitive to large sample size as stated in Chapter 2.

The ordinal statistics proposed by Merkle et al. (2014) could overcome the above shortcomings and locate violating parameter and level in an efficient way, assuming that the estimating model fits the data and only one changing level exists, as discovered by simulation studies. In the analyses described below, I will analyze the data using the ordinal statistics in the same manner as the simulation studies. Firstly, I will analyze this data set by 2PL model. Second, as stated in Simulation 1.2, because we are not sure if person hyperparameters change, a model estimating $\mu_l$ and
σ_l is considered. It is plausible that some parameter(s) demonstrating measurement invariance violations are actually due to person hyperparameters’ difference.

4.1 Parameter/Level Violation Detection using 2PL Model

In this section, a 2PL model is fitted first and measurement invariance is examined for all discrimination and difficulty parameters w.r.t. SES level. The results are shown in Figure 4.1. The first and second row shows $WDM_o$ and $\max LM_o$ statistic for simultaneously testing 18 items’ discrimination and difficulty parameters, respectively. The discrepancies in the violating level for both discrimination parameters (first column) and difficulty parameters (second column) indicate multiple parameters violate measurement invariance. In particular, both statistics for difficulty parameters demonstrate violation in each level. Further examination reveals that both $WDM_o$ and $\max LM_o$ $p$ value for almost all items’ (except item 1 and item 13) difficulty parameters is smaller than 0.05/18 (divided by 18 for Bonferroni correction). According to what is stated in Simulation 1.2, it may be caused by the change in person hyperparameters. Therefore, I will examine parameter measurement invariance in an extended model where $\mu_l$ and $\sigma_l$ are estimated.
Figure 4.1: Empirical fluctuation processes of the $WDM_o$ statistic (first row) and \( \text{max } LM_o \) (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using 2PL model
4.2 Parameter/Level Violation Detection using Extended Model

In this section, an extended model which estimates $\mu_l$ and $\sigma_l$ is specified. The result is shown in Figure 4.2. The panel arrangements are the same as Figure 4.1.

From Figure 4.2, we can see none of the statistics demonstrate measurement invariance violations. Besides, students ability means are increasing with the increase of SES, in the way as 0 (reference level), 0.54, 1.01, 1.26, 1.58, 2.25. Meanwhile, variances for each SES level stay very similar to the reference, in the way as 1 (reference level), 1.13, 1.17, 1.11, 1.26, 1.56. To simplify the model further, I fix the variance of each level to be the same as the reference level and only estimate mean for each level, the result is shown in Figure 4.3. It can be seen that there is still no measurement invariance violation, with the corresponding $p$-value for each panel as 0.11, 0.297, 0.054, 0.209. The level ability mean $\mu_l$ still increases with SES level in the same manner, such as 0 (reference level), 0.5, 0.94, 1.16, 1.46, 2.08.

In summary, I found that the positive correlation between SES and math achievement is due to the fact that students’ ability increase with the increase of SES. All parameters appear to fulfill the measurement invariance assumption after we take account of level mean $\mu_l$ changes. The level variance $\sigma_l^2$ changes can be ignored for simplicity in this data set. There is no biased item based on the current level arrangement.$^1$

---

$^1$The conclusion would change if other level arrangements were adopted, such as using different number of levels or applying different cut points on the SES. Thus it is recommended to define ordinal auxiliary variable’s level with substantive meaning in practice instead of maintaining equal sample size for better power.
Figure 4.2: Fluctuation processes of the $WDM_o$ statistic (first row) and max $LM_o$ (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using extended model.
Figure 4.3: Fluctuation processes of the $WDM_o$ statistic (first row) and max $LM_o$ (second row) for discrimination parameters (left panel) and difficulty parameters (right panel), using model only estimating group mean.
Chapter 5

Summary and concluding remarks

In this dissertation, I examined the recently proposed ordinal score-based tests’ abilities to locate violating parameters/levels by using PML estimation on 2PL model and extended 2PL model. For the parameter locating property, the tests have adequate power to isolate the violating parameter correctly, given the estimating model matches the data generating model. For the level locating property, the test statistics’ corresponding level is consistent with the changing level with decent percentages in 5000 replications, given there is only one change point. Applying ordinal score-based tests on IRT models estimated by PML provides a more general and flexible framework to detect DIF in IRT research. In the remainder of the last Chapter, I consider the tests’ application in various social science models and their extension to more complex scenarios.
5.1 Application

The factor-analytic framework and PML can provide easier alternatives to traditional difficult questions arising from IRT, such as DIF detection under the circumstances of multiple category responses, multidimensional ability structure and placing person ability/item parameter estimates obtained from multiple tests on the same scale.

5.1.1 Multiple Category Responses

The PML/factor analysis framework allows us to use the score-based tests in situations when the responses have multiple categories. In this situation, we may elect to use a Graded Response Model (GRM) (Samejima, 1969) instead of the 2PL. Abou El-Komboz, Zeileis, and Strobl (2014) used model-based recursive partitioning to detect DIF in the GRM. The unique contribution of our approach is that we locate the violating parameter(s) and level instead of only detecting whether DIF exists. Moreover, because of the modeling flexibility of factor analysis, researchers can adapt the model to the requirements of the data set (various constraining parameter(s) or more complex models) without worrying about carrying out the test. The test statistics and $p$-value will generate directly after researchers fit the model.

5.1.2 Multidimensional IRT

Multidimensionality is one cause of DIF(Millsap, 2005). However, it is difficult to test this hypothesis in IRT models. The challenge is caused by the MML estimation as mentioned in Chapter 2. PML provides a simpler estimation solution and can be applied easily to multiple-factor models. An extension to multidimensional models is
the full structural equation model (SEM).

5.1.3 Full SEM Approach

In practice, we often need to transform person parameters so that ability estimates are equivalent across different scales. This is called equating in IRT literature. For example, we need to equate test takers’ ability between old version and new version of GRE score. One common way of equating is linear equating. Suppose there are two vectors of person ability estimates obtained from scale A and scale B as \( \eta_A \) and \( \eta_B \) respectively. Assuming their standardized score are equal in the population, then transforming scale B score to scale A score is shown in the following:

\[
\frac{\eta_B - \mu_B}{\sigma_B} = \frac{\eta_A - \mu_A}{\sigma_A}
\]

\[
\eta_B = \eta_B \left( \frac{\sigma_A}{\sigma_B} \right) - \left( \frac{\sigma_A}{\sigma_B} \right) \mu_B + \mu_A,
\]

where \( \mu_A \) and \( \sigma_A \) stands for the mean and standard deviation of the sum score for scale A. The same notation rule applies for scale B.

However, DIF existing will complicate equating. Suppose female and male differ in scale A but not in scale B. Then \( \mu_A, \sigma_A \) will break down to \( \mu_{AM}, \sigma_{AF} \) and \( \sigma_{AM} \). The question is we should equate female and male separately or the original equating relationship based on the whole sample still holds. Dorans (2004) dealt this question by introducing new statistics within the frame work of test characteristic curve.

But with full SEM approach, we can conveniently test the equating relationship as coefficient between construct ACT and construct SAT, gender will be the auxil-
iary variable. We can fit the SEM model and test against gender directly without introducing more statistics.

5.2 Comparison to Other Models

To my knowledge, this is the first time that score-based tests were utilized in an extended 2PL model which accounts for group mean and variance differences. This extension is practically useful because it is a very restrictive assumption that group means and variances are the same. If violated, group mean/variance changes will be attributed to item parameter DIF. This is referred to as “false alarm” in the literature (Woods, 2009; ?, ?; W.-C. Wang & Yeh, 2003; Fischer, 1995; Kopf, Zeileis, & Strobl, 2015). Researchers have been using “anchor items” to deal with this problem (see Woods, 2009, for a review). The motivation can be explained in the following way: consider the Rasch model, with $i$ as subjects and $j$ as items

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\logit(p_{ij}) = \gamma_j + \eta_i,$$  \hspace{1cm} (5.3)

where $\gamma_j$ and $\eta_i$ are effects of item and subject, respectively. Now define $g_i \in \{0, 1\}$, which indicates the group to which subject $i$ belongs (we are assuming two groups). Imagine ways that $g_i$ could enter the Rasch model.

First, we could have

$$\logit(p_{ij}) = \gamma_j + \eta_i + \beta g_i.$$  \hspace{1cm} (5.4)

\footnote{level is traditionally referred to as group to ignore order information. I follow the traditional terminology in this section to refer to previous work.}
This is like a difference in factor means from group to group. People with \( g_i = 1 \) will be consistently better/worse across all items, as compared to people with \( g_i = 0 \).

Second, we could also have

\[
\logit(p_{ij}) = \gamma_j + \eta_i + \beta g_i + \delta_{g_i j}, \tag{5.5}
\]

where, for identification, we might take \( \delta_{0j} = 0 \) for all \( j \). The interaction terms \( \delta_{1j} \) appear confounded with items and subjects: we don’t know whether the item parameter (\( \gamma_j \)) differs by group, or whether the group parameter (\( \beta \)) differs by item. To solve this problem, we need to do two things: (1) estimate group mean and variance; (2) assume that there are some items that are invariant across groups; e.g., that \( \delta_{1j} = 0 \) for some \( j \). Under these two assumptions, we can separately estimate \( \beta \) and the \( \delta_{1j} \) for the remaining items. The items for which \( \delta_{1j} \) are assumed to be zero are called anchor items. Thus, the motivation for introducing anchor items is to differentiate items. Without any other information, however, algorithms developed to select anchor items need to make decisions about group mean change (Kopf et al., 2015). This “looping” nature of this problem makes it difficult to deal with even for only two groups.

By contrast, the extended 2PL model naturally differentiate items by \( \alpha_j \). Besides, estimating group means as parameters directly without introducing the interaction term makes modeling not confounded. For example, the worst scenario for difficulty parameter \( \gamma_j \) is that all difficulty parameters differ for people in different level to the same degree \( d_l \). Then the model becomes:

\[
\logit(p_{ij}) = (\gamma_j + d_l + \alpha_j \mu_l) + \sigma_l \alpha_j \eta_i. \tag{5.6}
\]
This is not a problem for extended 2PL model, because $\alpha_j$ is generally not equal to 1, parameter specificity will be retained. However, when $\alpha_j = 1$ (Rasch Model), this is a fatal problem, since there is no way to differentiate $d_l$ and $\mu_l$. Then difficulty parameter difference for all items $d_l$ would be counted as people’s ability difference $\mu_l$. Although this model has been considered in the DIF literature (De Ayala, 2009) [for a review], the traditional Maximum Likelihood estimations employed still makes it imperative to use “anchor items” as described in the previous paragraph. Whereas using PML as illustrated in this dissertation, we do not have to select anchor items.

Moreover, it is also an important note that group variances $\sigma_l$ differ. As illustrated in Chapter 2, when there is a single group, the variance is made up by $\lambda_j^2 \eta_i + \psi_j$, where $\eta_i \sim N(0,1)$. However, when multiple group exists, one group is chosen as reference and that group’s variance is fixed to 1 for identification, instead of constraining $\eta_i$’s variance. The other groups’ variances are estimated but we do not know if the group variance difference is caused by $\eta_i$ or $\psi_j$. Depending on constraining $\psi_j$ to be equal across group or not, the parameterization is called “delta” or “theta” (Muthén & Asparouhov, 2002). The “delta” parameterization attributes group variance differences to $\sigma_l$, as we did under this parameterization (adopted in this research). The effect of $\sigma_l$ is similar to the effect of $\mu_l$: ignoring estimation of $\sigma_l$ will cause false alarms, and estimating it will lower the power. However, under the “theta” parameterization, group difference is attributed to $\psi_j$, so we can not detect changes in $\sigma_l$.

Since more parameters are estimated in extended 2PL model compared to the traditional Rasch model or 2PL model. Larger sample sizes are required. As illustrated in Simulation 1.2, estimating 7 $\mu_l$ and 7 $\sigma_l^2$ requires the sample size to be increased
to 10 times to get similar power curves as Simulation 1.1 for 2PL model. It limits the tests’ application to areas where sample sizes are relatively small, such as clinical psychology. However, with the easier availability of big data, the tests should have more application area.

5.3 Further Development

5.3.1 Multi-change point

In this dissertation, I explored the two change points situation by observing “peaks” in the empirical statistical process. However, this method did not work perfectly. The “peak” method works only when the changing points are not adjacent to each other. Besides, the power is about 50%, much lower than the single change point situation. This is because CUSUM is less sensitive to parameter variation when the number of residuals gets larger (Chu, Hornik, & Kaun., 1995).

Another approach, MOSUM, sums a fixed number of residuals that move across the whole sample. The limiting process for the empirical MOSUM processes are the increments of a Brownian bridge. The difficulty of this approach is to choose the number or residuals (width of window size). If the width of window size is close to the duration of parameter (change point 2 corresponding observation percentage minus change point 1 corresponding observation percentage), the power is higher than CUSUM based tests (Chu et al., 1995). For the ordinal auxiliary variable, we can choose the window size as twice of the sample size of each level and detect change points in each window (assuming equal sample size in each level).
5.3.2 Multilevel Model

In educational research settings, students' responses to items are often not really independent, because they are nested in the class (taught by the same teacher) and school. Therefore, multilevel models are generally applied in this area. Score-based tests only rely on taking the derivative of the maximum likelihood function of each individual. As long as the individual derivative (analytical or approximation) can be specified, score-based tests can readily detect parameter invariance violations in the multilevel model.

5.3.3 Longitudinal Studies

Another kind of dependent data set is longitudinal/time series multivariate data. There are two sources of change for the appearing difference in multiple time points. One is the “real change” which can be reflected by the regression coefficients; the other is measurement invariance violation (although the latter is rarely formally tested). Researchers often implicitly assume that administering the same instrument ensures the same attribute is being assessed (Pitts, West, & Tein, 1996). This lack of examination is partly due to the inconvenience of conducting multiple LRTs, along with the high rejection rate (measurement invariance violation) for large sample sizes. The score-based tests are convenient to use for most of the IRT models and extended ones. Besides, in longitudinal studies or time series data set, observations are seldom independent. Score-based tests require us to estimate the variances of the score function, which can be biased by the inclusion of lagged dependent variable. Empirical results showed that the crucial key for these structural change tests to work well is to get a consistent estimate of the covariance matrix of the score function (or equivalently,
the partial derivative of the case-wise likelihood contributions w.r.t. the parameters \( \theta \). Dependent observations could “contaminate” this covariance matrix estimation. This could cause the tests have non-monotonic power or completely drop to zero.

A natural approach to deal with problem is to find a Heteroskedasticity and auto correlation consistent (HAC) covariance matrix estimator. Under this motivation, several methods were proposed, such as kernel HAC estimators with automatic bandwidth selection (Andrews, 1993) and weighted empirical adaptive variance estimators. However, the usual bandwidth selection method is not reliable according to the review (Perron, 2006). Specifically, the tests are very sensitive to the bandwidth used. A large one leads to properly sized tests in finite samples but with low power, and a small bandwidth leads to better power but large size distortions. Generally, HAC estimation would produce upward bias for the data-dependent bandwidth, which will cause non-monotonic power issue. This non-monotonic power feature is not desirable. Self-normalization might be able to deal with problem (Shao & Zhang, 2010).

5.4 Summary

In this dissertation, I extended the score-based tests to IRT models estimated by PML. This extension has advantages over traditional DIF detection methods, and the implementation is simpler with only estimation of null model (that assumes measurement invariance). Applied researchers in Psychology and Education could use these tests to examine measurement invariance in their own data sets conveniently. If some parameter/level demonstrates violation of measurement invariance, that item can be removed or different models can be specified according to the changing level
of the auxiliary variable.
Appendix A

Simulation 1.1: Estimated by MML

This Appendix demonstrates Simulation 1.1 Results estimated by MML. Details see Chapter 2, section Simulation 1.1.

Figure A.1 and Figure A.2 display power differences among statistics. Figure A.3 and Figure A.4 display power differences among sample sizes.
Figure A.1: Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\alpha_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 
Figure A.2: Simulated power curves for max $LM_o$, $WDM_o$, and $LM_{uo}$ across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\gamma_3$. $n = 960$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 

Violation Magnitude ($\gamma_3$) 

<table>
<thead>
<tr>
<th>$\gamma_3$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>m=4</td>
<td></td>
</tr>
<tr>
<td>m=8</td>
<td></td>
</tr>
<tr>
<td>m=12</td>
<td></td>
</tr>
</tbody>
</table>

maxLM_o, WDM_o, LM_uo
Figure A.3: Simulated power curves for observation 120, 480 and 960 of test statistic $WDM_\alpha$, across three levels of the ordinal variable $m$ and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), estimated by MML. The parameter violating measurement invariance is $\alpha_3$. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable $m$. 
Figure A.4: Simulated power curves for observation 120, 480 and 960 of test statistic \( WDM_o \), across three levels of the ordinal variable \( m \) and measurement invariance violations of 0–4 standard errors (scaled by \( \sqrt{n} \)), estimated by MML. The parameter violating measurement invariance is \( \gamma_3 \). Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable \( m \).
References


coefficient. Psychometrika, 44(4), 443–460.


VITA

I was born in Xi’an, Shaanxi Province, China. My undergraduate major was Biology (China Agriculture University), then I studied Applied Psychology in Peking University. In 2010, I went to University of Missouri to study Quantitative Psychology, along with Statistics.

After graduation, I will be a postdoc fellow in Psychology Department at Ohio State University. After this training, I plan to apply for a faculty position in quantitative Psychology/Education area at a research university.