Title
Detection of Differential Item Functioning in the Generalized Full-Information Item Bifactor Analysis Model

Author
Somerville, Jason Taro

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Detection of Differential Item Functioning in the Generalized Full-Information Item Bifactor Analysis Model

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Statistics

by

Jason Taro Somerville

2012

ABSTRACT OF THE DISSERTATION

Detection of Differential Item Functioning in the Generalized Full-Information Item Bifactor Analysis Model

by

Jason Taro Somerville

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2012

Professor Jan de Leeuw, Chair

In the field of psychometrics, there has been an increase in interest concerning the evaluation of fairness in standardized tests for all groups of participants. One possible feature of standardized tests is a group of testlets that may or may not contain differential item functioning (DIF) favorable to one group of participants over another. A testlet is a cluster of items that share a common stimulus. In this dissertation, a DIF detection method useful for testlet based data was developed and tested for accuracy and efficiency. The proposed model is an extension of the generalized full-information item bifactor analysis model. Unlike other IRT-based DIF detection models, the proposed model is capable of evaluating locally dependent test items and their potential impact on the DIF estimates. This assures the new capability of the bifactor DIF detection method that was not evident in previous methods. Item parameters were estimated using a maximum likelihood estimation (MLE) method producing expected a posteriori (EAP)

scores. Using the restrictions of a bifactor model, the dimensionality of integration can be analytically reduced and the efficiency can be increased. Following prior research regarding DIF on a PISA dataset, the proposed DIF model was applied to mathematics items of the Program for International Student Assessment (PISA) 2009 dataset to confirm the utility of the model. After the meaning of results to the PISA research community is conveyed, a simulation study was conducted to provide concrete evidence of the model's utility. Finally, limitations of this study from computational and practical standpoints were discussed, as well as directions for further research.

The dissertation of Jason Taro Somerville is approved.

Li Cai

Peter Bentler

Frederic Paik Schoenberg

Jan de Leeuw, Committee Chair

University of California, Los Angeles

2012

To my father Paul,

my mother Noriko,

and my brother Mike,

for their unwavering support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to thank the dissertation committee members: Dr. Peter Bentler, Dr. Li Cai, Dr. Jan de Leeuw, and Dr. Frederic Paik Schoenberg. Each of these professors provided me with valuable education in applied statistics during graduate-level classes. It goes without saying that they have played important roles concerning the backbone of my research. From the initial stages to the final completion of this dissertation project, I have been blessed with their sympathy and guidance.

Second of all, I would like to thank the staff members of the Department of Statistics: Glenda Jones, Jason Mesa, and Verghese Nanghera. These staff members were helpful in organizing the administrative matters such as paperwork and equipment necessary to fulfill the requirements for a Degree in Philosophy. It has been a pleasure to obtain support from a diligent group of staff members.

Third of all, I would like to thank all of the faculty members, staff members, and fellow graduate students in the Department of Statistics. They all gave me words of encouragement and statistical knowledge which contributed to the dissertation's completion. It was an honor to be a part of a department in which everybody was enthusiastic and intelligent.

Finally, I would like to thank my family, relatives, and friends. My parents were supportive in my pursuit of the PhD degree, in both emotional and financial aspects. My brother kept me optimistic with his positive personality. My relatives and friends around the world have done the same by keeping in touch with me and my family. It has been a joy to be connected with many people showing warm support.

VITA

2007	B.S., Statistics University of California, Santa Barbara Goleta, California
2008-Present	Teaching Assistant University of California, Los Angeles Los Angeles, California
2011	C.Phil., Statistics University of California, Los Angeles Los Angeles, California

HONORS AND AWARDS

Winter 2005, Fall 2005, Spring 2006	Dean's List University of California, Santa Barbara Goleta, California
2007-2008	Fellowship University of California, Los Angeles Los Angeles, California

CHAPTER 1

Introduction

Over the past few decades, there has been increased curiosity concerning how test scores are affected by the differences in the participants' abilities. In the field of psychometrics, as researchers fit complicated models to item response data, they attempt to meticulously evaluate the test items' consistency among the subgroups into which the participants could be partitioned. Detecting a test's bias, or lack thereof, can be valuable information for test developers who strive to design examinations capable of reliably measuring the academic ability of all students, regardless of the students' background. The purpose of this study is to develop and evaluate a new DIF detection model for test scores measured by the generalized full-information item bifactor analysis model.

This dissertation is a marriage of the bifactor model and the detection of differential item functioning (DIF). The former is a recently developed item response theory (IRT) model which assumes local dependence among items with provisions for additional random effects; the latter is an observation of a phenomenon occurring in some questionnaires or tests showing inconsistency. Despite the evidences of the bifactor model providing better fit to data than conventional models, there is still necessity of searching sources of variation in responses among subgroups. Hence, the combination of the latest model and bias detection can make helpful statements regarding the overall quality of tests in the field of education.

In the first part of this chapter, common concepts and key terms regarding DIF used in this study are introduced. Then, in the second part, the characteristics and benefits of the two-tier model are described in detail. Finally, the purpose and significance of this study with regards to real life situations are declared.

1.1 Differential Item Functioning

The fairness of the test items is an important aspect of the test to measure in order to evaluate its consistency. One of the most common approaches of doing so is an analysis of DIF, a natural phenomenon that can occasionally be observed when dealing with item response data. DIF shows an indication of diverse behavior by an item, showing evidence that people from different groups (such as genders and ethnicities) with the same latent trait have a different probability of answering the item correctly. The focal group is the group of participants on which the research is focused, while the reference group is a group to which the focal group is compared. If the items do not show significant DIF, they are invariant across groups of participants, which is an ideal situation since a test with no sign of bias is desirable in the field of psychometrics. Every student with the same ability should have an equal opportunity to choose the correct response to each item on a test.

The competence of one group of participants can be misrepresented by DIF present in an item. Here is one concept that explains the definition of DIF: "An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (Scheuneman, 1975). Another explanation is the following: "If each test item in a test had exactly the same item response function in every group, then people of the same ability or skill

would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased" (Lord, 1980). The goal is to confirm that the meaning which items attribute to the test is the same for all groups (Shepard, 1982).

DIF indicates that performance is different among subgroups; it is a slightly different concept from impact and bias. From the IRT point of view, DIF is present when different item characteristic curves (ICCs) can be identified from the different subgroups (Narayanan and Swaminathan, 1996). On the other hand, impact measures the difference in performance on an item between two groups, which comes from the difference in average ability of the groups (Dorans and Holland, 1993). Therefore, the impact measures how the difference in performance is affected by difference in ability; whereas DIF measures how the difference in performance emerges despite the lack of difference in ability. Also, item bias is different from DIF. If an item shows DIF, experts can implement an evaluation to see, in the social or content aspect, if the item favors one group over the other (Angoff, 1993). DIF is a condition that needs to be present for a biased item; however, the converse is not true.

The two major types of DIF are uniform DIF and non-uniform DIF. Uniform DIF occurs when one group performs uniformly better than the other group regardless of the ability level, which means there is no interaction between the ability level and the groups (Narayanan and Swaminathan, 1996). One the other hand, non-uniform DIF occurs when the magnitude and direction of the differential performance between two groups varies across the spectrum of the ability level. This phenomenon can be observed if the two ICCs intersect at some ability level. Some DIF detection methods are only capable of detecting uniform DIF, while other methods can find both uniform DIF and non-uniform DIF.

In order to execute the DIF evaluation smoothly, a "matching criterion" is essential for controlling the difference in ability between a focal group and a reference group. Matching criteria can be total scores on a test (as utilized in classical test theory), or they can be ability estimates (as utilized in IRT). Items exhibiting DIF would not be beneficial for the process of creating a medium for fair evaluation; therefore it is desirable to determine a matching criterion by using only DIF free items. An ideal matching criterion would be one that uses only DIF free items with help from a purification procedure (Dorans and Holland, 1993) that involves two steps. In the first step, known as the criterion refinement or purification step, items on the matching variable are analyzed for DIF, and any items that exhibit sizeable DIF are removed regardless of the DIF having a positive or negative sign. In the second step, the refined criterion is used for another DIF analysis of the same items and any other items excluded in the criterion refinement step.

1.2 Generalized Full-Information Item Bifactor Analysis Model

The generalized full-information item bifactor analysis model is a confirmatory item factor model which has features that are useful in psychometric research (Cai, Yang, and Hansen, 2011). One of the features is the flexibility while handling residual dependence of item responses without using copula functions (Braeken, Tuerlinckx, and De Boeck, 2007). Another feature is the maximum likelihood estimation, with proven accuracy and efficiency from demonstrations using data. Dimension reduction, yet another feature, can reduce the dimensionality of the latent variable, which gives the model a computational advantage over other models.

First, one must examine how the bifactor model makes the assumption allowing local dependence to be present among items. Let us consider a bifactor pattern for six items:

$$\begin{pmatrix} a_{10} & a_{11} & 0 & 0 \\ a_{20} & a_{21} & 0 & 0 \\ a_{30} & 0 & a_{32} & 0 \\ a_{40} & 0 & a_{42} & 0 \\ a_{50} & 0 & 0 & a_{53} \\ a_{60} & 0 & 0 & a_{63} \end{pmatrix}$$

in which the a's represent non-zero item slopes for the latent variable **θ**. The first column of item slopes is slope parameters for the primary dimension θ_0. All items load on the primary factor because it is constrained by the bifactor model. This constraint is effective for datasets in which items are intended to measure the same construct for all participants. The next three columns of item slopes are slope parameters for the specific dimensions θ_1, θ_2, and θ_3. The bifactor model permits items to load on at most one of the specific dimensions, and thus deep analysis can be completed by examining the primary dimension and the specific dimension.

The remaining two features – maximum likelihood estimation and dimension reduction - work well when used together. Its accomplishment is to estimate the item parameters accurately. The first step in doing so is to find the marginal distribution of the item responses. We first multiply the conditional probability of observed response **y** given the latent variables **θ** with the distribution of **θ** to obtain the joint distribution. Next, we integrate **θ** out of the joint distribution, and the end result is the marginal of **y**. As shown by Cai, Yang, and Hansen (2011), the marginal distribution of the item responses is expressed as follows:

$$f_a(\mathbf{y}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_a(y|\theta_0, \theta_1, \theta_2, \theta_3) f(\theta_0, \theta_1, \theta_2, \theta_3) d\theta_0 d\theta_1 d\theta_2 d\theta_3$$

where the subscript **a** is added to emphasize the fact that the unknown parameters depend on the conditional and marginal distributions on the unknown parameters.

Referring to the example illustrated with six items, all items depend on θ_0. However, the first two items are the only ones that depend on θ_1, the third item and the fourth item are the only ones that depend on θ_2, and the last two items are the only ones that depend on θ_3. For the 6 × 1 vector of item responses $\mathbf{y} = (y_1, \ldots, y_6)^t$, conditional independence implies

$$f_a(y|\theta_0, \theta_1, \theta_2, \theta_3) = f_a(y_1, y_2|\theta_0, \theta_1) f_a(y_3, y_4|\theta_0, \theta_2) f_a(y_5, y_6|\theta_0, \theta_3).$$

Hence, while finding the marginal distribution of **y**, one can integrate the two specific dimensions out of the joint distribution first, and then integrate θ_0 out so that marginal distribution of **y** is expressed as the following iterated integral:

$$f_a(\mathbf{y}) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f_a(y_1, y_2|\theta_0, \theta_1) f(\theta_1) d\theta_1\right] \times \left[\int_{-\infty}^{+\infty} f_a(y_3, y_4|\theta_0, \theta_2) f(\theta_2) d\theta_2\right] \times \left[\int_{-\infty}^{+\infty} f_a(y_5, y_6|\theta_0, \theta_3) f(\theta_3) d\theta_3\right] f(\theta_0) d\theta_0.$$

As a result of integration over the specific factors, the terms in the square brackets only depend on θ_0, which is integrated out in the final step. The marginal distribution can be approximated using quadrature approximation as

$$f_a(y) \doteq \sum_{q_0=1}^{Q} \left[\sum_{q_1=1}^{Q} f_a(y_1, y_2|X_{q_0}, X_{q_1}) W_{q_1}\right] \times \left[\sum_{q_2=1}^{Q} f_a(y_3, y_4|X_{q_0}, X_{q_2}) W_{q_2}\right] \times \left[\sum_{q_3=1}^{Q} f_a(y_5, y_6|X_{q_0}, X_{q_3}) W_{q_3}\right] W_{q_0},$$

where the integrand is evaluated at a set of Q discrete quadrature nodes (the X_q's) for each latent factor, with weights at each node equal to W_q. In this dissertation, rectangular quadrature is used, in which Q = 20 quadrature points equally spaced between -5 and 5 are necessary. The weights

are ordinates of normal densities evaluated at the corresponding quadrature nodes in the rectangular quadrature.

Rijmen, Vansteelandt, and De Boeck (2008) demonstrated that dimension reduction is efficient due to the conditional independence relations of the model being exploited during the E-step. These stem from a graphical model theory which can be described using a junction tree from the acyclic graph. In this theory, the structure of a transformed graph provides a factorization of the joint probability function of the manifest and latent variables, which are represented by nodes of the junction tree with compact clique state spaces. The goal of assessing the test fairness is achieved with fewer obstacles.

In the subsequent sections, some IRT models using the constraints of the bifactor model will be introduced. The models are a classical model for dichotomous response, a model for graded response, and a general partial credit model. In this dissertation, the IRT models being used are the first two models because the PISA 2009 data has binary items (correct/incorrect) which can be modeled by the first model and polytomous items (3 different answer choices) which can be modeled by the second model. Partial credit is not assumed is this data analysis.

While explaining the notation, the *i* subscript (for person *i*) and the *j* subscript (for item *j*) are dropped temporarily to avoid notational clutter. A generic item *j* is assumed to be scored in *K* categories. It is also assumed that each item loads on specific factor *s* in addition to the primary dimensions.

1.2.1 Classical Model for Dichotomous Response

This model is an extension of the 3-parameter logistic model (Reckase, 2009). The general dimension θ_0 and specific dimension θ_s determine the likelihood of the response being correct. Let the conditional probability of a correct response in this framework be

$$P(y = 1|\theta_0, \theta_s) = c + \frac{1 - c}{1 + exp\{-[d + a_0\theta_0 + a_s\theta_s]\}}$$

where *c* represents the guessing probability, *d* the intercept of the particular item, a_0 the slope with respect to the primary dimension, and a_s the slope with respect to specific dimension *s*. Due to the property of the items with dichotomous response, conditional probability of an incorrect response is the complement of the conditional probability of a correct response, such that the sum of probabilities equals 1. In this dissertation's data analysis, the guessing probability is set equal to zero since PISA is an educational assessment that does not involve high stakes, and therefore the phenomenon of guessing is not considered. If the item does not load on any specific dimension, then the term $a_s\theta_s$ disappears because the conditional probability solely depends on the primary dimension, as demonstrated in the following model modified for items loading solely on the primary dimension:

$$P(y = 1|\theta_0) = c + \frac{1-c}{1+exp\{-[d+a_0\theta_0]\}}.$$

1.2.2 A Model for Graded Response

The graded response model (Samejima, 1969) can be extended to cover the bifactor case. This model is similar to the model described by Muraki and Carlson (1995). Let *K* represent the number of graded categories in which responses are categorized and let $y \in \{0, 1, \ldots, K - 1\}$ represent the item response that belongs in one of the graded categories. Let the cumulative category response probabilities be

$$P(y \geq 1|\theta_0, \theta_s) = \frac{1}{1 + \exp\{-[d_1 + a_0\theta_0 + a_s\theta_s]\}},$$

$$\vdots$$

$$P(y \geq K - 1|\theta_0, \theta_s) = \frac{1}{1 + \exp\{-[d_{K-1} + a_0\theta_0 + a_s\theta_s]\}},$$

where $d_{1, \ldots,} d_{K-1}$ are a intercepts corresponding to each category, a_0 is the item slope on the primary dimension, and a_s is the item slope on the specific dimension *s*. The category response probability for category *k* is simply calculated by finding the difference between two neighboring cumulative probabilities

$$P(y = k|\theta_0, \theta_s) = P(y \geq k|\theta_0, \theta_s) - P(y \geq k + 1|\theta_0, \theta_s),$$

where the probability of the response belonging to the lowest category is defined by $P(y = 0|\theta_0, \theta_s) = 1 - P(y \geq 1|\theta_0, \theta_s)$ and the probability of the response belonging to the highest category is defined by $P(y = K - 1|\theta_0, \theta_s) = P(y \geq K - 1|\theta_0, \theta_s)$.

The graded response model utilizes a latent variable in a regression model that has the same structure as the proportional odds model. The proportional odds model is a regression model

used for ordinal data. It is an extension of the logistic regression model for dichotomous variables, allowing for more than two response categories. The proportional odds model, estimated using the maximum likelihood estimation method, applies to data meeting the proportional odds assumption which assumes any two pairs of outcome groups have the same relationship. The latent variable is treated as the unobserved random variable in a logistic regression model. An alternative nomenclature of the proportional odds assumption is the parallel regression assumption. Since any pair of outcome groups has the same relationship, only one set of coefficients is required, which explains the origin of the nomenclature. The proportional odds model uses the following logit function as described in Moustaki (2000):

$$\ln\left[\frac{\gamma_s(\theta)}{1-\gamma_s(\theta)}\right] = -(d_s + a\theta)$$

where s is the index of the category, and γ_s is the probability of a response in category s or lower. As documented in Mignani et al. (2005), the graded response model and the proportional odds model are similar to each other. In both of those models, the higher the value of the latent variable, the higher the probability of the corresponding response belonging in the highest category. The graded response model will be applied to the polytomous items in this dissertation's data analysis.

1.2.3 A General Partial Credit Model

The unidimensional generalized partial credit model (Muraki, 1992) can be extended to cover the bifactor case as well. Let K represent the number of ordinal categories in which responses are categorized and let $y \in \{0, 1, \ldots, K-1\}$ represent the item response that belongs in one of the ordinal categories. Adapting the notation in Thissen, Cai, and Bock (2010), let the conditional response probability for ordinal category $k = 0, \ldots, K-1$ be

$$P(y = k|\theta_0, \theta_s) = \frac{\exp\{T_k[a_0\theta_0 + a_s\theta_s] + d_k\}}{\sum_{l=0}^{K-1} exp\{T_l[a_0\theta_0 + a_s\theta_s] + d_l\}}$$

where T_k is the scoring function unique to its ordinal category k and d_k is the category intercept unique to its ordinal category k. The definitions of a_0 and a_s are identical to those in the previous models; the former is defined as the item slope of the general dimension and the latter is defined as the item slope of the specific dimension s. If the category intercepts are to be estimated from defining a multinomial logit, an identification restriction such that one of the d_k values equals zero is necessary. However, in this dissertation's data analysis, partial credit is not assumed to exist.

1.3 Purpose and Significance of Study

The objective of this study is to propose and examine a new statistical model to detect DIF for items fit to a bifactor model. Such DIF detection would be necessary since the bifactor model is accurate, efficient, and broadly applicable. Previous DIF detection extended to less flexible IRT models does not take the local dependencies of items into account, which causes item parameters and DIF magnitude to be estimated with bias. A biased estimate of DIF magnitude will be

misleading because of "artificial DIF," a by-product of inaccurate parameter estimates due to use of inappropriate IRT models (Angoff, 1993). Therefore, appropriate IRT models are essential for estimating item parameters and DIF more accurately. For this reason, this study proposes a new DIF detection model based on the bifactor model that has reliable accuracy, high efficiency, and a broad range of applicability.

In this study, marginal maximum likelihood estimation will be utilized to estimate parameters. Using the two-tier model restrictions, the dimensionality of integration can be reduced from the number of total dimensions to two.

A DIF detection model using the bifactor model needs to be developed since the bifactor model is versatile. The bifactor model can unify the dichotomous response model, the graded response model, and the general partial credit model in a single modeling framework (Cai, Yang, and Hansen, 2011). In addition to its versatility, the bifactor model has flexibility of allowing local item dependence (LID). Compared to a DIF detection model requiring items to be conditionally independent, the proposed DIF detection model is anticipated to estimate DIF magnitude more accurately and provide precise information about DIF magnitudes for studied items. Moreover, the maximum marginal likelihood estimation for this model only requires a two-dimensional integral. Therefore, there is no harm done in fitting the test items to a bifactor model.

For not only the statistics community, but for all stakeholders, this study is interesting and significant. This research is using one of the latest models developed in IRT, which has applications in the social sciences field. The results will be important because the expert in IRT will regard this research as an efficient way to see how two groups of participants can have

varied response patterns. Noticing the two groups' different response patterns to a test will lead to creating tests that contain items that are comprehensible for participants of both groups. All stakeholders are hoping to see fairness in test items.

CHAPTER 2

Literature Review

In this chapter, various DIF detection methods – some based on IRT – will be introduced. Advances in research and increase in knowledge can be observed while following the history of the procedures in which DIF can be detected. The history starts with the non-IRT-based DIF detection methods to locate the origins of DIF, followed by the IRT-based DIF detection methods to see how IRT plays a role on finding DIF. Lastly, the reasons regarding the choice of Wald test for the usage in this dissertation's data analysis will be explained.

2.1 Non-IRT Based DIF Detection Methods

A recollection of the earliest methods that are utilized to detect DIF between two groups in a sample would reveal the fundamentals of DIF detection. These methods are reliable, despite the fact that the latent variable is not utilized to estimate the probability of a correct response for either the focal group or the reference group.

2.1.1 Standardized P-Difference

The standardized P-difference (STD P-DIF) is the weighted mean difference between the proportions correct for the focal group and those for the reference group. It has been proposed as a DIF index by Dorans and Kulick (1986) and calculated using the following formula:

$$STD\ P\text{-}DIF = \sum_k n_{Fk}(p_{Fk} - p_{Rk}) / \sum_k n_{Fk}$$

where *n* is the number of examinees in a category, *p* is the proportion of examinees getting an item correct, subscript *k* is the score level in the matching criterion, subscript *R* indicates the statistic is for the reference group, and subscript *F* indicates the statistic is for the focal group. This measure is only sufficient under the Rasch model. Also, since STD P-DIF measures differences in difficulties between groups, it is only capable of measuring uniform DIF.

2.1.2 Mantel-Haenszel Method

The Mantel-Haenszel (M-H) method considers the difference in odds of having the correct response to an item between the focal group and the reference group. The statistic is

$$\alpha_{MH} = \frac{\sum_{k=1}^{s} n_{1rk} n_{0fk} / n_k}{\sum_{k=1}^{s} n_{0rk} n_{1fk} / n_k}$$

in which *s* is the number of score levels, n_k is the total number of people at score level *k*, n_{1rk} and n_{0rk} are the number of reference group people having a correct response and the number of reference group people having an incorrect response respectively, and n_{1fk} and n_{0fk} are the number of focal group people having a correct response and the number of focal group people having an incorrect response respectively. If the statistic is greater than 1, the odds imply that the item favors the focal group. A chi-square statistic can also be found:

$$\chi^2_{MH} = \frac{\left(\sum_{k=1}^{s} n_{1fk} - \sum_{k=1}^{s} E n_{1fk}\right)^2}{\sum_{k=1}^{s} Var(n_{1fk})}$$

where $E n_{1fk} = \frac{n_{1k} n_{fk}}{n_k}$ and $\mathrm{Var}(n_{1fk}) = \frac{n_{1k} n_{0k} n_{rk} n_{fk}}{n_k^2 (n_k - 1)}$

As explained by Zwick (1990), this measure works well in case all items are Rasch items and the investigated item is the only biased item; hence, it has similar limitations as the STD P-DIF.

2.1.3 Logistic Regression Method

The logistic regression method (Swaminathan and Rogers, 1990) has a clear advantage over the M-H method in terms of DIF detection because it is able to capture both uniform DIF and non-uniform DIF. It uses a matching criterion, a group indicator, and their interaction as predictors to estimate the logit of correct response. Another advantage of the logistic regression method over the M-H method is the usage of the ability variable to accurately assess every participant's ability to answer test items correctly. The model is as follows:

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_1\theta_j + \beta_2 G_j + \beta_3\left(\theta_j G_j\right)$$

in which P_{ij} is the probability of getting a correct answer on item i for participant j. β_0 is an intercept, β_1 is a regression coefficient for an ability of interest (θ_j), β_2 is a coefficient for grouping variable (G_j), and β_3 is a coefficient for an interaction effect between the ability and the grouping variable. The latent variable θ_j is measured on the same scale for both participants in the reference group and participants in the focal group. The grouping variable G_j is set to equal 0 for the reference group and 1 for the focal group. Comparing model fits between a model with and without the interaction coefficient will determine whether the item has non-uniform DIF or not. If β_3 is statistically significant, then we conclude that the item has non-uniform DIF. If β_2 is not statistically significant but β_3 is statistically significant, then we conclude that the item has non-uniform DIF. Given the item has non-uniform DIF, comparing model fits between a model with and without the grouping coefficient will determine whether the item has uniform DIF or not. If β_2 is statistically significant, then we conclude that the item has uniform DIF.

2.2 IRT-Based DIF Detection Methods

In this section we will discuss the DIF detection methods that utilize IRT in order to measure the performance difference between the reference group and the focal group, if the difference exists. The assumption made in these methods is that the probability of answering an item correctly can be predicted using a function of the latent variable.

2.2.1 SIBTEST Method

The Simultaneous Item Bias Test (SIBTEST) method (Shealy and Stout, 1993) is a non-parametric approach to DIF detection. SIBTEST is exclusive among other DIF detection methods since it is capable of detecting DIF in several items simultaneously as a unit. Initially, SIBTEST distinguishes between items measuring the ability (valid subtest) from items designated as DIF items (studied subtest). Then, it examines the difference in performance on an item in the studied subset, given the same level of ability determined by the valid subset. The amount of uniform DIF estimated by SIBTEST is as follows:

$$\hat{\beta}_U = \sum_{k=0}^{N} p_k (\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*)$$

in which p_k is the proportion of the focal group with score *k* on the valid subset, *N* is the maximum score of the valid subtest, $\bar{Y}_{Rk}^*$ is the adjusted mean score on the studied subtest for the reference group, and $\bar{Y}_{Fk}^*$ is the same statistic for the focal group.

Similarly, the amount of non-uniform DIF is as follows:

$$\hat{\beta}_C = \sum_{k=0}^{l-1} p_k(\bar{Y}_{Fk}^* - \bar{Y}_{Rk}^*) + \sum_{k=l+1}^{N} p_k(\bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*)$$

in which *l* is the valid subtest score at which tendency for a studied item or a studied subtest to favor one group switches toward the other group. As for classification of the DIF amount, if the absolute value of either a uniform DIF or a non-uniform DIF is greater than 0.088 the DIF is large; if the absolute value of either a uniform DIF or a non-uniform DIF is less than 0.059 the DIF is negligible (Roussos and Stout, 1996).

2.2.2 MIMIC Method

The Multiple Indicators Multiple Causes (MIMIC) confirmatory factor analysis model takes the following form:

$$y_i^* = \lambda_i \eta + \varepsilon_i$$

in which y_i^*= latent response variable *i* (when $y_i^* > \tau_i$, an observed variable, $y_i = 1$; τ_i, the threshold parameter, is related to item difficulty), η = latent trait, λ_i = factor loading for variable *i*, and ε_i = random error.

The factor analytic model for dichotomous item response data is equivalent to the normal ogive model which is described in detail by McDonald (1967) and Lord and Novick (1968). Muthen et al. (1991) extended this work for the MIMIC model, showing that the discrimination parameter, *a*, can be obtained using the value of λ from the MIMIC CFA model:

$$a_i = \frac{\lambda_i}{\sqrt{(1 - \lambda_i^2)\sigma_{\eta\eta}}}$$

where $\sigma_{\eta\eta}$ is the variance of the latent trait. Using the values of λ and τ, the difficulty parameter, *b*, can be obtained as follows:

$$b_i = \left[(\tau_i - \beta_i z_k)\lambda_i^{-1} - \mu_\eta\right]\sigma_{\eta\eta}^{-1/2}$$

where z_k is the group indicator with value 1 for focal group membership and value 0 for reference group membership, β_i is the measure of relationship between group and item response,

and μ_η is the mean of the latent trait. The relationship between the two models is further explained in Takane and de Leeuw (1987) and MacIntosh and Hashim (2003).

Capable of utilizing the anchoring method, the Multiple Indicators MIMIC method is one of the cutting-edge techniques for detecting DIF. Finch (2005) evaluated the MIMIC method and compared the performance with the Mantel-Haenszel method (Holland and Thayer, 1988), the Item Response Theory Likelihood Ratio Test (IRT-LRT) method (Wang and Yeh, 2003), and the Simultaneous Item Bias Test (SIBTEST) method (Shealy and Stout, 1993). The comparison established that the MIMIC method had the smallest Type I error rate and the highest power of DIF detection. Let us combine this method with anchoring to maximize the accuracy and minimize the error during item parameterization.

The MIMIC method with a pure short anchor (Shih and Wang, 2009) is a form of confirmatory factor analysis that can be extended to IRT models – such as the two-tier model - with external variables. Establishing a pure short anchor can eliminate the impact of DIF contamination because it is easier to locate one DIF-free item rather than many DIF-free items. In order to detect DIF effectively, a DIF-free-then-DIF Strategy (Wang, Shih, and Su 2007) is incorporated. Here is the algorithm:

1. Set Item 1 as the anchor and assess all other items for DIF. Using the MIMIC method with a 1-item anchor, obtain a DIF index for each studied item.

2. Set the next item as the anchor and assess all other items in test for DIF. Using the MIMIC method with a 1-item anchor, obtain a DIF index for each studied item.

3. Repeat Step 2 until the last item is set as the anchor.

4. Compute the mean absolute values of DIF indices for each item over all iterations and locate the desired number of items that have the smallest mean absolute values.

In order to evaluate the accuracy of the iterative MIMIC process (M-IT), the accuracy rate is compared with the chance level (probability of selecting the right items by chance). Usually, a high accuracy rate is required. After seeing the results, M-IT turns out to be successful in locating a set of up to four DIF-free items, since the accuracy rates are higher than the chance levels, even for small sample sizes. Overall, M-IT using a pure anchor yields a perfect rate of accuracy under most conditions while controlling the rate of Type 1 error well even with 40% DIF items. The efficiency of the method is also proven, since it can finish everything in a feasible computational time of 15 minutes, using the estimator of robust weighted least squares. However, it is desirable to use a different method using a full-information estimator for more accurate results.

2.2.3 IRT-LRDIF Method

The Item Response Theory Likelihood Ratio Test (IRT-LRT) is developed by Thissen, Steinberg, and Gerrard (1986) and further expanded by Thissen, Steinberg, and Wainer (1988). It allows for the comparison of model fit between the compact model and the augmented model, testing the null hypothesis that a vector of parameters equals zero. Although the item difficulty parameters are allowed to be different across the groups, the parameter estimates for a set of anchor items are constrained to be equal for both groups. The likelihood ratio statistic equals

$$LR = -2\ln\frac{L_0}{L_1}$$

in which L_0 is the likelihood for a model (M_0) in which item parameters for both groups are constrained equal and L_1 is the likelihood of a model (M_1) in which parameters take on different values in the two groups. Under the null hypothesis in which different groups have equal item difficulty parameters, *LR* is approximately χ^2 distributed with degrees of freedom equal to the number of parameters taking different values when comparing M_0 with M_1. If the χ^2 is statistically significant, then the model M_1 fits statistically better, indicating an item containing uniform DIF. Similarly, after allowing the item discrimination parameter to be different instead of the item difficulty parameter, existence of non-uniform DIF can be identified if the χ^2 in this scenario is statistically significant. In order to establish consistency of measurement among the focal group and reference group, parameter estimates need to be placed on the same scale. This common scale can be established by linking with anchor items (Millsap and Everson, 1993). Identifying a set of DIF-free anchor items is an important task with some challenges.

2.2.4 Wald Test

The Wald Test is a statistical test developed by Abraham Wald, which enables testing of the true value of the parameter based on the sample estimate. The value of the sample's parameter is compared to the hypothesized value. The null hypothesis is that there is no difference in parameters. If the Wald test is used to detect DIF, the parameter is the inter-group difference in item parameters. The Wald Statistic, which can use a full-information estimator, equals

$$\boldsymbol{p}'\boldsymbol{\Sigma}_{\boldsymbol{p}}^{-\mathbf{1}}\boldsymbol{p}$$

in which $\boldsymbol{p}$ is the vector of parameters and Σ_p is the covariance matrix (the variance and covariance of the aforementioned parameters). In this dissertation, the parameters are the difference in primary slope between groups and the difference in intercept between groups. Therefore, the corresponding Wald statistic equals

$$(a_{F0}-a_{R0} \quad d_F-d_R)'\begin{bmatrix} Var(a_{F0}-a_{R0}) & Cov(a_{F0}-a_{R0},d_F-d_R) \\ Cov(d_F-d_R,a_{F0}-a_{R0}) & Var(d_F-d_R) \end{bmatrix}^{-1}(a_{F0}-a_{R0} \quad d_F-d_R)$$

in which a_{F0} is the primary slope of the focal group, a_{R0} is the primary slope of the reference group, d_F is the intercept of the focal group, and d_R is the intercept of the reference group. This statistic is compared to a chi-squared value corresponding to the dimensionality of parameters (Harrell, 2001). This is not restricted to two-group comparisons. Such a comparison is also seen in the likelihood-ratio statistic, which is evidence that these two tests are interchangeable. In fact, it is proven that the Wald test and the likelihood-ratio test are asymptotically equivalent (Engle, 1984).

CHAPTER 3

Statistical Algorithms

This chapter introduces the statistical algorithms that are utilized in the analysis of the PISA 2009 data. The main purpose of these algorithms is to calculate the necessary measures with accuracy and efficiency, which leads to stating a valid conclusion while saving computation time.

3.1 Factor Analysis

Factor analysis is a statistical method that is employed to describe variability among observed variables in terms of factors also known as latent variables. The use of factor analysis in psychometrics was pioneered by Charles Spearman, an English psychologist who proposed the existence of a "general intelligence" (Spearman, 1904). In psychometrics, factor analysis is effective in providing mathematical models for the explanation of psychological theories of human ability and behavior. It is often associated with intelligence research discovering factors underlying human cognitive performance.

Thurstone (1947) introduced the common factor model, a linear model which was developed based on a statistical theory of abilities. Since then, statisticians analyzed linear covariance structures of the common factor model. One of the most significant foundations for factor

analysis is the normal theory maximum likelihood (Lawley and Maxwell, 1963), which is reexamined and improved by Jöreskog (1969). The "difficulty" factors are used to explain the varying endorsement probabilities among dichotomous test items. Factor analysis is conducted on item-level data, using nonlinear statistical methods. In the past decade, factor analysis has been a member of a large family of latent variable models that use hierarchical models formed using Bayesian statistics.

3.1.1 Traditional Factor Analysis Model

In the analysis of the PISA 2009 data, the traditional factor analysis model is used in order to maximally reproduce the correlations among variables. The items' true values and constructs are compared with the assumed values and constructs according to the researchers' understanding of the PISA 2009 data. In this case, the understanding is that the number of factors is far less than the number of variables. For *n* observed variables, the model is simply:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + u_jY_j \qquad (j = 1, 2, \ldots, n)$$

where *z* is the standard deviate of variable *j*, which can be described in terms of *m* common factors and a unique factor according to Harman (1976). The factors' coefficients are referred to as factor loadings, and have a value between -1 and 1. A coefficient close to 1 in absolute value indicates the importance of the factor.

3.1.2 EFA vs. CFA

The main purpose of the Exploratory Factor Analysis (EFA) is to discover the key factors in order to have a clear interpretation of factor patterns across samples. While conducting EFA, the

primary assumption is that any variable may have an association with any item. After the analysis is complete, the resulting factor loadings dictate the factor structure of the data.

When it comes to EFA, two issues of importance are the selection of the number of common factors and the rotation of the factors. For the first issue, it is common to set the number of eigenvalues that exceed 1.0 for the sample correlation matrix as the number of common factors. The scree test is an alternative method that identifies the last major discontinuity in the sequence from the plot of eigenvalues. Significance testing and model fit indices are reviewed in Bentler and Bonett (1980). The good-enough principle (Serlin and Lapsley, 1985) is useful when choosing the number of factors such that the model fit does not significantly improve with an addition of another factor. For the second issue, despite the fact that the oblique rotation is superior to the orthogonal rotation, both rotations should be attempted to guarantee the accuracy of the resulting factor model. If there is prior knowledge about the nature of factors, use of a target rotation is recommended according to Browne (2001). Ultimately, the interpretation by the researcher is essential in order to achieve satisfactory factor rotation results.

On the other hand, the main purpose of the Confirmatory Factor Analysis (CFA) is to determine if the number of factors and the factor loadings are consistent with what is being expected according to previous research. The expectations are regarding which items are being associated with which subset of variables. Factor structure obtained from the CFA is evaluated to see if the structure resembles the factor structure assumed from prior findings. An EFA model can solely be identified by setting the scale of the latent variables and not many other constraints, whereas a CFA model had additional constraints such as range restrictions and complex nonlinear dependence. A researcher specifies a large number of *a priori* zeros in the factor

loadings matrix while conducting CFA. However, the deficiency of such zeros in EFA would result in errors which must be fixed using oblique or orthogonal rotation. The convenience of CFA lies in the ability of simultaneously estimating models of different populations. In addition to this, user-defined restrictions can be imposed, which aids in studying factorial invariance. CFA is used in social research with applications in developing a test, such as an intelligence test, personality test, or a survey (Kline, 2010).

In this dissertation, the CFA is used since we already know that the items in the PISA 2009 data loads on one dimension measuring mathematics literacy. Because of this assumption, we will estimate the primary slope for each item without any constraints.

3.2 Benjamini-Hochberg Procedure

When conducting multiple significance tests, using single-inference procedures would result in an overly increased false positive (significance) rate. As noted in Williams et al. (1999), it has been found that the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is useful to correct for this kind of multiplicity (selection) effect. In fact, it has been proven that it has greater power than two procedures: the Bonferroni technique and the Hochberg procedure (Hochberg, 1988). The Benjamini-Hochberg procedure is derived after forming a new point of view on the problem of multiplicity, which is taking into account the proportion of errors among the rejected hypotheses. Such a proportion is called the false discovery rate (FDR), since Soriç (1989) identified a rejected hypothesis with a "statistical discovery."

According to Benjamini and Hochberg (1995), the FDR can be calculated in a scenario in which m null hypotheses, of which m_0 are true, are tested simultaneously. Let **R** be an observable random variable representing the number of hypotheses rejected. Random variables **S**, **T**, **U**, and

Table 3-1: Number of errors committed when testing *m* null hypotheses

Hypothesis	**Declared Insignificant**	**Declared Significant**	**Total**
True null hypothesis	U	V	m_0
False null hypothesis	T	S	$m - m_0$
Total	$\boldsymbol{m - R}$	**R**	$\boldsymbol{M}$

V are unobservable. If each individual null hypothesis is tested separately at level α, then **R** = **R**(α) is an increasing function of α. These random variables are depicted on Table 3-1.

The proportion of errors committed by falsely rejecting null hypotheses can be represented by the random variable **Q** = **V** / (**V** + **S**). We define the FDR Q_e as the expectation of **Q**:

$$Q_e = E(\boldsymbol{Q}) = E\left(\frac{\boldsymbol{V}}{\boldsymbol{V}+\boldsymbol{S}}\right) = E\left(\frac{\boldsymbol{V}}{\boldsymbol{R}}\right)$$

Consider testing *m* null hypotheses: $H_1, H_2, \ldots H_m$ and obtaining p-values $P_1, P_2, \ldots, P_m$ as a result. Let us re-arrange the p-values from smallest to largest to ensure $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ with the null hypothesis $H_{(i)}$ corresponding to $P_{(i)}$. Let *k* be the largest *i* for which

$$P_{(i)} \leq \frac{i}{m} q^*$$

where q* is a critical value of the proportion of errors represented by **Q**. The critical value is set equal to α (probability of making a Type I error). Then reject all $H_{(i)}$ for $i = 1, 2, \ldots, k$.

When $m_0 < m$, the FDR is smaller than or equal to the family-wise error rate (FWER), which is defined as P(**V** ≥ 1). As a result, any procedure that controls the family-wise error rate controls the FDR. If a procedure controls the FDR only, it can be less stringent and therefore having potential to gain power. The larger the number of the non-true null hypothesis is, the larger **S** tends to be. This large value of **S** causes the difference between FDR and FWER to be

large. Hence, the potential for increase in power is larger when more of the hypotheses are non-true.

A large simulation study in Benjamini and Hochberg (1995) compares the performances of the Benjamini-Hochberg procedure with some other Bonferroni-type procedures. The results of this simulation study show the power of the Bonferroni-type procedures decreasing when the number of hypotheses tested increases – the cost of multiplicity control. In addition to this, the results show the power of the Benjamini-Hochberg procedure is uniformly larger than that of other methods. Furthermore, the advantage increases as the number of non-true null hypotheses increases; the advantage also increases as *m* increases. Since the Benjamini-Hochberg procedure has a relative small loss of power as *m* increases, it is the most preferred method for simultaneously testing many null hypotheses while controlling the multiplicity effect.

The Benjamini-Hochberg procedure has been used in research applications such as reporting results from the National Assessment of Educational Progress (NAEP) as documented in Braswell et al (2001). When applied to research, this procedure can be easily implemented using spreadsheet software, as shown in Thissen et al (2002). The initial steps with regards to calculating p-values of the implementation in Thissen et al (2002) are skipped since the p-values are calculated after the Wald test. After the p-values of every item are calculated, the items are sorted using the p-value as the index in descending order. Then, an index number is given to each item, with 1 given to the item with the largest p-value, 2 given to the item with the second-largest p-value, and so on.

The index number is used to calculate the Benjamini-Hochberg critical value, which has the following formula:

$$\frac{(n - i + 1) * 0.05}{2 * n}$$

in which *n* is the number of items, 0.05 is the value of α, and *i* is the index number. If the p-value of an item is less than the Benjamini-Hochberg critical value, the direction of the difference is confidentially interpreted at the 0.025 level. Such an item is identified as containing a significant amount of DIF.

CHAPTER 4

Previous Applications of DIF in PISA

The PISA is a worldwide evaluation of 15-year-old students' achievement in mathematics literacy, reading literacy, and science literacy. Testing was first performed in 2000 and repeated every three years. Each test is divided into three sections of items according to the type of literacy being measured. There are more than 40 national versions written in official languages used in participating countries. PISA items are developed while attempting to maintain consistency among all national versions.

Research has already been conducted regarding DIF in PISA items. In this chapter, three research papers regarding DIF in PISA are introduced: Grisay et al. (2007), Xie and Wilson (2008), and Le (2009). Grisay et al. (2007) discusses a data analysis in which various national versions of PISA 2006 are verified for linguistic and cultural equivalence. In Xie and Wilson (2008) a linear logistic test model (LLTM) is applied to the mathematics items from PISA 2003 in order to explain the DIF more substantively. Le (2009) investigates the gender DIF across countries and test languages for science items in PISA 2009, providing a valuable contribution to the development of tests for international use. This collection of prior research has been inspiration for proposing a different method of evaluating DIF in a PISA dataset.

4.1 Translation Equivalence across PISA Countries

Due to the continuous increase in the number of countries participating in PISA, ensuring linguistic and cultural equivalence across the various national versions of the items in the assessment has become increasingly important. In Grisay et al. (2007), all national versions of the PISA 2006 are verified for equivalence against the English and French source versions developed by the PISA syndicate. This verification process and the empirical data analysis provide the information used in order to conclude whether the level of linguistic equivalence reached a globally acceptable standard in each of the participating countries. Grisay et al. (2007) focuses on methods of ensuring high levels of translation accuracy using appropriate criteria. In this dissertation, the data analysis investigates whether the level of fairness reached an acceptable standard in mathematics items in the PISA 2009 data.

4.2 Investigating DIF using a Linear Logistic Test Model

Xie and Wilson (2008) discusses a method of generalizing DIF by grouping of items sharing a common feature using the LLTM. This type of grouping is called a facet, and hence the particular version of LLTM used in Xie and Wilson (2008) is called the differential facet functioning (DFF) model. The DFF model helps to explain the DIF more substantively because the interaction effect is modeled at the facet level to yield a sound explanation of difference in performance between subgroups of participants.

Fischer (1973) developed the LLTM to estimate the effects of item properties instead of the items themselves. This model is often applied to test hypotheses regarding cognitive

operations utilized during solving mathematical problems. The logit expression of the LLTM is defined as follows:

$$logit = \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik},$$

where θ_n is the proficiency of person *n*, η_k is the difficulty parameter for item property *k*, *K* is the total number of item properties, and Q_{ik} is the indicator weight of item *i* on item property *k*. If item *i* belongs to item property *k*, Q_{ik} takes the value of 1, and 0 otherwise.

In order to investigate for DIF, a DFF term can be added to the logit expression of the LLTM as follows:

$$logit = \theta_n - \sum_{k=0}^{K} \eta_k Q_{ik} + Z_n \sum_{k=1}^{K} \gamma_k Q_{ik},$$

where Z_n is an indicator of person *n*'s group membership and γ_k is the DFF parameter for item property *k*. Let us use contrast coding for the indicator variable Z_n. If person *n* belongs to the reference group, Z_n takes the value of -1; otherwise, Z_n takes the value of 1.

The data under investigation in Xie and Wilson (2008) is the mathematics items from the PISA 2003 database. These items are developed along three domains: content, process, and situation. Under each domain, which is treated as a facet in the DFF model, there are multiple categories. The content domain has four categories: space and shape, change and relationship, quantity, and uncertainty. Reproduction, connections, and reflection are the three categories in the process domain. Lastly, the situation domain contains four categories: personal, educational or occupational, public, and scientific. This structure with two levels is a motivation for

applying the two-tier model, which contains primary dimensions and specific dimensions, to detect DIF in this dissertation which analyzes mathematics items from PISA 2009.

In Xie and Wilson (2008), data from countries that have a slight difference in mean performances on the mathematics scale are analyzed to minimize any overall country effect that may contribute to the difference in performance. For example, the performance of students in Japan is compared to the performance of students in Canada since the mean scores for the former group is slightly higher than those for the latter group. Results of this analysis indicate the Japanese students perform better than Canadian students overall. These results serve as an inspiration to investigate the PISA 2009 data in this dissertation for evidence of DIF between students in Malaysia and students in Singapore.

4.3 Investigating Gender DIF across Countries and Test Languages

Le (2009) investigates the effects of countries and test languages on gender DIF in science items from PISA 2006. The data is collected from 60 test language groups by 50 participating countries. An IRT method employing the partial credit model is used to detect the gender uniform DIF of each of the language groups and the whole international sample. When the closed response items are analyzed, the direction of DIF indicates that the items favor the males. These results of this data analysis provide the idea to use the Benjamini-Hochberg procedure described in Section 4.3 to evaluate the direction of DIF in PISA 2009 items in this dissertation.

CHAPTER 5

Proposed DIF Detection Method

In this chapter, the proposed DIF detection method which is used in the data analysis of this dissertation is explained. The constraints involved in fitting the bifactor model in the detection method are essential because they serve as a key component in computing between-group differences involved in the specific dimensions. In this data analysis, the specific dimensions are testlets, which are clusters of items that share a common stimulus. The proposed method involves 5 steps:

1. Fit the bifactor model to the set of items.
2. Do a sweep analysis.
3. Establish the anchor set of items containing DIF.
4. Test the items for DIF using the anchor set of items.
5. Identify the items containing DIF.

After those 5 steps are completed, the set of items can be divided into two categories: items that contain DIF and items that do not contain DIF.

5.1 Step 1: Fit the Bifactor Model

Step 1 involves fitting the bifactor model to the data that contains the items of interest. In this data analysis, the items of interest are the math items contained in the PISA 2009. 21 rectangular quadrature points between -5 and 5 are used and the convergence criterion is set to 0.001. While fitting the model, the following item parameters are constrained equal between groups: the discrimination parameter on the primary dimension (math literacy), the discrimination parameter on the item's corresponding specific dimension (testlet), and the item intercept(s). The mean vector and the covariance matrix of the latent variable for the focal group are freely estimated, while assuming that the counterparts for the reference group are the mean vector and covariance matrix corresponding to the standard normal distribution.

5.2 Step 2: Do a Sweep DIF Analysis

Step 2 will complete a sweep DIF analysis on the data fitted by the bifactor model. Conditional on the means and variances of the focal group estimated in Step 1, the item parameters (discrimination parameters and intercept parameters) and the corresponding covariance matrix will be estimated within each group. 21 rectangular quadrature points between -5 and 5 are used for estimation, and the convergence criterion is set to 0.001. Subsequently, the parameters and the covariance matrix will be used to conduct the Wald test for the difference between the parameter sets for each item across groups.

5.3 Step 3: Establish the Anchor Set

Step 3 will apply the Benjamini-Hochberg procedure to the results from Step 2 in order to establish the anchor set of items that do not contain DIF. All items contained in the anchor set have a p-value that is larger than the Benjamini-Hochberg critical value. This anchor set will be used in the subsequent test that detects DIF in the items excluded from the anchor set.

5.4 Step 4: Test the Items for DIF Using the Anchor Set

Step 4 will use the anchor set from Step 3 to test each of the candidate items (items excluded from the anchor set) for DIF. The Wald test for the difference between the parameter sets across groups will be conducted.

5.5 Step 5: Identify the Items Containing DIF

Step 5 will apply the Benjamini-Hochberg procedure to the results from Step 4 in order to identify the items that contain DIF. The completion of Step 5 will produce the final results of the proposed DIF detection method.

CHAPTER 6

Analysis of PISA 2009 Mathematics Items

In this chapter we discuss the analysis of the mathematics items of the PISA 2009 data motivated by prior research regarding DIF on a PISA dataset. The PISA 2009 data, available to the public via the Organization for Economic Cooperation and Development (OECD) website, contains the students' responses to a 2-hour paper-and-pencil assessment which was administered between September and November 2009. There are 35 mathematics items: 32 items have a binary response and 3 items have a polytomous response (categories 0, 1, and 2).

In this data analysis, cross-country DIF was examined between the two countries: Australia and New Zealand. These two countries have an education index of 0.993, which is among the highest in the world (United Nations Human Development Reports, 2008). Since the education indices for both countries are similar, we expect the overall achievement to be similar as well. The dataset contains responses from 14251 students in Australia and 4643 students in New Zealand. DIF analysis is suitable for comparing the item responses of students from Australia to item responses of students from New Zealand, followed by detecting each item of the test for any evidence of favoring one group over another.

6.1 DIF Analysis Using the Proposed DIF Detection Method

While glancing at the items, one can notice the existence of several testlets. Items belonging to the same testlet are asking a question related to the same data or same figure. This relationship is displayed in Figure 6-1. One can infer that the probability of correctly answering one of the items in a testlet has a high association with the probability of correctly answering the remaining items in the testlet. Therefore, one can take the testlets into account and apply the proposed DIF detection method to conduct a DIF analysis of the items.

6.1.1 Bifactor IRT Analysis

In the preliminary stage of data analysis, we are fitting the bifactor model to the math items contained in PISA 2009. For the 32 items with binary response, we fit the classical model for dichotomous response. For the 3 items with polytomous response, we fit the model for graded response. With the primary slope, specific slope, and intercept constrained equal between groups, the mean matrix and covariance matrix of the latent variable for the focal group are calculated relative to the counterparts for the reference group corresponding to the standard normal distribution. The means and the variances of the latent variable for the focal group along with the counterparts for the reference group are displayed in Table 6-1. The mean of the math literacy dimension is higher for the students of New Zealand (0.15) compared to the students of Australia (0.00), indicating higher achievement for the former group of students. The standard errors, which are entries of item parameter error variance-covariance matrices, are computed using the Supplemented EM algorithm (Cai, 2008).

Conditional on the means and variances of the latent variable displayed in Table 6-1, the item parameters (primary slope, specific slope, and intercept) for each item and the corresponding covariance matrix will be estimated within each group. However, the constraint is

placed such that the specific slopes for the focal group are equal to the specific slopes for the reference group. Table 6-2 displays the primary slopes and specific slopes of the 32 mathematics items with binary response, while Table 6-3 displays the intercepts. Table 6-4 displays the primary slopes and specific slopes of the 3 mathematics items with polytomous response, while Table 6-5 displays the intercepts. Between-group differences of primary slopes and intercepts are evident in the majority of items. The standard error of each item parameter is calculated using the Supplemented EM algorithm. Looking at item parameters of all 35 items, most of the standard errors of parameters are 0.20 or below.

.

Figure 6-1: Bifactor Model for the PISA 2009 Math Items

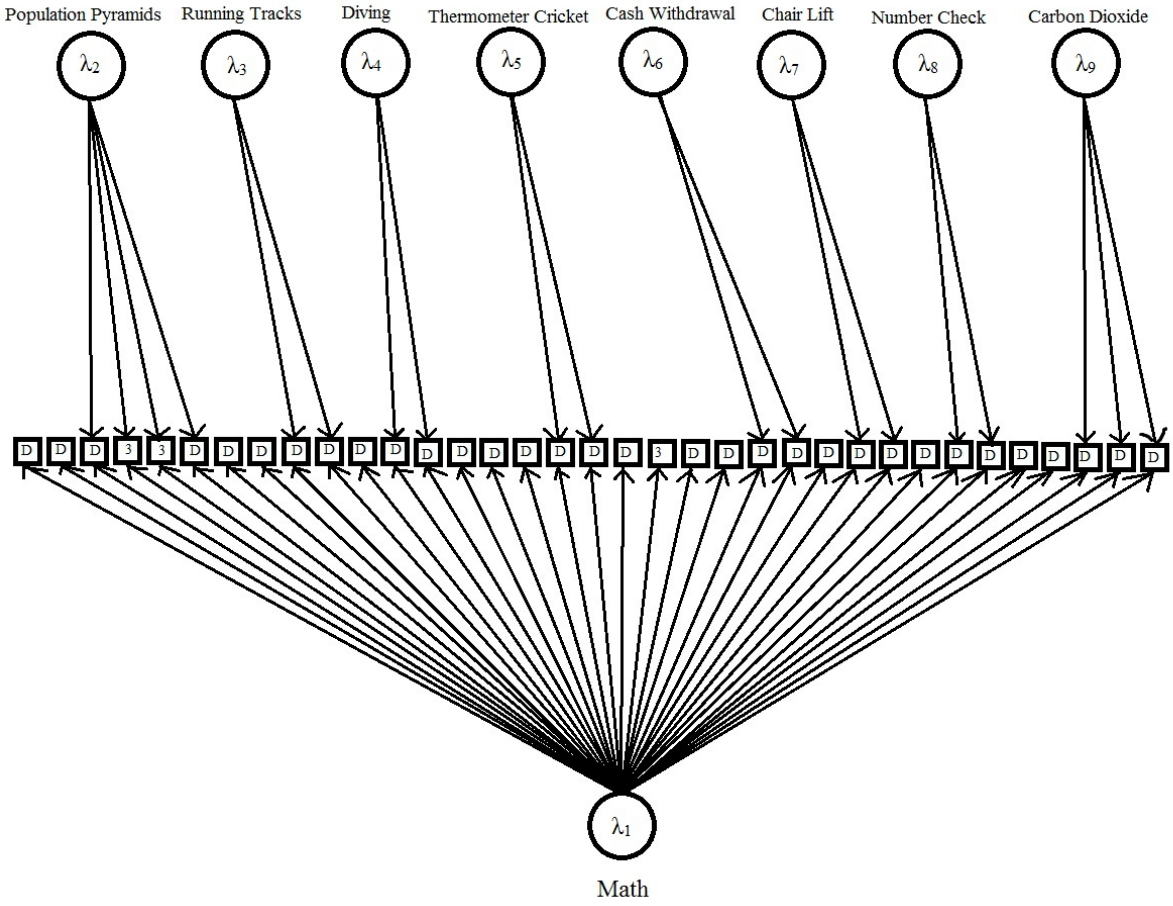

Table 6-1: PISA 2009 Factor Means and Variances (Standard Error)

	Math	**Population Pyramids**	**Running Tracks**	**Diving**	**Thermometer Cricket**	**Cash Withdrawal**	**Chair Lift**	**Number Check**	**Carbon Dioxide**
Means									
Australia	.00 (--)	.00 (--)	.00 (--)	.00 (--)	.00 (--)	.00 (--)	.00 (--)	.00 (--)	.00 (--)
New Zealand	.15 (.03)	-.04 (.08)	-.13 (.07)	.18 (.22)	-.04 (.10)	-.02 (.06)	-.08 (.10)	-.02 (.10)	-.07 (.05)
Variances									
Australia	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)	1.00 (--)
New Zealand	1.01 (.04)	.81 (.28)	.86 (.15)	.98 (2.02)	1.45 (.55)	.87 (.18)	.80 (.55)	1.21 (.62)	.62 (.15)

Note: Fixed parameters do not have standard errors

Table 6-2: Between-Group Comparison of Primary Slopes and Specific Slopes (Standard Error) of PISA 2009 Math Items with Binary Response

Item	Australia a_0	New Zealand a_0	Australia a_s	New Zealand a_s
M033Q01	.88 (.05)	.77 (.09)	0.00 (--)	0.00 (--)
M034Q01T	1.20 (.06)	1.14 (.09)	0.00 (--)	0.00 (--)
M155Q01	1.46 (.07)	1.65 (.13)	.99 (.15)	.99 (.15)
M155Q04T	1.01 (.05)	.77 (.08)	.40 (.09)	.40 (.09)
M192Q01T	1.53 (.07)	1.36 (.11)	0.00 (--)	0.00 (--)
M273Q01T	.94 (.05)	1.06 (.09)	0.00 (--)	0.00 (--)
M406Q01	3.75 (.16)	3.59 (.27)	2.70 (.27)	2.70 (.27)
M406Q02	4.64 (.20)	4.15 (.34)	2.70 (.27)	2.70 (.27)
M408Q01T	1.09 (.05)	1.00 (.08)	0.00 (--)	0.00 (--)
M411Q01	1.74 (.07)	1.84 (.14)	.29 (.15)	.29 (.15)
M411Q02	1.14 (.05)	1.28 (.10)	.29 (.15)	.29 (.15)
M420Q01T	1.28 (.06)	1.20 (.10)	0.00 (--)	0.00 (--)
M423Q01	.79 (.06)	.85 (.11)	0.00 (--)	0.00 (--)
M442Q02	1.86 (.08)	1.65 (.13)	0.00 (--)	0.00 (--)
M446Q01	1.58 (.07)	1.88 (.14)	0.82 (.12)	0.82 (.12)
M446Q02	2.12 (.13)	2.49 (.25)	0.82 (.12)	0.82 (.12)
M447Q01	1.28 (.06)	1.58 (.13)	0.00 (--)	0.00 (--)
M464Q01T	2.00 (.09)	2.01 (.17)	0.00 (--)	0.00 (--)
M474Q01	.61 (.04)	.78 (.09)	0.00 (--)	0.00 (--)
M496Q01T	1.80 (.09)	1.65 (.14)	1.44 (.10)	1.44 (.10)
M496Q02	1.43 (.07)	1.31 (.12)	1.44 (.10)	1.44 (.10)
M559Q01	1.23 (.06)	1.27 (.10)	0.00 (--)	0.00 (--)
M564Q01	.74 (.05)	.75 (.08)	.53 (.08)	.53 (.08)
M564Q02	.74 (.05)	.77 (.08)	.53 (.08)	.53 (.08)
M571Q01	1.65 (.07)	1.59 (.12)	0.00 (--)	0.00 (--)
M603Q01T	.94 (.05)	.93 (.09)	.61 (.10)	.61 (.10)
M603Q02T	1.82 (.09)	1.83 (.16)	.61 (.10)	.61 (.10)
M800Q01	.58 (.05)	.34 (.09)	0.00 (--)	0.00 (--)
M803Q01T	2.09 (.09)	2.33 (.19)	0.00 (--)	0.00 (--)
M828Q01	1.71 (.10)	1.61 (.13)	.77 (.10)	.77 (.10)
M828Q02	1.65 (.11)	1.41 (.21)	2.19 (.68)	2.19 (.68)
M828Q03	1.46 (.08)	1.31 (.11)	.57 (.08)	.57 (.08)

Note: Fixed parameters do not have standard errors

Table 6-3: Between-Group Comparison of Intercepts (Standard Error) of PISA 2009 Math Items with Binary Response

Item	Australia d	New Zealand d
M033Q01	1.49 (.05)	1.44 (.08)
M034Q01T	-.38 (.04)	-.20 (.08)
M155Q01	1.38 (.07)	1.40 (.11)
M155Q04T	.54 (.04)	.54 (.07)
M192Q01T	-.30 (.04)	-.05 (.08)
M273Q01T	.07 (.04)	.21 (.07)
M406Q01	-2.39 (.14)	-2.87 (.21)
M406Q02	-4.94 (.21)	-4.58 (.30)
M408Q01T	.30 (.04)	.10 (.07)
M411Q01	-.03 (.04)	.11 (.10)
M411Q02	-.04 (.04)	-.09 (.08)
M420Q01T	.78 (.04)	.70 (.08)
M423Q01	1.84 (.05)	1.85 (.09)
M442Q02	-.57 (.05)	-.89 (.10)
M446Q01	1.62 (.06)	1.58 (.11)
M446Q02	-3.99 (.14)	-4.18 (.27)
M447Q01	1.13 (.05)	1.45 (.10)
M464Q01T	-1.70 (.07)	-1.77 (.15)
M474Q01	1.10 (.04)	1.30 (.08)
M496Q01T	.46 (.06)	.49 (.10)
M496Q02	1.09 (.06)	1.02 (.10)
M559Q01	.68 (.04)	.78 (.08)
M564Q01	-.24 (.04)	-.24 (.07)
M564Q02	-.18 (.04)	-.27 (.07)
M571Q01	-.01 (.04)	-.14 (.09)
M603Q01T	-.42 (.04)	-.37 (.08)
M603Q02T	-1.34 (.06)	-1.43 (.13)
M800Q01	1.87 (.05)	1.91 (.08)
M803Q01T	-1.33 (.06)	-1.42 (.15)
M828Q01	-.59 (.06)	-.70 (.09)
M828Q02	.59 (.14)	.50 (.10)
M828Q03	-1.38 (.06)	-1.41 (.10)

Table 6-4: Between-Group Comparison of Primary Slopes and Specific Slopes (Standard Error) of PISA 2009 Math Items with Polytomous Response

Item	Australia a_0	New Zealand a_0	Australia a_s	New Zealand a_s
M155Q02D	1.43 (.06)	1.58 (.12)	.68 (.13)	.68 (.13)
M155Q03D	1.75 (.07)	2.02 (.16)	.42 (.12)	.42 (.12)
M462Q01D	1.60 (.10)	1.83 (.21)	0.00 (--)	0.00 (--)

Note: Fixed parameters do not have standard errors

Table 6-5: Between-Group Comparison of Intercepts (Standard Error) of PISA 2009 Math Items with Polytomous Response

Item	**Australia d_1**	**New Zealand d_1**	**Australia d_2**	**New Zealand d_2**
M155Q02D	1.69 (.06)	1.68 (.12)	.69 (.05)	.65 (.10)
M155Q03D	-1.32 (.06)	-1.49 (.14)	-2.91 (.09)	-3.40 (.20)
M462Q01D	-3.24 (.11)	-3.46 (.27)	-4.49 (.14)	-4.55 (.33)

6.1.2 DIF Detection

After determining that the math items loads on the bifactor model, it is feasible to complete a sweep DIF analysis. The primary slopes, specific slopes, intercepts, means and variances of the latent variable, and their corresponding covariance matrix of parameter estimates will be used to conduct the Wald test for the difference between the parameter sets for each item across groups. In the Wald test, the following null hypothesis is being tested:

$$H_0: [a_A, d_A] = [a_{NZ}, d_{NZ}]$$

where a_A and d_A represent the primary slope and intercept for the students from Australia respectively, and a_{NZ} and d_{NZ} represent the primary slope and intercept for the students from New Zealand. Let $\hat{a}_A$, $\hat{d}_A$, $\hat{a}_{NZ}$, and $\hat{d}_{NZ}$ be the maximum likelihood estimate of a_A, d_A, a_{NZ}, and d_{NZ}, respectively. The test statistic used for the joint difference between the parameters for the two groups is as follows:

$$\chi^2 = v' \Sigma^{-1} v$$

where v is $\left[\hat{a}_A - \hat{a}_{NZ}, \hat{d}_A - \hat{d}_{NZ}\right]$, Σ is the variance-covariance matrix of the differences between item parameter estimates. Table 6-6 shows the results of the Wald test. Subsequently, the Benjamini-Hochberg procedure is applied to the results of the Wald test in order to establish the anchor set of items that do not contain significant DIF. As seen in Table 6-7, all of the items except M442Q02 are included in the anchor set since their p-values are bigger than their Benjimini-Hochberg critical values. The fifth column of Table 6-7 displays three asterisks (***) for items in which the p-value is smaller than the critical value; otherwise, the space is left blank for items that satisfy the aforementioned criteria to be included in the anchor set.

Table 6-6: Results of the Wald Test on PISA 2009 Math Items under the Bifactor Model

Item	χ^2	p-value
M033Q01	1.3	0.5281
M034Q01T	4.3	0.1145
M155Q01	1.7	0.6385
M155Q02D	1.5	0.8182
M155Q03D	6.7	0.1529
M155Q04T	6.4	0.0940
M192Q01T	8.0	0.0182
M273Q01T	4.8	0.0912
M406Q01	8.2	0.0414
M406Q02	1.5	0.6750
M408Q01T	7.7	0.0215
M411Q01	2.4	0.5029
M411Q02	1.8	0.6215
M420Q01T	1.0	0.6068
M423Q01	0.3	0.8768
M442Q02	16.9	0.0002
M446Q01	4.8	0.1909
M446Q02	2.8	0.4239
M447Q01	8.8	0.0122
M462Q01D	3.2	0.3689
M464Q01T	0.3	0.8551
M474Q01	7.1	0.0293
M496Q01T	1.0	0.7923
M496Q02	1.1	0.7874
M559Q01	1.1	0.5851
M564Q01	0.0	0.9998
M564Q02	1.2	0.7573
M571Q01	2.1	0.3498
M603Q01T	0.4	0.9447
M603Q02T	0.6	0.8911
M800Q01	5.5	0.0631
M803Q01T	1.3	0.5272
M828Q01	2.1	0.5431
M828Q02	1.7	0.6390
M828Q03	2.2	0.5235

Table 6-7: Results of the Benjamini-Hochberg Procedure on PISA 2009 Math Items under the Bifactor Model

Item	p-value	Index	Critical value	Indicator
M564Q01	0.9998	1	0.0250	
M603Q01T	0.9447	2	0.0243	
M603Q02T	0.8911	3	0.0236	
M423Q01	0.8768	4	0.0229	
M464Q01T	0.8551	5	0.0221	
M155Q02D	0.8182	6	0.0214	
M496Q01T	0.7923	7	0.0207	
M496Q02	0.7874	8	0.0200	
M564Q02	0.7573	9	0.0193	
M406Q02	0.6750	10	0.0186	
M828Q02	0.6390	11	0.0179	
M155Q01	0.6385	12	0.0171	
M411Q02	0.6215	13	0.0164	
M420Q01T	0.6068	14	0.0157	
M559Q01	0.5851	15	0.0150	
M828Q01	0.5431	16	0.0143	
M033Q01	0.5281	17	0.0136	
M803Q01T	0.5272	18	0.0129	
M828Q03	0.5235	19	0.0121	
M411Q01	0.5029	20	0.0114	
M446Q02	0.4239	21	0.0107	
M462Q01D	0.3689	22	0.0100	
M571Q01	0.3498	23	0.0093	
M446Q01	0.1909	24	0.0086	
M155Q03D	0.1529	25	0.0079	
M034Q01T	0.1145	26	0.0071	
M155Q04T	0.0940	27	0.0064	
M273Q01T	0.0912	28	0.0057	
M800Q01	0.0631	29	0.0050	
M406Q01	0.0414	30	0.0043	
M474Q01	0.0293	31	0.0036	
M408Q01T	0.0215	32	0.0029	
M192Q01T	0.0182	33	0.0021	
M447Q01	0.0122	34	0.0014	
M442Q02	0.0002	35	0.0007	***

Using the anchor set, the candidate item that is excluded from the anchor set is tested for DIF using the Wald test for the difference between the parameter sets across groups. The null hypothesis is the same as the null hypothesis set for the previous Wald test:

$$H_0: [a_A, d_A] = [a_{NZ}, d_{NZ}]$$

The Benjamini-Hochberg procedure is applied to the results in order to determine if the item contains significant DIF or not. According to the results, there is detection of significant DIF in the candidate item. The final conclusion regarding the DIF detection analysis of PISA 2009 math items is that 1 out of 35 items shows significant DIF favoring the students of New Zealand over the students of Australia. After some change in design for the math item M422Q02, the 2009 PISA math items will be perfectly fair for two groups of students with similar education indices, such as the students of Australia and the students of New Zealand.

CHAPTER 7

Simulation Study

In order to provide concrete evidence of the model's utility, a simulation study is conducted. Throughout the simulation study, two measurements being recorded are Type I error rate and power. Type I error rate, alternatively noted as α, is the probability of mistakenly diagnosing an item as having DIF, while power is the probability of correctly diagnosing an item as having DIF. 80% of the items are designed to have no DIF, and therefore can serve as a basis to measure Type I error rate. The remaining 20% of the items are designed to have DIF, and thus is suitable as a medium in which power can be calculated.

As seen in Table 7-1, the simulation study employed variation of four factors: sample size per group, test length, magnitude of *a*-DIF, and magnitude of *d*-DIF. Each factor has two different values: 250 or 1000 participants, 10 or 40 items, 0.5 or 1.0 as the difference in the *a* parameter between groups, and 0.4 or 0.8 as the difference in the *d* parameter between groups. Despite the variation in factors, all items are dichotomous, the latent variable has a distribution $N(-0.3,0.8)$ for the focal group compared to $N(0, 1)$ for the reference group, and the sample sizes of both groups are equal. Responses to these items are generated according to the values of the four factors, parameters are estimated separately for the focal group and the reference group, and the difference in parameters is evaluated using the proposed DIF detection method.

Table 7-1: Simulation Study Design

Sample Size per Group	250, 1000 (x2)
Test Length	10, 40 (x2)
Magnitude of *a*-DIF	0.5, 1.0 (x2)
Magnitude of *d*-DIF	0.4, 0.8 (x2)
Total Number of Cells	2 x 2 x 2 x 2 = 16

For each of the 16 cells, the simulation procedure utilizing the proposed DIF detection method is run for 100 replications. In each cell, the Type I error rate and the power of the proposed DIF detection method is calculated. After 100 replications, an accurately estimated value of each statistic is obtained by calculating the arithmetic mean of the 100 values. There are some trends that could be observed while comparing values under varying conditions of the factors.

The estimated values for Type I error rate for all 16 cells are displayed in Table 7-2. As sample size increases, the Type I error rate increases since there are more responses with a chance of mistakenly diagnosing an item of having DIF. The increase of test length has a stabilizing effect on the Type I error rate. When the sample size is 250, the Type I error rate increases as the test length increases due to the lack of DIF detection from a relatively short test. On the other hand, when the sample size is 1000, the Type I error rate decreases as the test length increases since having more items will result in higher accuracy of DIF detection.

The estimated values for power for all 16 cells are displayed in Table 7-3. Sample size has the greatest influence on power. The estimated power is large when the sample size is large because there is increased evidence of DIF with increased sample size. Increasing the magnitude of *a*-DIF or *d*-DIF also induces an increase in power, since the DIF will be easily detected as the magnitude increases.

Overall, the results from Table 7-2 and Table 7-3 indicate that the results of the proposed DIF detection method are desirable as the sample size, test length, and DIF magnitude increases. The increase in these factors provide an opportunity for proper DIF detection for different reasons, and therefore the Type I error estimate is reduced to a reasonable value and the power estimate is increased to an ideal value of high standards.

Table 7-2: Estimated Type I Error Rates for Simulated Items

N	Test Length	*a*-DIF	*d*-DIF	Type I Error Rate
250	10	0.5	0.4	0.000
250	10	0.5	0.8	0.000
250	10	1.0	0.4	0.000
250	10	1.0	0.8	0.025
250	40	0.5	0.4	0.006
250	40	0.5	0.8	0.012
250	40	1.0	0.4	0.006
250	40	1.0	0.8	0.019
1000	10	0.5	0.4	0.150
1000	10	0.5	0.8	0.125
1000	10	1.0	0.4	0.125
1000	10	1.0	0.8	0.073
1000	40	0.5	0.4	0.006
1000	40	0.5	0.8	0.032
1000	40	1.0	0.4	0.013
1000	40	1.0	0.8	0.047

Note: Type I error rate is only calculated for items designated to not contain DIF.

Table 7-3: Estimated Power for Simulated Items

N	Test Length	*a*-DIF	*d*-DIF	Power
250	10	0.5	0.4	0.000
250	10	0.5	0.8	0.200
250	10	1.0	0.4	0.000
250	10	1.0	0.8	0.400
250	40	0.5	0.4	0.100
250	40	0.5	0.8	0.650
250	40	1.0	0.4	0.200
250	40	1.0	0.8	0.675
1000	10	0.5	0.4	1.000
1000	10	0.5	0.8	1.000
1000	10	1.0	0.4	1.000
1000	10	1.0	0.8	1.000
1000	40	0.5	0.4	0.075
1000	40	0.5	0.8	1.000
1000	40	1.0	0.4	0.225
1000	40	1.0	0.8	1.000

Note: Power is only calculated for item designed to contain DIF.

CHAPTER 8

Discussion

The main accomplishment was the development of a DIF detection method used with the generalized full-information item bifactor analysis model. Traditional DIF detection methods rely on the items to be conditionally independent of each other conditional on the latent trait, which is an assumption relaxed in the bifactor model. In this study the bifactor model assumes that there are closely related items in testlets, so DIF was detected accordingly.

In future research, there are some areas concerning DIF that could use further investigation. First, advancing from this study in which one primary dimension is examined, DIF detection methods can be applied to an IRT model with more than one primary dimension. An example of such a model is the two-tier full-information item factor analysis model (Cai, 2010) which permits cross-loadings of items on the primary factors, making it more flexible than the bifactor model. Second, since plenty has been accomplished in regards to showing the sensitivity of some DIF method to variations of parameter distributions, the use of DIF models to study change should be a topic of research. A substitution of "before" and "after" for "focal" and "reference" can provide a measure of likelihood of change (Wainer, 2010). There are many such opportunities in the future, since there is always a probability of observing a change in the educational system such as academic curriculum, testing, and assessments.

BIBLIOGRAPHY

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57(1),* 289-300.

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Braeken, J., Tuerlinckx, F., and De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika, 72,* 393-411.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B., and Johnson, M. (2001). *The nation's report card: Mathematics 2000.* NCES 2001-517. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111-150.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology,* 61, 309-329.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581-612.

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

Dorans, N. J., and Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., and Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Engle, R. F. (1984). Wald, Likelihood ratio, and Lagrange multiplier tests in econometrics. In M.D. Intirligator (Eds.), *Handbook of econometrics II* (pp. 796-801). New York, NY: Elsevier.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278-295.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 3,* 359-374.

Grisay, A., de Jong, J.H., Gebhardt, E., Berezner, A., and Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8(3),* 2007, 249-266.

Harman, H. H. (1976). *Modern factor analysis – Third edition revised.* Chicago, IL: The University of Chicago Press.

Harrell, F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis.* New York, NY: Springer-Verlag Inc.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple test of significance. *Biometrika, 75,* 800-803.

Holland, P. W., and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202.

Kline, R. B. (2010). *Principles and practice of structural modeling (3rd ed).* New York, New York: Guilford Press.

Lawley, D. N. and Maxwell, A. E. (1963). *Factor analysis as a statistical method.* London, UK: Butterworth.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9, 2,* 122-133.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (p. 212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

McDonald, R. P. (1967). *Nonlinear factor analysis.* Psychometric Monographs, No. 15.

MacIntosh, R., and Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis, *Applied Psychological Measurement, 27(5),* 372-379.

Mignani, S., Cagnone, S., Casadei, G., and Carbonaro, A. (2005). An item response theory model for student ability evaluation using computer-automated test results. In Vichi, M. Monari, P., Mignani, S., Montanari, A. (Eds.), *New Developments in Classification and Data Analysis* (pp. 325-332). Berlin, Germany: Springer-Verlag Inc.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias, *Applied Psychological Measurement, 17,* 297-334.

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement, 24(3),* 211-223.

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

Muraki, E., and Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73-90.

Muthen, B. O., Kao, C. F., and Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28(1),* 1-22.

Narayanan, P., and Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257-274.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Rijmen, F., Vansteelandt, K., and De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73,* 167-182.

Roussos, L. A., and Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33,* 215-230.

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

Scheuneman, J. D. (1975). *A new method of assessing bias in test items.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 106 359).

Serlin, R. C., and Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73-83.

Shealy, R. T., and Stout, W. F. (1993). A model-biased standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* (pp. 9-30). Baltimore: Johns Hopkins University Press.

Shih, C.-L., and Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33,* 184-199.

Soriç, B. (1989). Statistical "discoveries" and effect size estimation. *Journal of the American Statistical Association, 84,* 608-610.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15 (2),* 201-292.

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27,* 361-370.

Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables, *Psychometrika, 52,* 393-408.

Thissen, D., Cai, L., and Bock, R. D. (2010). The nominal categories item response model. In M. Nering and R. Ostini (Eds.), *Handbook of polytomous item response theory models: developments and applications* (pp. 43-75). New York, NY: Taylor and Francis.

Thissen, D., Steinberg, L., and Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thissen, D., Steinberg, L., and Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27,* 77-83.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: The University of Chicago Press.

United Nations Human Development Reports. (2008). *Human development indices* [Data file]. Retrieved from http://hdr.undp.org/en/media/HDI_2008_EN_Tables.pdf

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavorial Statistics, 35*,5-25.

Wang, W.-C., Shih, C.-L., and Su, Y.-H. (2007). *Establishing a common metric over groups for the assessment of differential item functioning.* Submitted for publication.

Wang, W.-C., and Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Williams, V.S.L., Jones, L.V., and Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42-69.

Xie, Y. and Wilson, Mark. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: an international testing context. *Psychological Science Quarterly, 50*, 2008 (3), 403-416.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.