ABSTRACT

Title of dissertation: TESTING DIFFERENTIAL ITEM FUNCTIONING BY REGULARIZED MODERATED NONLINEAR FACTOR ANALYSIS

Weimeng Wang, Doctor of Philosophy, 2022

Dissertation directed by: Professor Jeffrey R. Harring &
Associate Professor Yang Liu
Measurement, Statistics & Evaluation
Department of Human Development
and Quantitative Methodology
University of Maryland

Recent advancements in testing differential item functioning (DIF) have greatly relaxed restrictions made by the conventional multiple group item response theory (IRT) model with respect to the number of grouping variables and the assumption of predefined DIF-free anchor items. The application of the L_1 penalty in DIF detection has shown promising results in identifying a DIF item without a priori knowledge on anchor items while allowing the simultaneous investigation of multiple grouping variables. The least absolute shrinkage and selection operator (LASSO) is added directly to the loss function to encourage variable sparsity such that DIF parameters of anchor items are penalized to be zero. Therefore, no predefined anchor items are needed. However, DIF detection using LASSO requires a non-trivial model selection consistency assumption and is difficult to draw statistical inference. Given the importance of identifying DIF items in test development, this study aims to apply

the decorrelated score test to test DIF once the penalized method is used. Unlike the existing regularized DIF method which is unable to test the statistical significance of a DIF item selected by LASSO, the decorrelated score test requires weaker assumptions and is able to provide asymptotically valid inference to test DIF. Additionally, the deccorrelated score function can be used to construct asymptotically unbiased normal and efficient DIF parameter estimates via a one-step correction. The performance of the proposed decorrelated score test and the one-step estimator are evaluated by a Monte Carlo simulation study.

TESTING DIFFERENTIAL ITEM FUNCTIONING BY REGULARIZED MODERATED NONLINEAR FACTOR ANALYSIS

by

Weimeng Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Jeffrey R. Harring, Chair/ Advisor
Professor Yang Liu, Co-chair
Professor Gregory R. Hancock
Professor Takumi Saegusa
Professor Jade Wexler, Dean's Representative

Dedication

To my parents, from whom I have got unconditional love and support. I love you all dearly.

献给爱我支持我的父母。我爱你们。

Acknowledgments

To my parents and my grandparents, who love me unconditionally. Mom and dad, without your emotional and financial support, I would never pursue my dream alone in America or be the person I want to be. Grandpa, even if I will never see you again, you have raised me with all the love and kindness in the world. I know you would be very proud of me.

To my loving husband, Eddie: meeting you was the best thing that have ever happened to me since the pandemic. Thank you for all the little things you have done to make my academic life easier. You are my motivation to complete graduate school. I am so happy to build a family of our own with you.

I am deeply grateful to my understanding and compassionate academic advisor, Dr. Jeffrey Harring, who has encouraged, inspired, and mentored me for the past four years. Thank you, Dr. Harring. I appreciate all of your guidance and support through these hard years and consistent faith in me. I also want to thank Dr. Yang Liu, who has so effortlessly mentored and guided me. I could say without doubt that the majority of my technical skills and knowledge on modern test theory are gotten from you. Your scientific advice and knowledge have inspired me and will continue to have influence on me. I also want to thank my dissertation committee, Drs. Gregory Hancock, Takumi Saegusa, Jade Wexler, who have provided insightful discussions and suggestions, which enriched my dissertation.

I would like to express my sincere appreciation for the EDMS family. To my peers and friends: thank you all for your emotional and intellectual support.

To other faculty members at EDMS: I am extremely grateful to all the academic training, mentorship, and support.

Lastly, my gratitude also goes to my colleagues at the U.S. Food and Drug Administration (FDA). Thank you, Drs. Weiya Zhang and Laura Lee Johnson, for granting me maximum flexibility at work. Without your support, I could never finish my dissertation so efficiently. I also want to recognize my team lead, Dr. Lili Garrard. Lili, your are the best team lead anyone could imagine. Thank you for always speaking up and advocating for me. To my team members, Monica Morrell, Marian Strazzeri, and Xin Yuan: I am deeply grateful to you all for taking more responsibilities at work allowing me to concentrate on my dissertation. Your companionship throughout this whole journey has been invaluable.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables vii

List of Figures viii

List of Abbreviations and Symbols ix

1 Introduction 1
- 1.1 Statement of the Problem . . . 1
- 1.2 Purpose of the Study . . . 6

2 Literature Review 8
- 2.1 A General Definition of Differential Item Functioning . . . 8
- 2.2 Existing IRT DIF Detection Procedures . . . 15
  - 2.2.1 Asymptotic Hypothesis Tests . . . 15
    - 2.2.1.1 Conventional Hypothesis Tests . . . 16
  - 2.2.2 Anchor Item Issue . . . 20
  - 2.2.3 Empirical Anchor Selection Method . . . 22
    - 2.2.3.1 Anchor Selection Strategies . . . 25
    - 2.2.3.2 Robust Regression Methods . . . 26
    - 2.2.3.3 Regularization Methods . . . 27
- 2.3 Statistical inference for the L_1 penalty . . . 35
  - 2.3.1 Sub-model Inference . . . 37
    - 2.3.1.1 Sample-splitting Technique . . . 38
    - 2.3.1.2 Exact Post-selection Inference . . . 39
  - 2.3.2 Full Model Inference . . . 41
  - 2.3.3 Summary . . . 42

3 Theory 45
3.1 A General Theory for Decorrelated Score Function 45
3.2 Extension to the Penalized EM Algorithm to Detect DIF 49
3.2.1 Penalized Expectation-Maximization Algorithm 50
3.2.2 M-step Optimization 52
3.2.3 Selection of λ 53
3.2.4 Decorrelated Score Tests to Detect DIF 53

4 Monte Carlo Simulations 55
4.1 Study Design 55
4.1.1 The True Data Generation Model 58
4.1.2 Estimation 61
4.2 Evaluation Criteria 64

5 Results 66
5.1 Results: Hypothesis Testing 66
5.2 Results: Parameter Recovery 72
5.3 Results: Standard Error Estimates 86

6 Discussion and Conclusion 97
6.1 Summary 97
6.2 Future Studies 101

A The Coordinate Descent Algorithm 107

B Derivatives of the Observed Fisher Information 109

C Assumptions 111

References 114

List of Tables

2.1 Summary of Anchor Schemes 24

4.1 Model Parameters of the True Data Generating Model 59
4.2 Fixed λ for Each Condition 62

List of Figures

2.1 Conditional item characteristic functions of DIF items using the MNLFA model with one covariate . . . 12
2.2 Item characteristic curves and latent distributions of DIF and non-DIF items with DIF-free anchors and DIF anchors . . . 21

4.1 Conditional probability of endorsing an item using the true data generation parameters . . . 60

5.1 Type I error results of incorrectly detecting a DIF item under the null condition . . . 68
5.2 Power results of correctly detecting a DIF item under the alternative condition . . . 70
5.3 False detection rate results of incorrectly detecting a DIF item under the alternative condition . . . 72
5.4 Bias of item parameters . . . 76
5.5 Average bias of a-DIF parameters . . . 77
5.6 Average bias of d-DIF parameters . . . 78
5.7 Bias of population parameters. . . . 79
5.8 Density plot of a non-zero effect . . . 80
5.9 Variance of item parameters . . . 81
5.10 Average variance of a-DIF parameters . . . 83
5.11 Average variance of d-DIF parameters . . . 84
5.12 Variance of population parameters. . . . 85
5.13 Standard error recovery of item parameters . . . 89
5.14 Average standard error recovery of a-DIF parameters . . . 90
5.15 Average standard error recovery of d-DIF parameters . . . 91
5.16 Standard error recovery of population parameters. . . . 92
5.17 Relative efficiency of the standard error estimates for the item parameters . . . 94
5.18 Average relative efficiency of the standard error estimates for a-DIF and d-DIF parameters . . . 95
5.19 Relative efficiency of the standard error estimates for population parameters. . . . 96

List of Abbreviations and Symbols

DIF differential item functioning.

EM Expectation-Maximization.

IRT item response theory.

LASSO least absolute shrinkage and selection operator.

ML maximum likelihood.

MNLFA the moderated nonlinear factor analysis.

Chapter 1: Introduction

1.1 Statement of the Problem

A major goal of educational and psychological assessment is to create interpretable scores on a relevant construct being measured. A well-constructed test score representing a latent trait (e.g., proficiency level), should be reliable and valid. More importantly, inferences, providing individual clinical diagnoses or supporting the efficacy of a teaching method, drawn from the test score are sufficiently supported. One major threat to the validity of the intended use of a score generated from an instrument is the lack of measurement invariance or existence of *differential item functioning* (DIF). Sometimes, characteristics of a test may introduce unintended systematic score differences between individual test takers or subgroups of examinees of the same ability level, which can result in invalid score interpretation. The psychometric phenomenon, DIF or a lack of measurement invariance, occurs when equally capable examinees with different backgrounds have different probabilities of endorsing an item. Recently, the DIF analysis, originated as a tool to increase test validity in the educational measurement (Holland & Thayer, 1988), has drawn growing interests in the social science and health care research (e.g., Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Millsap, Gunn, Everson,

& Zautra, 2015; Orlando & Marshall, 2002). For example, females are found to be more likely to choose the "I cry easily" item as compared to their equally depressed male counterparts (Steinberg & Thissen, 2006). Had this item not been identified as a DIF item, depression levels of women patients could have been spuriously overestimated. Regardless of the field that DIF analysis is applied too, the existence of DIF items may favor or disfavor certain subgroups of examinees and thus jeopardizes the fairness of a test and prevents valid inferences from its scores. Therefore, DIF detection should be conducted for routine operations of psychological and educational assessment to ensure valid interpretations of test scores.

Item response theory (IRT), a collection of mathematical and statistical models, was—and still remains—the foundational modeling system in educational assessment (Embretson & Reise, 2000). Recently, IRT has been applied to many substantive domains including personality, psychpathology, and clinical outcome assessments (e.g., Reise & Revicki, 2014, patient or clinician reported outcomes). One advantage of an IRT model in conducting a DIF analysis is that it offers a formal statistical framework for assessing DIF. The core idea of an IRT model is that item responses within a test covary due to the underlying continuous latent construct that is being measured (Steinberg & Thissen, 2006). The relation between item responses and the underlying latent construct is defined by *item response functions* or *trace lines* (Lazarsfeld, 1950, p. 363). Further, each item trace line takes a specific functional form as a function of the latent construct governed by item parameters. Each value of the trace line represents the conditional probability of endorsing a particular response category given the latent construct. Ideally, the

trace line or item function should be the same for all examinees regardless of their group membership to ensure valid score interpretation. Otherwise, "it is clear that the item is biased." (Lord, 1980, p. 212)

Despite the popularity and advantages of DIF detection using IRT models, the inherent assumptions therein make practical applications of DIF analyses challenging. For example, IRT DIF detection methods usually assume the existence of *predefined* DIF-free items[1], also known as anchor items. Typically, DIF-free items are needed to anchor the latent scale so that DIF can be distinguished from the between group difference in latent trait distributions. Using the aforementioned depression example, if more women choose "I cry easily" than men do, it is difficult to distinguish whether it is because men and women response differently to this specific item (i.e., DIF) or because of the gender difference in depression (i.e., Steinberg & Thissen, 2013, latent distribution difference). Therefore, correctly specifying anchor items is crucial in correctly identifying a DIF item. Violations to the DIF-free anchor assumption may lead to inflated false detection rates in finding DIF items (W.-C. Wang, 2004; W.-C. Wang & Yeh, 2003; Woods, 2009a; Woods, Cai, & Wang, 2013). A practical obstacle is locating DIF-free anchor items without a priori information. Also, how many anchor items should be included to maintain adequate power but not increase the risk of anchor item contamination (i.e., incorrectly including DIF items as a subset of anchors) remains to be an open question. Another challenge is detecting DIF items associated with multiple grouping variables. Conventional IRT DIF detection methods typically investigate DIF with

[1]See W. Wang, Liu, and Liu (2022) for an exception.

respect to one categorical grouping variable (e.g., gender, ethnicity, SES) at a time. However, it is impossible to investigate the complex nature of DIF due to interconnected background characteristics (M. Liu, 2017) and thus it is unlikely to truly uncover the source of DIF (Shea, 2013). Failing to incorporate the interdependence of multiple grouping covariates may negatively impact the validity of score interpretation and harm the test fairness (McGraw, Lubienski, & Strutchens, 2006). For instance, gender, ethnicity, and social class related DIF analyses have been widely studied in educational standardized tests (e.g., SAT, GRE) for equal rights and test fairness considerations (Coley, 2001). A large number of these studies investigate DIF effects without considering the interconnection between the selected grouping variables. Nevertheless, ethnicity-related DIF could potentially only exist within a specific gender, SES or language group but spuriously exhibit as an ethnicity-related DIF. Had multiple grouping variables and their interaction effects not been investigated together, incorrect DIF items could be found in any of these related grouping variables. Due to the increasing necessity of creating fair and equitable test scores across all subgroups of examinees, multiple person characteristics such as SES, gender, and ethnicity should be studied simultaneously in a DIF analysis.

As a solution to these practical difficulties encountered by conventional IRT DIF methods, regularization has been applied to IRT models to detect DIF items. In particular, the use of the L_1 penalty or the least absolute shrinkage and selection operator (LASSO) has shown promising results in DIF detection due to its efficiency in variable selection (e.g., Bauer, 2017; Belzak & Bauer, 2020; Magis, Tuerlinckx, & De Boeck, 2015; Tutz & Schauberger, 2015). Specifically, the L_1 penalty has been

successfully applied to the moderated nonlinear factor analysis (MNLFA) model to identify DIF items (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020). Such a modeling framework offers greater flexibility to detect DIF for multiple grouping variables simultaneously. Specifically, item parameters and population distribution parameters, upon applying suitable link functions, can be expressed as linear or non-linear functions of multiple person characteristics. Consequently, DIF detection can be treated as a variable selection problem—non-zero coefficients are only assigned to covariates that cause DIF. The goal of the penalty procedure is to encourage sparseness as much as possible with respect to DIF parameters such that DIF parameters of anchor items are penalized to be zero while those of DIF items are non-zero. Hence, no predefined anchor items are needed as items with DIF parameters penalized to be zero are used as anchors.

Despite the flexibility of the L_1 penalty for DIF detection, several concerns need to be addressed carefully. First, the accuracy of DIF detection based on the L_1 penalty is contingent on a non-trivial assumption. In order for LASSO to select the right covariate for each item, and thus, items exhibiting true DIF, a variable selection consistency condition is needed (Zhao & Yu, 2006). In the current case, failing to meet the condition could result in mistakingly flagging non-DIF items or miss real DIF items even in large samples. Second, with a finite sample size especially those encountered in the social sciences, inferential statistics such as confidence interval estimates and p-values are extremely important to differentiate a true DIF item from a sampling error. For example, a common practice in DIF detection in educational and psychological assessments is effect size reporting after DIF is detected. In

this case, an uncertainty measure may prove to be helpful in deciding whether a DIF effect size is truly different from zero. Nevertheless, it is difficult to draw statistical inference based on the LASSO-type estimator due to shrinkage resulting from penalization (Belzak & Bauer, 2020; Huang, 2018; Lindstrøm & Dahl, 2020; Tutz & Schauberger, 2015). Lastly, LASSO creates biased parameter estimates and non-normal limiting distribution (Fu & Knight, 2000), which makes subsequent analyses difficult (Chen, Bauer, Belzak, & Brandt, 2021; Fan & Li, 2001). For example, DIF effect sizes, item parameter estimates, and standard error estimates cannot be directly used to quantify uncertainty in latent score estimates (Y. Liu & Yang, 2018).

1.2 Purpose of the Study

The current study aims to fill the gap in the literature to make inferential claims once LASSO is used. Specifically, a decorrelated score test (Ning & Liu, 2017) is proposed to detect DIF for binary response data with multiple covariates after the L_1 penalty is used. The decorrelated score test could potentially be extended to accommodate continuous and discrete response data and multidimensional latent variables. However, the simplest and the most commonly used case was considered to better demonstrate the extension of the decorrelated score test to detect DIF. Unlike the existing regularized DIF method, the decorrelated score test does not require the variable selection consistency assumption and is able to provide valid inference on DIF effects. Specifically, a sparse score vector with respect to the

focal parameter is estimated consistently so that the resulting score test statistic has an asymptotically normal reference distribution. Additionally, an asymptotic unbiased estimator can be constructed using a one-step debiased estimator using the decorrelated score function.

A Monte Carlo simulation study will be conducted to examinee the finite sample behavior of the decorrelated score test under the null condition without DIF items and the alternative condition with a mix of DIF and non-DIF items. Its performance in controlling the Type I error rate, establishing sufficient power, controlling the false detection rate in identifying a DIF item under different DIF-related conditions are compared with three methods: (1) regularization method based on LASSO selection only, (2) a naive model refitting method (Belzak & Bauer, 2020), and (3) the oracle solution assuming known anchors. Moreover, the efficacy of each method is evaluated in terms of hypothesis testing, parameter recovery, and standard error estimates.

Chapter 2: Literature Review

2.1 A General Definition of Differential Item Functioning

Lord (1980, p. 212) stated "if an item has a different item response function for one group than for another, it is clear that the item is biased." Although nowadays DIF is not equivalent to item bias (Cole, 1993), the existence of DIF items certainly needs a thorough review in case of item bias. Mathematically, DIF can be defined as follows. Considering a categorical random variable Y_{ij} representing the item response from person i, $i = 1, \cdots, n$ to item $j \in \mathcal{J} = \{1, \ldots, J\}$, define the conditional item response function, denoted as $f_j(y|\theta, \mathbf{x}) = P(Y_{ij} = y|\theta_i = \theta, \mathbf{x}_i = \mathbf{x})$, as the probability of endorsing a particular answer conditional on the person's latent ability level $\theta_i = \theta \in \mathbb{R}$ and person covariate vector $\mathbf{x}_i \in \mathbb{R}^K$. More broadly, item j exhibits DIF if

$$f_j(y|\theta, \mathbf{x}) \neq f_j(y|\theta, \tilde{\mathbf{x}}), \tag{2.1}$$

for some $\mathbf{x} \neq \tilde{\mathbf{x}} \in \mathbb{R}^K$. Equation 2.1 indicates that DIF exists when the conditional item response function f_j differs for any $\mathbf{x} \neq \tilde{\mathbf{x}}$ after controlling for the same level of latent ability. In other words, measurement invariance holds for item j, only when

$f_j(y|\theta, \mathbf{x}) = f_j(y|\theta, \tilde{\mathbf{x}})$ for all $\mathbf{x} \neq \tilde{\mathbf{x}}$. Given the definition of DIF denoted in Equation 2.1, a more general modeling framework under the IRT models, MNLFA, is firstly reviewed followed by the demonstration of special cases of the model that can be used for DIF detection. The connection between the MNLFA with DIF detection is discussed in detail.

Typically, a unidimensional IRT model specifies the discrete item responses by a unidimensional and continuous latent variable θ with a specific functional form f_j. For example, the two-parameter logistic (2PL, Birnbaum, 1968) model assumes that the item response for item j from person i, (i.e., $Y_{ij} \in \{0, 1\}$) follows a Bernoulli distribution ($Y_{ij}|\theta_i \sim \text{Bern}(P_{ij}(\theta_i))$), where P_{ij} denotes the probability of endorsing item j for person i. Incorporating the person covariate vector $\mathbf{x}_i$ into the conditional item response function (cIRF) to affect both the item response and the latent distribution offers greater flexibility in testing DIF across different levels of both continuous and categorical variables. It also permits testing DIF for multiple grouping variables as well as their interaction effects (e.g., gender $\times$ age) at the same time. Using the same notation, the cIRF f_j is written as

$$
\begin{aligned}
f_j(Y_{ij}|\theta, \mathbf{x}) &= P_{ij}^{Y_{ij}}(1 - P_{ij})^{1-Y_{ij}}, \text{ where} \\
P_{ij} = P(Y_{ij} = 1|\theta_i = \theta, \mathbf{x}_i = \mathbf{x}) &= \frac{1}{1 + \exp[-\alpha_j(\mathbf{x}) - \beta_j(\mathbf{x})\theta]},
\end{aligned} \tag{2.2}
$$

where $\alpha_j(\mathbf{x})$ and $\beta_j(\mathbf{x})$ are the item intercept and item slope functions, respectively, which can be further expressed as functions of the person covariate vector $\mathbf{x} \in \mathbb{R}^K$. Latent ability θ_i is assumed to be normally distributed: $\theta_i \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$, in

which the latent population parameters: mean (μ) and variance (σ^2) can also be expressed as functions of the person covariate vector $\mathbf{x}$.

Various model constraints have been applied to the model for identification purposes (e.g., Bauer & Hussong, 2009; Glas, 1998; Moustaki, 2003). Because the moderated non-linear factor analysis (MNLFA, Bauer & Hussong, 2009) subsumes most commonly used parametric IRT models as special cases, this general form is used in this study. Relations with more traditional IRT models (e.g., 2PL, Birnbaum, 1968; multiple group IRT, R. D. Bock & Zimowski, 1997; multiple-indicator multiple-cause, MIMIC, model, Jöreskog & Goldberger, 1975) are discussed after the general modeling framework is demonstrated.

MNLFA was originally proposed by Bauer and Hussong (2009) to conduct the integrative data analysis. Although the original model can handle multidimensional latent distributions and various types of item responses (e.g., ordinal responses), the current review only discusses the parameterization that fits the cIRF in Equation 2.2. Specifically, with binary item responses assuming a unidimensional latent variable, MNLFA models latent population parameters as

$$\mu(\mathbf{x}) = \boldsymbol{\gamma}^\top \mathbf{x} \quad \sigma^2(\mathbf{x}) = \exp(\boldsymbol{\delta}^\top \mathbf{x}), \tag{2.3}$$

where K-dimensional vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ represent effects of the person covariate vector $\mathbf{x}$ on the latent mean and variance on a logarithmic scale. These two vectors are further referred to as latent population parameters. Also, item intercept and item

slope functions are expressed as

$$\alpha_j(\mathbf{x}) = d_j + \boldsymbol{\beta}_{0j}^\top \mathbf{x}, \tag{2.4}$$

$$\beta_j(\mathbf{x}) = a_j + \boldsymbol{\beta}_{1j}^\top \mathbf{x}, \tag{2.5}$$

where d_j and a_j are the item intercept and item slope parameters. The K-dimensional vectors $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_{1j}$ stand for person covariate effects on the item intercept and item slope parameter, respectively. To detect DIF for item j, a composite hypothesis test is typically conducted to test against the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{\beta}_{0j} = \boldsymbol{\beta}_{1j} = \mathbf{0}. \tag{2.6}$$

As can be seen, $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_{1j}$ are of particular interests in DIF detection and represent d-DIF and a-DIF parameters, respectively. Figure 2.1 visualizes different types of DIF items generated from a MNLFA model. The panels in the first column are 3-dimensional cIRFs, on which there are three colored lines indicating conditional probabilities of endorsing an item at three fixed values of θ and x. The corresponding 2-dimensional plot with the same color is shown in the panels of column 2 and column 3, respectively. Top, middle, and bottom panels visualize one item with a-DIF and d-DIF (i.e., $\beta_0 \neq 0$ and $\beta_1 \neq 0$), d-DIF only (i.e., $\beta_0 \neq 0$ and $\beta_1 = 0$), and a-DIF only (i.e., $\beta_0 = 0$ and $\beta_1 \neq 0$).

If there is no DIF for all $j \in \mathcal{J}$ (i.e., $\boldsymbol{\beta}_{0j} = \mathbf{0}$, and $\boldsymbol{\beta}_{1j} = \mathbf{0}$), $\boldsymbol{\gamma} \equiv \mathbf{0}$, and $\boldsymbol{\delta} \equiv \mathbf{0}$, the MNLFA model is equivalent to a single group 2-PL model (Birnbaum,

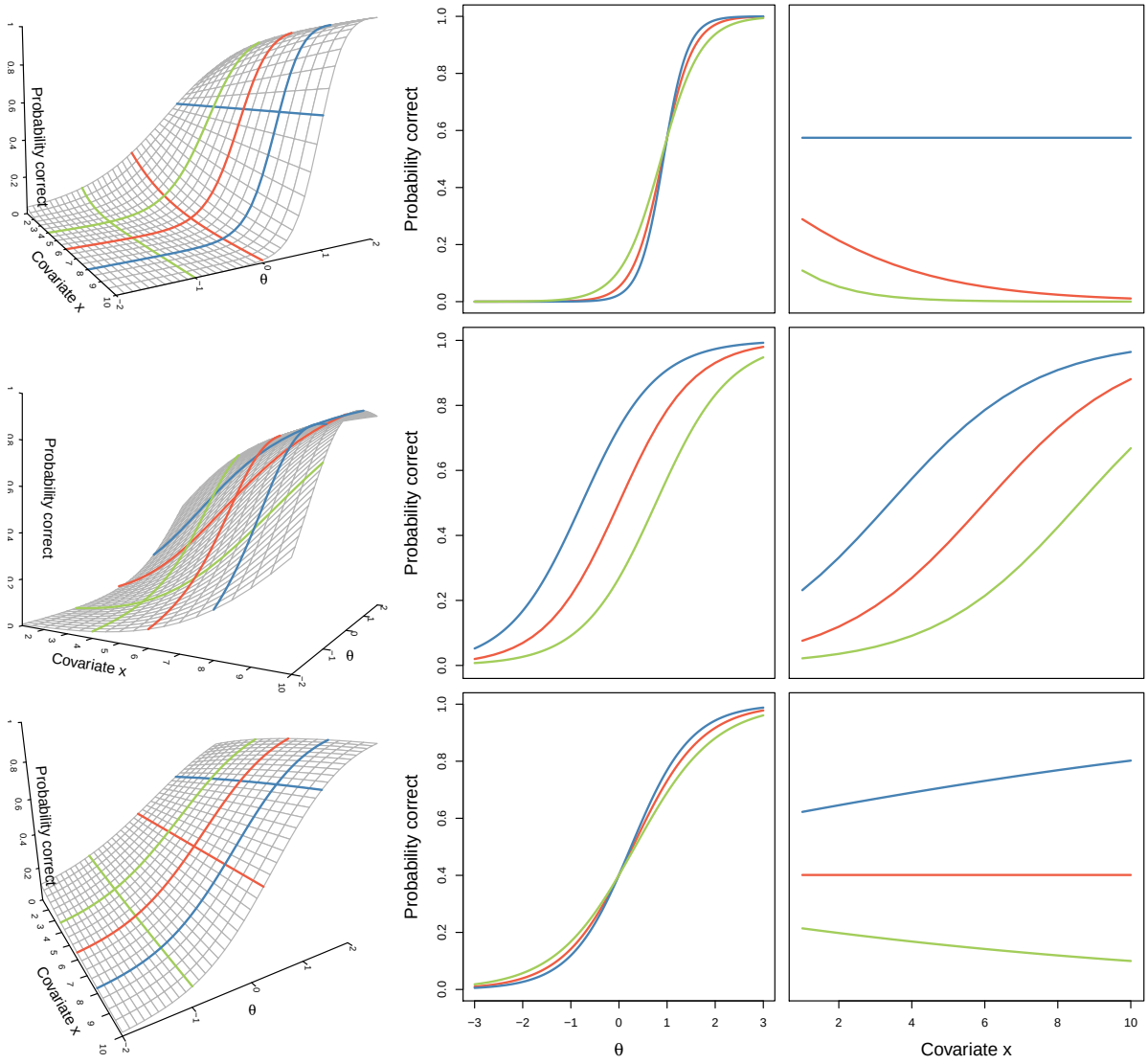


Figure 2.1. Conditional item characteristic functions (cICFs) of different types of DIF items. Each row of the graphical table standards for one type of item (from top to bottom: an item with both a-DIF and d-DIF, an item with d-DIF only, and an item with a-DIF only). The three-dimensional perspective plots on the far left of the plot are the true cICFs, three lines with three fixed values of x and θ each. Lines with same colors are showed in two-dimensional plots.

1968). Similarly, if person covariate vector $\mathbf{x}$ only contains discrete covariates, the MNLFA model is constrained to be a multiple group IRT model (R. D. Bock &

Zimowski, 1997). Another two special cases of the MNLFA model exist in the literature that are commonly used for testing a-DIF and d-DIF. One is the multiple-indicator multiple-cause (MIMIC) model (Jöreskog & Goldberger, 1975; Muthén, 1989). Unlike the MNLFA, a MIMIC model only allows person characteristics to influence the mean of the latent continuous variable thereby allowing for group mean differences. Latent ability variance is constrained (i.e., $\sigma^2(\mathbf{x}) \equiv \mathbf{1}$). Also, the original MIMIC model does not allow for interaction between the person covariate and the latent ability and thus it is unable to detect a-DIF (i.e., $\beta_j(x) \equiv \beta_j$). Another special case is the MIMIC with interaction model proposed by Woods and Grimm (2011). The MIMIC-interaction model extended the MIMIC model by adding interactions between covariates and latent distributions, and thus, the model can be used for testing a-DIF (i.e., $\beta_j(x)$ is expressed in Equation 2.5). However, unlike the MNLFA, the MIMIC-interaction model still constrains the variance of the latent ability from varying across groups by constraining $\sigma^2(\mathbf{x}) \equiv \mathbf{1}$.

Typically, a minimal constraint is needed to identify the MNLFA model. This requires that there exists at least one item for each column of $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_{1j}$ is zero/anchor. For a test with J binary response items assuming conditional independence of the random vector $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iJ})^\top$, the marginal likelihood function $f(\mathbf{y}_i|\mathbf{x}_i)$ of an observed individual response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{1J})^\top$ assuming conditional independence among item responses given the latent variable and person

covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})^\top$ is expressed as

$$f(\mathbf{y}_i|\mathbf{x}_i) = \int \prod_{j=1}^{J} f_j(y_{ij}|\theta_i, \mathbf{x}_i)\phi(\theta_i|\mathbf{x}_i)d\theta_i, \tag{2.7}$$

in which $\phi(\cdot)$ is the probably density function of a normal distribution governed by the population parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Assuming that the item response vectors $\mathfrak{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n)^\top$ of sample size n is independent and identically distributed (i.i.d), the sample likelihood function is written as

$$f_n(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{x}_i), \tag{2.8}$$

where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)^\top$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ represent observed data matrices. Subsequently, the corresponding marginal log-likelihood is expressed as

$$\ell_n(\boldsymbol{\xi}; \mathbf{Y}|\mathbf{X}) = \frac{1}{n} \log f_n(\mathbf{Y}|\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{x}_i), \tag{2.9}$$

where $\boldsymbol{\xi} = (a_1, d_1, \boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{11}^\top, \ldots, a_J, d_J, \boldsymbol{\beta}_{0J}^\top, \boldsymbol{\beta}_{1J}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)^\top$ is a collection of model parameters. Typically, item parameter are estimated using the (marginal) maximum likelihood (ML) estimation due to its asymptotic properties (e.g., consistency and asymptotic normality) under mild regularity conditions. Two types of optimization schemes are typically utilized to compute estimates: (1) a variant of Newton's algorithm (R. D. Bock & Lieberman, 1970; Haberman, 1988) and (2) the Expectation-Maximization (EM) (EM, R. Bock & Aitkin, 1982) algorithm. Optimizing log-likelihood in Equations 2.7 and 2.8 using a variant of Newton's algorithm requires

integrating out the unobserved latent variable via a numerical approximation (e.g., adaptive quadrature). Moreover, the optimization procedure involves computation of the gradient (score) and sometimes the Hessian matrix. However, as the number of items increases (e.g., 12 items in Belzak & Bauer, 2020; Bauer et al., 2020), the algorithm is computationally inefficient. Alternatively, the EM algorithm can be applied to latent variable models by treating the latent variable as missing and maximizing the complete sample log-likelihood. In doing so, EM algorithm transforms the large test level maximization problem into an item-by-item maximization problem, which greatly reduces the computational difficulty induced by the large number of items and the increasing number of covariates. In addition, the EM algorithm is not sensitive to starting values as is Newton's method.

2.2 Existing IRT DIF Detection Procedures

Section 2.1 has introduced a flexible modeling framework to detect DIF. However, greater flexibility oftentimes associates with more complexity. One of the challenges with DIF detection resulting from model complexity is the procedure to conduct hypothesis tests. In the following section, asymptotic hypothesis tests are reviewed followed by their disadvantages inherent in the IRT DIF detection method.

2.2.1 Asymptotic Hypothesis Tests

Given the parametric definition of DIF items in Equations 2.1 and 2.2, testing DIF can be considered as a nested model comparison problem, in which the null

hypothesis that item j has no DIF in Equation 2.6 is equivalent to imposing a constraint function $\mathbf{c} : \mathbb{R}^q \to \mathbb{R}^{q-p}$ on the alternative space, where p and $q\,(q > p)$ are the total number of model parameters under the null and alternative model, respectively. For example, to test whether item j has DIF, $q - p = 2K$. In general, the alternative model under the conventional DIF detection method is usually a model with minimal constraints where item parameters of anchor items are constrained to be the same across groups. Thus, an anchor is usually set at the item level instead of the item parameter level. However, as mentioned above, to identify an MNLFA model, at least one item per column of $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_{1j}$ needs to be fixed at 0. As such, anchors do not need to be set at the item level. However, for practical considerations, researchers usually assume there exists at least one DIF-free anchor item to align latent scales. Following a more traditional definition of an anchor, an anchor set can be defined as

$$\mathcal{A} = \{j = 1, \ldots, J : f_j(y|\theta, \mathbf{x}) = f_j(y|\theta, \tilde{\mathbf{x}}) \text{ for all } \mathbf{x} \neq \tilde{\mathbf{x}}\}. \tag{2.10}$$

Then, standard asymptotic tests in the nested model comparison context can be used to detect DIF one item at a time.

2.2.1.1 Conventional Hypothesis Tests

Common asymptotic tests to test DIF include the likelihood ratio test (IRT LRT, Thissen, Steinberg, & Wainer, 1993), the Wald test (e.g., Lord, 1980), and the score test (e.g., Glas, 1998). Under suitable regularity conditions, the three types

of asymptotic tests are equivalent with a common limiting distribution χ^2_{q-p} with degrees of freedom equal to the number of constraints imposed to the unrestricted model. These tests are briefly summarized here.

IRT LRT (Thissen et al., 1993) or what is commonly referred to as IRT-LR-DIF, involves the comparison of the log-likelihood functions of two models, a restricted model (or a compact model) and an unrestricted (or an augmented model). The test statistic is

$$\mathrm{T}_{\mathrm{LR}} = 2(\ell_n(\hat{\boldsymbol{\xi}}; \mathbf{Y}|\mathbf{X}) - \ell_n(\tilde{\boldsymbol{\xi}}; \mathbf{Y}|\mathbf{X})), \tag{2.11}$$

where $\tilde{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\xi}}$ are ML estimators under the null and alternative hypothesis, respectively. To test DIF for a specific item, the restricted model constrains item parameters of the testing item to be the same across groups of interests whereas the unrestricted model estimates item parameters of the testing item freely while keeping the rest of model parameters to be the same as the unconstrained version. This process is repeated one item at a time. However, one possible disadvantage of this approach is the computational burden due to repeated model fitting if the fitted model is complex.

In contrast, the Wald test (e.g., Langer, 2008; Woods, 2009b; Woods et al., 2013) reduces the computation burden by fitting only an unconstrained model. The test statistic for the Wald test takes the following form

$$\mathrm{T}_{\mathrm{W}} = \hat{\mathbf{c}}^\top \left[\dot{\hat{\mathbf{C}}} \hat{\boldsymbol{\mathcal{I}}}_n^{-1} \dot{\hat{\mathbf{C}}}\right]^{-1} \hat{\mathbf{c}}, \tag{2.12}$$

where $\hat{\mathbf{c}}$ is the component evaluated at the unrestricted/ alternative ML estimator $\hat{\boldsymbol{\xi}}$, $\dot{\mathbf{C}}$ is the Jacobian matrix $\dot{\mathbf{C}} = \dot{\mathbf{C}}(\boldsymbol{\xi}) = \frac{\partial \mathbf{c}}{\partial \boldsymbol{\xi}^T}$, and $\hat{\boldsymbol{\mathcal{I}}}_n$ denote any consistent estimator of the Fisher information. For example, if item j is tested for DIF and item 1 is used as an anchor in a two group setting, item parameter vector is shown as $\boldsymbol{\xi}_j = (a_j, d_j, \beta_{0j}, \beta_{1j})^\top$ and the model parameter vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_2^\top, \cdots, \boldsymbol{\xi}_J^\top, \gamma, \delta)^\top$ is a $q = 4J - 2$ dimensional vector, then $\mathbf{c}(\boldsymbol{\xi})$ can be constructed as the following

$$\mathbf{c}(\boldsymbol{\xi}) = \begin{bmatrix} 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a_2 \\ d_2 \\ \beta_{02} \\ \beta_{12} \\ \vdots \\ a_j \\ b_j \\ \beta_{0j} \\ \beta_{1j} \\ \vdots \\ a_J \\ b_J \\ \beta_{0J} \\ \beta_{1J} \\ \gamma \\ \delta \end{bmatrix} = \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} = \mathbf{0}. \tag{2.13}$$

Variations of the Wald test arise with various estimation methods for $\hat{\boldsymbol{\xi}}$ and subsequently $\hat{\boldsymbol{\mathcal{I}}}_n(\hat{\boldsymbol{\xi}})$. Typically, $\boldsymbol{\xi}$ is estimated by maximizing the marginal likelihood shown in Equations 2.8 and 2.9, which greatly reduces item parameter estimation bias and minimizes the underestimation of the standard error estimates. Additionally, various algorithms in approximating the large sample information matrix has greatly improved the utility of the Wald test to detect DIF. For example, the observed information matrix can be calculated using the supplemental expectation maximization algorithm (Cai, 2008) to form the improved Wald statistic (e.g., Cai, Thissen, & du Toit, 2011; Langer, 2008).

Unlike the Wald test and the likelihood ratio test, the score test does not require fitting the alternative model (i.e., unconstrained model). This can be of great value when the alternative model is considerably difficult to estimate. Among the three different testing procedures, the score test has been underutilized probably due to the inaccessibility in existing commercial psychometrics software packages (e.g., IRTPRO, Cai et al., 2011, Flexmirt, Cai, 2017, and `mirt` Chalmers, 2012). Nevertheless, DIF detection based on the score test has been extended to various IRT models (e.g., Glas, 1998; Glas & Falcón, 2003; Glas, 1999). Specifically, the score test statistic can be formalized as

$$\mathrm{T_S} = s(\tilde{\boldsymbol{\xi}}; \mathbf{Y}|\mathbf{X})^\top \boldsymbol{\mathcal{I}}_n(\tilde{\boldsymbol{\xi}})^{-1} s(\tilde{\boldsymbol{\xi}}; \mathbf{Y}|\mathbf{X}), \tag{2.14}$$

where the score function $s(\boldsymbol{\xi}; \mathbf{Y}|\mathbf{X}) = \frac{\partial \ell_n(\boldsymbol{\xi}; \mathbf{Y}|\mathbf{X})}{\partial \boldsymbol{\xi}}$ is evaluated at parameter estimates of the null model.

2.2.2 Anchor Item Issue

Despite the common utility of these asymptotic hypothesis tests to detect DIF items, in practice, the accuracy of detecting a DIF item is contingent on the availability of anchor items (also referred to as the matching variable to align the latent scale as previously mentioned in Chapter 1). By definition, anchor items are supposed to be DIF-free to avoid the inflated false detection rate in detecting a DIF item. Without DIF-free anchor items, DIF detection procedures are subject to a renown circularity issue (Berger & Tutz, 2016; Shih & Wang, 2009; Yuan, Liu, & Han, 2021). That is DIF analysis on a set of items of interest depends on the potentially contaminated anchor items. The circularity issue is demonstrated using simulated data with sample size $1,000$. Figure 2.2 displays latent distributions (the first column) and item characteristic curves (ICCs, second and third columns) of a DIF-free item and a DIF item. Item parameters are calibrated from a multiple group 2PL model. The top panel shows results of models with correctly specified anchor items whereas the bottom panel shows results of models with incorrectly specified anchor items. When anchor items are indeed DIF-free, the latent distribution of the second group (blue line) can be correctly estimated as compared to the true latent distribution of the second group (dashed blue line). Additionally, ICCs of the DIF-free item and the DIF item drawn from estimated item parameters (solid line) are overlapped with their respective true ICCs (dashed lines). However, when DIF items are falsely included as anchors, estimated and true latent distributions of the second group are considerably different from each other, which can create

spurious between-group discrepancies in the estimated item parameters for non-DIF items. Consequently, the false detection rate of a DIF test may be seriously inflated. Similarly, the between group difference of DIF item parameters might be artificially decreased, which can cause low power in identifying a DIF item. In other words, DIF detection depends on DIF free anchors, which themselves are subject to anchor contamination.

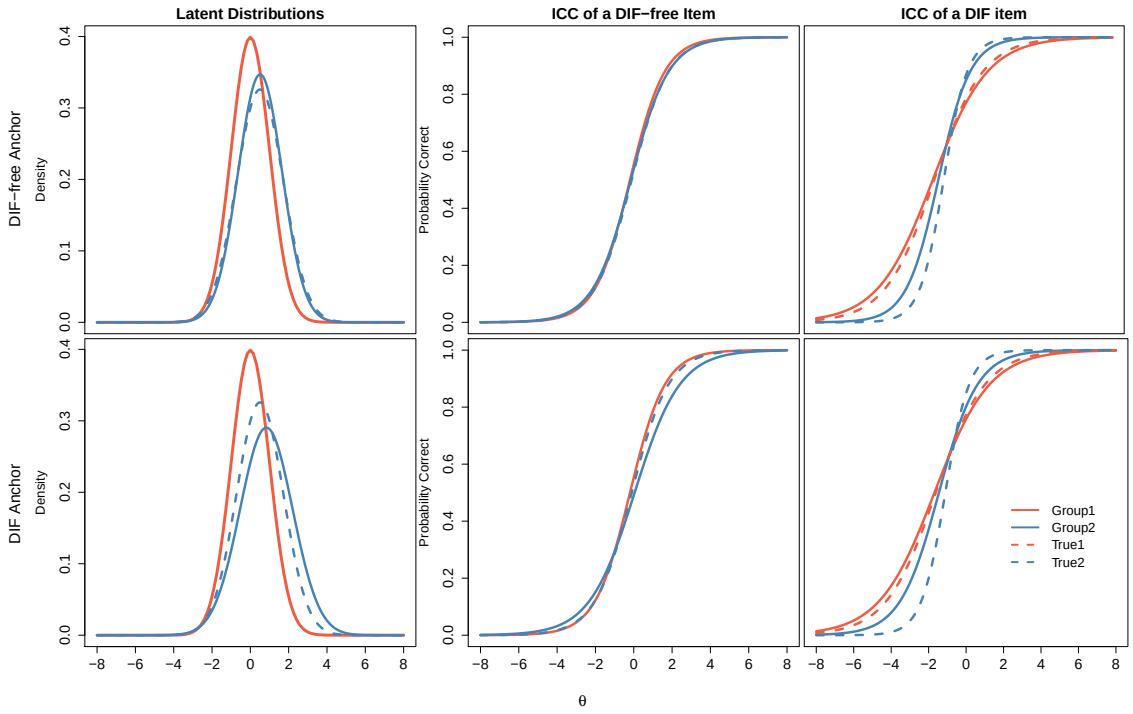


Figure 2.2. Latent distributions (first column) and item characteristic curves (ICCs, second and third column) drawn from true (dashed line) and estimated (solid line) item parameters calibrated using a two-group 2-parameter logistic (2PL) model with DIF-free anchor items (Upper panel) and anchor item contamination (lower panel) scenarios. Binary response data were simulated using a two-group 2PL model using the `mirt` package with 10 items, three of which are DIF items. A total of 1,000 item responses are generated in each group. The latent distribution of group 1 (red line) follows a standard normal distribution and that of group 2 (blue line) follows $\mathcal{N}(\mu = 0.5, \sigma^2 = 1.5)$.

2.2.3 Empirical Anchor Selection Method

Given the importance of identifying a set of anchor items a priori in DIF detection, it is crucial to take careful considerations of the anchoring scheme. There are four anchor schemes commonly used in the literature to ensure the assumption of DIF-free anchors is met. These anchor schemes are summarized in Table 2.1. Three of these anchor schemes do not depend on an anchor purification procedure, which adopts an iterative process as summarized in Section 2.2.3.1. The rest of the anchor schemes differ by existence of a priori knowledge of anchor items. The first anchor scheme is the constant anchor, which treats a fixed number of items, preferably ranging from four to six, as anchor items. Thissen et al. (1993) suggested to include "designated anchors" in the data collection design when new test items are pretested for the sole purpose of being anchors. However, as noted by the authors, the selection of designated anchors is complicated which could potentially vary by the subpopulations of interest or a particular (sub)test. Another concern with the fix-length anchor method is the length of the anchor set. An appropriate length has to maintain sufficient statistical power in detecting DIF items while also reducing the risk of anchor contamination. Studies have shown that increasing the percentage of anchor items can increase the power of detecting DIF items but also suffers from a high risk of violating the DIF-free assumption of anchor items (W.-C. Wang & Yeh, 2003; W.-C. Wang, 2004; Woods, 2009a). If these fix-length anchors are not created by design, anchor item selection strategies can facilitate DIF tests. Another anchor scheme that can be used without a priori knowledge of anchors is

the "all-other" method (see, e.g., Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; W.-C. Wang, Shih, & Sun, 2012), which treats all items except the testing item as anchors. Lastly, the all-item anchor scheme essentially treats all items as anchors by first estimating the mean and variance of latent distributions for the groups in comparison. Then, conditional on the estimated mean and variance of the latent distribution, a DIF test is conducted for each item. However, the latter two methods are not recommended as they only work when there are no DIF items or only very few items have DIF. Several studies have found that when DIF items are unbalanced (DIF items unanimously favor one group), the false alarm rate of identifying a DIF item can be inflated (W.-C. Wang, 2004; W.-C. Wang & Yeh, 2003; Woods, 2009a; Woods et al., 2013).

Table 2.1: Summary of Anchor Schemes

Anchor Schemes	Anchor Selection Strategy	Steps
Constant Anchor	Yes/ No	Items are tested for DIF one at a time by using a fixed-length of item as anchors.
All-other	No	Each item is tested for DIF using all remaining items as anchor items.
All-item	No	Latent means and variances of the latent ability for the two groups are estimated first by treating all items as anchors. Then each item is test for DIF conditional on the estimated mean and variance from the initial step.
Iterative forward and backward	Yes	*Forward:* DIF tests are conducted by looping through each item, and items are marked as DIF or non-DIF items. Then, the loop is rerun treating non-DIF items as anchors until no further non-DIF items are added in as anchors. The order of anchors is determined by an anchor selection strategy. *Backward:* DIF tests are conducted by looping through each item by treating all-other item as anchors, and items are marked as DIF and non-DIF items. Then, the loop is rerun by teasing out DIF items from the anchor set. The loop stops until no item changes its status. The DIF test order is determined by the anchor selection strategy.

When there is not a priori knowledge on DIF-free anchors, potential candidate items can be investigated empirically. These methods include ad hoc anchor selection strategies (for an overview see Kopf, Zeileis, & Strobl, 2015; Shih & Wang, 2009), robust regression methods (W. Wang et al., 2022), and regularization methods (e.g., Belzak & Bauer, 2020; Tutz & Schauberger, 2015).

2.2.3.1 Anchor Selection Strategies

In general, the anchor item selection strategy (e.g., Kopf et al., 2015; W.-C. Wang et al., 2012; W.-C. Wang & Su, 2004; Woods, 2009b) aims to find candidate anchor items via preliminary DIF tests to decide a final anchor set. Then, conditional on the final anchor set, DIF tests can be conducted for the rest of the items. These anchor selection strategies are used along with one of the anchor schemes, which require an explicitly defined anchor set (see Table 2.1). For example, the constant anchor scheme requires a fixed number of anchor items to align the scale of the latent distributions. To identify which specific candidate items can be used as anchors, anchor item selection strategies determine the order of the candidate items to be used as anchors based on different criterion functions (e.g., rank by the magnitude of the test statistics using the all-other anchor scheme). However, the estimation of the candidate model is also computationally demanding. This method also lacks theoretical justification. Alternatively, robust regression methods and regularization methods identify DIF items directly without explicitly specifying anchor items. These two approaches are discussed in detail in the following section.

2.2.3.2 Robust Regression Methods

Unlike most of the DIF detection methods mentioned above, robust regression methods are based on separate calibration where item responses from the two groups in comparison are calibrated separately using the same measurement model assuming that the latent variable follows a standard normal distribution. Then a reference line is determined by regressing one set of item parameters onto the other, preferably using robust regression methods. Finally, test statistics can be formularized as residuals from the regression line determined by the majority of the items. To demonstrate, W. Wang et al. (2022) considered the two group 2PL model with $\theta_i^{(1)} \sim \mathcal{N}(0,1)$ and $\theta_i^{(2)} \sim \mathcal{N}(\mu, \sigma^2)$. For the second group, further let $\theta_i^{(2)} = \sigma\underline{\theta}_i^{(2)} + \mu$, where $\underline{\theta}_i^{(2)} \sim \mathcal{N}(0,1)$. Subsequently, the correspondingly transformed discrimination and difficulty parameters in the second group satisfy $\underline{a}_j^{(2)} = \sigma a_j^{(2)}$ and $\underline{b}_j^{(2)} = \frac{b_j^{(2)} - \mu}{\sigma}$, implying that item parameters of the two groups, when separately calibrated, fall on a straight line if items are DIF-free. As a result, any deviation of the item from the line indicates DIF. Additionally, the authors proposed three types of test statistics along with their sampling distributions based on the ordinary least square method, the least trimmed square method, and the Tukey's Bisquare regression method. As is shown in their simulation studies, the false detection rate can be well controlled using the later two robust estimators even when there is a mix of DIF and non-DIF items. It is worth to mention that based on the fact that item parameters from separate calibration fall on the same line, Yuan et al. (2021) proposed a graphical approach to identify the reference line governed by a subset of item parameters in

the Rasch model framework. Despite its promising results in detecting DIF without prior knowledge on anchor items, this method so far is limited to binary response data and can only detect DIF items across two groups. Additional research is needed to test DIF for continuous variables or the multi-group comparison scenario.

2.2.3.3 Regularization Methods

Regularization methods, on the other hand, can detect DIF items without predefined anchors and across multiple grouping variables simultaneously. Typically regularization DIF detection methods, which are referred as reg-DIF in the sequel, rely on a more flexible modeling framework where item parameters can be influenced by person covariates as expressed in Equations 2.4 and 2.5. Then, regression coefficients of grouping covariates can be viewed as DIF parameters. Therefore, DIF detection can be treated as a variable selection problem, in which all regression coefficients are penalized to be 0 if an item is DIF-free. Various penalties (e.g., L_1 penalty) have been successfully applied to the DIF detection setting using Rasch-type models (Magis et al., 2015; Tutz & Schauberger, 2015), partial credit models (Schauberger & Mair, 2020), and SEM models (Belzak & Bauer, 2020; Huang, 2018; Liang & Jacobucci, 2020). The goal of this procedure is to encourage spareness as much as possible with respect to DIF parameters such that DIF parameters of a certain item are penalized to be zero indicating an anchor item while items with non-zero DIF parameters exhibit some degree of DIF. Consequently, no predefined anchor items are needed.

To illustrate how anchor items and DIF items are identified simultaneously, let's dive into the model estimation with penalty. The objective function (i.e., penalized negative sample log-likelihood) is defined as

$$p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X}) = -\ell_n(\boldsymbol{\xi}; \mathbf{Y}|\mathbf{X}) + \sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j), \tag{2.15}$$

where $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{0j}^\top, \boldsymbol{\beta}_{1j}^\top)^\top$ denotes all DIF parameters, $\ell_n(\boldsymbol{\xi}; \mathbf{Y}|\mathbf{X})$ is the sample log-likelihood expressed in Equation 2.9 and $p_\lambda(\cdot)$ is a penalty function gauged by the penalty weight λ. Then the local minimizer of $p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X})$ with respect to $\boldsymbol{\xi}$

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X}) \tag{2.16}$$

can be obtained by the penalized EM algorithm. Different penalties $p(\cdot)$ exist in the literature and some of which are more appropriate than others in the DIF detection context. For example, two commonly used forms of $p(\cdot)$ are the L_1 penalty or LASSO penalty in which $p_\lambda(\boldsymbol{\beta}_j) = \lambda\|\boldsymbol{\beta}_j^\top\|_1$ where $\|\|_1$ indicates the L_1 norm that sums absolute values of all elements and the L_2 penalty or the Ridge penalty where $p_\lambda(\boldsymbol{\beta}_j) = \lambda\|\boldsymbol{\beta}_j\|_2 = \lambda(\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j)$. The Ridge penalty shrinks DIF parameters toward 0 but not exactly 0. Therefore, it might not be an ideal approach to identify DIF items. As a comparison, the LASSO penalty, a more suitable approach to identify DIF items, can shrink DIF parameters directly to zero. Additionally, a combination of L_1 and L_2 penalties known as elastic net can also be used for DIF detection (Liang & Jacobucci, 2020), which is expressed as $p_{\lambda,\tau}(\boldsymbol{\beta}_j) = (1-\tau)\|\boldsymbol{\beta}_j\|_2 + \tau\|\boldsymbol{\beta}_j\|_1$.

Other penalty functions that have been proved to be effective in identifying a DIF item includes the minimax concave penalty (MCP, Zhang, 2010) and the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001), which are defined as follows:

$$\text{MCP: } p(\beta_j, \lambda, \tau) = \sum_{i=1}^{2k} \begin{cases} \lambda|\beta_i| - \frac{\beta_i^2}{2\tau} & \text{if } |\beta_i| \leq \lambda\tau \\ \frac{1}{2}\lambda^2\tau & \text{if } \lambda\tau < |\beta_i| \end{cases} \tag{2.17}$$

$$\text{SCAD: } p(\beta_j, \lambda, \tau) = \sum_{i=1}^{2k} \begin{cases} \lambda|\beta_i| & \text{if } |\beta_i| \leq \lambda \\ -\frac{|\beta_i|^2 + \lambda^2 - 2\lambda\tau|\beta_i|}{2(\tau-1)} & \text{if } \lambda < |\beta_i| \leq \lambda\tau \\ \frac{\lambda^2(\tau+1)}{2} & \text{if} \lambda\tau < |\beta_i| \end{cases}. \tag{2.18}$$

Originally, these two penalty functions were proposed to reduce bias induced by the shrinkage procedure with the L_1 penalty and were shown to have oracle properties (i.e., the estimator asymptotically performs equally well as the ML estimator assuming anchors are known). However, the extra tuning parameter τ produces a non-concave penalized log-likelihood function, which could be computational challenging. For example, even though L_1 is a special case of SCAD and MCP functions as $\tau \to \infty$, the second tuning parameter could make the estimation more challenging especially with latent variable models (Jacobucci, Grimm, & McArdle, 2016; Belzak, 2021). Belzak (2021) found that MCP can result in different parameter estimates with different starting values especially when the sample size is small and the effect size of DIF is large. Therefore, LASSO is commonly used for its variable selection

capability and computational simplicity in DIF detection.

As can be seen from the objective function, the tuning parameter, λ, regulates the magnitude of the penalty term $p(\cdot)$ and should be carefully selected. Increasing the λ encourages sparsity and will result in less DIF items while decreasing λ leads to more DIF items until the MNLFA model is no longer identified. As a result, determining the optimal λ value is critical in finding the best fitting model and thus correct DIF items. Typically, λ is selected using a pre-defined grid of λ values using some model selection criteria (e.g., cross-validation or information indices). To avoid the model identification issue, a viable solution is to start with a relatively large λ value and gradually decreases the magnitude of λ. Decreasing λ value will possiblly improve model fit by allowing DIF parameters to be non-zero until one of the model fit indexes select the optimal model. Note this model selection process is similar to the DIF method using the multiple group IRT model with the drop down anchor scheme, which starts from the most constrained model with all items function as anchors allowing latent distribution of the second group to be freely estimated. However, the regularization approach is more elegant in that the solution path of the model parameters are continuous weighted by the λ values.

Despite the good properties of using LASSO for DIF detection, multiple research studies have documented several issues. First, the accuracy of DIF detection is contingent on its variable selection consistency which requires a non-trivial assumption (Li & Jacobucci, 2021). Before applying LASSO to DIF detection, it is imperative to know how well LASSO selection is related to the truth. Typically, this is done by investigating the variable selection consistency to ensure that the right

model is selected under a sufficiently large sample. Zhao and Yu (2006) have shown that there exists a sufficient and necessary *irrepresentable condition*, which states the relation between the relevant (i.e., $\mathbf{X}_s$) and irrelevant variables (i.e., $\mathbf{X}_{s^c}$) in the true model, for variable selection consistency in linear regression models, where $s = \{j : \beta_j^* \neq 0\}$ and s^c denotes the complement of s. Further, let $\boldsymbol{\beta}_s$ and $\mathbf{X}_{s^c}$ be the corresponding subvector or submatrix. The condition states that there exists a positive vector η such that

$$|C_{s^c,s}C_{s,s}^{-1}\text{sign}(\boldsymbol{\beta}_s)| \leq \mathbf{1} - \eta, \tag{2.19}$$

where $C_{s,s} = \frac{1}{n}\mathbf{X}_s^\top\mathbf{X}_s$,$C_{s^c,s} = \frac{1}{n}\mathbf{X}_{s^c}^\top\mathbf{X}_s$. Failing to meet this condition, LASSO could select the wrong model. However, assumptions needed to establish selection consistency in DIF detection has yet been identified. For example, Belzak (2021) investigated the performance of the Reg-DIF method using the LASSO penalty in detecting DIF items using a simulation study. The false detection rate in identifying a DIF item increased a pervasive amount when the sample size increased from 500 (false detection rates= 0.03 and 0.1 in small and large DIF effect size conditions) to $2,000$ (false detection rates = 0.09 and 0.14 in small and large DIF effect size conditions) with only three person covariates in the true data generating model controlling for the rest of the manipulating factors. These contradictory results from large sample theory indicate that the model selection consistency assumption might be violated[1]. Second and more importantly, even if the Reg-DIF method using LASSO achieves

[1]or it could also be that the penalized EM algorithm Belzak (2021) in updating the M-step is incorrect.

variable selection consistency, its finite sample performance remains largely uninvestigated. Although the L_1 penalty has been proved to be very effective to produce sparsity, empirical data analyses with finite samples still observed very small effect sizes with DIF, which leaves the DIF detection decision inconclusive if the sample size is not sufficiently large. For example, Belzak (2021) investigated DIF for items of the delinquent behavior instrument using the National Longitudinal Study of Adolescent to Adult Health sample data. Comparing with the IRT LRT, the Reg-DIF method has detected additional two items with very small DIF effect sizes (i.e., for item D9 "steal $>$ \$50": effect sizes of age and male on item intercept are 0.08 and -0.07, respectively; for item D13 "steal $<$ \$50": the effect size of male on item intercept is 0.01). It is likely that these small effect sizes are caused by sampling errors and failing to take into account the sampling error may result in inflated false detection rates. For the same reason, researchers (e.g., Liang & Jacobucci, 2020; Lindstrøm & Dahl, 2020) have recommended using regularization to detect DIF as an exploratory method only and suggest that the Reg-DIF method should never replace traditional hypothesis tests for DIF detection. Considering that social science research seldom has extremely large samples, inferential statistics such as confidence interval estimates and p-values can be extremely valuable in distinguishing a true DIF item from a sampling error. Unfortunately, standard statistical inference tools such as standard error estimates and p-values cannot directly be computed using conventional ways due to biased parameter estimates resulting from the shrinkage from penalization (Belzak, 2021; Huang, 2018; Lindstrøm & Dahl, 2020; Tutz & Schauberger, 2015). Consequently, the penalized EM algorithm results in biased

DIF effect sizes and unreliable uncertainty measures.

As an easy and naive remedy of the issues above, a two-step approach is typically applied to detect DIF items using the Reg-DIF method. Specifically, step 1 fits a MNLFA model using the penalized EM algorithm to select relevant DIF items. In step 2, the MNFLA is refitted using marginal ML estimator assuming anchors selected by LASSO in step 1 (e.g., Bauer et al., 2020; Belzak & Bauer, 2020; Belzak, 2021; Tutz & Schauberger, 2015). Despite the simple way of mitigating the bias, model refitting using the marginal ML estimation is not theoretical rigorous and is unable to draw inference if DIF effects are penalized to zero. Roughly speaking, there are two problems with inferences drawn from model refitting which focuses on the sub-model selected by LASSO. The following notation is introduced to facilitate the presentation.

Given an index set $\mathcal{M} \subseteq \{1, \ldots, K\}$, define the column of the matrix $\mathbf{X}$ as $\mathbf{X}_{\mathcal{M}}$. Then, denote symbols with a superscript in a parenthesis as any values relate to the sub-model. For example, $\boldsymbol{\beta}_{0j}^{(\mathcal{M})}$ represent regression coefficients in the sub-model with $\mathbf{X}_{\mathcal{M}}$ only. Also, any quantities followed by (λ) and with a hat (e.g., $\hat{\boldsymbol{\beta}}_j(\lambda)$) indicate LASSO estimates using the full model. Finally, symbols with superscript $*$ indicate the population parameter of the true full data generating model. Then, the model refitting basically performs estimation and inference based on the sub-model containing only person covariates indexed by $\mathcal{M} = \hat{\mathcal{M}}$, which is selected by the same data. Thus, instead of making inference on the full model defined in Equation

2.2, target parameters are in the sub-model shown below

$$\text{logit}P(Y_{ij}=1|\theta_i=\theta,\mathbf{X}_{i,\mathcal{M}}=\mathbf{x}_{\mathcal{M}},\mathbf{X}_{i,\mathcal{M}'}=\mathbf{x}_{\mathcal{M}'})=-(d_j+a_j\theta+\mathbf{x}_{\mathcal{M}}\boldsymbol{\beta}_{0j}^{(\mathcal{M})}+\mathbf{x}_{\mathcal{M}'}\boldsymbol{\beta}_{1j}^{(\mathcal{M}')}\theta), \tag{2.20}$$

where $\mathcal{M}'\subseteq\{1,\cdots,K\}$ is another index set which could differ from $\mathcal{M}$. The first problem arises when the true regression coefficients in the sub-model are not the same as the coefficients in the full model (i.e., $\boldsymbol{\beta}_{0j}^{(\hat{\mathcal{M}})}\neq\boldsymbol{\beta}_{0j,\mathcal{M}}^{*}$ or $\boldsymbol{\beta}_{1j}^{(\hat{\mathcal{M}})}\neq\boldsymbol{\beta}_{1j,\mathcal{M}}^{*}$) which often might not be. Taking the intercept DIF for instance, when LASSO is used for model selection, we let $\hat{\mathcal{M}}\equiv\text{supp}(\hat{\boldsymbol{\beta}}_{0j}(\lambda))\equiv\{k:\hat{\beta}_{0jk}\neq 0\}$. The support of LASSO estimates does not necessarily equal the true non-zero set of $\boldsymbol{\beta}_{0j}$ unless variable selection consistency is met and the sample size is considerably large. Secondly, when $\mathcal{M}$ is inferred by data, inferential statistics on $\boldsymbol{\beta}_{0j}^{(\mathcal{M})}$ derived from classical statistical theory (e.g., Cai, 2008; Yuan, Cheng, & Patton, 2014) are no longer valid (see e.g., Berk, Brown, Buja, Zhang, & Zhao, 2013; Meinshausen, Meier, & Bühlmann, 2009) due to an implicit assumption that model selection is non-adaptive (Fithian, Sun, & Taylor, 2014) required by the classical theory. Intuitively, valid inference may be drawn if the true model is selected with a high probability. However, given the large sample property of variable selection consistency, the randomness from model selection due to the finite sample could invalid any inferential statistics. Chen et al. (2021) investigated the standard error accuracy using the naive model refitting procedure to detect DIF and found that the naive approach can understand the SEs by 50% as compared with the empirical standard errors. More rigorous approaches

are needed to quantify the uncertainty of $\boldsymbol{\beta}_{0j}^{(\mathcal{M})}$ in Equation 2.20 given that $\mathcal{M} = \hat{\mathcal{M}}$ is selected based on data.

2.3 Statistical inference for the L_1 penalty

While these inferential difficulties have prevented us from conducting hypothesis tests at the item level or quantifying the uncertainty of DIF effect sizes once LASSO is applied to the MNLFA model, there exists an extensive amount of recent literature on performing valid inference that overcomes the two difficulties summarized above in linear and generalized linear modeling frameworks. However, these methods seldom have been applied to latent variable models using the L_1 penalty (see an exception in a recent discussion on post-selection inference for structure equation models, Huang, 2020). The following section summarizes recent advancements in statistical inference after LASSO is used in the linear modeling framework, which then leads to discussion on the possibility of extending them to a latent variable modeling framework using the penalized EM algorithm expressed in Equation 2.15. Ideally, statistical inferential methods that can be extended to help DIF detection possess the following characteristics: (1) within the linear and generalized linear modeling framework, it is theoretically justified[2] , (2) the method can be easily applied to the penalized EM algorithm where the loss function is not globally convex, 3) it has a justified and computational efficient hypothesis test to identify DIF at item and individual parameter levels even if the focal DIF parameters are penal-

[2]For this reason bootstrap methods within the penalized likehood estimator are excluded from this review.

ized to zero, and 4) the method provides unbiased estimator which is preferably asymptotic efficient so that the DIF effect size and its uncertainty measure can be estimated directly. Note that in the current setting, 3) and 4) are critically important as parameters penalized to be zero are anchors. It is important for the inferential method to provide some intuition on the quality of anchors.

Now, shifting the focus to linear regression models. Consider a multiple linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ is an $n \times p$ design matrix, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\top$ is a vector of $i.i.d$ random variables each of which has a mean of 0 and variance of σ^2, $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. Similarly as before, an index set $\mathcal{M} \subseteq \{1, \ldots, p\}$ represents any sub-models. Generally speaking, these inferential methods of the LASSO estimator can be categorized into two classes including a sub-model view and a full-model view. The distinction between the two are critical in defining the meaning of the coefficients in the sub-model and identifying the target of the inference. Berk et al. summarized the distinction between the two as

Sub-model view. Each sub-model stands on its own and thus has its own parameters (i.e., the regression coefficient β_j of the same covariate X_j in different sub-models have different true parameters). The full model is just a special case of a sub-model when $\mathcal{M} = \{1, \cdots, p\}$. The deselected parameters are not zero but non-existent. Therefore, the target parameter of interest is $\boldsymbol{\beta}^{(\mathcal{M})}$, which is usually different from $\boldsymbol{\beta}^*_M$. Note that the true regression coefficients in the sub-model is defined as $\boldsymbol{\beta}^{(\mathcal{M})} \equiv (\mathbf{X}^\top_{\mathcal{M}}\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}^\top_{\mathcal{M}}\boldsymbol{\mu} = \arg\min_{\beta'} \|\boldsymbol{\mu} - \mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}'\|_2$, where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{Y})$. The reason for focusing on the selected model is that all models are wrong and one

cannot have access to all related variables.

Full model view. Parameters of the full model describe the true data generating mechanism of the response variable. The deselected parameters are zeros. Regression coefficients of the selected variables are estimated by fixing the deselected parameters to be zero. Inference focuses on β_j^*.

2.3.1 Sub-model Inference

Recall that under the sub-model view, each sub-model stands on its own so each variable appears in 2^{p-1} models. Thus, in total, there are $p^{2^{p-1}}$ possible well-defined population parameters:

$$\{\beta_j^{(\mathcal{M})} : \mathcal{M} \subseteq \{1, \ldots, p\},\, j \in \mathcal{M}\}. \tag{2.21}$$

However, post-selection inferences can only be drawn for the parameter $\beta_j^{(\hat{\mathcal{M}})}$ in the selected model defined by $\hat{\mathcal{M}}$, which can be an issue if $j \notin \mathcal{M}$. For instance, suppose a confidence interval $C_j^{(\hat{\mathcal{M}})}$ is constructed for $\beta_j^{(\hat{\mathcal{M}})}$ at nominal level α

$$P(\beta_j^{(\hat{\mathcal{M}})} \in C_j^{(\hat{\mathcal{M}})}) \geq 1 - \alpha. \tag{2.22}$$

The event inside the probability is not well-defined because $\beta_j^{\mathcal{M}}$ is not defined if $j \notin \mathcal{M}$. As a solution, a confidence interval can be constructed if and only if model

$\mathcal{M}$ is selected and hence $C_j^{(\hat{\mathcal{M}})}$ is defined as

$$P(\beta_j^{(\hat{\mathcal{M}})} \in C_j^{(\hat{\mathcal{M}})} | \hat{\mathcal{M}} = \mathcal{M}) \geq 1 - \alpha. \tag{2.23}$$

To build a conditional interval, the following two approaches can be used.

2.3.1.1 Sample-splitting Technique

One way is to conduct sample-splitting to break the conditional relation between model selection and hypothesis testing. For example, Wasserman and Roeder (2009) proposed to split the data into two parts and conduct model selection use one and hypothesis testing using the other. Although this sample-splitting technique is easy to perform and can asymptotically control the Type I error rate, an arbitrary split can result in drastic changes if the split were conduct differently (Meinshausen et al., 2009). Additionally, splitting data into halves can potentially reduce the power in detecting a true effect due to the reduced sample size. In the same vein, Meinshausen et al. (2009) extended the single split idea to multiple sample-splits and then aggregated the resulting p-values from multiple splits, which increases the power while reducing the false inclusion rate as compared with the single split procedure. Nevertheless, sample splitting could be relatively less practical to be extended to the Reg-DIF method considering the estimation challenges and the sample size requirement with latent variable models. Moreover, aggregating results (e.g., confidence interval, p-values, or standard error estimates for the estimated effect sizes) especially for those covariates that are not included in the model across multiple

splits is challenging.

2.3.1.2 Exact Post-selection Inference

Alternatively, sample-splitting can be avoided by directly characterizing the conditional distribution of the LASSO-selected estimator given the event that some covariates have been selected. Lee, Sun, Sun, and Taylor (2016) and Tibshirani, Taylor, Lockhart, and Tibshirani (2016) have studied the geometry of LASSO and characterize the event of $\hat{\mathcal{M}} = \mathcal{M}$ as a union of polyhedra, which characterizes the conditional distribution of the LASSO-estimator under linear and generalized linear modeling frameworks. Specifically, the conditional distribution of the regression coefficient given the selected model is truncated Gaussian. Subsequently, a test statistic can be formularized and confidence interval can be constructed by inverting the cumulative distribution function. An interactive visualization of LASSO partitions into polyhedra based on the selected model is presented in Lee et al. (2016) and Harris (2014). However, the downside of the exact post-selection inference is that the validity of the confidence interval to ensure an exact coverage rate of $1-\alpha$ is based on a fixed tuning parameter (i.e., λ). The randomness induced by optimizing λ can create an infinite confidence interval (Taylor & Tibshirani, 2018). A simulation study by Huang (2020) reported that 0.2% of the confidence intervals constructed based on the polyhedral method are infinite.

Alternatively, Berk et al. (2013) proposed a simultaneous inference approach which considers the event for all $j \in \hat{\mathcal{M}}$. Post-selection inference yields valid infer-

ence by considering all possible model selection procedures that could have produced the given sub-model. Thus, the method controls for the familywise error rate for all $j \in \hat{\mathcal{M}}$ as

$$P(\beta_j^{(\hat{\mathcal{M}})} \in C_j^{(\hat{\mathcal{M}})} \text{ for all } j \in \hat{\mathcal{M}}) \geq 1 - \alpha. \tag{2.24}$$

However, as noted by the author, controlling for the familywise error rate could result in conservative selection results and wider confidence interval length. Additionally, in order to satisfy Equation 2.24 for any $\hat{\mathcal{M}}$, a selection procedure dependent constant needs to be estimated. This constant functions the same as a critical value with a known sampling distribution. As can be imagined, the estimation of the constant is non-trivial. Although Berk et al. (2013) provide a numerical approximation of the constant, the derivation depends on the particular linear equation. Extension to latent variable models can be challenging (Huang, 2020).

To summarize, post-selection inference from the sub-model view could add great value to quantify the uncertainty of the estimated model parameters using the penalized EM algorithm if the technical and computational difficulties are resolved. Nevertheless, as mentioned in the beginning of Section 2.3, special interests in drawing inference on anchors direct the attention to the full model view discussed below.

2.3.2 Full Model Inference

Notably different from the sub-model view, statistical inference on the full model or the true data generating model might be more useful in quantifying DIF effect sizes and their uncertainties especially when the target DIF effect size is penalized to be zero. Methods of this type usually depend on a debiased step or desparsifying step to construct asymptotically unbiased estimators for $\boldsymbol{\beta}$ so that asymptotic normality holds. For example, the debiased estimator can take the following form

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) + \frac{\hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda))}{n} \tag{2.25}$$

where $\hat{\boldsymbol{\beta}}(\lambda)$ is a vector of LASSO estimates that minimizes the average sum of square residuals with the L_1 penalty and $\hat{\Theta}$ is some form of approximation to the inverse of the crossproduct of the design matrix (i.e., $(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}$). It then can be proved that the debiased parameter in Equation 2.25 converges to a normal distribution with the mean at the true value and the variance covariance matrix as a function of Θ (Javanmard & Montanari, 2013, 2014; Van de Geer, Bühlmann, Ritov, & Dezeure, 2014; Zhang & Zhang, 2014). However, most of the debiased methods mentioned here are often limited to linear models, which makes the extension to the IRT models difficult where the loss function is more complicated than the sum of squares loss. The only exception is Van de Geer et al.'s debiased method which utilizes the inversion of the Karush–Kuhn–Tucker (KKT) condition to construct

the debiased version of the parameter estimates in which LASSO estimators are estimated using the penalized ML estimation method.

Different from those Wald-type tests, Ning and Liu (2017) proposed a decorrelated score test which directly conducts a hypothesis test based on a decorrelated score function. Additionally, an asymptotically unbiased parameter estimator can be obtained using a one-step correction with the decorrelated score function. Compared with the other debiased estimators which can be tested using Wald-type tests, the author argues that the score test can be more powerful with small sample sizes. What's more appealing of this method is its possibility to be extended to DIF detection with IRT models due to its general theorems derived based on the penalized M-estimator. Unlike the Wald-type debiased LASSO estimator, the theoretical results of Ning and Liu (2017) can be applied to the class of penalized M-estimators of which the penalized marginal ML estimator used for IRT models is a special case. This point will be revisited and discussed in much detail in Chapter 3. It should be noted that the one-step debiased estimator using the decorrelated score function is eventually the same as the debiased LASSO estimator of Van de Geer et al. (2014).

2.3.3 Summary

In a nutshell, Chapter 2 reviewed existing IRT DIF detection methods within which conventional hypothesis tests (i.e., the LRT, the Wald test, and the score test) are commonly used regardless of the specific IRT model. Issues such as anchor item identification inherent in the IRT DIF detection method can result in spurious

results. Incorrect anchors may cause inflated Type I error rates in falsely identifying DIF items and are likely to reduce the power to identify true DIF items. Empirical anchor selection approaches such as anchor selection strategies, methods using robust regression models, and regularization methods are possible solutions to identify true anchors. DIF detection using regularization offers the most general solution of the three in the sense that it can detect DIF items that function differently across multiple person characteristics simultaneously without predefined anchor items. An additional benefit of this method is that it could possibly detect DIF items caused by the interconnection between grouping variables that could otherwise be ignored by other single covariate DIF methods. Unfortunately, the difficulty in drawing inference limits its usage as a primary method to detect DIF items (Liang & Jacobucci, 2020, 2020). Current naive solution of the issue by model refitting using the ML estimator after the LASSO selection is likely to be wrong and hence should not be used (Huang, 2020).

Due to the paucity of DIF literature on the issue, the second half of the review focuses on inferential approaches using the LASSO estimator in the general statistical literature. The goal is to find theoretical justified inferential procedures that can be applied to the regularized DIF approach. These inferential methods are categorized into the sub-model view and the full model view and their primary distinction is the inference target. Although sub-model view methods (e.g., sample-splitting methods and exact post selection inference methods) are valid solutions, the current proposal adopts the full model view due to its ability to draw inference on parameters of the deselected variables. The deselected variables or more

accurately deselected parameters function as anchors in the DIF detection context. Inference on anchors essentially provides an evaluation of the quality of the anchors. Additionally, the decorrelated score test is derived from a more general framework as compared with other debiased methods. For these reasons, the current study focuses on extension the decorrelated score test and its one-step debiased estimator to the MNLFA framework with the LASSO penalty to detect DIF items.

Chapter 3, hereafter, introduces the decorrelated score test and its one-step debiased estimator developed by Ning and Liu (2017). Additionally, the general theory and its assumptions in characterizing the limiting distribution of the descorelated score test are established. Lastly, application of the general theory into the MNLFA modeling framework to detect DIF will be discussed in greater detail.

Chapter 3: Theory

3.1 A General Theory for Decorrelated Score Function

Considering n independent and identically distributed multivariate vectors, $\mathfrak{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n)^\top$, following a statistical model $\mathcal{P} = \{P_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in \Xi\}$, where $\boldsymbol{\xi}$ is a $d-$dimensional vector of unknown parameters and $\Xi \subset \mathbb{R}^d$ is the parameter space. Partition model parameters $\boldsymbol{\xi}$ into $\boldsymbol{\xi} = (\boldsymbol{\psi}, \boldsymbol{\eta}^\top)^\top$, where $\boldsymbol{\psi}$ is the d_0-dimensional vector of focal parameters and $\boldsymbol{\eta}^\top$ stands for $d_1 = (d - d_0)-$dimensional vector of nuisance parameters. Given the loss function $\ell(\boldsymbol{\xi}, \mathbf{Y}) = -\ell_n(\boldsymbol{\xi}, \mathbf{Y})$, define $\boldsymbol{\mathcal{I}} = \mathbb{E}_{\boldsymbol{\xi}}(\nabla^2 \ell(\boldsymbol{\xi}, \mathbf{Y}))$. Let $\boldsymbol{\xi}^*$ be the true value of $\boldsymbol{\xi}$. Similarly, denote $\boldsymbol{\mathcal{I}}^* = \mathbb{E}_{\boldsymbol{\xi}^*}(\nabla^2 \ell(\boldsymbol{\xi}^*, \mathbf{Y}))$. Note that for the rest of the study, d is fixed. Let's first consider the case when the model is identified. For example, an MNLFA model with sufficiently large subset of known anchors. The asymptotic normality of the ML estimator ($\hat{\boldsymbol{\xi}} = \text{argmin}_{\boldsymbol{\xi} \in \Xi} \ell(\boldsymbol{\xi}, \mathbf{Y})$) follows from a first-order Taylor series expansion of the score function at $\boldsymbol{\xi}^*$,

$$\mathbf{0} = \nabla_{\boldsymbol{\xi}} \ell(\hat{\boldsymbol{\xi}}, \mathbf{Y}) = \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}^*, \mathbf{Y}) + \nabla^2_{\boldsymbol{\xi}, \boldsymbol{\xi}} \ell(\boldsymbol{\xi}^*, \mathbf{Y})(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) + R_n$$

where the remainder term is of the form $R_n = \frac{1}{2}(\hat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}})^T \nabla^3_{\boldsymbol{\xi},\boldsymbol{\xi}} \ell(\bar{\boldsymbol{\xi}}, \mathbf{Y})(\hat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}})$ and $\bar{\boldsymbol{\xi}}$ is in between $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}^*$. By rearranging the equation,

$$\begin{aligned}
&\sqrt{n}\,(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) = -\sqrt{n}\,\boldsymbol{\mathcal{I}}^{*-1} \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}^*, \mathbf{Y}) + \sqrt{n} R_n \\
\Leftrightarrow\ &\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^* \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* \end{pmatrix} = -\sqrt{n} \begin{pmatrix} \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi},\boldsymbol{\psi}} & \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi},\boldsymbol{\eta}} \\ \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\eta},\boldsymbol{\psi}} & \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\eta},\boldsymbol{\eta}} \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{\boldsymbol{\psi}} \ell(\boldsymbol{\xi}^*, \mathbf{Y}) \\ \nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{\xi}^*, \mathbf{Y}) \end{pmatrix} - \sqrt{n} R_n,
\end{aligned}$$

where subscripts of $\boldsymbol{\mathcal{I}}^*$ are corresponding partitions of the matrix. Drop $\mathbf{Y}$ from the rest of the notations. As $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta},\boldsymbol{\eta}}$ is invertible, using the block inverse formula, it follows that

$$\sqrt{n}\,(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) = -\sqrt{n}\,\boldsymbol{\mathcal{I}}^{*-1}_{\boldsymbol{\psi}|\boldsymbol{\eta}} s(\boldsymbol{\xi}^*) + \sqrt{n} R_n, \quad R_n = o_p(1/\sqrt{n})$$

where $\mathbf{s}(\boldsymbol{\xi}^*) = \nabla_{\boldsymbol{\psi}} \ell(\boldsymbol{\xi}^*) - \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi}\boldsymbol{\eta}} \boldsymbol{\mathcal{I}}^{*-1}_{\boldsymbol{\eta}\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{\xi}^*)$ and $\boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi}|\boldsymbol{\eta}} = \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi}\boldsymbol{\psi}} - \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\psi}\boldsymbol{\eta}} \boldsymbol{\mathcal{I}}^{*-1}_{\boldsymbol{\eta}\boldsymbol{\eta}} \boldsymbol{\mathcal{I}}^*_{\boldsymbol{\eta}\boldsymbol{\psi}}$ are the efficient score and efficient information, respectively. Note that efficient score can be interpreted as the projection of the $\nabla_{\boldsymbol{\psi}} \ell(\boldsymbol{\xi})$ to the orthogonal complement of the score function with respect to the nuisance parameters (Vaart, 1998). Under H_0 : $\boldsymbol{\psi}^* = \mathbf{0}$, it holds that $n \nabla_{\boldsymbol{\psi}}(\mathbf{0}, \hat{\boldsymbol{\eta}})^\top [\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\psi}|\boldsymbol{\eta}}]^{-1} \nabla_{\boldsymbol{\psi}}(\mathbf{0}, \hat{\boldsymbol{\eta}}) \xrightarrow{\mathcal{D}} \chi^2_{d_0}$, where $\hat{\boldsymbol{\eta}} = \text{argmin}_{\boldsymbol{\eta}} \ell(\mathbf{0}, \boldsymbol{\eta})$ is constrained parameter estimates. It is straightforward to see that the estimated efficient score ($\mathbf{s}(\mathbf{0}, \hat{\boldsymbol{\eta}})$) equals $\nabla_{\boldsymbol{\psi}}(\mathbf{0}, \hat{\boldsymbol{\eta}})$ due to the fact that $\nabla_{\boldsymbol{\eta}} \ell(\mathbf{0}, \hat{\boldsymbol{\eta}}) = \mathbf{0}$.

However, when the model is not identified, $\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta},\boldsymbol{\eta}}$, in general, is not invertible. Intuitively, there exists redundant parameters in the model. A natural extension of the above case when the model is not identified is estimating an "efficient score"

so that the influence of entries of the redundant parameters in the nuisance score is minimized. Following the same logic as the efficient score, a sparse score can be estimated by projecting $\nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\xi})$ to the orthogonal complement of a low-dimensional subspace spanned by the non-redundant part of the nuisance score vector. Therefore, a sparse vector/matrix is needed to find the best sparse linear combination of $\nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\xi})$ to approximate $\nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\xi})$. Mathematically, the decorrelated score, or the extension of the efficient score when the model is not identified, has no correlation with the score function with respect to the nuisance score (i.e., $\mathbb{E}(\mathbf{s}(\boldsymbol{\xi})^\top \nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\xi})) = \mathbf{0}$). Geometrically, the decorrelated score has the same interpretation which is the projection of $\nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\xi})$ to the orthogonal complement of the linear space spanned by the nuisance score function $\nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\xi})$. Following this intuition, define the decorrelated score function as

$$\mathbf{s}(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\xi}) - \mathbf{W}^\top \nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\xi}), \text{ where } \mathbf{W}^\top = \boldsymbol{\mathcal{I}}_{\boldsymbol{\psi}\boldsymbol{\eta}}\boldsymbol{\mathcal{I}}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1} \in \mathbb{R}^{d_0 \times d_1}. \tag{3.1}$$

To find the projection of $\nabla_{\boldsymbol{\psi}}\ell(\boldsymbol{\xi})$ to the orthogonal complement of the linear space span by the nuisance score function $\nabla_{\boldsymbol{\eta}}\ell(\boldsymbol{\xi})$, we can estimate the sparse $\mathbf{W}$ using Algorithm 1. As can be seen, the key is to estimate a sparse matrix $\hat{\mathbf{W}} = (\hat{\mathbf{W}}_{*1}, \cdots, \hat{\mathbf{W}}_{*d_0})$ column by column to construct the decorrelated score $(\mathbf{s}(\boldsymbol{\psi}, \boldsymbol{\eta}))$ so that the additional and redundant parameters do not influence the decorrelated score.

Algorithm 1 Estimated the decorrelated score function

Require: Negative sample log-likelihood $\ell(\boldsymbol{\psi}, \boldsymbol{\eta})$, penalty function $p_\lambda()$, and tuning parameters λ and λ'.

1: Estimate $\hat{\boldsymbol{\xi}}$ using penalized ML as in equation 2.15 and partition $\hat{\boldsymbol{\xi}}$ into $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\psi}}^\top, \hat{\boldsymbol{\eta}}^\top)^\top$

2: Estimate $\mathbf{W}$ column by column

$$\hat{\mathbf{W}}_{*j} = \arg\min_{\mathbf{w}_j} \frac{1}{2n} \sum_i^n \{\nabla_{\psi_j} \ell_i(\hat{\boldsymbol{\xi}}) - \mathbf{W}_{*j}^\top \nabla_{\boldsymbol{\eta}} \ell_i(\hat{\boldsymbol{\xi}})\}^2 + p_{\lambda'}(\mathbf{W}_{*j}) \tag{3.2}$$

3: Calculate the estimated descorrelated score function using

$$\hat{\mathbf{s}}(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}}) = \nabla_{\boldsymbol{\psi}} \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}}) - \hat{\mathbf{W}}^\top \nabla_{\boldsymbol{\eta}} \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}}) \tag{3.3}$$

return $\hat{\mathbf{s}}(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}})$

Further the Dscore test statistic can be constructed

$$\hat{T}_{Dscore} = n\hat{\mathbf{s}}(\mathbf{0}, \boldsymbol{\eta})^\top [\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\psi}|\boldsymbol{\eta}}]^{-1} \hat{\mathbf{s}}(\mathbf{0}, \boldsymbol{\eta}), \tag{3.4}$$

where $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\psi}|\boldsymbol{\eta}} = \nabla^2_{\boldsymbol{\psi},\boldsymbol{\psi}} \ell(\hat{\boldsymbol{\xi}}) - \hat{\mathbf{W}}^\top \nabla^2_{\boldsymbol{\eta},\boldsymbol{\psi}} \ell(\hat{\boldsymbol{\xi}})$. Under some technical assumptions[1], it can be proved that

$$\hat{T}_{Dscore} \xrightarrow{\mathcal{D}} \chi^2_{d_0}. \tag{3.5}$$

Moreover, a one-step asymptotical unbiased estimator $\tilde{\boldsymbol{\psi}}$ can be constructed using a single Newton step as

$$\tilde{\boldsymbol{\psi}} = \hat{\boldsymbol{\psi}} - \hat{\boldsymbol{\mathcal{I}}}^{-1}_{\boldsymbol{\psi}|\boldsymbol{\eta}} \hat{\mathbf{s}}(\hat{\boldsymbol{\xi}}) \tag{3.6}$$

[1]These assumptions are documented in Appendix C and are not yet verified in latent variable models. The intuition of the assumption verification is discussed in Appendix C

and with assumptions C.0.1 to C.0.4 documented in Appendix C, Ning and Liu (2017) showed that

$$\sqrt{n}(\tilde{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\mathcal{I}}^{*-1}_{\boldsymbol{\psi}|\boldsymbol{\eta}}). \tag{3.7}$$

Subsequently, the $(1-\alpha)\times 100\%$ confidence interval can be constructed for a linear combination of $\boldsymbol{\psi}^*$ (i.e., $\mathbf{c}^\top\boldsymbol{\psi}^*$ where $\mathbf{c}^\top$ is a d_0-dimensional constant vector) as $[\tilde{\boldsymbol{\psi}} - n^{-1/2}\Phi^{-1}(1-\alpha/2)(\mathbf{c}^\top\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\psi}|\boldsymbol{\eta}}\mathbf{c})^{-1/2}, \tilde{\boldsymbol{\psi}} + n^{-1/2}\Phi^{-1}(1-\alpha/2)(\mathbf{c}^\top\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{\psi}|\boldsymbol{\eta}}\mathbf{c})^{-1/2}]$.

3.2 Extension to the Penalized EM Algorithm to Detect DIF

So far, the discussion has been focused on the approximation of the sparse matrix $\mathbf{W}$ and the decorrelated score function $\hat{\mathbf{s}}(\boldsymbol{\psi}, \hat{\boldsymbol{\eta}})$. In the following section, penalized ML estimation will be discussed in detail to obtain the initial parameter estimates ($\hat{\boldsymbol{\xi}}$, i.e., step 1 of Algorithm 1). Recall that the penalized sample log-likelihood (previously defined in Equation 2.15) can be written as

$$p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n}\log f(\mathbf{y}_i|\mathbf{x}_i) + \sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j),$$

in which $f(\mathbf{y}_i|\mathbf{x}_i)$ is the marginal likelihood of $\mathbf{y}_i$ written in Equation 2.7 and $p_\lambda(\boldsymbol{\beta}_j) = \lambda||\boldsymbol{\beta}_j||_1$ is the L_1 penalty.

3.2.1 Penalized Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm (R. Bock & Aitkin, 1982) is used to find the local minimizer of $p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X})$ with respect to $\boldsymbol{\xi}$. The basic idea of the penalized EM algorithm is to repeatedly approximate the upper bound of the penalized negative sample log-likelihood $p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X}) \leq q_n(\boldsymbol{\xi}; \boldsymbol{\xi}^{(r)}), r = 0, 1, 2, \cdots$ defined in Equation 3.8 at each iteration (E-step) and obtain the optimizer (M-step). As a result, the alternation between the E-step and the M-step produces a sequence of parameter updates $\boldsymbol{\xi}^{(r)}$. The final parameter estimates $\hat{\boldsymbol{\xi}}$ is referred to as the penalized ML estimator. A critical property of the EM algorithm is that the parameter estimate updates the penalized negative sample log-likelihood in a non-increasing fashion (i.e., $p_n(\boldsymbol{\xi}^{(r+1)}; \mathbf{Y}, \lambda|\mathbf{X}) - p_n(\boldsymbol{\xi}^{(r)}; \mathbf{Y}, \lambda|\mathbf{X}) \leq q_n(\boldsymbol{\xi}^{(r+1)}, \boldsymbol{\xi}^{(r)}) - q_n(\boldsymbol{\xi}^{(r)}, \boldsymbol{\xi}^{(r)}) \leq 0$ where $q_n(\boldsymbol{\xi}, \boldsymbol{\xi}^r)$ can be viewed as the upper bound of $p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda|\mathbf{X})$). Specifically, at iteration r, $q_n(\boldsymbol{\xi}; \boldsymbol{\xi}^{(r)})$ is defined as

$$
\begin{aligned}
q_n(\boldsymbol{\xi}; \boldsymbol{\xi}^{(r)}) &= \mathbb{E}_{(\theta_i|\boldsymbol{\xi}^{(r)}, \mathbf{Y}, \mathbf{X})} \left[p_n(\boldsymbol{\xi}; \mathbf{Y}; \theta_i|\mathbf{X}) \right] && (3.8) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{(\theta_i|\boldsymbol{\xi}^{(r)}, \mathbf{y}_i, \mathbf{x}_i)} \left[\log f(\boldsymbol{\xi}; \mathbf{y}_i; \theta_i|\mathbf{x}_i) \right] + \sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j) && (3.9) \\
&= -\frac{1}{n} \sum_{i}^{n} \int f(\theta_i|\boldsymbol{\gamma}^{(r)}, \boldsymbol{\delta}^{(r)}, \mathbf{y}_i, \mathbf{x}_i) \Big(\log \left(f(\mathbf{y}_i|\theta_i, \boldsymbol{\xi}, \mathbf{x}_i) f(\theta_i|\boldsymbol{\xi}, \mathbf{x}_i) \right) \Big) d\theta_i \\
&\quad + \sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j) && (3.10)
\end{aligned}
$$

Dropping $\sum_{j=1}^{J} p_\lambda(\boldsymbol{\beta}_j)$ for simplicity, the first term of the right hand side (RHS) equals

First term of the RHS

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{i=1}^{n}\int f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)\Big(\sum_{j=1}^{J}\log f_j(\mathbf{y}_{ij}|\theta_i,\boldsymbol{\xi}_j,\mathbf{x}_i)+\log f(\theta_i|\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x}_i)\Big)d\theta_i\\
&= -\frac{1}{n}\sum_{i=1}^{n}\int f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)\Big(\sum_{j=1}^{J}\log f_j(\mathbf{y}_{ij}|\theta_i,\boldsymbol{\xi}_j,\mathbf{x}_i)\Big)d\theta_i\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)\Big(\log f(\theta_i|\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x}_i)\Big)d\theta_i\\
&= -\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\int f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)\Big(\log f_j(\mathbf{y}_{ij}|\theta_i,\boldsymbol{\xi}_j,\mathbf{x}_i)\Big)d\theta_i\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)\Big(\log f(\theta_i|\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x}_i)\Big)d\theta_i.
\end{aligned}
$$

Further, denote $\theta_{iq}, q = 1, \cdots, Q$ and w_{iq} as quadrature nodes and weights for person i, respectively. The intractable integral in the $q_n(\boldsymbol{\xi};\boldsymbol{\xi}^{(r)})$ can be approximated by summations on this quadrature grid as

$$
\begin{aligned}
q_n(\boldsymbol{\xi};\boldsymbol{\xi}^{(r)}) \approx -\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{q=1}^{Q} e_{iq}^{(r)}\Big(\log f_j(\mathbf{y}_{ij}|\theta_{iq},\boldsymbol{\xi}_j,\mathbf{x}_i)\Big)+\sum_{j}^{J} p_\lambda(\boldsymbol{\beta}_j) \qquad (3.11)\\
-\frac{1}{n}\sum_{i=1}^{n}\sum_{q=1}^{Q} e_{iq}^{(r)}\Big(\log f(\theta_{iq}|\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x}_i)\Big), \qquad (3.12)
\end{aligned}
$$

where $e_{iq}^{(r)} = f(\theta_i|\boldsymbol{\gamma}^{(r)},\boldsymbol{\delta}^{(r)},\mathbf{y}_i,\mathbf{x}_i)$ is the posterior probability of θ_i at iteration r, which is approximated by

$$
e_{iq}^{(r)} = \frac{\prod_{j=1}^{J} f_j(y_{ij}|\theta_{iq},\boldsymbol{\xi}_j^{(r)},\mathbf{x}_i)w_{iq}}{\sum_{q'=1}^{Q}\prod_{j=1}^{J} f_j(y_{ij}|\theta_{iq'},\boldsymbol{\xi}_j^{(r)},\mathbf{x}_i)w_{iq'}}. \qquad (3.13)
$$

Finally, the Bock-Aitkin EM algorithm with the L_1 penalty can be achieve using the following steps until convergence is reached:

E-STEP. Compute the posterior weights $e_{iq}^{(r)}$, $i = 1, \ldots, n$, $q = 1, \ldots, Q$;

M-STEP.

For latent population parameters, compute

$$(\boldsymbol{\gamma}^{(r+1)}, \boldsymbol{\delta}^{(r+1)})' = \arg\min_{\boldsymbol{\gamma}, \boldsymbol{\delta}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{q=1}^{Q} e_{iq}^{(r)} \left(\log f(\theta_{iq} | \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{x}_i) \right) \right\} \quad (3.14)$$

For parameters for item j, compute

$$\boldsymbol{\xi}_j^{(r+1)} = \arg\min_{\boldsymbol{\xi}_j} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{q=1}^{Q} e_{iq}^{(r)} \left(\log f_j(y_{ij} | \theta_{iq}, \boldsymbol{\xi}_j, \mathbf{x}_i) \right) + p_\lambda(\boldsymbol{\beta}_j) \right\}. \quad (3.15)$$

3.2.2 M-step Optimization

As was previously shown, at each iteration $r+1$, the M-step optimizes $\boldsymbol{\xi}^{(r+1)} = \arg\min_{\boldsymbol{\xi}} p_n(\boldsymbol{\xi}; \mathbf{Y}, \lambda | \mathbf{X})$, which can be split into two optimization problems shown in Equations 3.14 and 3.15. These two equations update the population parameters (i.e., $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$) and item parameters, respectively. Equation 3.14 can be obtained by a Newton-type optimizer. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is used in the current study. In contrast, Equation 3.15 for each item j needs to be handled separately due to the non-differentiability of the L_1 penalty. Finding $\boldsymbol{\xi}_j^{(r+1)}$, $j = 1, \ldots, J$ in the M-step amounts to a conditional density estimation problem in a sample size nQ with weights $e_{iq}^{(r)}$. Specifically, maximizing the unpenalized conditional density can be achieved by solving the reweighted least

square problem. With the L_1 penalty, the coordinate descent algorithm (Friedman, Hastie, & Tibshirani, 2010) can be used to solve the penalized weighted least square problem with pseudo data of a sample size nQ. Details of the coordinate descent algorithm is provided in the Appendix A

3.2.3 Selection of λ

To determine the optimal value of λ from a predefined grid, information criteria or cross-validation can be used. To avoid the computation burden resulting from fitting models repeatedly, the current study uses the Bayesian Information Criterion (BIC, Schwarz, 1978) to select the optimal λ. Previous studies (e.g., Bauer & Hussong, 2009; Belzak & Bauer, 2020; Belzak, 2021) have shown that BIC performs relatively well in correctly identifying DIF items and outperforms other information criteria (i.e., AIC, Akaike, 1974) in controlling the Type I error rate using the Reg-DIF method.

3.2.4 Decorrelated Score Tests to Detect DIF

After obtaining the initial parameter estimate using the penalized EM algorithm, the decorrected score test can be constructed to detect DIF items following steps in Algorithm 1. Specifically, $\hat{\mathbf{W}}$ (Equation 3.2) and the test statistic $\hat{T}_{Dscore}$ (Equation 3.5) are calculated using the score function (i.e., the gradient of the loss function $\ell(\boldsymbol{\xi}, \mathbf{Y}|\mathbf{X}) = -\frac{1}{n}\sum \log f(\mathbf{y}_i|\mathbf{x}_i)$) and the Fisher information. To differentiate from the decorrelated score function $\mathbf{s}(\boldsymbol{\xi})$, $\vec{\mathbf{s}}$ is used to define the gradient of the

loss function. The observed Fisher information matrix can be approximated using Louis' formula (Louis, 1982)[2] as

$$\hat{\boldsymbol{\mathcal{I}}}(\boldsymbol{\xi}) = \frac{\partial^2 \ell(\boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^\top}\bigg|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}} = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{E}[\mathbf{H}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)] + \sum_{i=1}^{n} \mathbb{E}[\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)^\top] - \sum_{i=1}^{n} \mathbb{E}[\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)]\,\mathbb{E}[\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)]^\top\right), \tag{3.16}$$

where

$$\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i) = \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\xi}} \tag{3.17}$$

$$\mathbf{H}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i) = \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^\top} \tag{3.18}$$

are the complete data score vector and Hessian matrix of observation i. The expectation in Equation 3.16 is taken with respect to the posterior distribution of θ_i shown in Equation 3.13 (i.e., $f(\theta_i|\mathbf{y}_i) = f(\mathbf{y}_i, \theta_i)/f(\mathbf{y}_i)$). The gradient and Hessian of $\log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)$ can be found in Appendix B.

[2]The last term (i.e., $\sum_{i=1}^{n} \mathbb{E}[\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)]\,\mathbb{E}[\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)]^\top$) is 0 when evaluated at the ML estimate.

Chapter 4: Monte Carlo Simulations

In order to investigate the finite sample behavior of the proposed Dscore test in testing DIF under different conditions, a Monte Carlo simulation study is conducted. Performance of the proposed Dscore test is evaluated in comparison to three methods: (1) the Wald test assuming known and correctly specified anchors, (2) the Reg-DIF method (i.e., results based on the LASSO selection only), and (3) and the naive model refitting approach (Belzak & Bauer, 2020). The efficacy of each procedure will be investigated by assessing hypothesis tests, parameter recovery, and estimates of standard errors (SEs). The simulation study design and evaluation criteria are discussed in Section 4.1 and Section 4.2, respectively. All computations were performed in R (R Core Team, 2020).

4.1 Study Design

Binary response data were generated from a unidimensional MNLFA model under two conditions—with or without DIF items. Two factors were manipulated including (1) the total sample size ($n = 500, 1,000, \& \ 2,500$) and (2) the percentage of DIF items (0%, 25%, & 50%). As a result, there were $3 \times 3 = 9$ fully crossed conditions. These factors were selected due to their relevance to the DIF detection

mechanism of the decorrelated score test. Moreover, actual values of the manipulated factors were chosen to conform to real-world data analytic scenarios and align with previous methodological studies (e.g., Bauer, 2017; Belzak & Bauer, 2020).

1. Three levels of the sample size condition (500, 1000, and 2,500) representing small, medium, and large sample sizes are typically used in both the applied social science research (e.g., Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004; Scott et al., 2010) and methodology investigation of DIF methods in the two-group scenario (e.g., Bolt, Hare, Vitale, & Newman, 2004; Magis, Béland, Tuerlinckx, & De Boeck, 2010; W.-C. Wang & Yeh, 2003; Woods, 2009a; Woods & Grimm, 2011). It is anticipated that as the sample size increases the performance of the Dscore test improves in the sense that the true positive rate will eventually approach to $1-$ the nominal level while the false positive rate remains at the nominal level.

2. The proportion of DIF items may greatly impact the performance of the DIF test especially when the proportion of DIF items is large. As the proportion of DIF items increases, the probability of selecting an incorrect anchor may increase. Therefore, both power and the false positive rate in identifying a DIF item can be negatively influenced. The selected values reflect what is typically observed in DIF studies (e.g., W.-C. Wang & Yeh, 2003; Woods, 2009a).

3. The effect size of DIF can also influence the performance of DIF tests as large signal can be easily detected by the Reg-DIF method whereas detecting the

small DIF effect size is more challenging. This seems to suggest that DIF items with larger effect sizes can be more easily detected. However, larger DIF effect sizes have been found to negatively influence the false positive rate using traditional multiple-group IRT methods (e.g., W.-C. Wang & Yeh, 2003; W. Wang et al., 2022) if there is no prior knowledge on anchor items. As a comparison, the Reg-DIF method may outperform the traditional IRT method in controlling the false positive rate without predefined anchor items (Belzak & Bauer, 2020). DIF items within each alternative condition are carefully designed to include large and small DIF parameters. Due to the limited DIF research on the topic, model parameters are generated based on a real data analysis using the UK normative sample data of the Revised Eysenck Personality Questionnaire (EPQ-R, Eysenck, Eysenck, & Barrett, 1985)[1]. Initial analysis results using the EM algorithm with anchors selected using the Reg-DIF method demonstrated a similar DIF parameter trend as is manipulated by Belzak (2021). That is the effect size for a-DIF is typically smaller than that of d-DIF and the effect size of a binary covariate is larger than that of a continuous variable. Effect sizes of d-DIF (i.e., β_{0j}) for continuous and categorical grouping variables range from 0.1 to 0.6 and 0.3 to 1.1, respectively and effect sizes of a-DIF (i.e., β_{1j}) for continuous and categorical grouping variables range from 0.1 to 0.7 and 0.1 to 1.1, respectively. However, based on a trial run, the power of detecting a DIF item is nearly identical across different methods and are all larger than 0.9 even when the sample size is

[1]We are grateful to Dr. Paul Barrett for granting us access to the data

small ($n = 500$). A decision is made to decrease the DIF effect size by half to differentiate power across the selected DIF methods. True data generating values of these DIF items are described in details in the following section.

4.1.1 The True Data Generation Model

The binary response for examinee i for item j, denoted $Y_{ij} \in \{0, 1\}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$, was generated from a MNLFA model (see Equation 2.2). The total number of items was fixed at $J = 12$. Model parameters of the data generating model are tabulated in Table 4.1. Specifically, item discrimination and intercept parameters are the ML estimates of item parameters using the EPQ-R data with five items identified as anchors. Three person covariates are considered in the study which mimic gender, age, and their product. Specially, the dichotomous grouping variable is generated from a Bernoulli distribution with success probability of .5 (i.e., $x \sim \text{Bern}(0.5)$). Then, for those in category 1, age is generated from $\mathcal{N}(0.2, 1)$. Otherwise, age is generated from $\mathcal{N}(0, 1)$, which creates a correlation between age and gender of 0.1. An interaction effect is then created by multiplying age and gender. By including an interaction effect, the correlations between the interaction effect and the two variables is large ($> .6$). Furthermore, three types of DIF items were generated. Under the 25% DIF condition, items 1 to 3 are DIF items representing items with large, median, and small DIF effect sizes. When the proportion of DIF items is increased to 50%, items 4 to 6 are added to the DIF set and their DIF effect

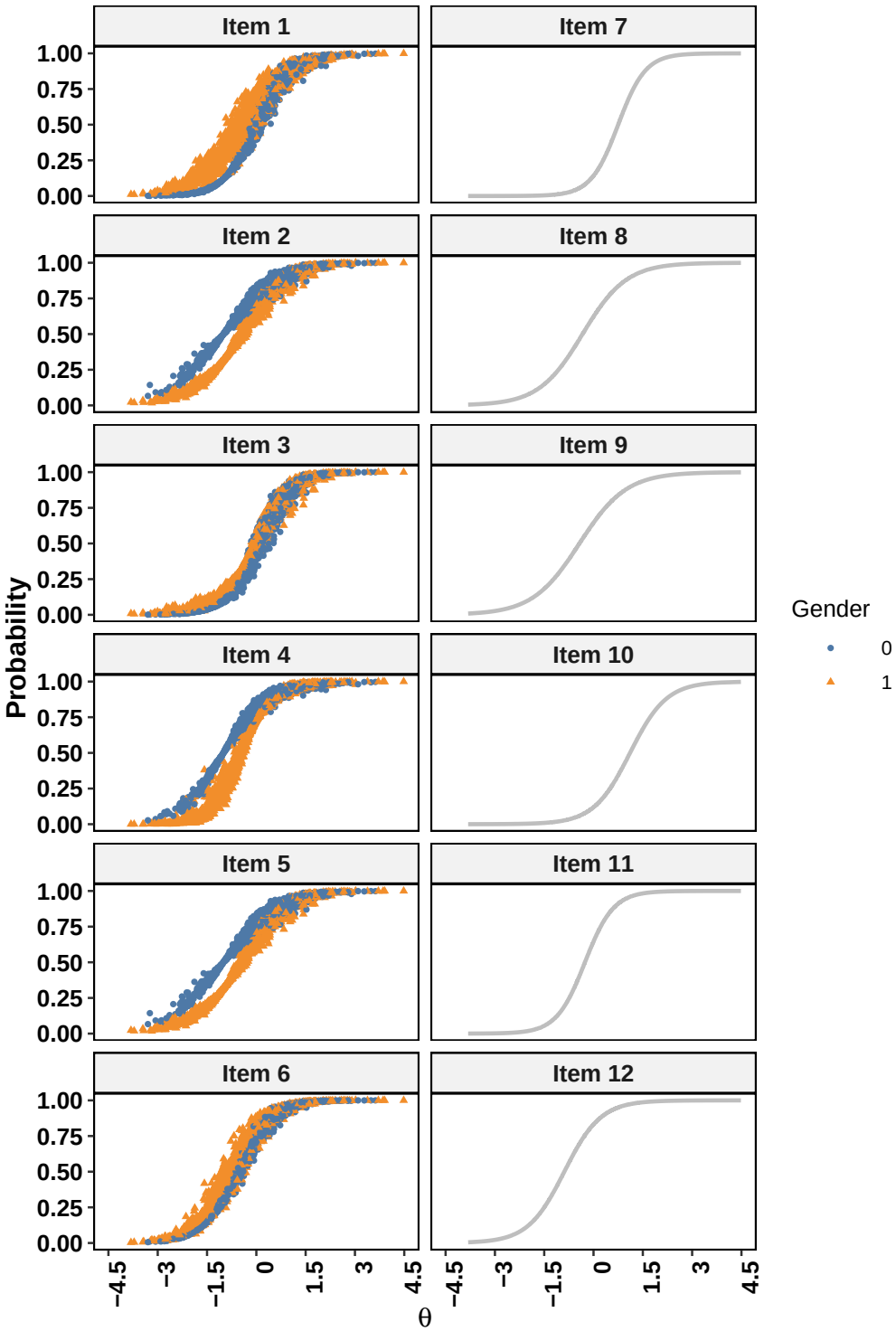
sizes replicate those of items 1-3.

Table 4.1: Model Parameters of the True Data Generating Model

Item	d_j	a_j	a-DIF Age β_{11}	a-DIF Gender β_{12}	a-DIF Product β_{13}	d-DIF Age β_{01}	d-DIF Gender β_{02}	d-DIF Product β_{03}
1	0.00	2.00	0.20	0.50	0.20	0.20	-0.50	-0.20
2	1.20	1.20	-0.20	-0.50	0.00	-0.20	0.25	0.00
3	-0.20	2.00	-0.25	0.25	0.10	-0.15	-0.25	-0.15
4*	1.50	1.50	0.20	-0.50	-0.20	0.20	0.50	0.20
5*	1.20	1.20	-0.20	-0.50	0.00	-0.20	0.25	0.00
6*	1.10	1.90	-0.25	0.25	-0.10	-0.15	-0.25	0.15
7	-1.80	2.40	0.00	0.00	0.00	0.00	0.00	0.00
8	0.50	1.50	0.00	0.00	0.00	0.00	0.00	0.00
9	0.60	1.40	0.00	0.00	0.00	0.00	0.00	0.00
10	-2.00	1.80	0.00	0.00	0.00	0.00	0.00	0.00
11	0.60	2.30	0.00	0.00	0.00	0.00	0.00	0.00
12	1.60	1.80	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{\gamma}$				$(-0.2, -0.2, -0.2)^\top$				
$\boldsymbol{\delta}$				$(-0.1, 0.3, 0.1)^\top$				

Note. * *indicates that effect sizes of these items under the 25% DIF condition are 0.*

To visualize the generated items, Figure 4.1 displays the true probability of each person endorsing items with the condition of 50% DIF items and $n = 2,500$. Plots on the left correspond to DIF items whereas those on the right visualize non-DIF items. Gender influence on the item response can be seen from different colors and shapes. Lastly, the latent variable follows a normal distribution $\mathcal{N}(\boldsymbol{\gamma}^\top \mathbf{x}, \boldsymbol{\delta}^\top \mathbf{x})$ conditional on $\mathbf{x}$, and similarly true population parameter values are shown in Table 4.1, which are also enlightened by real data. A total of 500 replications for each condition are implemented to ensure that the 95% normal approximation confidence band at the nominal level 0.05 to be $[0.031, 0.069]$.



e

Figure 4.1. Conditional probability of endorsing an item. 2,500 item responses to all 12 items were generated using the moderated non-linear factor analysis model with three covariates. Items 1 to 6 are DIF items.

4.1.2 Estimation

For each replication under each condition, the following four methods are fitted to the binary response data to test DIF at the item level and the parameter level. MNLFA models are estimated using penalized maximum likelihood for the Reg-DIF method and the Dscore method or ML for the refit method and oracle solution. For both estimation methods, integration with respect to the latent variable is approximated by a 49-point Gauss–Hermite quadrature. The four methods are discussed in detail here.

First, each item is tested for DIF based on LASSO selection only (i.e., the Reg-DIF method). The MNLFA model fitted with all three grouping covariates is estimated using the penalized EM algorithm. For each replication, a series of penalty values are tested in a descending order starting from a λ value that penalizes all DIF effects to zero for all items. To save computation time, model parameters from the larger λ value serve as a "warm start" for succeeding runs with smaller λ. The starting value will be set to the optimal convergence rate (i.e., $\sqrt{1/n}$). In practice, the range and the granularity of λ needs to be evaluated case by case. However, in a simulation study this is not possible. As the selection of λ depends on the sample size and the dimension of covariates, a pilot study is conducted for each sample size condition to find the range and granularity of the λ. The pilot study shows that decreasing the $\sqrt{n}\lambda$ value from 1 to 0.25 at an interval of 0.1 or 0.05 should be sufficient. Starting from the aforementioned starting point (i.e., $\sqrt{1/n}$), a series values tested at a descending order so that BIC decreases until a turning

point. Finally, to select the optimal tuning parameter under each replication, BIC is used. However, the pilot study found that BIC calculated based on the penalized EM parameter estimates always select a relatively smaller λ value when the DIF effect size is large, which results in too many FDR. This phenomena persists even when sample size is large $n = 2,500$ or $n = 5,000$. We suspect that the less optimal λ selection is due to parameter bias due to the regularization. This can be verified that when the λ value is selected by the BIC value calculated based on the maximum likelihood estimates (i.e., the model refitting method), the FDR rate is more controlled especially under the large sample size condition. Therefore, a decision was made to fixed λ at an estimated optimal value (i.e., $c\sqrt{1/n}$ where the constant c is estimated by $\hat{c} = \sqrt{n}\sum_{r=1}^{R}(\hat{\lambda}_r)/R$). The conditions with $n = 500$ were used to estimate the constant. Results show that $\hat{c} = 0.8291, 0.6883, 0.5727$ when the number of DIF items is 0, 3, and 6, respectively. Therefore, λ is fixed at the estimated optimal value for each condition shown in Table 4.2. For the reg-DIF method, an item does not exhibit DIF if both $\boldsymbol{\beta}_{0j}$ and $\boldsymbol{\beta}_{1j}$ are zero vectors. Otherwise, the item is considered as a DIF item.

Table 4.2: Fixed λ for Each Condition

	Number of DIF items		
Sample size	0	3	6
500	0.04	0.03	0.03
1,000	0.03	0.02	0.02
2,500	0.02	0.01	0.01

Note. Values are rounded to two decimal places.

Next, the naive model refitting method can be applied by refitting the same MNLFA model with the anchors selected using method 1. This model is estimated

using the marginal ML estimation method with the EM algorithm. The marginal likelihood function can be approximated using the same configuration as in the penalized EM algorithm. The convergence tolerance for the log-likelihood change for the penalized EM algorithm and the M-step are set to be 10^{-4} and 10^{-6}, respectively. The maximum number of iterations is set to be 500. Additionally, the Dscore test is conducted after the initial run (method 1) using the penalized EM algorithm. Details of the Dscore test was described in Chapter 3 and thus are not repeated here. A critical step to conduct the Dscore test is to estimate $\hat{\mathbf{W}}$ as displayed in Equation 3.2. As the theory only requires that λ and λ' are of the same rate, λ' will be set to the selected λ value in the initial parameter estimation step to speed up the computation and save time. Ning and Liu (2017) and Fang, Ning, and Liu (2017) have found that the decorrelated score test is not sensitive to λ' and both fixed $\lambda' = 0.5\sqrt{\log d/n}$ in their simulation studies. Furthermore, asymptotic unbiased parameter estimates can be estimated using the one-step bias correction (see Equation 3.6). Note that although the focal parameter can be multidimensional or unidimensional (i.e., the debias step can be conducted at the item level or parameter level), the current study investigates the one-step bias correction by parameter type. For instance, for each item, the bias correction treats $(a_j, d_j, \boldsymbol{\beta}_{0j}, \boldsymbol{\beta}_{1j})^\top$ as the focal parameters and everything else as nuisance parameters. The population parameter estimates are corrected at the same time by treating all item parameters as nuisance parameters.

Lastly, to compare the performance of the Dscore test with the oracle solution assuming anchors are known, Wald tests are performed to test DIF at the item-level. For each replication and each condition, item 11 and item 12 are treated as

anchors. Then, a Wald test is conducted one item at a time using Equation 2.12. In addition, final model parameter estimates and their corresponding standard errors are estimated using anchors selected by the Wald test.

4.2 Evaluation Criteria

The comparative inferential performance of the four methods is evaluated in terms of (1) hypothesis testing in testing a DIF item, (2) parameter recovery, and (3) recovery of standard errors. For hypothesis tests, rejection rates at α level 0.05 are used to investigate the Type I error rate, the false positive rate (FPR), and power.

To evaluate the parameter recovery, parameter bias (shown in Equation 4.1) and variance of model parameter estimates (shown in Equation 4.2) are calculated. Particularly, focal parameters are DIF parameters including a-DIF and d-DIF parameters of zero and non-zero effects, item parameters (i.e., a_j and d_j), and population parameters.

$$\text{Bias} = \sum_{r=1}^{R} \hat{\xi}^r / R - \xi^* \tag{4.1}$$

$$\text{Variance} = \frac{\sum_{i=1}^{R} (\hat{\xi}^r - \sum_{i=1}^{R} \hat{\xi}^r / R)^2}{R}. \tag{4.2}$$

Lastly, standard error estimates are evaluated by comparing the square root of the mean of the variance of parameter estimates against the empirical standard errors (i.e., Monte Carlo standard deviation of the parameter estimates). Addition-

ally, relative efficiency of the one-step debiased estimator as compared with the oracle solution (i.e., ξ_o) is calculated as follows.

$$\text{Relative efficiency} = \frac{\sum_{r=1}^{R} \text{VAR}(\hat{\xi}^r)/R}{\sum_{r=1}^{R} \text{VAR}(\hat{\xi}_o^r)/R} \tag{4.3}$$

Chapter 5: Results

5.1 Results: Hypothesis Testing

In this section, the Type I error rate, false positive rate (FDR), and power of detecting a DIF item is investigated. To visualize the comparative performance, the empirical rejection rate at the nominal level 0.05 was plotted under (1) the null condition when there is no DIF items (i.e., Type I error), (2) the alternative condition when there is a mix of DIF and DIF-free items for DIF items only (i.e., Power), and (3) the alternative condition when there is a mix of DIF and DIF-free items for DIF-free items only (i.e., FDR).

Type I Error Rate

Figure 5.1 shows the Type I error rate of detecting a DIF item under the null condition when there are no DIF items. The horizontal dashed lines show the 0.05 nominal level and the horizontal dotted lines represent the 95% normal approximation confidence band at the nominal level 0.05 across 500 replications. If the rejection rate falls within the 95% confidence band (i.e., 95% confidence interval $=[0.031, 0.069]$), the specific method is considered to have well-controlled Type I error rate of incorrectly detecting a DIF item. Larger values than the upper bound of the confidence band at the nominal level 0.05 represent over-rejecting the null

hypothesis whereas smaller values than the lower bound of the confidence band indicate under-rejecting the null hypothesis.

Overall, the Reg-DIF, model refit method, and decorrelated score test showed controlled Type I error under all manipulated sample size conditions as the rejection rate of all these methods fell within the 95% confidence band. The rejection rate of the Reg-DIF and the model refit methods were smaller than 0.02 across all sample size conditions, which were smaller than the rejection rate of the oracle solution and the decorrelated score test. Of note, the performance of the oracle solution by treating the last two items as anchors was expected. Under the small sample size ($n = 500$) condition, the Type I error rate was inflated (rejection rate ranged from 0.05 to 0.11) as compared to all other methods in comparison. Results could be improved by choosing a different set of anchor items or increasing the number of anchor items. As sample size increased, the Type I error rate of the oracle solution ranged from 0.03 to 0.07, which fell within the 95% confidence band.

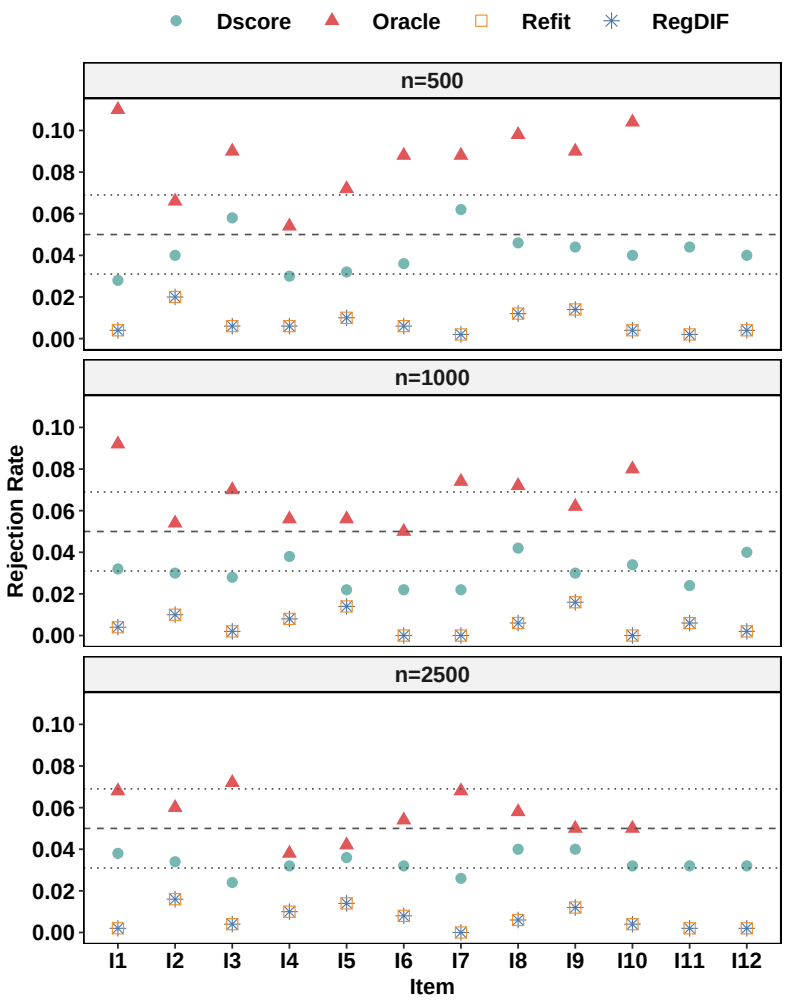


Figure 5.1. Type I error results of incorrectly detecting a DIF item under the null condition when there is no DIF items. Different methods are displayed in different colors and shapes. Two dotted horizontal line shows the 95% normal-approximation confidence band at the nominal level 0.05 (horizontal dashed line). The oracle solution only performed on item 1 to item 10 as the last two items are treated as anchors.

Power

Figure 5.2 shows the power of detecting DIF items under different conditions. The power can be negatively impacted by the sample size while more robust to the number of true DIF items. When the sample size was small, the power of detecting a DIF item was as low as 0.20 for the smallest DIF effect size item (i.e.,

item 3 with 3/12 true DIF items). As the sample size increased, all methods had at least 0.80 probability of identifying a true DIF item. Also, as expected, items with larger DIF effect size were relatively easier to be identified while those with smaller DIF effect sizes were less likely to be identified especially when the sample size was small. For example, when $n = 500$ and the first three items were DIF items, the rejection rate was as low as 0.20 for the smallest DIF effect size item (i.e., item 3) but as high as 0.80 for the largest DIF effect size item (i.e., item 1). Overall, the decorrelated score test was more powerful than the Reg-DIF and the model refit methods across all conditions. The difference was more obvious under more challenging conditions where the sample size was small. For example, when $n = 500$ and there were three DIF items, the power ranged from 0.38 to 0.85 and from 0.18 to 0.80 using the decorrelated score test and the Reg-DIF or model refit methods respectively. The performance of the Reg-DIF method and the model refit method were nearly identical when the total number of DIF items was small. The difference between the two methods was more obvious when the number of DIF items was large. Specifically, the model refit method was slightly less powerful than Reg-DIF by design. The reason was that once an item was identified as an anchor item, it's DIF parameters were excluded in the refitted model. In other words, this item can no longer be identified as a DIF item.

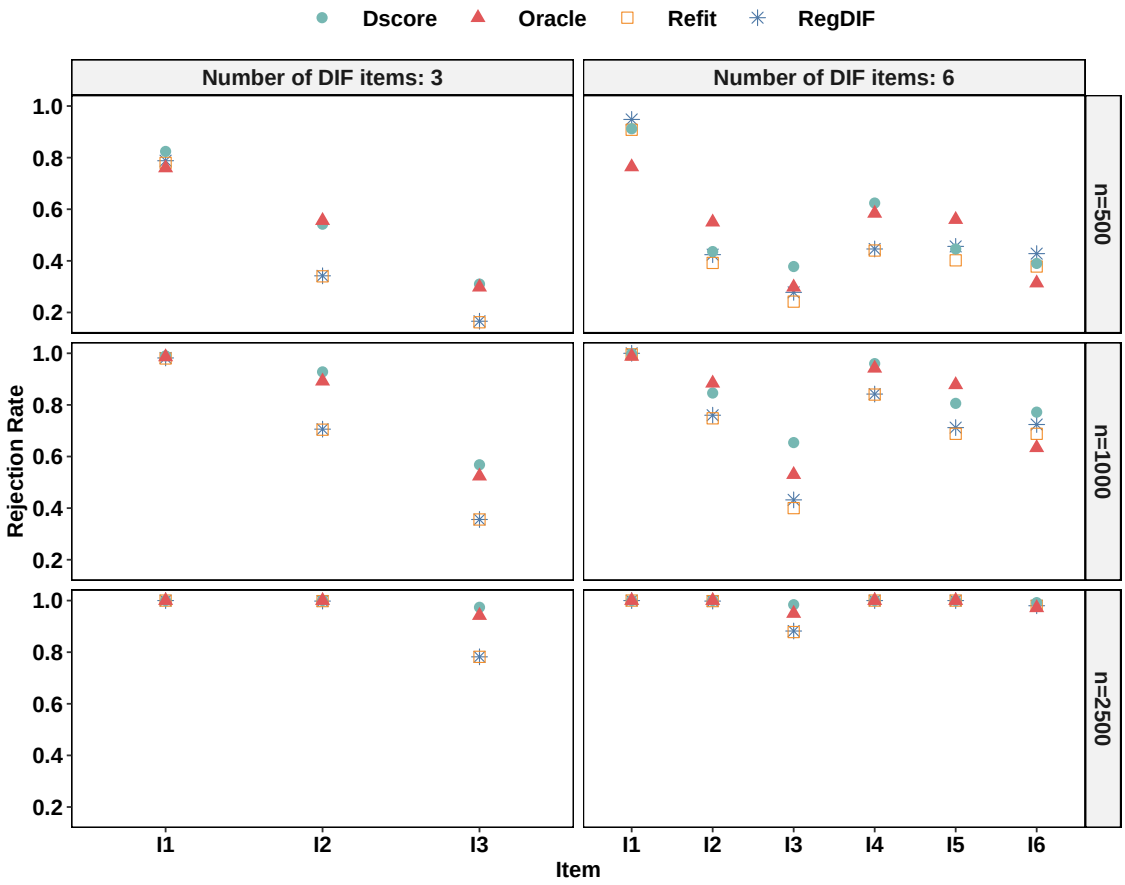


Figure 5.2. Power results of correctly detecting a DIF item under the alternative condition when there is a mix of DIF and DIF-free items. Different methods are displayed in different colors and shapes. For each method, the empirical rejection rate under the nominal level 0.05 are calculated for each DIF item under each condition. The column shows the condition when the number of DIF items is 3 or 6 out of 12 items. Each row represents a different sample size condition. The reference dashed line shows the empirical rejection rate = 0.05

False Detection Rate

Finally, results for incorrectly detecting an anchor item as a DIF item are visualized using the empirical rejection rate for the true anchor item under alternative conditions. As displayed in Figure 5.3, the performance of the decorrelated score test was outstanding as the FDR of the decorrelated score test ranged from 0.03 to 0.07, which fell within the 95% normal-approximation confidence band, indicating

controlled FDR. Conversely, the Reg-DIF and model refit method under-rejected the null hypothesis when the number of DIF items was small (i.e., FDR $\leq$ 0.03 when 3 out of 12 items are DIF items). This is understandable, as previously shown that the power of identifying the true DIF item was generally less than the decor-related score test and the oracle solution. However, when the number of DIF items was large, the Reg-DIF method over-rejected the null hypothesis (e.g., FDRs for item 8 and item 9 were approximately 0.14). As previously illustrated, the model refit method was more conservative in rejecting a null hypothesis as compared to the Reg-DIF method. Intuitively, it made sense that the FDR was more controlled as compared to the Reg-DIF method. Nevertheless, it still over-rejected the null hypothesis when the sample size was not sufficiently large. For instance, when the sample size was 1000 and item 1 to item 6 were the true DIF items, the FDR for item 7 using the model refit method was approximately 0.09, which was larger than the upper bound of the confidence band of the nominal level of 0.05. As the sample size increased, the FDR using the model refit method fell within the 95% confidence band at the nominal level 0.05.

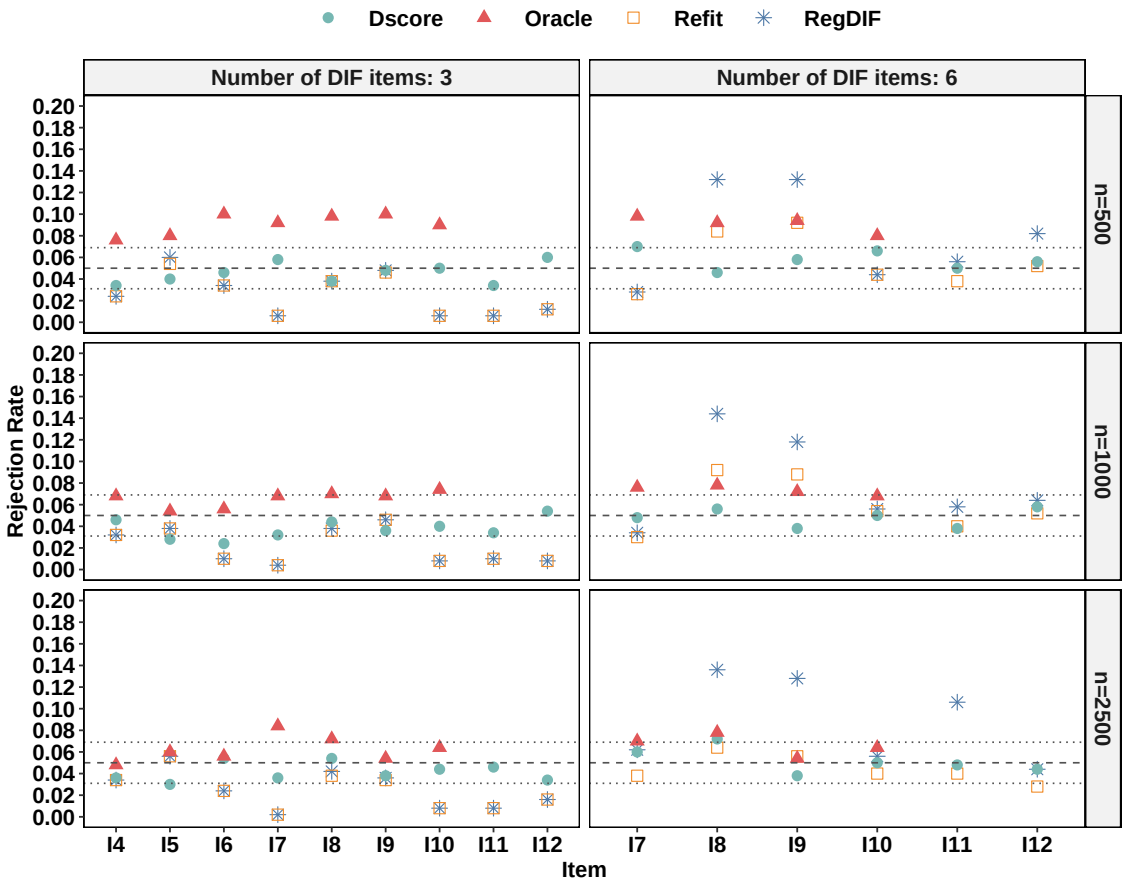


Figure 5.3. False detection rate results of incorrectly detecting a DIF item under the alternative condition when there is a mix of DIF and DIF-free items. Different methods are displayed in different colors and shapes. For each method, the empirical rejection rate at the nominal level 0.05 are calculated for each DIF item under each condition. Two dotted horizontal line shows the 95% normal approximation confidence band for the nominal level 0.05 (horizontal dashed lines). The column shows the condition when the number of DIF items is 3 or 6 items out of 12 items. Each row represents a sample size condition.

5.2 Results: Parameter Recovery

To investigate parameter recovery, the bias and variance of the parameter estimates using each method were computed. Bias and variance of the item parameter estimates including item slope (a_j), item intercept (d_j), a-DIF parameters ($\boldsymbol{\beta}_{1j}$),

d-DIF parameters ($\boldsymbol{\beta}_{0j}$), and population parameters ($\boldsymbol{\gamma}, \boldsymbol{\delta}$) using different methods are displayed in Figure 5.4 to Figure 5.12. Model parameters were estimated using the penalized EM algorithm for the Reg-DIF method, the EM algorithm for the model refit method, and the one-step debiased estimate for the decorrelated score method, respectively. For the oracle solution, the MNLFA model was re-estimated using the EM algorithm using anchors selected by the Wald test. For all methods other than the one-step debiased estimator using the decorrelated score function, DIF parameters penalized to be 0 or not included in the final model were treated as 0 to compute bias and variance of an estimator.

Bias

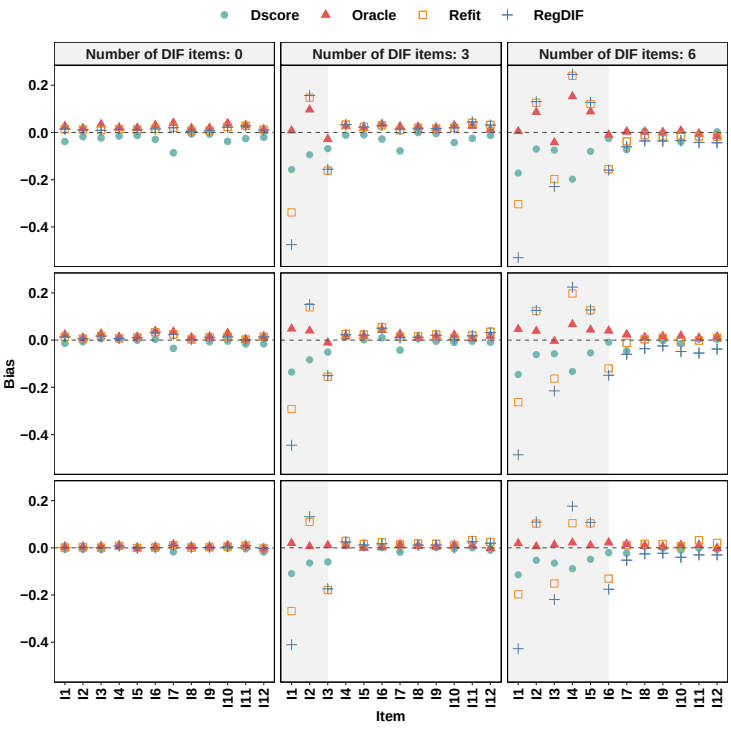
When there were no DIF items, all methods recovered model parameters well with bias in general ≤ 0.05. This was understandable because when there was no DIF, the Type I error rate, as previously shown, was well-controlled. The penalized EM performs similarly as the EM since in most cases λ was large enough to penalize all DIF parameters to be 0. Even if this did not happen, the model refit method was always fitting the correct model under the null condition.

When there were a mix of DIF and non-DIF items, the performance of different methods varies. The Reg-DIF method often resulted in biased item parameter estimates and DIF parameters for the true DIF items (i.e., items 1 to 3 in the 3 DIF item condition and item 1 to 6 in the 6 DIF item condition) due to the shrinkage. The bias for the item parameters for the anchor items was relatively small. For example, while the bias for item slopes of the DIF items ranged from -0.53 to 0.17, that of the anchor items ranged from -0.06 to 0.04. More importantly,

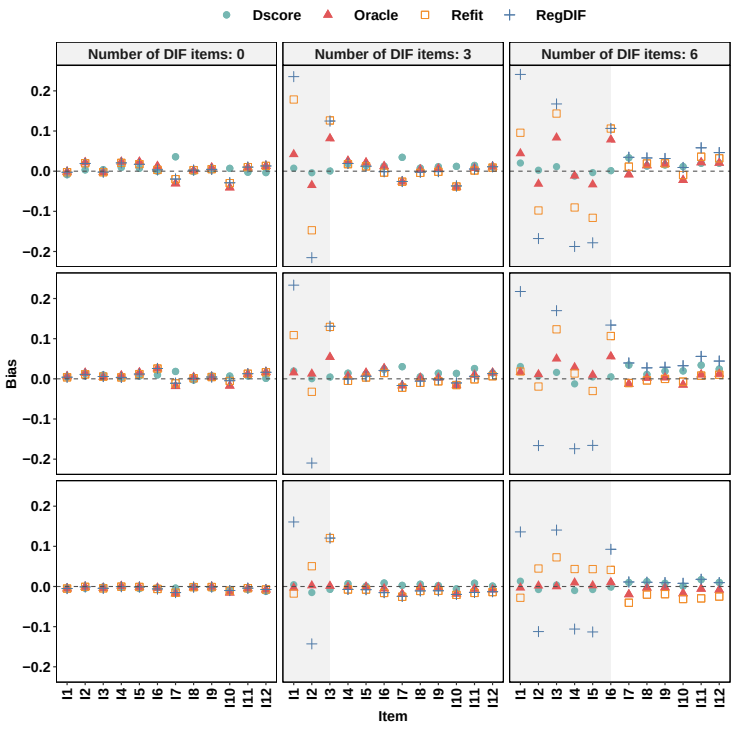
bias, although decreased, remained relatively large as much as 0.4 (e.g., for item 1) when sample size was large ($n = 2500$). Compared with the Reg-DIF method, the model refit method reduced bias for the true DIF items by 3% to 92% depending on the item. Interestingly, bias reduction for item slope parameters of DIF items was less obvious than for those of the item intercept parameters. This can be seen from Figure 5.9 that when sample size was relatively small ($n = 500$ or $n = 1,000$), the crosses (representing the Reg-DIF method) and the squares (representing the model refit method) were nearly overlapped for the item slopes whereas the squares were relatively closer to the 0 bias reference line. Note that bias remained relatively large for the a-DIF parameter even sample size was large (see last row of the Figure 5.5) due to failing to select the right DIF-effect. It was more clear when the sampling distribution of non-zero DIF parameters was investigated (see Figure 5.8). As can be seen, the sampling distribution of the non-zero DIF effect (blue line) using the model refit method was bi-modal with one mode at 0 indicating that the effect was not selected by the LASSO and another non-zero mode.

As a comparison, the one-step debiased parameter estimate using the decorrelated score function performed remarkably well. The bias using the one-step debiased estimator based on the decorrelated score function ranged from -0.17 to -0.05 and from -0.015 to 0.03 for item slopes and item intercepts of the true DIF items, respectively. Similarly, it was found that the recovery of the item intercepts was better than that of the item slopes. DIF parameters were recovered well as the bias ranged from -0.03 to 0.06 across conditions. Its advantages in reducing bias due to shrinkage were more obvious as compared with the model refit method

under smaller sample size or more DIF items conditions. Under the most difficult condition (i.e., small sample size and large number of DIF items), it performed equally well as compared to the oracle solution. More importantly, as the sample size increased, the bias decreased. Finally, as for the population parameters, bias all of methods under all conditions ranged from -0.05 to 0.05. However, it did not mean that penalized EM estimator was unbiased. Given the small effect size of the population parameters ($\boldsymbol{\gamma} = (-0.2, -0.2, -0.2)^\top$ and $\boldsymbol{\delta} = (-0.1, 0.3, -0.1)^\top$), bias equaled 0.05 and was still notably large. As can be seen, bias still remained especially for the effect on the population variance parameter. Unexpectedly, the one-step debiased parameter estimator using the decorrelated score function did not seem to differ much from the Reg-DIF method.



(a) Bias of the item slope parameter



(b) Bias of the item intercept parameter estimates

Figure 5.4. Bias of item parameters: (a) item slopes and (b) item intercepts. Different methods are displayed in different colors and shapes. The column shows the condition when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a sample size condition. DIF items are shown in the grey shaded area.

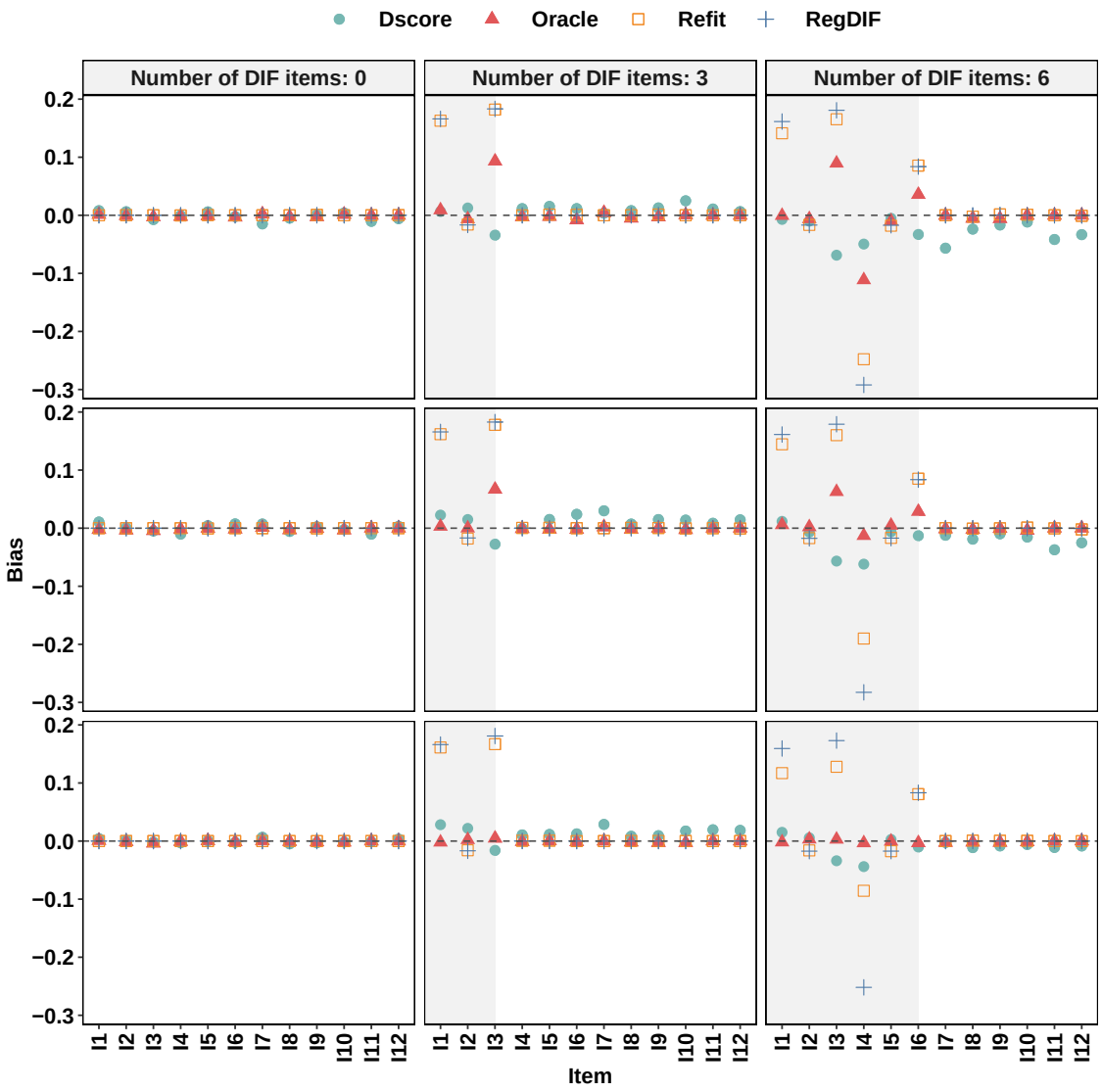


Figure 5.5. Average bias of a-DIF parameters. Different methods are displayed in different colors and shapes. The column shows the condition when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a sample size condition. DIF items are shown in the grey shaded area.

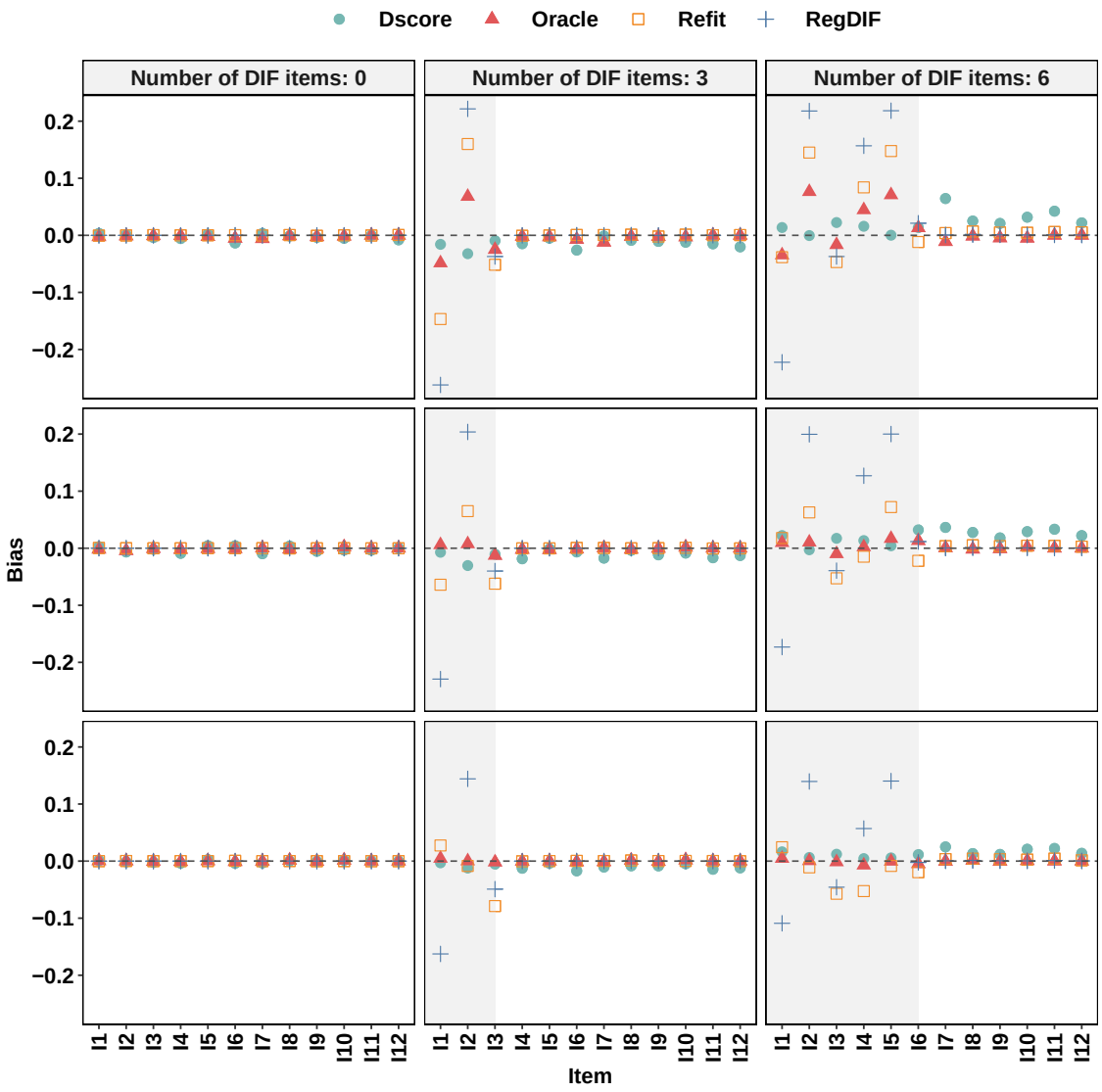


Figure 5.6. Average bias of d-DIF parameters. Different methods are displayed in different colors and shapes. The column shows the condition when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a sample size condition. DIF items are shown in the grey shaded area.

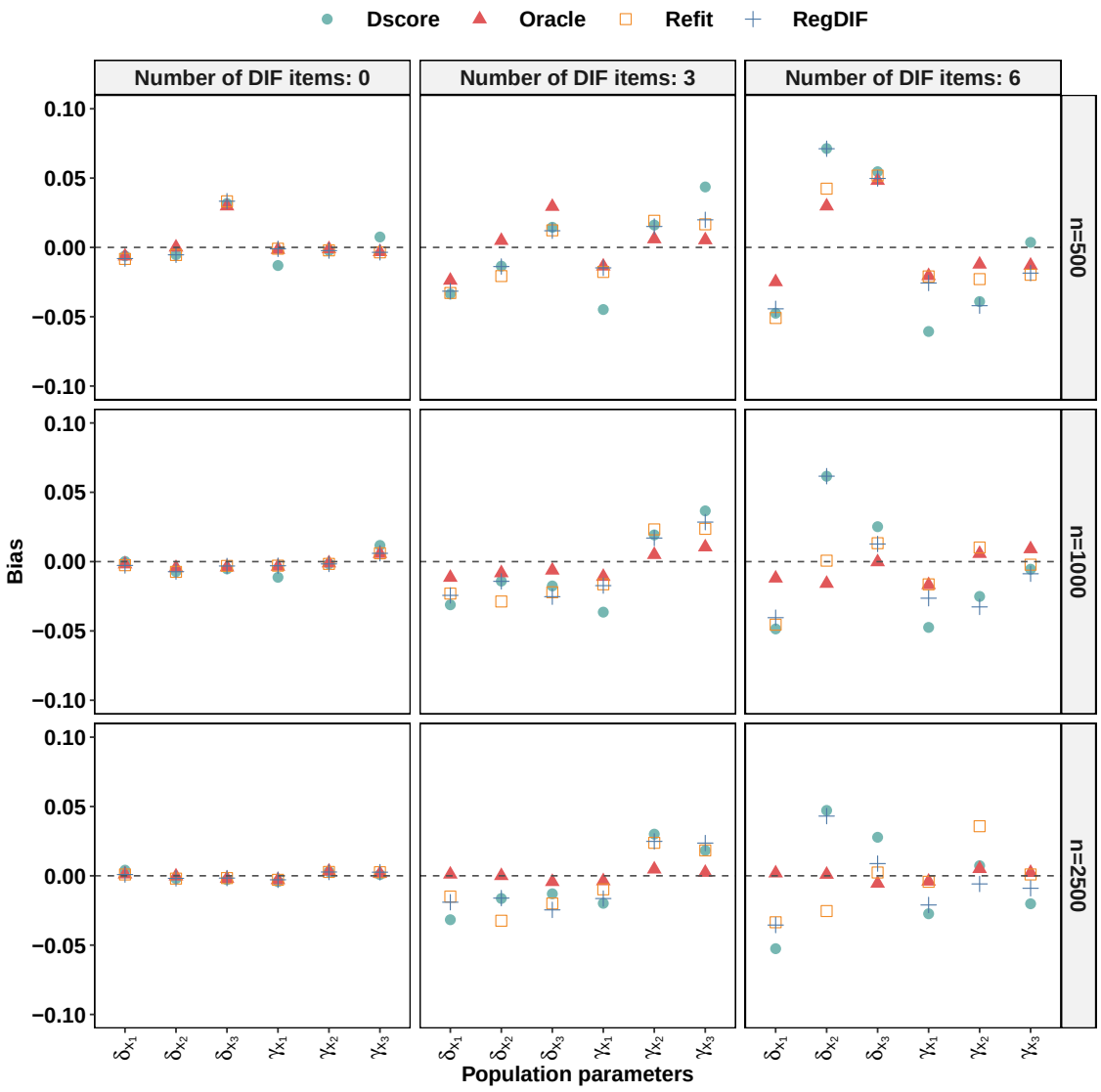


Figure 5.7. Bias of population parameters. Different methods are displayed in different color and shape. The column shows the condition when the number of DIF items is 0, 3 or 6 items out of 12 items. Each row represents the sample size condition.

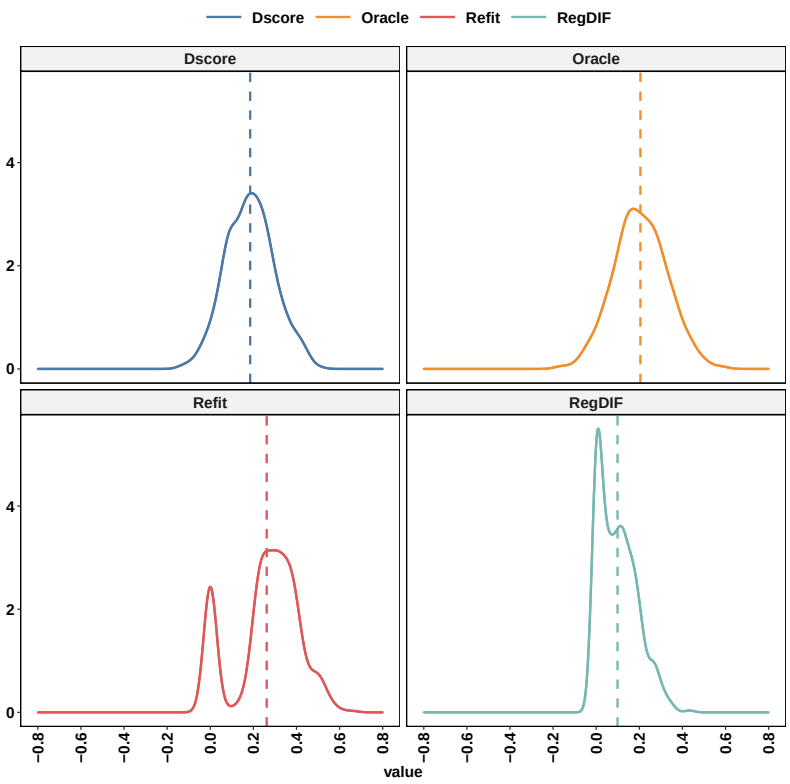


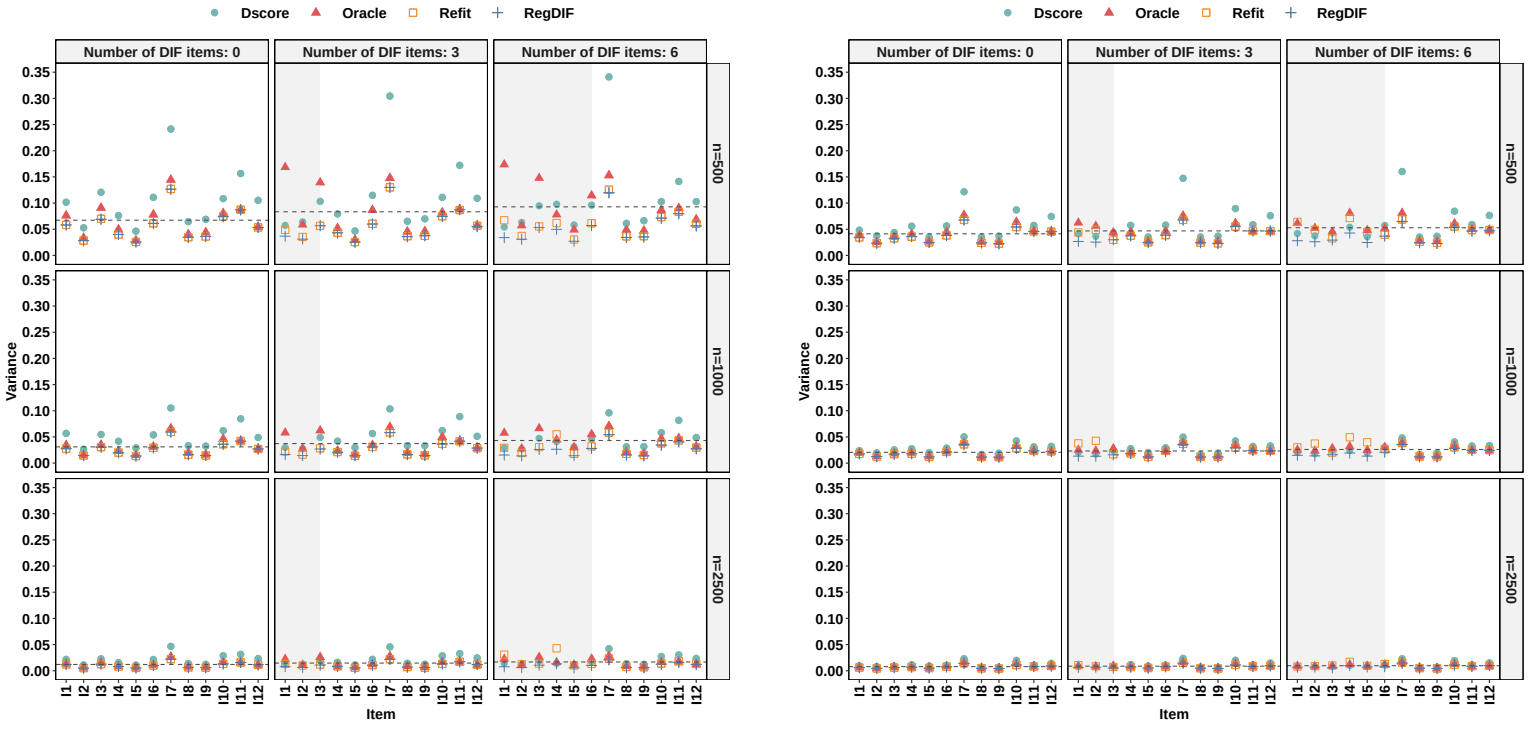
Figure 5.8. Density plot of a non-zero DIF parameter by different method. The vertical reference lines indicate the means of the parameter estimates of each effect for each method. The example is plotted using the sample size $2,500$ and first 3-item DIF condition. The non-zero DIF effect is the continuous DIF effect of the first item.

Variance of Parameter Estimates

Variances of the model parameter estimates across 500 replications were calculated for each method and were summarized by the model parameter types including item parameters, a-DIF parameters, d-DIF parameters, and population parameters.

Figure 5.9 presents the variance of item parameters. Large values indicate more uncertainty of the parameter estimates while smaller values indicate less uncertainty. As the sample size increases, variance should, in general, decrease. Penalized EM parameter estimates (i.e., Reg-DIF) in general produced less variance but at the

sacrifice of more bias. Compared with the penalized EM estimator, the model refit item parameter estimates of the DIF items introduced slightly larger variability but comparable variability for the non-DIF items. The one-step debiased estimator had more variability as compared with the other two methods when the sample size was small. As sample size increased, variances of the item parameter estimates using all methods became similarly small. Of note, the sampling distribution of the item slope and item intercept parameter estimates of item 7 for the one-step debiased estimator was positively skewed under small sample size conditions.



(a) Variance of the item slope parameter estimates

(b) Variance of the item intercept parameter estimates

Figure 5.9. Variances of item parameter estimates: (a) item slopes and (b) item intercepts. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. DIF items are shown in the grey shaded area. The grey dashed reference line displays the mean variance across all items of the oracle solution for each condition to be used as a benchmark.

As for the DIF parameter estimates (see Figure 5.10 and Figure 5.11), the Reg-DIF method was the least variable method among all. Similarly as the performance in the item parameters, the model refit method produced similarly variable a-DIF parameter estimates but notably more variable d-DIF parameter estimates as compared to the Reg-DIF especially for the DIF items. Lastly, the one-step debiased parameter estimator using the decorrelated score function was the most variable estimator for the a-DIF parameter as compared to all other methods especially when the sample size was small. However, it was less variable for the d-DIF parameter for the DIF item as compared with the model refit method.

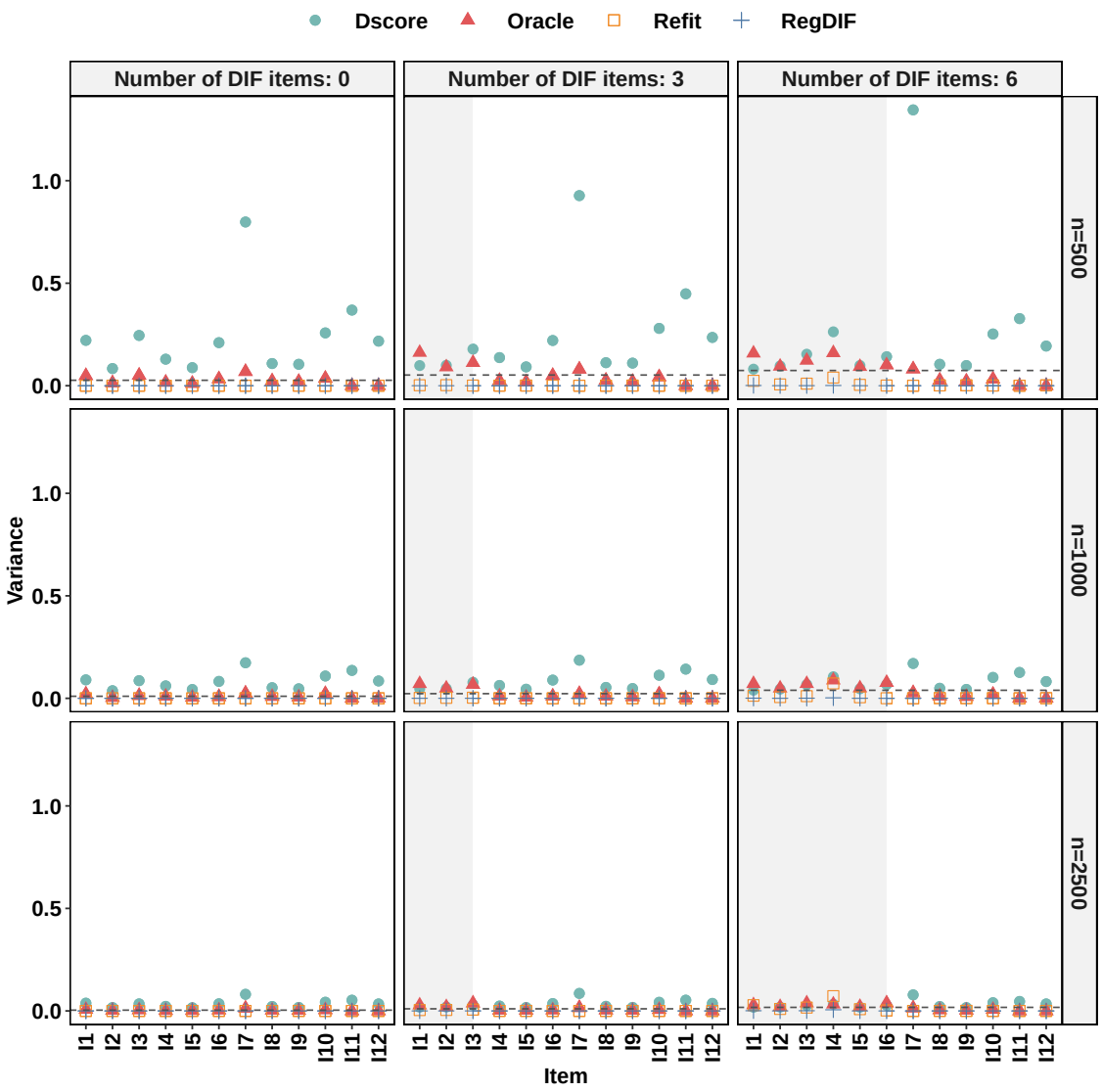


Figure 5.10. Average variance of a-DIF parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays the mean variance across all items of the oracle solution for each condition to be used as a benchmark. DIF items are shown in the grey shaded area.

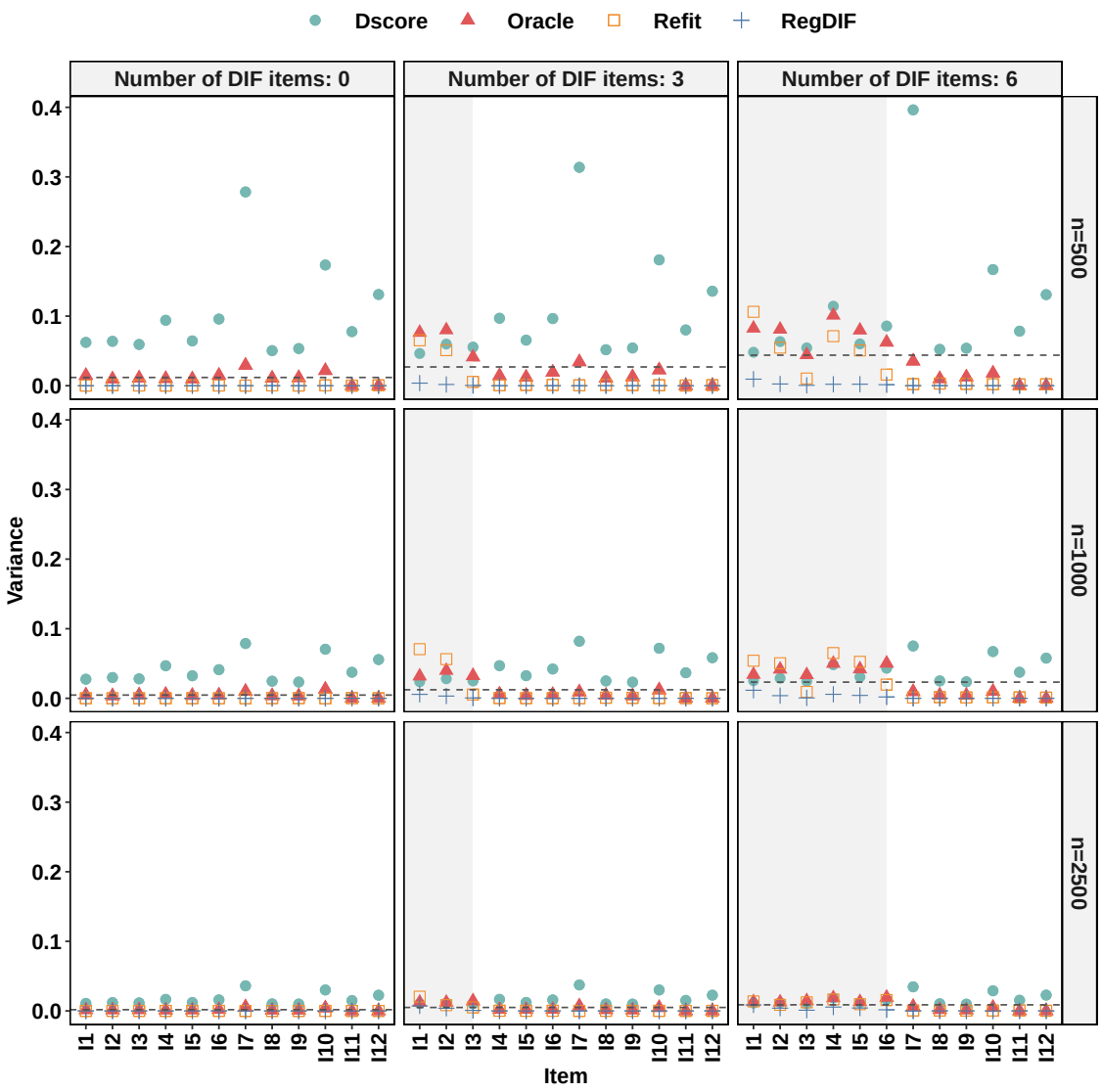


Figure 5.11. Average variance of d-DIF parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays the mean variance across all items of the oracle solution for each condition to be used as a benchmark. DIF items are shown in the grey shaded area.

Finally, Figure 5.12 visualizes variances of the population parameters. The three methods: Reg-DIF, decorrected score function bias correction, and the model refit method show similar uncertainty in population parameter estimates. Compared

with the oracle solution, all three methods showed less variability in $\hat{\boldsymbol{\delta}}$ but similar variability in $\hat{\boldsymbol{\gamma}}$. As the sample size increased, not only the variability in population parameter estimates decreased but also the differences between methods decreased.

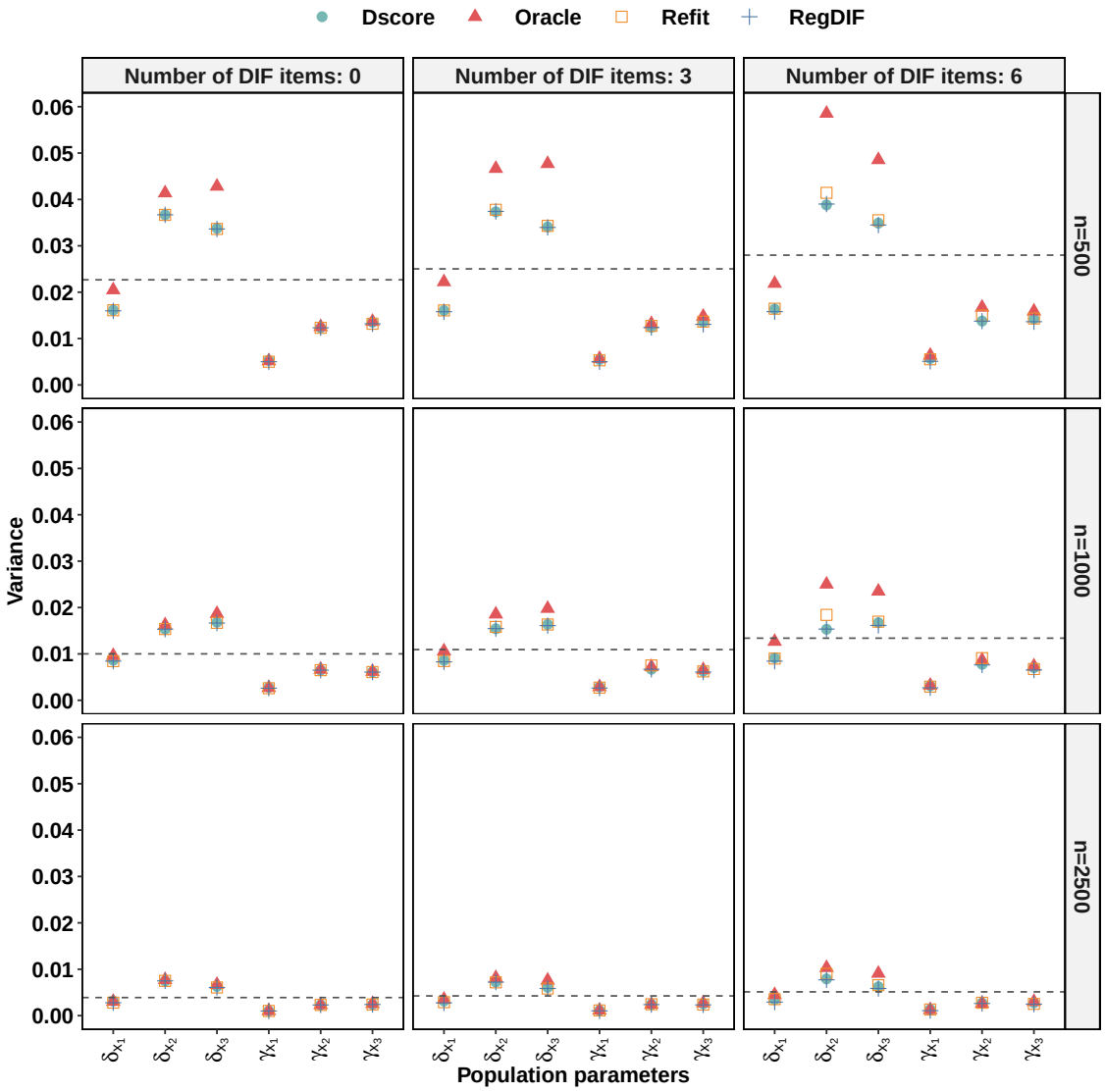


Figure 5.12. Variances of population parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays the mean variance across all items of the oracle solution for each condition to be used as a benchmark.

5.3 Results: Standard Error Estimates

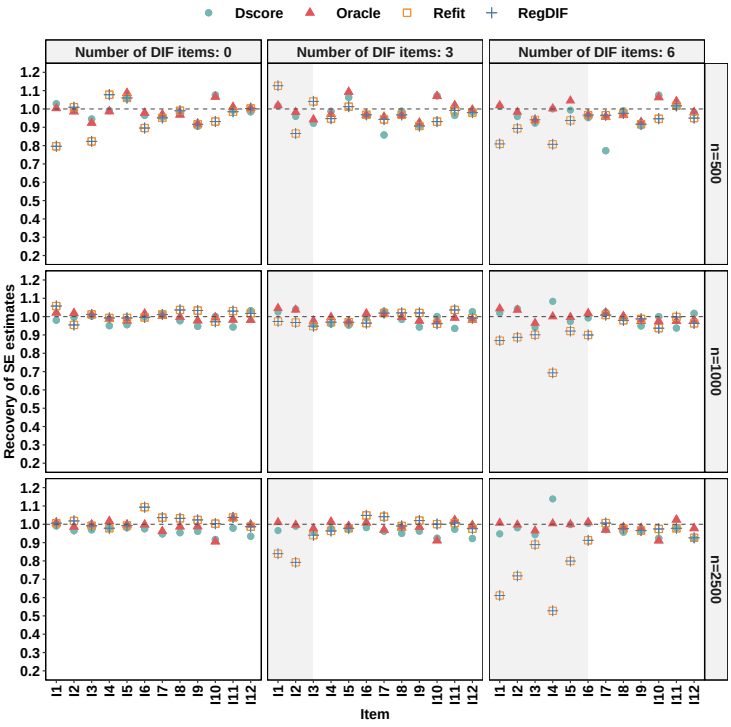
To evaluate the standard error (SE) estimates, recovery of SE and relative efficiency were computed. The recovery of SE was calculated as the ratio between the square root of the mean of the variance of parameter estimates over the empirical standard deviation of the Monte Carlo parameter estimates. If the ratio is closer to 1, the method has a more valid uncertainty measure. Efficiency was calculated as the ratio between the mean of the variance of parameter estimates over the mean of the variance of the oracle parameter estimates. One caveat when calculating the SE using the model refit method is that there is no valid SE estimates for DIF parameters penalized to be 0. If the Reg-DIF method did not select a specific DIF effect for a specific replication, the corresponding DIF covariate was excluded from the model refit method. In such cases, SE estimates of the unselected DIF effect was treated as 0. These arbitrary 0 SEs would artificially decrease the average of the estimated SEs across replications and thus should be interpreted with caution. Alternatively, omitting DIF effects penalized to be 0 often exclude too many DIF effects across replications resulting in instability of the empirical SEs and inaccurate recovery SE measure[1]. Fortunately, the interpretation of the recovery of SE for the item parameter was not impacted. To avoid the same issue, the oracle solution presented onward were based on the SEs of the MNLFA model parameter estimates treating the last two items as anchors.

Standard Error Recovery

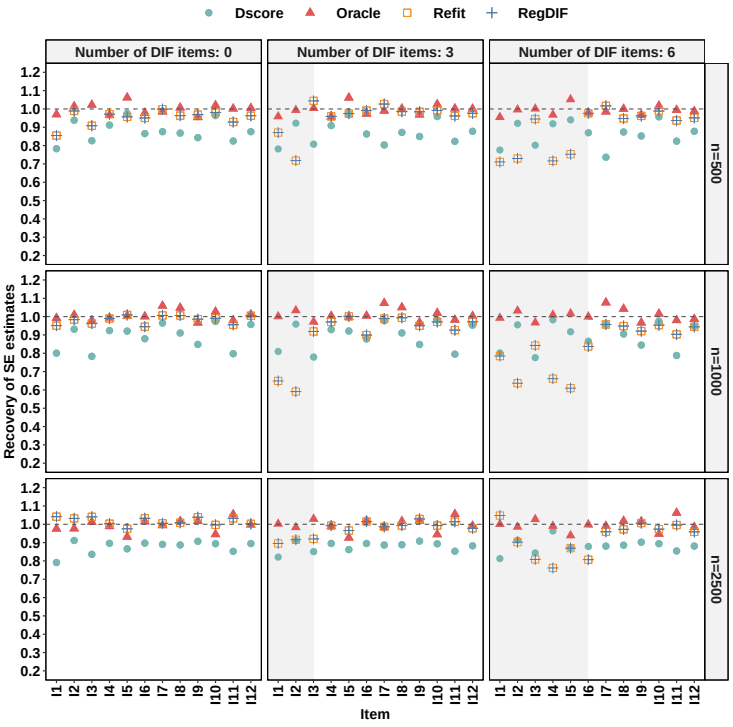
[1]Recovery of SEs measure for the model refit method based on non-zero DIF effects only has strange large values. Similar findings were reported from Chen et al. (2021)

The SE recovery for the item parameters, a-DIF, d-DIF, and population parameters are visualized in Figure 5.13 to 5.16. As there are no valid SE estimates from the penalized EM alogrithm, the Reg-DIF method currently presented in the following figures were based on the SE estimates using the model refit method. For item parameters, the model refit method tended to underestimate the SEs of item parameters for the DIF items while recovered SEs of item parameters for the anchor items well. The ratio between the average SEs and the empirical SEs can be as low as 0.55 for the DIF item when the sample size was large and the number of DIF items was large while SE ratio was approximately 1 for the anchor items. In contrast, the one-step debiased estimator using the decorrelated score function recovered the SE of the item slope parameter well but underestimated the SE of the item intercept parameter. The SE ratios for item slope parameters using the one-step debiased estimator were approximately 1 but were in general around 0.8 to 0.9 for item intercept parameters for both DIF and anchor items. The recovery of SEs of DIF parameters were displayed in Figure 5.14 and Figure 5.15. Note that as some DIF parameters of certain items were never selected by the Reg-DIF method leaving the recovery of SEs unavailable for the model refit method. This is actually one of the disadvantages of using the model refit method that inferences can only be drawn on the selected DIF parameters. Again, the model refit method tended to underestimate SEs of DIF parameters as much as 80% for both the a-DIF and d-DIF parameters. The underestimation of the SEs of DIF effects could contribute to incorrectly detecting an anchor item as a DIF item as was shown previously shown in Figure 5.3. As a comparison, the decorrelated score method can not only

recover SEs of non-zero DIF effects but also recover zero DIF effects. For a-DIF and d-DIF, SEs ranged from 0.95 to 1.08 and 0.90 to 1.02, respectively. Note that it appeared that SEs of a-DIF and d-DIF effects for item 7 were underestimated. However, given that variances of a-DIF and d-DIF parameters for item 7 were strangely large due to the skewness of the sampling distribution, it made sense that SEs of DIF parameters of item 7 seem to be underestimated. Lastly, all methods reached comparable and better SEs of interaction effects on population means and variances (i.e., $\gamma_{x_3}, \delta_{x_3}$) than their main effects. The differences between methods were more obvious under the large sample size condition. The model refit method in general underestimated SEs of population parameters as much as 10% as compared to standard deviations of parameter estimates over 500 replications. However, the one-step debiased estimator overestimated SEs of effects of the continuous covariate on population parameters while underestimated SEs of effects of the categorical covariate on population parameters. This observation was more severe when the sample size is large.



(a) Standard error recovery of the item slope parameter



(b) Standard error recovery of the item intercept parameter

Figure 5.13. Standard error recovery of item parameters: (a) item slopes and (b) item intercepts. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays ratio of 1 indicating perfect recovery.

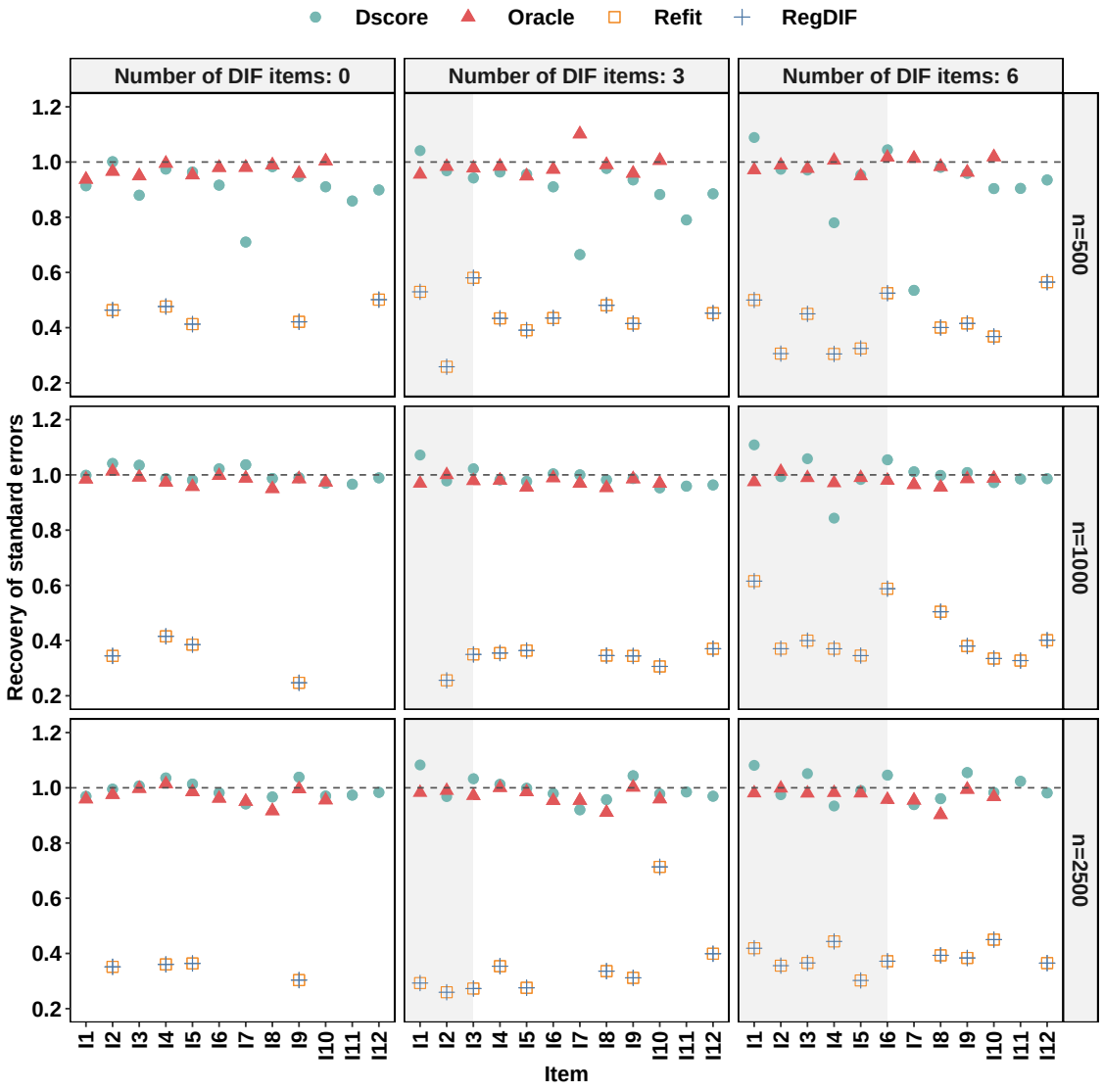


Figure 5.14. Average standard error recovery of a-DIF parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays ratio of 1 indicating perfect recovery.

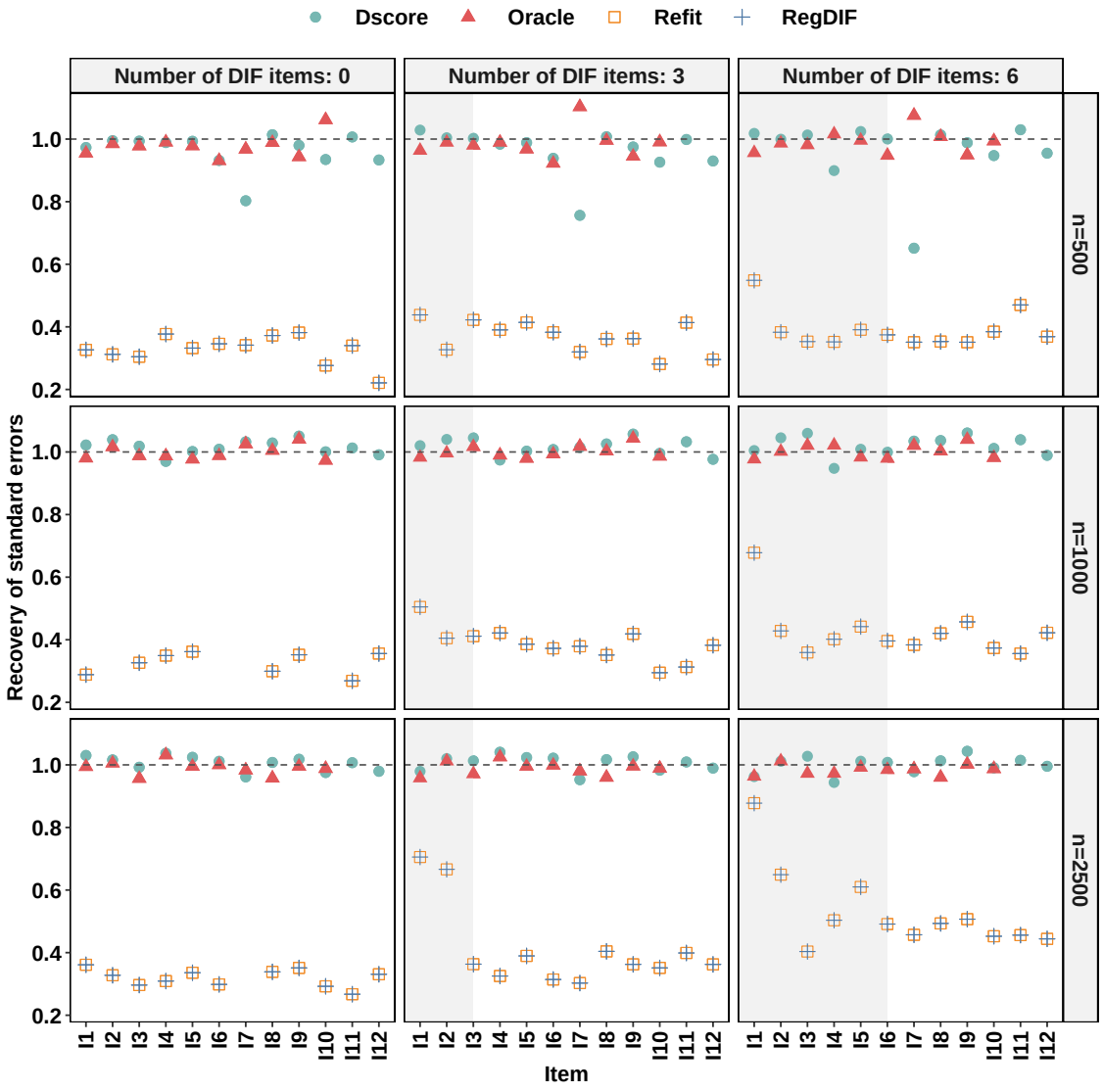


Figure 5.15. Average standard error recovery of d-DIF parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays ratio of 1 indicating perfect recovery.

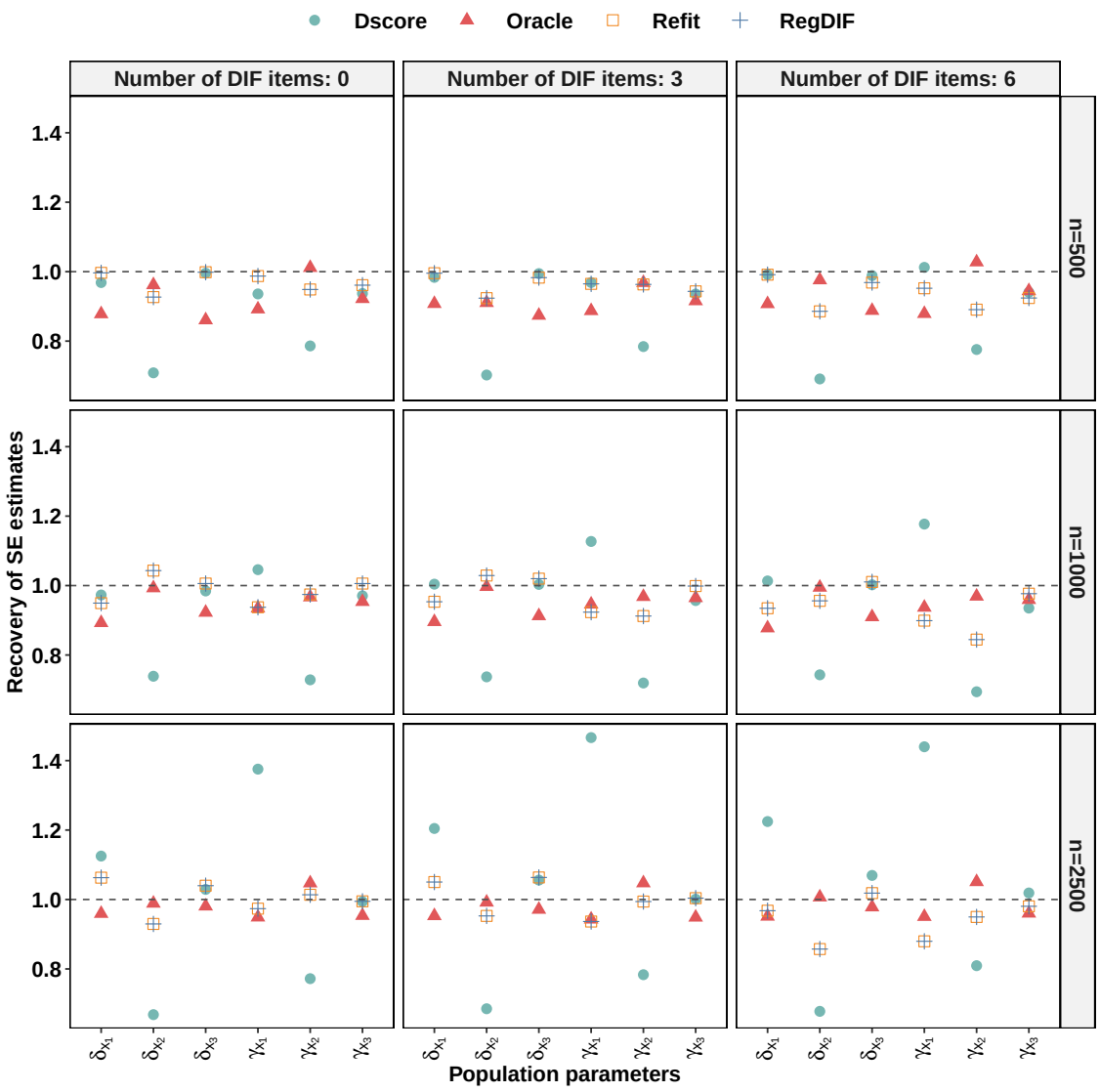


Figure 5.16. Standard error recovery of population parameters. Different methods are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition. The grey dashed reference line displays ratio of 1 indicating perfect recovery.

Relative Efficiency

The measure of relative efficiency is meaningful when SEs are recovered relatively well. Therefore, the discussion is only concentrated on the one-step debiased estimator using the decorrelated score function. Figure 5.17 to Figure 5.19 display

the relative efficiency of the one-step debiased estimator as compared to the oracle solution by treating the last two items as anchors. In general, the one-step debiased parameter estimates appeared to be more efficient than the oracle solution for item slope parameters especially for the DIF items. However, this did not mean that the one-step debias estimator was a better estimator than the ML estimator. Depending on the number of anchor items fitted with the oracle solution, results could change. This could be seen for the last two items where the relative efficiency was larger than 1. Because the last two items were treated as anchor items so that the oracle solution was more precise. The relative efficiency of the item intercept parameter should not be overly interpreted due to the fact that the SE of the item intercept parameter was underestimated. An interesting observation was that the one-step debiased estimator produces smaller standard errors for the a-DIF and d-DIF parameters in general as compared with the oracle solution with only two known anchors. A similar finding was reported in Zhang and Zhang (2014) which studies in the context of linear models. Lastly, the relative efficiency for the population parameter should not be interpreted as the SEs were not recovered well using the decorrelated score function.

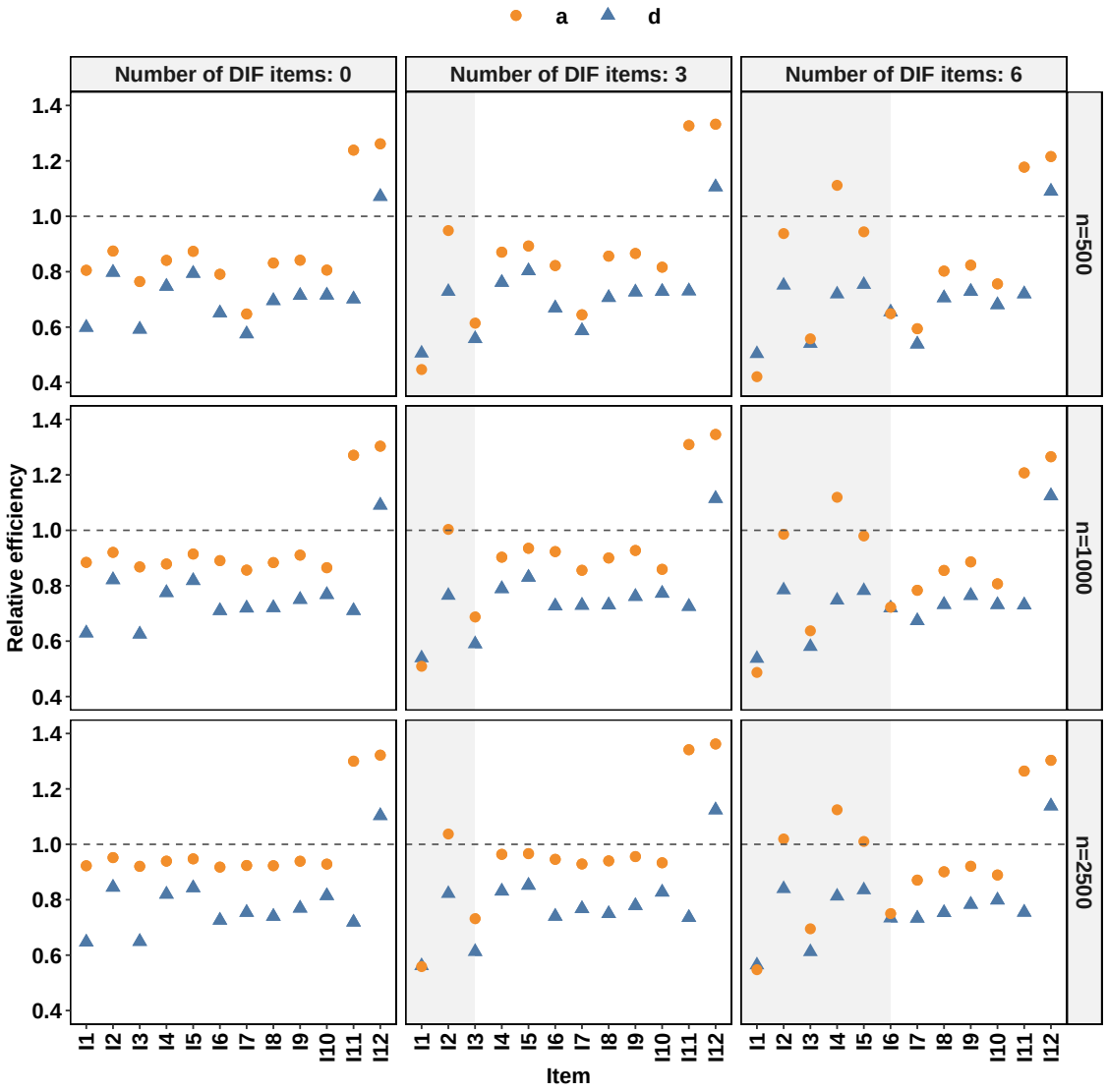


Figure 5.17. Relative efficiency of the standard error estimates for the item parameters. Different parameter types are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition.

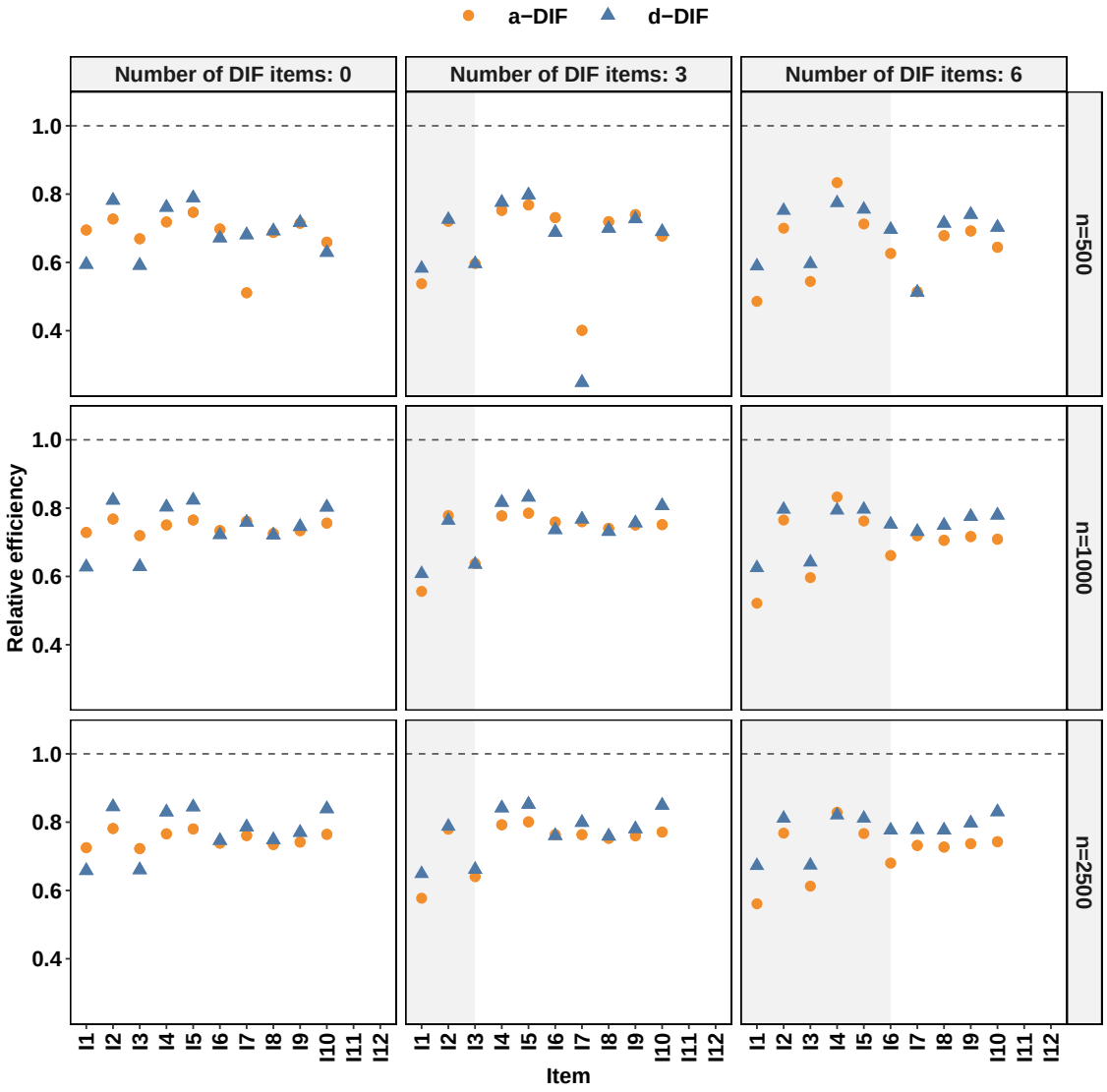


Figure 5.18. Average relative efficiency of the standard error estimates for a-DIF and d-DIF parameters. Different parameter types are displayed in different colors and shapes. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition.

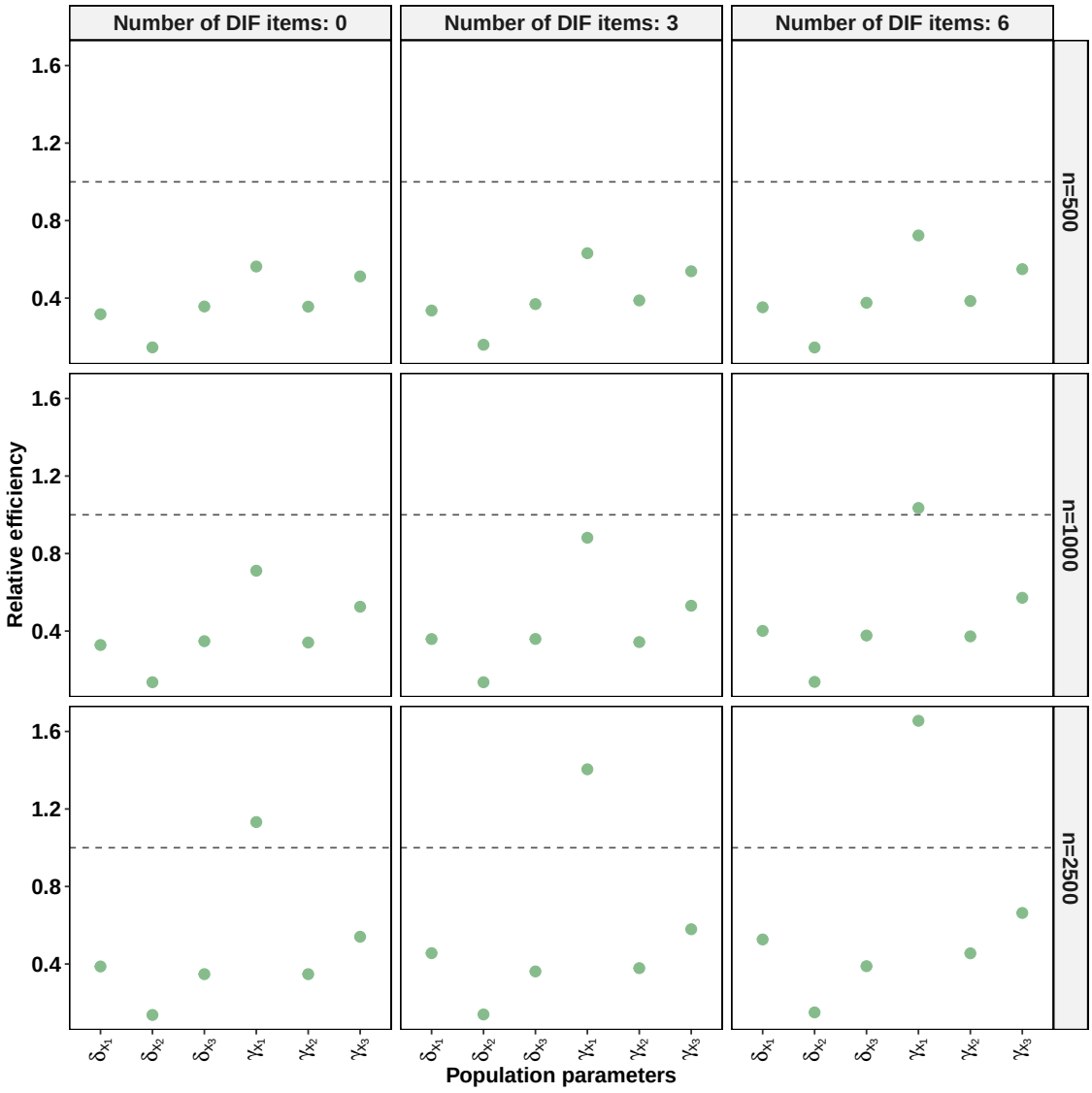


Figure 5.19. Relative efficiency of the standard error estimates for population parameters. The column shows conditions when the number of DIF items is 0, 3 or 6 out of 12 items. Each row represents a specific sample size condition.

Chapter 6: Discussion and Conclusion

6.1 Summary

One of the fundamental issues and remaining challenges inherent in DIF detection is finding the correct anchor items. Although recent development in the DIF detection literature (such as using regularization in the MNLFA modeling framework) has shown promising results in detecting DIF items without predefined anchor items, issues such as inflated false detection rate and inability to draw valid inference still remain.

The goal of the current study is to apply the decorrelated score test to test DIF items based on the L_1- penalized maximum likelihood solution under the MNLFA modeling framework. Unlike DIF detection based on regularization only, the decorrelated score test is valid for all DIF and DIF-free items across all covariates. Additionally, it has been shown that the decorrelated score function can be further used to construct an asymptotically unbiased and efficient estimator. Furthermore, the simulation study has shown the comparative performance of the decorrelated score test, DIF detection using regularization only, the model refit method, and the Wald test assuming the correct anchor items in hypothesis testing and parameter recovery.

Overall, the decorrelated score is advantageous than the other two anchor-free DIF detection methods in three aspects. First, the decorrelated score test shows more statistical power in detecting a DIF item while maintaining controlled false detection rate even compared with the oracle solution with two anchor items available. As summarized and highlighted by previous studies (Belzak & Bauer, 2020; Jacobucci et al., 2016), the Reg-DIF method and the model refit method often result in inflated false detection rate in identifying a true DIF item especially when the sample size is small. Similar to these studies, the current simulation study also shows inflated false detection rate especially under the most difficult yet realistic conditions (i.e., when the sample size is small and the number of DIF items is large). However, one difference from the previous study is that the model refit method exhibits good asymptotic behavior in detecting a DIF item while Bauer et al. (2020) and Belzak (2021) report exacerbated inflated FDR when sample size increases. As was previously discussed, the current simulation study shows that the model refit method has similar performance as the decorrelated score test in correctly identifying a true DIF item while controlling for the Type I error when the sample size is large. Belzak and Bauer (2020) tried to explain the unexpected asymptotic behavior of the model refit method in their simulation study and concluded that it is probably due to the soft-thresholding within the M-step of the penalized EM algorithm. However, the author was vague about the reason for elevated FDR when sample size increases. One possibility is that BIC calculated using penalized ML estimator is not a good λ selector. The current study avoids the model selection issue by fixing the λ to its estimated rate. Due to the conflicted performance of the

model refit method, the decorrelated score test is the preferred approach to identify DIF items when the anchors are unknown or there is not sufficiently large number of anchors. The model refit method can be trustworthy when the sample size is large and λ is selected accurately.

Another important finding from the current simulation study is that the one-step debiased estimator using the decorrelated score function yields less biased item and DIF parameter estimates. Specifically, as compared with the model refit method, the one-step debiased estimator recovers the item slopes, item intercepts, a-DIF, and d-DIF parameters well and sometimes even comparable to the oracle solution. As expected, the model refit method reduces the bias for model parameters estimated from the penalized EM algorithm. However, bias still remains. This is expected as LASSO seems to be more sensitive to detecting d-DIF rather than the a-DIF, a point that we return to at the end of this section, it is likely that a-DIF parameters are likely to be mistakenly unselected. Accordingly, the model refit method fits an incorrect model and, therefore, bias remains. Given the less biased item and DIF parameter estimates, it is recommended to use the one-step debiased estimator to produce reliable point estimates without fitting an additional MNFLA model using EM. However, less bias often indicates more parameter variability. This is evidenced by the fact that there are more uncertainty in the item parameter and DIF parameter estimates using the one-step debiased estimator based on the decorrelated score function. Another interesting observation that is alluded to earlier is that consistent with the model refit method, the one-step debiased estimator always recover item intercept parameters better as compared to item slope parame-

ters. This finding is also aligned other studies (see Bauer et al., 2020; Belzak, 2021, for examples) that d-DIF parameters are in general recovered better than a-DIF parameters, which could potentially explain the superiority of the item intercept parameter recovery. Consequently, a-DIF and item slope parameters have better initial parameter estimates to be plugged into the bias correction function.

Lastly, the decorrelated score test is able to draw reliable statistical inference on a-DIF and d-DIF parameters for zero and non-zero effects, respectively, where the model refit method may fail or is not able to generate one when the DIF effect is zero. As discussed earlier, SEs cannot be computed if DIF effects are penalized to be zero. To this end, the one-step debiased estimator is the only method considered in the present study that can produce reliable statistical inference for the zero DIF effects. As for the remaining DIF effects that are not penalized to be zero, the model refit method often underestimate the SEs of the DIF effects (Chen et al., 2021; Huang, 2018) leading to incorrect inferences. This may be of concern, if substantive researchers are interested in testing a-DIF and d-DIF separately. If so, the decorrelated score test offers promising and valid statistical test to identify DIF at both the item and the parameter levels.

Nevertheless, the simulation study also suggests one critical issue of the one-step debiased estimator based on the decorrelated score function. As summarized in in Chapter 5, parameter estimates and standard error estimates of unpenalized parameters especially the population parameters are ill-behaved. For example, unlike other model parameters, the one-step debiased estimator failed to reduce the bias in the population parameter estimate caused by the shrinkage in the initial penal-

ized parameter. Contrary to the theory, SEs of population parameters are also not recovered well from the estimated efficient information. SEs of population parameters are either over-estimated or under-estimated whereas SEs of the item intercepts are underestimated. Issues arise if the research question concerns the differences in population means and variances. Perhaps the model refit method can be used when the number of DIF items is small. However, it remains to be biased and the SE of the population mean is still underestimated when the number of DIF items is large. To this end, there is no optimal solution. As the population parameter estimate usually is not the goal of the DIF-related methodological research, parameter and SE recoveries are typically not reported in the regularized DIF literature. Findings from the current simulation cannot be verified with previous research. The debias using the decorrelated score function in the general statistical literature also did not report parameter recoveries for unpenalized parameters. Some insights in how to improve the performance of the one-step debiased estimator in recovering the population parameter are provided in the next section.

6.2 Future Studies

Although the proposed decorrelated score test and the one-step debiased estimator have shown promising results in detecting DIF items without predefined anchor items and provided valid uncertainty measures for DIF effects, limitations and issues still remain to be addressed by future studies.

First, the simulation study has suggested that the one-step debiased estima-

tor does not seem to reduce the bias of the penalized ML estimator and that the associated SEs are unreliably estimated using the efficient information. Given that the ill recovery with the uncertainty measure also happens with unpenalized item parameters, it is conjectured that additional penalty weights might be needed for the population parameter or a more careful selection of penalty weights is needed. Some preliminary investigation suggests that if the population parameter is penalized in the initial L_1 penalization stage, the bias correction using the decorrelated score function shows better performance. Future studies are needed to examine the behavior of the decorrelated score function when population parameters are penalized. Alternatively, finding the optimal λ' for the population parameter for each iteration might also improve the unpredicted behavior of SEs. In the current simulation study, λ' is set to the fixed λ based on the estimated rate as the theory suggests $\lambda \asymp \sqrt{1/n}$ and $\lambda' \asymp \sqrt{1/n}$ are approximately the same rate to ensure the l_2 error bound for initial parameter estimates. It might worth to explore the influence of λ' on the SE recovery of the unpenalized parameter. Although Ning and Liu (2017) and Fang et al. (2017) have reported that the performance of the decorrelated score test is insensitive to the actual λ' value in the high-dimensional regression model and proportional hazard model, it is not clear whether λ' will impact the one-step debiased estimator especially when there are unpenalized parameters in the model.

Second, as λ is critically important for the initial parameter estimates and the accuracy of DIF detection using the Reg-DIF method and the model refit method, future research is encouraged to investigate different model selection critera in addition to BIC. As mentioned in Chapter 4, a pilot study found that BIC calculated

based on the penalized EM parameter estimates always selects a relatively smaller λ value when the DIF effect size is relatively large and thus results in too many false positives. The current study used a fixed λ value estimated from the rate calculated from the $n = 500$ condition based on the ML estimator which potentially avoids the issue of poor performance of BIC. This may also explains the relatively smaller inflated false detection rate and power as compared with other Reg-DIF studies (e.g., Bauer et al., 2020; Belzak & Bauer, 2020). Given the sparse literature on model selection accuracy in the regularized DIF framework, future research is needed to find viable and efficient penalty selection approaches such as adaptive data-dependent penalty selection (e.g., Chichignoud, Lederer, & Wainwright, 2016).

Third, future research is needed to investigate the impact of different penalty functions on the performance of the decorrelated score test in DIF detection and the asymptotic behavior of the debiased parameter estimator. The general theory provided in Ning and Liu (2017) can be directly applied to many penalty functions (e.g., adaptive LASSO, Zou, 2006). Our simulation study along with the other studies (e.g., Belzak & Bauer, 2020) showed similar differential performance in recovering different parameters and SEs. In addition, applying different penalty weights for a-DIF and d-DIF may yield better initial parameter estimates and thus improve the performance of the one-step debiased estimator.

Fourth, although the original theorem from Ning and Liu (2017) discussed the high-dimensional case where the number of covariate is no longer fixed but grows potentially larger than the sample size, the current study limits the discussion to the low-dimensional case assuming that the number of covariates is fixed. Efforts can

be made to extend to the high-dimensional setting where the number of grouping variables exceeds the sample size. For example, with the development of computer adaptive tests, more person-level data such as process data might be of interests to be used as covariates.

Fifth, the performance of the proposed decorrelated score test can be investigated under more variety of conditions such as when the inherent model assumptions are violated. The current simulation study assumes the same linear functional form impacts the person covariate vector on the item intercept and item slope in the same manner. However, in reality, the influence of the person covariate on the item characteristics can follow a different functional form and that item responses can be categorical. Understanding the behavior of the decorrelated score test when the DIF effect is misspecified can inform applied research. Although it is, in general, recommended to find the best fitting model including the functional form before testing DIF, knowing how robust the decorrelated score test is in terms of power and false detection rate in identifying a DIF item when the model is misspecified will be beneficial. If recovery of a DIF effect is of interest, the functional form of the item slope and item intercept functions can be also be approximated using basis expansion. Moreover, the proposed decorrelated score test in testing DIF items can be extended to other popular item response models (e.g., graded response model, GRM, Samejima, 1969). Similarly, the MNLFA model specified in the simulation study assumes conditional independence on the latent ability. This is a strong assumption that in many testing scenarios is likely violated. For instance, items could be nested within the same item context and thus have residual correlation even

after controlling for the latent construct. In such a case, it should be straightforward to accommodate multidimensional latent variables, such as a fully correlated or multidimensional factors model.

Lastly, DIF effect size measures can be critically important when deciding to remove, modify, or keep a flagged DIF item. In practice it might be more desirable or more realistic to flag items with large DIF impact, which makes effect size reporting more essential in DIF detection. Latent regression models such as MNLFA models provide a natural DIF effect size measure (i.e., the person covariate effect on the item slope $\boldsymbol{\beta}_{1j}$ and item intercept $\boldsymbol{\beta}_{0j}$). However, more thorough and careful study on the meaningful cut-off values are needed to facilitate decision making. Alternatively, item level or scale level effects can also be helpful and can be extended to the MNLFA modeling framework. For example, average unsigned difference (see Woods, 2011) which calculates weighted differences in the expected response functions between the focal and reference groups can be extended to the MNLFA model for selected values or levels of person covariate to evaluate the magnitude of the DIF effect at the item level. In addition to the item level DIF impact, the differential test function (DTF) index (Roju, Van der Linden, & Fleer, 1995) or the expected total test score difference due to DTF (Stark, Chernyshenko, & Drasgow, 2004) can be computed to inform the overall DIF effect on the scale level.

In sum, the proposed decorrelated score test and its one-step debiased estimator based on the regularized moderated nonlinear factor analysis model offers a promising solution to the practical obstacles encountered by conventional IRT methods. Valid inference at the DIF parameter level opens doors for more complicated

substantive research questions such as investigating the complex nature of DIF due to interconnection of background characteristics.

Appendix A: The Coordinate Descent Algorithm

As discussed in Section 3.2.2, updating item parameter $\boldsymbol{\xi}_j^{(r+1)}$ for item j at iteration $(r+1)$ amounts to solving a penalized weighted least square problem with pseudo data of sample size nQ. Specifically, the pseudo person covariate matrix $\underline{\mathbf{x}}$ can be constructed as $\underline{\mathbf{x}} = \mathbf{x} \otimes 1_Q \in \mathbb{R}^{nQ \times K}$, where 1_Q is a Q-dimensional vector with 1s and $\otimes$ denotes the Kronecker product. Similarly, the pseudo item response vector for item j can be construct using $\underline{\mathbf{y}}_j = \mathbf{y}_j \otimes 1_Q \in \mathbb{R}^{nQ}$. Also, let $\boldsymbol{\theta} = (\vec{\theta}_1', \cdots, \vec{\theta}_n')^\top$ where $\vec{\theta}_q = (\theta_{i1}, \cdots, \theta_{iQ})^\top$ and $\boldsymbol{\theta} \in \mathbb{R}^{nQ}$. The conditional probability of getting item j right given the latent ability θ_i can be evaluated as

$$p^{(r)}(\underline{y}_i = 1 | \theta_{iq}, \underline{\mathbf{x}}_i, \boldsymbol{\xi}_j^{(r)}) = \frac{1}{1 + \exp\left[-(a_j^{(r)}\theta_i + d_j^{(r)} + \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j}^{(r)} + \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j}^{(r)}\theta_i)\right]}. \tag{A.1}$$

Then, the negative weighted conditional log-likelihood in Equation 3.15 (the first term of the equation) can be approximated as follows after taking the 2^{nd} order Taylor expansion at the current value $\boldsymbol{\xi}^{(r)}$

$$\frac{1}{2n} \sum_{i=1}^{nQ} w_i (z_i - d_j - a_j - \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j} - \theta_i \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j})^2 + C(\boldsymbol{\xi}_j^{(r)})^2, \tag{A.2}$$

where

$$z_i = d_j^{(r)} + a_j^{(r)} + \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j}^{(r)} + \theta_i \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j}^{(r)} + \frac{\underline{y}_i - p^{(r)}}{p^{(r)}(1-p^{(r)})} \tag{A.3}$$

$$w_i = e_{iq}^{(r)} p^{(r)}(1-p^{(r)}), \tag{A.4}$$

in which $\underline{y}_i$ and θ_i is the i^{th} element of $\underline{\mathbf{y}}_j$ and $\boldsymbol{\theta}$, respectively. $C(\cdot)$ is a constant evaluated at the current value. Then, the coordinate descent to solve the penalized weighted least-square can be achieved using the following pseudo code

Algorithm 2 Coordinate Descent in Logistic Regression with Lasso Penalty

Initialize $\boldsymbol{\xi}_j^{(r)}$ using random values
repeat
  for $m = 1, \ldots, K$ **do**
    calculate z_i as in Equation A.3
    calculate w_i as in Equation A.4
    update $d_j^{(r+1)} = \frac{\sum w_i(z_i - a_j^{(r)} - \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j}^{(r)} - \theta_i \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j}^{(r)})}{\sum w_i}$
    update $a_j^{(r+1)} = \frac{\sum w_i(z_i - d_j^{(r)} - \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j}^{(r)} - \theta_i \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j}^{(r)})}{\sum w_i}$
    update $\boldsymbol{\beta}_{0jm}^{(r+1)} = \frac{\mathcal{S}\left(\sum w_i \underline{x}_{im}(z_i - a_j^{(r)} - d_j^{(r)} - \underline{\mathbf{x}}_{i,-m}^\top \boldsymbol{\beta}_{0j,-m}^{(r)} - \theta_i \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{1j}^{(r)}, \lambda\right)}{\sum w_i \underline{\mathbf{x}}_{im}^2}$
    update $\boldsymbol{\beta}_{1jm}^{(r+1)} = \frac{\mathcal{S}\left(\sum w_i \underline{x}_{im}(z_i - a_j^{(r)} - d_j^{(r)} - \underline{\mathbf{x}}_i^\top \boldsymbol{\beta}_{0j}^{(r)} - \theta_i \underline{\mathbf{x}}_{i,-m}^\top \boldsymbol{\beta}_{1j,-m}^{(r)}, \lambda\right)}{\sum w_i \underline{\mathbf{x}}_{im}^2}$
    $\triangleright$ $i = 1, \ldots, nQ$
    $\triangleright$ $\mathcal{S}(a, b)$: soft thresholding operator
  end for
until converged.

Appendix B: Derivatives of the Observed Fisher Information

The first and second derivatives of the $\log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)$ with respect to $\boldsymbol{\xi}$ in Equation 3.16 can be expressed as follows. For the gradient $\vec{\mathbf{s}}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i) = (\vec{\mathbf{s}}_1, \vec{\mathbf{s}}_2, \ldots, \vec{\mathbf{s}}_J, \vec{\mathbf{s}}_{\boldsymbol{\gamma}}, \vec{\mathbf{s}}_{\boldsymbol{\delta}})^\top$, each of the element is expressed as follows

$$\begin{aligned}
\vec{\mathbf{s}}_j &= \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\xi}_j} \\
&= \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial (d_j, a_j, \boldsymbol{\beta}_{0j}, \boldsymbol{\beta}_{1j})^\top} \\
&= (y_{ij} - p_{ij}, (y_{ij} - p_{ij})\theta_i, (y_{ij} - p_{ij})\mathbf{x}_i^\top, (y_{ij} - p_{ij})\mathbf{x}_i^\top \theta_i)^\top,
\end{aligned}$$

in which $p_{ij} = f(y_{ij} = 1|\theta_i, \mathbf{x}_i)$ (see Equation 2.2).

$$\begin{aligned}
\vec{\mathbf{s}}_{\boldsymbol{\gamma}} &= \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\gamma}} \\
&= \frac{(\theta_i - \mathbf{x}_i^\top \boldsymbol{\gamma})\mathbf{x}_i^\top}{\exp(\mathbf{x}_i^\top \boldsymbol{\delta})}
\end{aligned}$$

and

$$\begin{aligned}
\vec{\mathbf{s}}_{\delta} &= \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\delta}} \\
&= \frac{(\theta_i - \mathbf{x}_i^\top \boldsymbol{\gamma})^2 \mathbf{x}_i^\top}{2\exp(\mathbf{x}_i^\top \boldsymbol{\delta})} - \frac{\mathbf{x}_i^\top}{2}.
\end{aligned}$$

For Hessian matrix $\mathbf{H}(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i) = \text{diag}(\mathbf{H}_j, \mathbf{H}_{\boldsymbol{\gamma},\boldsymbol{\delta}})$ in which $\mathbf{H}_J$ is a block diagonal matrix defined as

$$\mathbf{H}_J = \frac{\partial \log f(\boldsymbol{\xi}, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^\top} \tag{B.1}$$

$$= \begin{bmatrix} \mathbf{H}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{H}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{H}_J \end{bmatrix}. \tag{B.2}$$

$\mathbf{H}_j = \frac{\partial \log f(\boldsymbol{\xi}_j, \theta_i, \mathbf{y}_i|\mathbf{x}_i)}{\partial \boldsymbol{\xi}_j \partial \boldsymbol{\xi}_j^\top}$ is the Hessian matrix with respect to item parameters of item j written as

$$\mathbf{H}_j = \begin{bmatrix} p_{ij}(p_{ij}-1) & p_{ij}(p_{ij}-1)\theta_i & p_{ij}(p_{ij}-1)\mathbf{x}_i^\top & p_{ij}(p_{ij}-1)\mathbf{x}_i^\top\theta_i \\ p_{ij}(p_{ij}-1)\theta_i & p_{ij}(p_{ij}-1)\theta_i^2 & p_{ij}(p_{ij}-1)\mathbf{x}_i^\top\theta_i & p_{ij}(p_{ij}-1)\mathbf{x}_i^\top\theta_i^2 \\ p_{ij}(p_{ij}-1)\mathbf{x}_i & p_{ij}(p_{ij}-1)\mathbf{x}_i\theta_i & p_{ij}(p_{ij}-1)\mathbf{x}_i\mathbf{x}_i^\top & p_{ij}(p_{ij}-1)\mathbf{x}_i\mathbf{x}_i^\top\theta_i \\ p_{ij}(p_{ij}-1)\mathbf{x}_i\theta_i & p_{ij}(p_{ij}-1)\theta_i^2\mathbf{x}_i^\top & p_{ij}(p_{ij}-1)\mathbf{x}_i\mathbf{x}_i^\top\theta_i & p_{ij}(p_{ij}-1)\mathbf{x}_i\mathbf{x}_i^\top\theta_i^2 \end{bmatrix}.$$

The Hessian matrix with respect to the population parameter ($\mathbf{H}_{\boldsymbol{\gamma},\boldsymbol{\delta}}$) is defined as

$$\mathbf{H}_{\boldsymbol{\gamma},\boldsymbol{\delta}} = \begin{bmatrix} \frac{-\mathbf{x}_i\mathbf{x}_i^\top}{\exp(\mathbf{x}_i^\top\boldsymbol{\delta})} & \frac{-(\theta_i-\mathbf{x}_i^\top\boldsymbol{\gamma})\mathbf{x}_i\mathbf{x}_i^\top}{\exp(\mathbf{x}_i^\top\boldsymbol{\delta})} \\ \frac{-(\theta_i-\mathbf{x}_i^\top\boldsymbol{\gamma})\mathbf{x}_i\mathbf{x}_i^\top}{\exp(\mathbf{x}_i^\top\boldsymbol{\delta})} & \frac{-(\theta_i-\mathbf{x}_i^\top\boldsymbol{\gamma})^2\mathbf{x}_i\mathbf{x}_i^\top}{2\exp(\mathbf{x}_i^\top\boldsymbol{\delta})} \end{bmatrix}.$$

Appendix C: Assumptions

As mentioned before, the general theory of the decorrelated score test and its one-step estimator are established under some assumptions (Ning & Liu, 2017). These assumptions need to be verified to guarantee the asymptotic properties of the Dscore test and the asymptotic unbiased estimator as shown in Equations 3.5 and 3.7. The following four assumptions are briefly summarized and the verification will be provided in the final dissertation.

Assumption C.0.1 (Consistency conditions for initial parameter estimates)**.** *For some sequence* $a_1(n)$ *and* $a_2(n)$ *converge to* 0 *as* $n \to \infty$*, it holds*

$$\lim_{n\to\infty} P_{\boldsymbol{\xi}^*}(\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1 \lesssim a_1(n)) = 1 \tag{C.1}$$

$$\lim_{n\to\infty} P_{\boldsymbol{\xi}^*}(\|\hat{\mathbf{W}} - \mathbf{W}^*\|_1 \lesssim a_2(n)) = 1, \tag{C.2}$$

where $\|\|_1$ stands for the L_1 operator norm of a matrix (e.g., $\|\mathbf{A}\|_1 = max_{1\le j\le m}\sum_{i=1}^{n}|a_{ij}|$) and $\lesssim$ denotes that the left side is less than or equal to the right hand side times some constant $C > 0$. Although estimation consistency of LASSO has been studied in linear and generalized linear models (Knight & Fu, 2000; Ning & Liu, 2017), it has never been studied in the latent variable modeling framework. As the loss function or the negative sample log-likelihood of MNLFA model is not strictly convex,

the proof of parameter estimation consistency can be non-trivial. Moreover, as mentioned by Ning and Liu (2017), there can be extra difficulty in bounding $\|\hat{\mathbf{W}} - \mathbf{W}^*\|$, as $\hat{\mathbf{W}}$ depends on $\hat{\boldsymbol{\xi}}$.

Assumption C.0.2 (Concentration of the gradient and Hessian)**.** *Let* $\mathbf{V}^* = (\mathbf{I}_{d_0 \times d_0}, -\mathbf{W}^{*\top})^\top$, *then assume*

$$\|\nabla \ell(\boldsymbol{\xi}^*)\|_\infty = \mathcal{O}_p(\sqrt{\log d/n}) \tag{C.3}$$

$$\|\mathbf{V}^{*\top} \nabla^2 \ell(\boldsymbol{\xi}^*) - \mathbb{E}_{\boldsymbol{\xi}^*}(\mathbf{V}^{*\top} \nabla^2 \ell(\boldsymbol{\xi}^*))\|_\infty = \mathcal{O}_p(\sqrt{\log d/n}), \tag{C.4}$$

where $\|\|_\infty$ *of a vector* $\vec{A}$ *for example indicates* $\|\vec{A}\|_\infty = \max_{1 \le i \le d} |a_i|$ *and of a matrix is the maximum absolute row sum of the matrix (i.e.,* $\|\mathbf{A}\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^m |a_{ij}|$ *). In the low dimensional setting, it might be sufficient to prove under some finite moment assumptions on the gradient and Hessian matrix.*

Assumption C.0.3 (Local smoothness on the loss function)**.** *Let* $\hat{\boldsymbol{\xi}}_0 = (\mathbf{0}, \hat{\boldsymbol{\eta}}^\top)^\top$, $\hat{\mathbf{V}} = (\mathbf{I}_{d_0 \times d_0}, -\hat{\mathbf{W}}^\top)^\top$, *and* $\mathbf{V}^* = (\mathbf{I}_{d_0 \times d_0}, -\mathbf{W}^{*\top})^\top$. *For both* $\check{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_0$ *and* $\check{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}$, *it holds that*

$$\|\mathbf{V}^{*\top}\{\nabla \ell(\check{\boldsymbol{\xi}}) - \nabla \ell(\boldsymbol{\xi}^*) - \nabla^2 \ell(\boldsymbol{\xi}^*)(\check{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)\}\|_\infty = o_p(n^{-1/2}) \tag{C.5}$$

$$\|(\hat{\mathbf{V}} - \mathbf{V}^*)^\top (\nabla \ell(\check{\boldsymbol{\xi}}) - \nabla \ell(\boldsymbol{\xi}^*))\|_\infty = o_p(n^{-1/2}). \tag{C.6}$$

This assumption implicitly assumes that the $\ell(\boldsymbol{\xi})$ is second-order differentiable. The verification of the assumption should be straightforward if the loss function is a quadratic form of $\boldsymbol{\xi}$.

Assumption C.0.4 (Convergence of the score function)**.** *Let* $\boldsymbol{\Sigma}^* = \lim_{n\to\infty} Var(n^{1/2}\nabla\ell(\boldsymbol{\xi}^*))$. *Then the score function holds that*

$$\sqrt{n}\nabla\ell(\boldsymbol{\xi}^*)^\top \mathbf{V}^*(\mathbf{V}^{*\top}\boldsymbol{\Sigma}^*\mathbf{V}^*)^{-1}\mathbf{V}^{*\top}\nabla\ell(\boldsymbol{\xi}^*) \xrightarrow{\mathcal{D}} \chi^2_{d_0}. \tag{C.7}$$

This assumption can be established by verifying the Lindeberg's condition, which is a sufficient condition for a sequence of independent random variables.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526.

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125.

Belzak, W. (2021). *Using regularization to evaluate differential item functioning among multiple covariates: A penalized expectation-maximization algorithm via coordinate descent and soft-thresholding* (Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690.

Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. *Journal of Educational and Behavioral Statistics*, *41*(6), 559–592.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 802–837.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). MA: Addison-Wesley.

Bock, R., & Aitkin, M. (1982). Marginal maximum likelihood estimation of item parameters. *Psychometrika*, *47*(3), 369–369.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, *35*(2), 179–197.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In *Handbook of modern item response theory* (pp. 433–448). Springer, New York, NY.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment*, *16*(2), 155.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented

EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 309–329.

Cai, L. (2017). flexmirt 3.51: Flexible multilevel multidimensional item analysis and test scoring [computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]. Lincolnwood, IL: Scientific Software International.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the center for epidemiological studies depression (CES-D) scale: an item response theory analysis. *Medical Care*, 281–289.

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2021). Advantages of spike and slab priors for detecting differential item functioning relative to other bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–18.

Chichignoud, M., Lederer, J., & Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research*, *17*(1), 8162–8181.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25–29). New York & London: Routledge.

Coley, R. J. (2001). Differences in the gender gap: Comparisons across racial/ethnic groups in education and work. policy information report.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the mini-mental state examination. *Medical Care*, S134–S142.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, *6*(1), 21–29.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fang, E. X., Ning, Y., & Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(5), 1415–1437.

Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from https://www.jstatsoft.org/v033/i01 doi: 10.18637/jss.v033.i01

Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

Glas, C. A. (1998). Detection of differential item functioning using lagrange multiplier tests. *Statistica Sinica*, *8*(3), 647-667.

Glas, C. A. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*(3), 273–294.

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

Haberman, S. J. (1988). A stabilized newton-raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, 193–211.

Harris, N. (2014). *Visualizing lasso polytope geometry.* https://www.naftaliharris.com/blog/lasso-polytope-geometry/.

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

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499–522.

Huang, P.-H. (2020). Postselection inference in structural equation modeling. *Multivariate Behavioral Research*, *55*(3), 344–360.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566.

Javanmard, A., & Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in neural information processing systems* (pp. 1187–1195).

Javanmard, A., & Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, *60*(10), 6522–6554.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345–355.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). Department of Psychology, University of North Carolina; Chapel Hill.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, S. Star, & J. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York: Wiley.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927.

Li, X., & Jacobucci, R. (2021). Regularized structural equation modeling with stability selection. *Psychological Methods*.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 722–734.

Lindstrøm, J. C., & Dahl, F. A. (2020). Model selection with lasso in multi-group structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 33–42.

Liu, M. (2017). *Differential item functioning in large-scale mathematics assessments: Comparing the capabilities of the rasch trees model to traditional approaches* (Unpublished doctoral dissertation). University of Toledo.

Liu, Y., & Yang, J. S. (2018). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*, *83*(2), 333–354.

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

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 226–233.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in naep mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, *37*(2), 129–150.

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681.

Millsap, R. E., Gunn, H., Everson, H., & Zautra, A. (2015). Using item response

theory to evaluate measurement invariance in health-related measures. In S. P. Reise & D. A. Revicki (Eds.), .

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 337–357.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.

Ning, Y., & Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, *45*(1), 158–195.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a spanish translation of the ptsd checklist: detection and evaluation of impact. *Psychological Assessment*, *14*(1), 50.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reise, S. P., & Revicki, D. A. (2014). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Roju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). Irt-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Richmond, VA: Psychometric Society.

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*(1), 279–294.

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

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Qquality of Life Outcomes*, *8*(1), 1–9.

Shea, C. A. (2013). *Using a mixture IRT model to understand english learner performance on large-scale assessments*. University of Massachusetts Amherst.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184–199.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*(3), 497.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402.

Steinberg, L., & Thissen, D. (2013). Item response theory. In J. S. Comer & P. C. Kendall (Eds.), *The oxford handbook of research strategies for clinical psychology* (pp. 336–373). Oxford University Press.

Taylor, J., & Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, *46*(1), 41–61.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*(1), 21–43.

Vaart, A. W. v. d. (1998). Semiparametric models. In *Asymptotic statistics* (p. 358–432). Cambridge University Press. doi: 10.1017/CBO9780511802256.026

Van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202.

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics*, *0*(0), 10769986221109208. Retrieved from https://doi.org/10.3102/10769986221109208 doi: 10.3102/10769986221109208

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113-144. Retrieved from https://doi.org/10.1207/s15324818ame1702_2

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498.

Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics*, *37*(5A), 2178.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57.

Woods, C. M. (2009b). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57.

Woods, C. M. (2011). Dif testing for ordinal items with poly-sibtest, the mantel and gmh tests, and irt-lr-dif when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement*, *35*(2), 145–164.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547. Retrieved from https://doi.org/10.1177/0013164412464875

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361.

Yuan, K.-H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for mles of item parameters in irt. *Psychometrika*, *79*(2), 232–254.

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: Qq plots and graphical test. *Psychometrika*, *86*(2), 345–377.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.