USING REGULARIZATION TO EVALUATE DIFFERENTIAL ITEM FUNCTIONING AMONG MULTIPLE COVARIATES: A PENALIZED EXPECTATION-MAXIMIZATION ALGORITHM VIA COORDINATE DESCENT AND SOFT-THRESHOLDING

William C. M. Belzak

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology and Neuroscience (Quantitative).

Chapel Hill 2021

Approved by:
Daniel J. Bauer
Kathleen M. Gates
Andrea M. Hussong
Yufeng Liu
Naim U. Rashid
David Thissen
ABSTRACT

William C. M. Belzak: Using Regularization to Evaluate Differential Item Functioning Among Multiple Covariates: A Penalized Expectation-Maximization Algorithm via Coordinate Descent and Soft-Thresholding
(Under the direction of Daniel J. Bauer)

Testing for differential item functioning (DIF) has undergone rapid statistical developments in recent years. Namely, the moderated nonlinear factor analysis (MNLFA) model allows for simultaneous testing of DIF in multiple categorical and continuous covariates (e.g., age, gender, ethnicity, etc.). Recent work has also implemented a LASSO regularization approach to identify DIF and select anchor items for model identification. Although regularized MNLFA provides greater flexibility to evaluate DIF, less development has been made in efficiently estimating model parameters. Most previous implementations of MNLFA have directly maximized the observed marginal likelihood function, which limits the method to only a few items and covariates. Additionally, penalization in the MNLFA model has only been performed outside of the optimization routine, which results in a non-standard method for setting estimates to zero. To overcome these difficulties, I introduce a penalized expectation-maximization (EM) algorithm that efficiently estimates many more item parameters than previous implementations and performs regularization during optimization. I extend the regularized MNLFA model to include not just soft-thresholding for LASSO penalization, but also firm-thresholding for the MCP approach. A Monte Carlo simulation study and an empirical data analysis evaluates this new algorithm, comparing the LASSO and MCP approaches against previous work. Finally, a discussion of future research directions concludes the dissertation.
To the many who dedicated their professions to advancing the science of psychological measurement,
I am indebted.
ACKNOWLEDGEMENTS

This dissertation would not exist without the unconditional love and support from my parents, Byron and Kathleen, and from my brother, Ben. You have given me the most precious gift any child and brother can receive: a loving and stable home.

To my fiancé, Brooke: you have been my rock ever since we met. Much of my confidence and success in graduate school is because of you. I am also thankful to have a new family to call my own. Cassandra, Helen, Dylan, Eugena, Dean, and especially, Darren: thank you for opening your arms to me. Although I wish you were still here, Darren, to celebrate the union of our families and to appreciate this accomplishment, I know you were proud of Brooke and I; I know I am proud to call you my future father-in-law.

Thank you to my friends in the Thurstone Psychometric Laboratory. You made graduate school a joyful experience. I am also indebted to my Raleigh and Carrboro crew, who have always provided a home away from home when I needed it the most.

Finally, I am most grateful to have had an advisor and professional mentor that so effortlessly guided my development as a scientist and psychometrician. I can say without a doubt that my graduate training would have been far less efficient without you, Dan. Thank you for all your help, encouragement, and support these last four years. I also want to acknowledge Dave, my “grand-advisor” (as Dan liked to call you), as well as my committee members, Katie, Andrea, Yufeng, and Naim, and my other faculty co-mentors, Patrick and Ken, for your sound expertise and astute advice on this work and my other research and professional endeavors.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................................................................................ ix

LIST OF FIGURES ........................................................................................................................................................................ x

LIST OF ABBREVIATIONS ................................................................................................................................................................... xii

CHAPTER 1: BACKGROUND .................................................................................................................................................................... 1

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

Moderated Nonlinear Factor Analysis ........................................................................................................................................ 4

Measurement Invariance and Differential Item Functioning ......................................................................................................... 4

Model Specification ........................................................................................................................................................................... 5

Likelihood Function .......................................................................................................................................................................... 8

DIF Detection ............................................................................................................................................................................... 10

Regularized Differential Item Functioning .................................................................................................................................. 10

Model Selection .............................................................................................................................................................................. 12

Previous Research ......................................................................................................................................................................... 12

Estimation Issues ........................................................................................................................................................................... 14

Observed Likelihood ................................................................................................................................................................. 14

Parameter Selection .................................................................................................................................................................... 15

CHAPTER 2: ALGORITHM .................................................................................................................................................................. 18

Penalized Expectation-Maximization ........................................................................................................................................... 18

Expected Value of Complete Data Log-Likelihood ......................................................................................................................... 19
LIST OF TABLES

Table 1 - Item Parameter Estimates ................................................................. 29
Table 2 - Impact Parameter Estimates ............................................................ 30
Table 3 - Reg-DIF Final Model Results ........................................................... 31
Table 4 - Population Parameter Values for Monte Carlo Simulation Study ........ 34
Table 5 - Items Evaluated for DIF in the Empirical Example ............................ 68
Table 6 - Reg-DIF Parameter Estimates for Non-Violent Delinquency Using BIC 73
Table 7 - Reg-DIF Parameter Estimates for Violent Behavior Factor BIC ............ 75
LIST OF FIGURES

Figure 1 - LASSO and MCP False Positive Rates Using BIC ................................................. 42
Figure 2 - LASSO and MCP False Positive Rates Using AIC ................................................. 43
Figure 3 - MCP False Positive Rates for Three Values of \( \gamma \) Using BIC ................................. 45
Figure 4 - Penalized EM (LASSO and MCP) and Observed Marginal Likelihood (LASSO) False Positive Rates Using BIC ............................................................... 48
Figure 5 - LASSO and MCP True Positive Rates Using BIC ...................................................... 50
Figure 6 - LASSO and MCP True Positive Rates Using AIC ...................................................... 51
Figure 7 - MCP True Positive Rates for Three Values of \( \gamma \) Using BIC ........................................ 53
Figure 8 - Penalized EM (LASSO and MCP) and Observed Marginal Likelihood (LASSO) True Positive Rates Using BIC ............................................................... 55
Figure 9 - LASSO and MCP True Positive and False Positive Rates for Conditions with Many Items and Many Covariates Using BIC ................................................. 57
Figure 10 - LASSO and MCP True Positive and False Positive Rates for For Intercept and Slope DIF Effects in Conditions with Many Items and Many Covariates Using BIC ...................................................................................... 58
Figure 11 - LASSO and MCP True Positive and False Positive Rates for Conditions with Many Items and Many Covariates Using AIC .................................................. 59
Figure 12 - LASSO and MCP Mean Squared Error for DIF Effects that Were Present in the Population ............................................................................................................. 61
Figure 13 - LASSO and MCP Mean Squared Error for DIF effects that Were Present in the Population and Were Correctly Identified ......................................................... 62
Figure 14 - LASSO and MCP Mean Squared Error for DIF Effects that Were Not Present in the Population ................................................................. 65

Figure 1A - LASSO and MCP False Positive Rates for Intercepts Only Using BIC ............... 107

Figure 2A - LASSO and MCP True Positive Rates for Intercepts Only Using BIC ............... 108

Figure 3A - LASSO and MCP False Positive Rates for Slopes Only Using BIC .................... 109

Figure 4A - LASSO and MCP True Positive Rates for Slopes Only Using BIC .................... 110
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Selection and Shrinkage Operator</td>
</tr>
<tr>
<td>MCP</td>
<td>Minimax Concave Penalty</td>
</tr>
<tr>
<td>MNLFA</td>
<td>Moderated Nonlinear Factor Analysis</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NLMIXED</td>
<td>Nonlinear-Mixed Effects (SAS Procedure)</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>
CHAPTER 1: BACKGROUND

Introduction

A prerequisite for making sound scientific inferences is to evaluate whether our measurements are equally valid across observations. For instance, we may want to measure depression among a diverse sample of people, including those who are male and female, black and white, and young and old. If individuals interpret one or more items on a depression scale differently based on their background characteristics (e.g., a young black female compared to an elderly white male), the measurement of depression may be confounded and biased. One example of this occurs with the item “cries easily”, which is endorsed at higher rates by women relative to men for the same underlying level of depression (Steinberg & Thissen, 2006). This phenomenon is referred to as differential item functioning (DIF) or a lack of measurement invariance and can lead to distorted outcomes, such as women presenting with artificially higher depression relative to men. Ultimately, DIF threatens the credibility of our scientific inferences if left unevaluated and unmitigated.

One primary method for evaluating measurement invariance is to statistically test a psychometric scale for DIF. Historically, such tests have been conducted for only one discrete background characteristic at a time, such as gender or ethnicity, because statistical and estimation limitations have prevented more complex evaluations of DIF. More recently, however, statistical advances have allowed for the simultaneous testing of DIF among multiple characteristics. Specifically, the moderated nonlinear factor analysis (MNLFA) model (Bauer, 2017; Bauer & Hussong, 2009) allows for DIF as a simultaneous function of both continuous
and categorical covariates, as well as different item response distributions and nonlinear constraints on the latent distribution. This welcomed flexibility, however, brings about new problems, including how to optimally identify DIF among many covariates, as well as how to efficiently estimate large numbers of model parameters. Recent work on the former has shown promising results for the use of $L_1$ (LASSO) regularization in identifying DIF (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020). In particular, LASSO for DIF detection exhibits fewer Type I errors without much decrement in power compared to using likelihood ratio tests in an iterative fashion (Thissen, Steinberg, & Wainer, 1993). While research is actively evaluating regularization for DIF testing, the latter problem of estimation has not been sufficiently addressed yet.

The first difficulty of estimation relates to the form of the likelihood function that is maximized. Specifically, the original implementation of MNLFA involved directly maximizing the observed marginal likelihood function (Bauer & Hussong, 2009). This requires integrating out the (unobserved) latent variable via numerical methods (i.e., adaptive quadrature) and using multivariate optimization techniques to obtain point estimates and standard errors for the model parameters. Although this estimation technique is flexible and used successfully in other modeling frameworks (i.e., mixed effects modeling), obtaining estimates via direct marginalization becomes impractical with more than 12 items or so (Bock & Aitkin, 1981; Bock & Lieberman, 1970). The second estimation difficulty relates to the inclusion of the $L_1$ penalty in the likelihood function. To avoid the non-differentiability of the absolute value penalty, previous research efforts fixed parameters to zero once the DIF estimates crossed an arbitrarily small threshold, thereby performing parameter selection outside of the optimization procedure (Bauer et al., 2020; Belzak & Bauer, 2020). This procedure contrasts with soft-thresholding (Tibshirani,
1996), which occurs within optimization and solves the non-differentiability problem in a theoretically sound way. These estimation limitations – namely, directly maximizing the marginal likelihood and thresholding outside of the optimization routine – impede both methodological and applied researchers from actively investigating and using regularization to test for DIF among multiple covariates in the MNLFA model.

This dissertation overcomes these challenges by pursuing four main goals. The first two goals involve developing an estimation method that (1) leverages the EM algorithm (Bock and Aitkin, 1981; Dempster, Laird, & Rubin, 1977) to allow for many item parameters to be estimated, while also (2) incorporating coordinate descent optimization to perform soft-thresholding (Friedman, Hastie, & Tibshirani, 2010; Tibshirani, 1996). The last two goals then involve evaluating this penalized EM algorithm by (3) comparing it with the observed marginal likelihood approach and examining other penalty approaches through an extensive simulation study, and (4) using an empirical example to showcase the regDIF R package (R Core Team, 2019; a beta version of which has been placed on GitHub), which implements the Reg-DIF procedure using the penalized EM algorithm. In the following lines, I accomplish these goals by first providing background information on the MNLFA model and the Reg-DIF procedure, and subsequently giving mathematical details on the penalized EM algorithm. Next, I provide details on the simulation study and empirical example, and thoroughly examine the results while providing a tutorial of the regDIF R package. Finally, I discuss the results and their implications for evaluating DIF, while also providing future directions for research.
Moderated Nonlinear Factor Analysis

Bauer and Hussong (2009) introduced MNLFA to solve methodological problems of integrative data analysis (IDA), a technique that pools data across independent studies through the harmonization of similar psychometric measures (e.g., depression measured by two different scales). MNLFA statistically evaluates whether harmonized scales yield equivalent scores between studies and within any given study (e.g., girls and boys in a single study). More recently, Bauer (2017) showed that MNLFA generalizes DIF testing from a single categorical background characteristic (e.g., study status) to multiple continuous and categorical characteristics (e.g., study status, gender, race, age, etc.). This modeling advance greatly expanded the psychometric toolbox for applied researchers, not only for those who seek to combine different studies, but also for those who wish to evaluate a scale for multiple possible sources of measurement bias. In the following, I provide details on the model and likelihood function, as this information provides the mathematical foundation for the penalized EM algorithm. This sets the stage for describing an alternative method of DIF evaluation: namely, using regularization to select DIF effects. First, however, I more formally define measurement invariance to emphasize important considerations for evaluating DIF.

Measurement Invariance and Differential Item Functioning

Measurement invariance is defined as

\[ f(Y|\theta, X) = f(Y|\theta), \]

where \( f(\cdot) \) is a multivariate response distribution, \( Y \) is a \( N \times J \) matrix of \( J \) item responses for \( N \) individuals, \( \theta \) is an \( N \times 1 \) vector of latent scores, and \( X \) is a \( N \times K \) matrix of \( K \) observed exogenous covariates (Millsap, 2011). Equation 1 states that measurement invariance occurs when all of the observed item responses are conditionally independent of one or more observed
covariates controlling for the latent variable. In other words, there are no external influences on the item responses after accounting for identical levels on the latent construct.

In contrast, when measurement invariance does not hold such that \( f(Y|\theta, X) \neq f(Y|\theta) \), DIF is present because one or more item responses are not conditionally independent of the external covariates. At this point, it is important to distinguish DIF from impact, or differences in the latent distribution as a function of the exogenous covariates. That is, individuals with different levels of the exogenous covariates may on average have higher or lower levels of the latent construct than other individuals, but there still may be biases in the item response distributions as a function of those covariates. The goal of measurement invariance / DIF testing is thus not to equate individuals on the latent construct but rather to evaluate whether the item responses are affected by the observed covariates over and above the influence of the latent variable (Millsap, 2011).

Finally, Equation 1 assumes a unidimensional model such that only one latent variable affects the item responses. In a multidimensional model, however, we would have \( \theta \) be a matrix of \( N \times R \) scores for \( R \) latent variables. Measurement invariance would then be defined as \( f(Y|\theta, X) = f(Y|\theta) \). Although the MNLFA model can incorporate multidimensionality, I focus on unidimensional models for DIF testing in the following description of the MNLFA framework. This not only follows most common applications of DIF evaluation, but it is also computationally straightforward to evaluate DIF on a single latent dimension.

**Model Specification**

The base MNFLA model parallels the generalized linear model framework (McCullagh & Nelder, 1989) by specifying the response distribution for each item and defining a linear predictor and link function to relate the expected value of each response distribution to the latent
variable and observed covariates. For instance, a binary item would be distributed Bernoulli or binomial (expected value is a probability); a continuous item would be distributed normal (expected value is the mean); a count item would be distributed negative binomial or Poisson (expected value is a rate), etc.

To specify the MNFLA model, the linear predictor $\eta_{ij}$ for person $i$ and item $j$ is first defined as

$$
\eta_{ij} = \nu_{ij} + \lambda_{ij} \theta_i,
$$

(2)

where $\nu_{ij}$ and $\lambda_{ij}$ are the intercept and slope that relate the latent factor, $\theta_i$, to the item response. This parameterization is different than most IRT models because I, following Bauer and Hussong (2009), include an $i$ subscript on the item intercept and slope, implying that each individual has a unique linear predictor value. This allows for the item parameters to be specified as deterministic functions of person-covariates to permit DIF, which will be shown later. Furthermore, in Equation 2 I assume scalar values for $\eta_{ij}$ and $\nu_{ij}$. In the case of a categorical item response (e.g., Disagree: -1; Neutral: 0; Agree: 1), however, the linear predictor and intercept would be vector valued to reflect the cumulative logits and $C - 1$ location parameters. The latent factor is typically assumed distributed as $\theta_i \sim \text{Normal}(\alpha_i, \psi_i)$ with latent mean $\alpha_i$ and variance $\psi_i$.

Next, the link function is written as

$$
g(\mu_{ij}) = \eta_{ij},
$$

(3)

where $g(\mu_{ij})$ is a function that transforms $\mu_{ij}$, the expected value of item $j$ for person $i$, to the range of the linear predictor $\eta_{ij}$. I may also define $\mu_{ij}$ in terms of the inverse link function, written as $\mu_{ij} = g^{-1}(\eta_{ij})$. Finally, the choice of the link function / inverse link function is determined by the item response distribution. For instance, if we assume the observed item
responses $y_{ij}$ are distributed as $y_{ij} \sim \text{Bernoulli}(\mu_{ij})$, where $\mu_{ij}$ is the conditional probability of person $i$ endorsing or correctly answering item $j$, we can use the logit link function

$$\log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = \eta_{ij}$$

(4)

or, equivalently, the inverse link function

$$\mu_{ij} = \frac{1}{1 + \exp(-\eta_{ij})}.$$  

(5)

We could also use the probit link function for Bernoulli-distributed item responses, although both functions often give similar results. To simplify the presentation, I use Bernoulli-distributed items and the logit link throughout this paper. Other common response distributions and link functions include the normal distribution with the identity link, where $g(\mu_{ij}) = \mu_{ij}$, and the negative binomial or Poisson distributions with the log link, where $g(\mu_{ij}) = \log(\mu_{ij})$.

In summary, these three components – the linear predictor, the link function, and the response distribution – define the base MNLFA model. What permits testing of DIF within MNLFA is that the base model parameters are then permitted to be deterministic functions of exogenous covariates.

First, for the latent distribution parameters, impact is modeled as

$$\alpha_i = \alpha_0 + \mathbf{x}_i^T \alpha,$$

(6)

$$\psi_i = \psi_0 \exp(\mathbf{x}_i^T \psi).$$

(7)

Here, the baseline parameters $\alpha_0$ and $\psi_0$ capture the latent mean and variance for person $i$ when all covariates equal zero. Typically, the MNLFA model sets $\alpha_0 = 0$ and $\psi_0 = 1$ to identify the model; I use that convention from here forward. The parameter vectors $\alpha$ and $\psi$ represent the effects of the covariates $\mathbf{x}_i$ on the latent mean and variance (i.e., impact). Note that a log-linear
expression is used for $\psi_i$ to constrain the variance to be positive. Second, for the item parameters, DIF is modeled as

$$v_{ij} = v_{0j} + x_i^T v_{1j},$$

$$\lambda_{ij} = \lambda_{0j} + x_i^T \lambda_{1j}. \tag{8}$$

Like the impact equations, the baseline intercept $v_{0j}$ and baseline slope $\lambda_{0j}$ respectively capture the severity (location) and discrimination (strength) of the item response as it relates to the latent variable when all covariates are zero. The vectors of intercept DIF parameters $v_{1j}$ and slope DIF parameters $\lambda_{1j}$ represent the differences in severity and discrimination from the baseline intercept and slope as a function of the exogenous covariates.\(^1\) Putting Equations 2-9 together, the complete MNLFA model for binary items is defined as

$$\mu_{ij} = \frac{1}{1 + \exp(-[(v_{0j} + x_i^T v_{1j}) + (\lambda_{0j} + x_i^T \lambda_{1j})] \theta_i))}, \tag{10}$$

$$\theta_i \sim \text{Normal}(x_i^T \alpha, \exp(x_i^T \psi)). \tag{11}$$

To estimate the model parameters, the likelihood function must be defined next.

**Likelihood Function**

The observed marginal log-likelihood function of the MNLFA model may be written

$$\log L(y_i | x_i) = \log \left( \prod_{i=1}^{N} \int f(y_{ij} | \theta_i, x_i; \omega_j) \, d\theta_i \right), \tag{12}$$

such that

$$\phi(\theta_i | x_i; \xi) = \frac{1}{\sqrt{2\pi\psi_i}} \exp \left( -\frac{[\theta_i - \alpha_i]^2}{2\psi_i} \right). \tag{13}$$

\(^1\) The MNLFA model simplifies to a two-group IRT model if $x_i$ is a vector of dummy or effect-coded values indicating group membership. This would make $v_{1j}$ and $\lambda_{1j}$ the difference in intercept and slope between both groups (i.e., difference relative to the reference group); in other words, $v_{1j}$ and $\lambda_{1j}$ represent DIF effects.
\[ f(y_{ij}|\theta_i, x_i; \omega_j) = \mu_{ij}^{y_{ij}}[1 - \mu_{ij}]^{1-y_{ij}}, \quad (14) \]

where \( \phi(\theta_i|x_i; \xi) \) is the conditional normal distribution of the latent variable \( \theta_i \) governed by the impact parameters \( \xi = \text{vec}[\alpha_i, \psi_i] \); \( f(y_{ij}|\theta_i, x_i; \omega_j) \) is the conditional item response function for item \( j \) governed by the baseline item parameters and DIF parameters \( \omega_j = \text{vec}[v_{0j}, v_{1j}, \lambda_{0j}, \lambda_{1j}] \), where \( y_{ij} \) is an indicator variable representing whether person \( i \) endorsed or correctly answered item \( j \); and \( y \) combines \( \xi \) and \( \omega_j \) for all \( j \) into a single vector of model parameters to be optimized given the observed item responses \( y_i \) and observed covariates \( x_i \).

Furthermore, because we do not have access to the latent scores, we integrate over \( \theta_i \) for \( N \) observations to obtain the observed data log-likelihood. But given that the integral cannot be solved analytically, numerical methods are used in practice to approximate the integral. For instance, we may choose a finite number of grid points over the range of \( \theta_i \), weight the grid points by the expected density of the distribution at each point, and sum over the weighted points. This is referred to as quadrature and will be shown in more detail later.

Early implementations of MNLFA directly maximized the log-likelihood function shown in Equation 12.\(^2\) This involves obtaining the gradient vector and sometimes the Hessian matrix (or an approximation of it) of \( y \), setting the log-likelihood equation to zero, and updating the parameter estimates using the gradient (and possibly the Hessian) in each iteration of the multivariate optimization procedure until convergence. I discuss the computational difficulty of this approach in more detail later, particularly as the number of items increases beyond 12 or so.

---

\(^2\) More recent implementations of MNLFA with Mplus software use an accelerated EM algorithm to obtain parameter estimates (Muthén & Muthén, 2017). This makes the estimation process considerably faster than directly maximizing the observed likelihood in NLMIXED (SAS Institute, 2018). It is unclear, however, what likelihood function is being used with the EM algorithm in Mplus. Furthermore, it is not currently possible to implement an \( L_1 \) penalty in Mplus.
For now, it suffices to say that the likelihood function in Equation 12 has most often been used to obtain maximum likelihood estimates for the MNLFA model.

**DIF Detection**

Although the MNLFA model provides impressive flexibility to evaluate DIF among multiple covariates, the task of identifying which DIF parameters should be present in the model is non-trivial. Common approaches for DIF detection include using likelihood ratio tests (Thissen et al., 1993), Wald tests (Langer, 2008; Lord, 1977; 1980; Woods, Cai, & Wang, 2013), and score tests / modification indices (Oort, 1998; Steenkamp & Baumgartner, 1998), all of which perform relatively well when there are only a few scale items with DIF. When more items exhibit DIF, however, these inferential methods tend to break down because untenable assumptions must be made about model identification constraints (Bauer et al., 2020; Belzak & Bauer, 2020). That is, some items must be assumed invariant to identify the model, and when large amounts of DIF are present in the scale, faulty assumptions about which items are anchors frequently lead to incorrect sampling distributions for DIF test statistics and higher-than-nominal Type I error rates. Fortunately, our research group has proposed an alternative approach which does not rely on these untenable assumptions but instead uses a penalty function in the log-likelihood to shrink some DIF parameters to zero (Bauer et al., 2020). This method is referred to as regularized differential item functioning or Reg-DIF.

**Regularized Differential Item Functioning**

Reg-DIF uses a LASSO (least absolute shrinkage and selection operator) penalty to simultaneously identify DIF parameters that improve model fit and select anchor items that do not express DIF. The LASSO penalty function is simply appended to the MNLFA log-likelihood as
\[ \log L(\mathbf{y}_i | \mathbf{y}_i, \mathbf{x}_i)_{\text{Reg}} = \log L(\mathbf{y}_i | \mathbf{y}_i, \mathbf{x}_i) - \tau(\| \mathbf{v}_1 \|_1 + \| \lambda_1 \|_1), \]

where \( \log L(\mathbf{y}_i | \mathbf{y}_i, \mathbf{x}_i)_{\text{Reg}} \) is the regularized log-likelihood function that is maximized; \( L(\mathbf{y}_i | \mathbf{y}_i, \mathbf{x}_i) \) is unchanged from Equation 12; \( \tau \) is a tuning parameter defined by the researcher; and \( \| \cdot \|_1 \) is the \( L_1 \) norm, which sums the absolute values of the affected parameters. The penalty function is applied to the DIF parameters – namely, \( \mathbf{v}_{1j} \) and \( \lambda_{1j} \) for all \( j \), or the differences in intercepts and slopes for all items – because the goal is to identify which exogenous covariates affect the item responses after controlling for the latent variable. In other words, the penalty rewards sparsity of DIF, thus helping to identify which items should be anchored (i.e., have the same parameter estimates) across individuals. The penalty does this by multiplying the sum of the absolute-valued parameters by the tuning parameter. This is then subtracted from the MNLFA log-likelihood function. As the tuning parameter \( \tau \) increases, some DIF parameters shrink towards zero because they become too costly to maintain in the penalized likelihood function. In the limit as \( \tau \) grows large, all DIF parameters shrink to and become zero; that is, no DIF is present in the model when \( \tau \) is big enough. Conversely, if \( \tau = 0 \), the regularized log-likelihood simplifies to the MNLFA log-likelihood. It is important to clarify, however, that when \( \tau = 0 \) or is close to zero, this results in a fully saturated model with DIF specified on all items and no anchor items. Thus, the model is under-identified when \( \tau = 0 \) or approaches zero. While previous work on the regularized MNLFA model dealt with this identifiability problem by using a Moore-Penrose inverse (i.e., see Barata & Hussein, 2012), the current work solves this issue by starting with a large penalty that removes all DIF from the model to begin with, and then gradually decreases the magnitude of \( \tau \), permitting the inclusion of DIF parameters that sufficiently improve fit, until
the optimal model can be selected. This approach selects the optimal model well before the penalty approaches zero and the model becomes under-identified.

**Model Selection**

The main task of performing LASSO regularization is to vary the magnitude of $\tau$ across a range of non-negative values and to choose the optimal model via some selection criteria. In particular, fit indices (e.g., AIC, BIC) and cross-validation (e.g., leave-one-out, k-fold) are commonly used to identify the optimal degree of penalization (Hastie, Tibshirani, Friedman, 2017). Our research group has previously used AIC and BIC for choosing $\tau$ due to its computational ease. Although $k$-fold cross validation has shown to perform well in other regularization research (Hastie et al., 2017), it is not clear whether this approach is computationally feasible for Reg-DIF at the current moment. As such, this will be discussed in more detail later. I turn to highlighting previous research efforts next.

**Previous Research**

Research on using regularization to evaluate DIF has shown promising results relative to other methods. In particular, Tutz and Schuberger (2015) implemented LASSO regularization for DIF testing among multiple covariates by using the one-parameter logistic IRT model wherein item slopes are constrained to equality. They showed that regularization outperformed other methods of DIF detection (i.e., likelihood ratio tests and Wald tests) in terms of true and false positives, particularly in small sample sizes. Magis, Tuerlinckx, and De Boeck (2015) also applied a LASSO penalty to a one-parameter logistic model and found better control of Type I error compared to other DIF evaluation procedures. This was especially the case when the proportion and magnitude of DIF were both large. More recently, our research group (Bauer et al., 2020; Belzak & Bauer, 2020) implemented regularization for DIF testing in the two-
parameter logistic IRT model, where item slopes are freely estimated, and showed that LASSO regularization outperformed the iterative likelihood ratio test method. In line with previous research, when the proportion and magnitude of DIF were large and sample size increased, Reg-DIF resulted in fewer Type I error rates without large decrements in power. In contrast to Tutz and Schauburger and Magis et al., however, both of whom evaluated regularization for DIF testing using both information criteria and cross-validation, we only evaluated information criteria (i.e., BIC and AIC) in our research because of computational limitations. We also used an alternative method for setting DIF parameter estimates to zero as the penalty increased. That is, for DIF parameters approaching zero, we set these parameters equal to zero after crossing an arbitrarily small threshold. In other words, this thresholding approach was performed outside of the optimization routine rather than within, as is commonly done in most other applications. It is unclear whether this non-standard method of thresholding may have affected the performance of the approach.

Other research has also found that different penalty functions besides LASSO may be effective for identifying DIF. In particular, Huang (2018) used the minimax concave penalty (MCP) function in a multiple-group structural equation model to evaluate DIF. Huang found that the MCP function performed well when BIC was used for model selection. As expected, larger sample sizes and bigger DIF effects led to better performance with this penalty, including higher rates of true positive and lower rates of false positives. Furthermore, when compared to the LASSO penalty, MCP most notably reduced estimator bias in the penalized effects by including an additional tuning parameter. Although LASSO is a special case of MCP, this second tuning parameter makes MCP a non-convex function. This can complicate the model estimation process which relies on convex functions to identify the optimal solution; however, it can also lead to
better performance in evaluating DIF relative to LASSO due to its de-biasing nature. Given that our research group has only evaluated LASSO for the evaluation of DIF in MLNFA models, it is worth exploring whether an alternative penalty function such as MCP may outperform LASSO. This will be explored more later.

Having described regularization in the MNLFA model, model selection criteria, and previous research on Reg-DIF, I now examine estimation issues that have significantly hindered previous efforts of using and evaluating Reg-DIF for the MNLFA model.

**Estimation Issues**

There are two main issues that prevent methodological and applied researchers from fully using and evaluating Reg-DIF with many scale items and exogenous covariates. They include (1) the form of the observed likelihood function that is optimized, and (2) the way in which parameter selection via the $L_1$ penalty is performed.

**Observed Likelihood**

The first main problem is that our previous implementations of Reg-DIF in the MNLFA model have most often maximized the observed marginal log-likelihood function shown in Equation 12. Although this technique works well when the number of parameters is relatively small, it becomes intractable as the number of parameters grows large. In fact, we were limited to only 12 items in our evaluations of Reg-DIF (Bauer et al., 2020; Belzak & Bauer, 2020), presumably because large $M \times M$ information or Hessian matrices ($M$ is the number of parameters in $\mathbf{\theta}$) must be inverted multiple times in each iteration of direct maximization of the observed marginal likelihood. This inevitably led to prohibitive computational time. The same problem was also identified by Bock and Lieberman (1970), who used Newton-Raphson to obtain item parameter estimates via marginal maximum likelihood. To overcome this
computational limitation, Bock and Aitken (1981) introduced an application of the EM algorithm to simplify the large model-wide maximization problem into smaller item-by-item maximization steps. The first aim of this dissertation is to pursue a similar solution for the problem of efficiently estimating parameters in the MNLFA model.

**Parameter Selection**

Another issue of estimation is the method in which DIF parameters are selected from the model. Namely, to perform regularization there must be some mechanism of thresholding parameter estimates to zero once \( \tau \) is larger than the corresponding estimate. This is a difficult task because the \( L_1 \) norm introduces points of non-differentiability to the log-likelihood, thereby jeopardizing the feasibility of optimization. Non-differentiability occurs because the partial derivative of the log-likelihood function with respect to a given focal \(^3\) parameter is undefined at zero, as there is a non-differentiable “kink” in the absolute value function of the parameter (e.g., \( \frac{d|\tau|}{d\tau} \) is undefined when \( \tau = 0 \)). To circumvent this problem, the *soft-thresholding operator* (Donoho & Johnstone, 1994; Tibshirani, 1996) is used to define the partial derivative of the log-likelihood when the focal parameter equals zero – that is, when \( \tau \) becomes larger than the corresponding parameter estimate. Furthermore, to implement soft-thresholding, optimization typically occurs at the parameter level rather than the model level, in large part because univariate optimization is more computationally efficient in regularization contexts compared to multivariate optimization (Hastie et al., 2017). One method often implemented in the context of \( L_1 \) regularization is therefore coordinate-wise descent (Friedman, Hastie, Hofling, & Tibshirani, 2007). Coordinate-wise descent is a univariate optimization technique that updates parameters

---

\(^3\) “Focal” signifies a parameter that is present in the \( L_1 \) penalty.
one-at-a-time while holding all other parameters constant. This procedure uses the first and second-order partial derivatives (or an approximation thereof) to update estimates (e.g., Newton-Raphson), and then cycles through the parameters until some convergence criterion is met. In addition, because penalty values are often supplied to the regularization routine in successive values (i.e., $\tau$ in Equation 15), each model provides warm starting values for the next penalized model. This is often referred to as cyclical coordinate descent and has been shown to work quite well with regularization, yielding solutions much faster than conventional convex optimizers (Friedman et. al., 2007).

Despite the utility of cyclical coordinate descent, previous work on Reg-DIF has been limited by the type of software available for performing optimization. That is, only standard convex optimizers have been readily available for implementing Reg-DIF, and in turn, has prevented the use of soft-thresholding for dealing with non-differentiability of the likelihood function. As a workaround, we performed thresholding outside of the optimization procedure in our prior work. Specifically, we defined an arbitrarily small threshold value that, if the penalized estimate became smaller than the threshold value, the penalized parameter would be set to zero. Although this ad hoc method gets around the points of non-differentiability, albeit in a non-standard manner, it is possible that using soft-thresholding, which makes direct use of the first and second derivatives, will lead to more optimal selection of DIF parameters while using $L_1$ regularization. Thus, the second aim of this dissertation is to implement and evaluate soft-thresholding for performing Reg-DIF.

In sum, to overcome the computational and estimation limitations which have hampered previous applications of Reg-DIF, I develop an alternative algorithm for fitting MNLFA with a penalty function in this dissertation. In particular, I introduce a penalized EM algorithm with
coordinate-wise descent to implement different types of regularization in the MNLFA model. This approach not only allows for many more item responses to be modeled at a time, but also more covariates to be tested for DIF while using proper thresholding techniques and alternative penalty functions. I describe the mathematical details of this algorithm next.
CHAPTER 2: ALGORITHM

This section first reformulates the MNLFA log-likelihood into a missing data problem, showing how the EM algorithm can be used to estimate parameters more efficiently. I then describe how the penalty is applied to this new likelihood function.

Penalized Expectation-Maximization

Bock and Aitken (1981) show that marginal maximum likelihood estimation of a 2-parameter IRT model may be reformulated as a missing data problem, where the latent scores are considered missing from the analysis. In doing so, the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is leveraged to obtain parameter estimates in the 2-parameter IRT model with much greater efficiency than directly maximizing the marginal log-likelihood. More precisely, the EM algorithm cycles between two steps until convergence:

(1) the E-step, which computes the expected value of the complete data log-likelihood given provisional estimates; and

(2) the M-step, which optimizes the parameter estimates based on the values produced in the E-step, in turn providing new provisional estimates to the E-step.

I use this approach to transform the MNLFA log-likelihood in Equation 12 into a missing data problem as well. I thus define the complete data log-likelihood function for the MLNFA model as

$$\log L(\mathbf{y} | \mathbf{y}_i, \mathbf{x}_i, \theta_i)_C = \log \left( \prod_{l=1}^N \phi(\theta_l | \mathbf{x}_i; \xi) \prod_{j=1}^J f(y_{ij} | \theta_l, \mathbf{x}_i; \omega_j) \right).$$

(16)
Notice that in Equation 16 I am not integrating out the latent variable because I assume knowledge of the missing data. Hence, the complete data log-likelihood is a combination of the missing data, $\theta_i$, and the observed data, $y_{ij}$ and $x_i$. Because I do not have access to the missing latent scores, however, maximizing $\log L(\mathbf{y}|\mathbf{Y})_C$ is not possible. Instead, the EM algorithm makes use of the complete data log-likelihood in a different way.

**Expected Value of Complete Data Log-Likelihood**

Rather than dealing with the intractable complete data log-likelihood function, an alternative approach is to maximize a surrogate function for $L(\mathbf{y}|\mathbf{Y})_C$, specifically by taking the expected value of the complete data log-likelihood with respect to the missing data $\theta_i$. This may be written as follows:

$$Q(\mathbf{y}|\mathbf{Y}^{(t)}) = \mathbb{E}[\log L(\mathbf{y}_i; x_i; \theta_i)_C | \mathbf{y}_i, \mathbf{x}_i]$$

$$= \mathbb{E} \left[ \log \left( \prod_{i=1}^{N} \phi(\theta_i|x_i; \xi) \prod_{j=1}^{J} f(y_{ij}|\theta_i, x_i; \omega_j) \right) \right]$$

$$= \int \left[ \sum_{i=1}^{N} \log[\phi(\theta_i|x_i; \xi)]P(\theta_i|y_i, x_i; \mathbf{y}^{(t)}) \right.$$

$$+ \sum_{i=1}^{N} \sum_{j=1}^{J} \left. \log[f(y_{ij}|\theta_i, x_i; \omega_j)]P(\theta_i|y_i, x_i; \mathbf{y}^{(t)}) \right] d\theta,$$

where $Q(\mathbf{y}|\mathbf{Y}^{(t)})$ is the surrogate function to maximize; $\mathbf{y}^{(t)}$ is the vector of provisional model estimates at iteration $t$; and $P(\theta_i|y_i, x_i; \mathbf{y}^{(t)})$ is the posterior distribution of the missing data $\theta_i$.

I make use of Bayes’ theorem to compute the posterior distribution of $\theta_i$ shown in Equation 19:
\[
P(\theta_i | y_{ij}, \mathbf{x}_i ; \mathbf{y}^{(t)}) = \frac{P(\mathbf{y}_{ij}, \theta_i ; \mathbf{y}^{(t)})}{P(\mathbf{y}_{ij} ; \mathbf{y}^{(t)})} = \frac{\prod_{j=1}^{J} f(y_{ij} | \theta_i, \mathbf{x}_i ; \omega_j^{(t)}) \phi(\theta_i | \mathbf{x}_i ; \xi^{(t)})}{\int \prod_{j=1}^{J} f(y_{ij} | \theta_i, \mathbf{x}_i ; \omega_j^{(t)}) \phi(\theta_i | \mathbf{x}_i ; \xi^{(t)}) d\theta_i}, \tag{20}
\]

where \(P(\mathbf{y}_{ij}, \theta_i ; \mathbf{y}^{(t)})\) is the joint distribution of \(y_{ij}\) and \(\theta_i\), or, in other words, the likelihood \(\prod_{j=1}^{J} f(y_{ij} | \theta_i, \mathbf{x}_i ; \omega_j^{(t)})\) multiplied by the prior \(\phi(\theta_i | \mathbf{x}_i ; \xi^{(t)})\); and \(P(\mathbf{y}_{ij} ; \mathbf{y}^{(t)})\) is the marginal distribution of \(y_{ij}\), which ensures the posterior is a proper probability distribution. Although Equation 20 cannot be directly computed because we are missing \(\theta_i\), I instead use quadrature points in place of \(\theta_i\). I thus rewrite the posterior as

\[
P(Z_q | y_{ij}, \mathbf{x}_i ; \mathbf{y}^{(t)}) = \frac{\prod_{j=1}^{J} f(y_{ij} | Z_q, \mathbf{x}_i ; \omega_j^{(t)}) W_q}{\sum_{q=1}^{Q} \prod_{j=1}^{J} f(y_{ij} | Z_q, \mathbf{x}_i ; \omega_j^{(t)}) W_q}. \tag{21}
\]

In Equation 21, \(Z_q\) is a real number in the range of the probability density of the latent variable \(\theta_i\); for instance, I may define a set of, say, 51 equally spaced values ranging from \(-6\) to \(6\), with \(Z_q\) representing a specific point in the interval. \(W_q\) is the corresponding weight of the probability density for each \(Z_q\), that is, a set of normalized ordinates: \(\phi(Z_q | \mathbf{x}_i ; \xi^{(t)})\). I then sum the quadrature points to approximate the integral in the denominator of Equation 21.

Similar to Equation 21, I use quadrature points in place of the latent scores in the conditional latent variable distribution, \(\phi(\theta_i | \mathbf{x}_i ; \xi)\), and the \(J\) conditional item response distributions, \(\prod_{j=1}^{J} f(y_{ij} | \theta_i, \mathbf{x}_i ; \omega_j^{(t)})\). This allows me to rewrite the expected complete data log-likelihood in Equation 19 as a summation of \(Q\) quadrature points (as well as \(N\) individuals and \(J\) items):
\[
Q(\mathbf{y}|\mathbf{y}^{(t)}) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \log[\phi(Z_q|x_i; \xi^{(t)})]P(Z_q|y_i, x_i; \mathbf{y}^{(t)})
\]

\[
+ \sum_{i=1}^{N} \sum_{q=1}^{Q} \sum_{j=1}^{J} \log[f(y_{ij}|Z_q, x_i; \varphi_j^{(t)})]P(Z_q|y_i, x_i; \mathbf{y}^{(t)}).
\]

The first set of summations in Equation 22 represents the contribution of the conditional latent variable distribution. The second set of summations includes \( J \) conditional item response distributions. Note that in Equation 22 I interchange the order of summations between the \( N \) individuals and \( Q \) quadrature points (from Equation 19) following Bock and Aitken (1981).

The form of the likelihood in Equation 22 not only shows that the latent variable parameters can be estimated separately from the item parameters, but also, and more importantly, that maximization may be performed independently for each item. This greatly simplifies the optimization procedure for large numbers of items and many DIF covariates. That is, Equation 22 intimates the main advantage of using the EM algorithm, namely, that the large \( M \times M \) maximization problem simplifies to \( J \) smaller maximization steps of size \( m_j \times m_j \), plus 1, where \( M \) is the total number of parameters in the MNLFA model; \( m_j \) is the number of parameters for item \( j \); and “plus 1” indicates an additional M-step for the latent variable parameters. This simplification occurs because the EM algorithm produces increasingly better estimates of the unobserved (latent) scores for each individual, allowing use of these latent score estimates as predictors in separate logistic regression models to obtain consistent and unbiased parameter estimates. In contrast, the observed likelihood approach must integrate out the missing data and optimize over all parameters at once. This becomes increasingly burdensome as the number of parameters grows. Thus, the EM algorithm provides a significantly more scalable way to estimate many item parameters in IRT models compared to the observed likelihood approach.
Lastly, to ensure the log-likelihood has an optimal solution, I show in Appendix A that Equation 22 is indeed concave. That is, the EM algorithm converges to a stable point given the components that make up the likelihood function in Equation 22.

**Penalty Function**

To incorporate penalization into the EM algorithm, I append a penalty function to the expected value of the complete data log-likelihood function. This allows for regularization (i.e., selection) of the DIF parameters. It is written

\[
Q(y|Y^{(t)})_{\text{Reg}} = Q(y|Y^{(t)}) - P(v_1^{(t)}, \lambda_1^{(t)}),
\]

where \(Q(y|Y^{(t)})_{\text{Reg}}\) is the regularized surrogate function, \(Q(y|Y^{(t)})\) is identical to Equation 22, and \(P(v_1^{(t)}, \lambda_1^{(t)})\) is a generic penalty function applied to the intercept and slope DIF effects. \(P(v_1^{(t)}, \lambda_1^{(t)})\) may be defined using variety of different penalty functions, including the \(L_1\) norm (LASSO) and the minimax concave penalty (MCP), among others. I focus on these two penalties in this dissertation because they have been shown to perform well in both latent variable and logistic regression applications (Huang, 2018). I define the LASSO and MCP functions in greater detail below.

Since \(Q(y|Y^{(t)})\) in Equation 22 is in part a summation of item response distributions multiplied by the posterior distribution of the latent variable, applying a penalty function to \(Q(y|Y^{(t)})_{\text{Reg}}\) also factors out to the item level. I therefore subtract the penalty function from each item independently of all other items; effectively, this performs regularization of DIF. For instance, I maximize the surrogate function for item \(j\) as
\[
Q_j(y_j|y_j^{(t)})_{\text{Reg}} = \sum_{i=1}^{N} \sum_{q=1}^{Q} \log[f(y_{ij}|Z_q, x_i; \theta_j^{(t)})]P(Z_q|y_i, x_i; y_j^{(t)}) \nonumber
\]
\[ - P(\psi_{1j}^{(t)}, \lambda_{1j}^{(t)}), \]

such that the function \(P(\psi_{1j}^{(t)}, \lambda_{1j}^{(t)})\) penalizes the DIF parameters for item \(j\) only. This is repeated for all item responses, given the posterior values computed in the E-step.

Now that the penalty has been subtracted from the surrogate log-likelihood function in Equation 23, I show how coordinate descent and soft-thresholding is implemented to perform Reg-DIF.

**Coordinate Descent**

Coordinate descent is a simple optimization procedure that finds the minimum of the negative log-likelihood function with respect to each model parameter at a time, while all other model parameters are held constant. This procedure repeats across every parameter until convergence is complete. To implement coordinate descent, I use a second-order Taylor series expansion of the function defined in Equation 24.

**Quadratic Approximation**

The updating rule for a parameter estimate employs a second-order Taylor series expansion or a quadratic approximation of \(Q_j(y_j|y_j^{(t)})\) as a function of a single penalized parameter in \(y_j\), or a second-order Taylor series expansion of the non-regularized surrogate function for item \(j\). Below, I use \(\zeta_j\) to represent a generic DIF parameter on item \(j\). The Taylor series expansion with respect to \(\zeta_j\) is then
\[ Q_j(\mathbf{y}_j|\mathbf{y}^{(t)}_j)_{\text{Reg}} \]

\[ \approx Q_j(\mathbf{y}_j, \zeta_j^*|\mathbf{y}^{(t)}_j, \zeta_j^{(t)}_j) + \partial_{\zeta_j} Q_j(\mathbf{y}_j, \zeta_j^*|\mathbf{y}^{(t)}_j, \zeta_j^{(t)}_j)(\zeta_j - \zeta_j^*) \]

\[ + \frac{1}{2} \partial^2_{\zeta_j} Q_j(\mathbf{y}_j, \zeta_j^*|\mathbf{y}^{(t)}_j, \zeta_j^{(t)}_j)(\zeta_j - \zeta_j^*)^2 - P(\mathbf{v}_{1j}^{(t)}, \lambda_{1j}^{(t)}). \]

In Equation 25, \( \mathbf{y}_j \) is now the vector of model parameters that excludes the active parameter \( \zeta_j \) and which are held constant at their current value; \( \zeta_j^* \) is the point at which the surrogate function is approximated with respect to \( \zeta_j \); and \( \partial_{\zeta_j} Q_j(\cdot) \) and \( \partial^2_{\zeta_j} Q_j(\cdot) \) represent the first and second partial derivatives of the surrogate function with respect to \( \zeta_j \), which are derived in Appendix B.

In effect, the goal of approximating \( Q_j(\mathbf{y}_j, \zeta_j|\mathbf{y}^{(t)}_j, \zeta_j^{(t)}_j) \) with a quadratic function is to optimize (get an updated estimate of) \( \zeta_j \), which I denote as \( \hat{\zeta}_j \), using a more tractable formulation of the surrogate function. Given that we want to include the penalty function to regularize DIF parameters, however, we also need an updating rule that handles the points of non-differentiability which occur due to the LASSO and MCP functions, conveniently shown as \( P(\mathbf{v}_{1j}^{(t)}, \lambda_{1j}^{(t)}) \) in Equation 25. I therefore use soft-thresholding to implement LASSO regularization and firm-thresholding to implement MCP regularization (Zhang, 2010). These techniques are described next.

**Soft-Thresholding**

Soft-thresholding is a mathematical operation that determines the value of the parameter estimate based on the relative magnitude of the penalty and the optimized/updated parameter estimate (Donoho, 1995). For a generic, updated estimate denoted as \( z \), along with penalty parameter \( \tau \), soft-thresholding is defined as
\[
S(z, \tau) = \text{sign}(z)(|z| - \tau)_+ = \begin{cases} 
    z - \tau & \text{if } z > 0 \text{ and } \tau < |z|, \\
    z + \tau & \text{if } z < 0 \text{ and } \tau < |z|, \\
    0 & \text{if } \tau \geq |z|. 
\end{cases}
\] (26)

Equation 26 states that if the penalty parameter \(\tau\) is greater than or equal to the updated estimate \(z\), the value of \(S(z, \tau)\) equals zero (Friedman, Hastie, & Tibshirani, 2010). Otherwise, the penalty is either subtracted from or added to the updated estimate depending on whether \(z\) is greater than or less than zero. More precisely, \(z\) refers to the first partial derivative of the quadratic approximation of the surrogate function set to zero. Equivalently, \(z\) is an estimate of the generic parameter (e.g., DIF effect) that has been optimized or updated (here, using coordinate descent) but which has not yet been penalized. In short, soft-thresholding is simply the function that shrinks the updated parameter. With respect to this quadratic approximation in Equation 25, I thus use the soft-thresholding operator to update the DIF parameter estimates according the LASSO function. For instance, I take the first partial derivative of the quadratic approximation with respect to \(\zeta_j\) (excluding the penalty function) and set it equal to zero to get updated estimate, that is, \(z\). This is shown as

\[
z = \zeta_j \approx \zeta_j^* - \frac{\partial \zeta_j Q_j(Y_j, \zeta_j^*|Y_j^{(t)}, \zeta_j^{(t)})}{\partial^2 \zeta_j Q_j(Y_j, \zeta_j^*|Y_j^{(t)}, \zeta_j^{(t)})},
\] (27)

which is then used to update the estimate via soft-thresholding as

\[
\zeta_j^{(t+1)} = S(\zeta_j, \tau).
\] (28)

In Equation 28, \(\zeta_j^{(t+1)}\) is the updated shrunken estimate that is used in the next E-step. This procedure generalizes to all DIF parameter estimates, one at a time, holding all other parameters constant. Notably, this is the same soft-thresholding update given by Sun, Chen, Liu, Ying, and Xin (2016), who developed a penalized EM algorithm for the purpose of selecting slopes in a
multidimensional IRT model. Finally, for the parameter estimates that are not penalized—namely, the latent variable parameters and the baseline item parameters—we still use the quadratic approximation of the surrogate function but instead just use the Newton-Raphson algorithm to update estimates (i.e., Equation 27).

**Firm-Thresholding**

To apply the MCP regularization function to the DIF parameters, I use the firm-thresholding operator, defined as

\[
F(z, \tau, \gamma) = \begin{cases} 
\frac{\gamma}{\gamma-1} S(z, \tau) & \text{if } \gamma \tau \geq |z|; \\
z & \text{if } \gamma \tau < |z|,
\end{cases}
\]

(28)
pursuant to \( \gamma > 1 \). In Equation 28, \( \gamma \) is not to be confused with the vector of model parameters; rather, \( \gamma \) is a second tuning parameter that relaxes the amount of penalization done by \( S(z, \tau) \) as \( \tau \) decreases, where \( \tau \) and \( z \) are defined the same as before. Specifically, the relaxation parameter \( \gamma \) reduces the coefficient bias that occurs when performing LASSO penalization. As \( \gamma \) approaches 1, the firm-thresholding operator more quickly converges to or “jumps” to the nearly unbiased estimate (Zhang, 2010). Although it is desirable to identify the nearly unbiased estimate quickly, the tradeoff is that optimization becomes unstable. That is, the MCP function is known to be non-convex, particularly as \( \gamma \) approaches 1. As \( \gamma \) approaches infinity, however, firm-thresholding transforms into the soft-thresholding operator and thus becomes LASSO regularization; hence, this is why the soft-thresholding operator is incorporated into the MCP function. The goal is to identify \( \gamma \) so that the solution is nearly unbiased but also yields a global maximum solution. Given that tuning of \( \tau \) is required, where many values are typically used to identify the optimal degree of penalization, \( \gamma \) is often fixed at a specific value to avoid a grid search across both \( \tau \) and \( \gamma \). The ncvreg R package, for example, sets \( \gamma = 3 \) because it was found
to work well in a variety of regression settings (Breheny & Huang, 2011). In this dissertation, I pilot test a variety of γ values for performing Reg-DIF using the MCP function and compare a select few.

Having described both the penalization EM algorithm and the coordinate descent procedure, I now turn to a simulated example to demonstrate the equivalency between the new estimation approach and previous implementations of Reg-DIF with the MNLFA model.

**Proof of Concept**

The goal of this preliminary analysis is to demonstrate that my newly developed penalized EM algorithm gives similar results compared to previous implementations of Reg-DIF, as one would expect. To accomplish this goal, I use a simulated dataset that originates from a larger study comparing scores generated from MNLFA models versus more traditional scoring models, particularly as they might be applied in an integrative data analysis setting (Curran, Cole, Bauer, Rothenberg, Hussong, 2018). The subset of data considered here consists of six Bernoulli distributed item responses from 500 individuals that load on a single normally distributed latent factor. Three exogenous covariates have large DIF effects on two of the six item responses (i.e., Items 4 and 5 have DIF), in addition to large effects on the latent mean and small effects on the latent variance (i.e., impact). I refer to these covariates as study, gender, and age. Study and gender were effect coded (−.5, .5) whereas age was continuous and standardized. This means that all covariates had variances equal to 1, ensuring they would be penalized equally. To identify the model, I set the baseline latent mean and variance to 0 and 1, respectively, and constrained all DIF parameters on Item 1 to be zero (making this a designated anchor item). Although neither the new nor previous approach for conducting Reg-DIF necessarily requires an a priori designation of anchor items, doing so here allows me to compare
the estimators both when there is no penalty applied (i.e., the tuning parameter is set to zero) as well as when the penalty value is selected to optimize the BIC of the model.

**Baseline Comparison**

I first compare the parameter estimates from the EM algorithm against the estimates from the observed marginal likelihood approach, setting the tuning parameter equal to zero (i.e., no penalty is applied). This allows me to isolate any differences that may be due to the form of the likelihood rather than the manner in which estimates are set to zero with the penalty. I developed the R package regDIF to implement the EM algorithm, whereas I used the nonlinear mixed effects procedure in SAS (NLMIXED, SAS Institute, 2018) to maximize the observed marginal likelihood. Tables 1 and 2 provide item and latent variable parameter estimates. As expected, the results show nearly identical correspondence between both estimation approaches. Specifically, the largest divergence in parameter estimation was only .006, which was likely due to differences in convergence criteria. Tighter convergence options on both methods (not shown here) yielded even greater correspondence in estimates.
Table 1. Item Parameter Estimates

<table>
<thead>
<tr>
<th>Item</th>
<th>Parameter</th>
<th>EM</th>
<th>Observed ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\nu_{01}$</td>
<td>-0.534</td>
<td>-0.533</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{01}$</td>
<td>1.023</td>
<td>1.023</td>
</tr>
<tr>
<td></td>
<td>$\nu_{11}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\nu_{21}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\nu_{31}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{11}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{21}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{31}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>$\nu_{02}$</td>
<td>-0.843</td>
<td>-0.843</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{02}$</td>
<td>1.060</td>
<td>1.059</td>
</tr>
<tr>
<td></td>
<td>$\nu_{12}$</td>
<td>-0.078</td>
<td>-0.078</td>
</tr>
<tr>
<td></td>
<td>$\nu_{22}$</td>
<td>-0.070</td>
<td>-0.070</td>
</tr>
<tr>
<td></td>
<td>$\nu_{32}$</td>
<td>-0.051</td>
<td>-0.051</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{12}$</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{22}$</td>
<td>0.145</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{32}$</td>
<td>-0.150</td>
<td>-0.151</td>
</tr>
<tr>
<td>3</td>
<td>$\nu_{03}$</td>
<td>-1.243</td>
<td>-1.242</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{03}$</td>
<td>1.375</td>
<td>1.374</td>
</tr>
<tr>
<td></td>
<td>$\nu_{13}$</td>
<td>-0.237</td>
<td>-0.237</td>
</tr>
<tr>
<td></td>
<td>$\nu_{23}$</td>
<td>0.301</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>$\nu_{33}$</td>
<td>-0.156</td>
<td>-0.156</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{13}$</td>
<td>-0.066</td>
<td>-0.067</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{23}$</td>
<td>-0.311</td>
<td>-0.311</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{33}$</td>
<td>-0.636</td>
<td>-0.636</td>
</tr>
<tr>
<td>4</td>
<td>$\nu_{04}$</td>
<td>-1.711</td>
<td>-1.717</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{04}$</td>
<td>1.656</td>
<td>1.655</td>
</tr>
<tr>
<td></td>
<td>$\nu_{14}$</td>
<td>0.195</td>
<td>0.195</td>
</tr>
<tr>
<td></td>
<td>$\nu_{24}$</td>
<td>-0.096</td>
<td>-0.096</td>
</tr>
<tr>
<td></td>
<td>$\nu_{34}$</td>
<td>0.056</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{14}$</td>
<td>-0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{24}$</td>
<td>0.089</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{34}$</td>
<td>-0.791</td>
<td>-0.791</td>
</tr>
<tr>
<td>5</td>
<td>$\nu_{05}$</td>
<td>-1.776</td>
<td>-1.776</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{05}$</td>
<td>1.583</td>
<td>1.583</td>
</tr>
<tr>
<td></td>
<td>$\nu_{15}$</td>
<td>-3.25</td>
<td>-3.25</td>
</tr>
<tr>
<td></td>
<td>$\nu_{25}$</td>
<td>-0.840</td>
<td>-0.841</td>
</tr>
<tr>
<td></td>
<td>$\nu_{35}$</td>
<td>0.693</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{15}$</td>
<td>-0.025</td>
<td>-0.026</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{25}$</td>
<td>-0.591</td>
<td>-0.591</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{35}$</td>
<td>-2.12</td>
<td>-2.11</td>
</tr>
<tr>
<td>6</td>
<td>$\nu_{06}$</td>
<td>-2.026</td>
<td>-2.025</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{06}$</td>
<td>1.510</td>
<td>1.509</td>
</tr>
<tr>
<td></td>
<td>$\nu_{16}$</td>
<td>-0.202</td>
<td>-0.202</td>
</tr>
<tr>
<td></td>
<td>$\nu_{26}$</td>
<td>0.187</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>$\nu_{36}$</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{16}$</td>
<td>0.014</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{26}$</td>
<td>-0.234</td>
<td>-0.234</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{36}$</td>
<td>-0.484</td>
<td>-0.485</td>
</tr>
</tbody>
</table>

Note. $\nu_{0j}$ and $\lambda_{0j}$ are baseline intercept and slope estimates for item $j$. $\nu_{kj}$ and $\lambda_{kj}$ are DIF intercept and slope estimates for covariate $k$ on item $j$. EM refers to the Expectation-Maximization algorithm. Observed ML refers to the observed marginal likelihood approach. regDIF (EM estimation) used Guass-Hermite quadrature with 51 points, and the convergence criterion was set to $1 \times 10^{-7}$ (i.e., change in the sum of the absolute values of the point estimates from iteration $t$ to $t+1$). In contrast, NLMIXED (Observed ML approach) used adaptive quadrature with 21 points, and the convergence criterion was set to $1 \times 10^{-11}$.

This example thus shows that the EM algorithm yields identical estimates to the observed marginal likelihood approach using the same conditions, a necessary step before evaluating it using regularization.
Table 2. Impact Parameter Estimates

<table>
<thead>
<tr>
<th>Impact</th>
<th>Parameter</th>
<th>EM</th>
<th>Observed ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\alpha_1$</td>
<td>.542</td>
<td>.542</td>
</tr>
<tr>
<td></td>
<td>$\alpha_2$</td>
<td>-.100</td>
<td>-.100</td>
</tr>
<tr>
<td></td>
<td>$\alpha_3$</td>
<td>.865</td>
<td>.865</td>
</tr>
<tr>
<td></td>
<td>$\psi_1$</td>
<td>.167</td>
<td>.169</td>
</tr>
<tr>
<td></td>
<td>$\psi_2$</td>
<td>.038</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>$\psi_3$</td>
<td>.333</td>
<td>.334</td>
</tr>
</tbody>
</table>

Note. $\alpha_k$ and $\psi_k$ are mean and variance impact estimates for covariate $k$. See note in Table 1 for the convergence criteria used for both methods.

**Reg-DIF Procedure**

I now examine the extent to which the penalized EM algorithm with soft thresholding identified similar DIF effects compared to the observed marginal likelihood approach. I use the Bayesian Information Criterion (BIC) to identify the optimal tuning parameter $\tau$ and, in effect, the final model results. In Table 3, I show which effects were identified by each procedure.

For this data set, the penalized EM algorithm yielded a more parsimonious solution than the observed marginal likelihood approach, at least for intercept DIF. Specifically, the penalized EM algorithm correctly identified intercept DIF effects for gender and study on Item 5, whereas the observed marginal approach identified all three covariates as exhibiting DIF on Item 5 (i.e., incorrectly for age). The penalized EM algorithm also correctly identified Item 2 as not exhibiting DIF while the observed marginal approach incorrectly identified DIF for Item 2. In contrast, however, the penalized EM algorithm incorrectly identified study as having DIF on Item 4, whereas the observed marginal approach correctly identified DIF by age on Item 4. Finally, the penalized EM algorithm correctly identified slope DIF on Item 5 for gender, but
incorrectly identified slope DIF on Item 4 for study. No slope DIF was identified using the observed marginal approach.

Table 3. Reg-DIF Final Model Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Item</th>
<th>Covariate for which DIF was identified</th>
<th>Intercept DIF</th>
<th>Loading DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>2</td>
<td>Study</td>
<td>Study</td>
<td>Study</td>
</tr>
<tr>
<td>EM</td>
<td>4</td>
<td>Gender, Study</td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td>5</td>
<td>Gender, Study</td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Observed ML</td>
<td>2</td>
<td>Age</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Observed ML</td>
<td>4</td>
<td>Age</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Observed ML</td>
<td>5</td>
<td>Age, Gender, Study</td>
<td>Age, Gender, Study</td>
<td></td>
</tr>
</tbody>
</table>

Note. Bold covariates represent DIF effects that were generated in the data (i.e., true DIF).

These findings suggest there may be performance differences between these estimation approaches, most likely because different methods were used to threshold estimates. That is, I incorporated soft-thresholding to shrink the DIF effects within the penalized EM algorithm. In contrast, the observed marginal likelihood approach did not use soft-thresholding but rather set DIF estimates to zero once they passed an arbitrarily small threshold value (due to the penalty). Beyond noting these differences between the two approaches, it is not possible to make more definitive conclusions based on a single data set. A more rigorous evaluation of the penalized EM algorithm, with more conditions and replications, will provide greater insight into whether there are any consistent performance differences between the two approaches. In the next chapter, I describe the Monte Carlo simulation study that I conducted to evaluate the penalized EM approach for Reg-DIF more thoroughly.
CHAPTER 3: EVALUATION

Monte Carlo Simulation Study

To evaluate the penalized EM algorithm more systematically, I ran an extensive Monte Carlo simulation study using a subset of previously generated data (Curran, Cole, Bauer, Hussong, & Gottfredson, 2016). Although these researchers evaluated scoring procedures in the MNLFA framework under various conditions of DIF, which is not the focus here, these data provided a useful benchmark for identifying DIF and a degree of consistency between studies. I also extended this study to include a larger number of scale items and more exogenous covariates. Not only do I compare my findings with those obtained using the observed marginal likelihood in previous work (Bauer et al., 2020), in which our group uses a smaller subset of these data, but I also include conditions where Reg-DIF would be impossible to perform without using the penalized EM algorithm.

Design Factors

In this study, I evaluated four of the five study factors originally studied by Curran et al. (2016), each with two levels and 500 replications per cell. In addition, I evaluated two different penalty functions. The five factors thus included were:

1. **Sample size**: 500 or 2000 observations;
2. **Number of items**: 6 or 12 items;
3. **Proportion of DIF**: \( \frac{1}{3} \) or \( \frac{2}{3} \) of items exhibit DIF;
4. **Magnitude of DIF**: small or large, as measured by an adaptation of the weighted “area between the curves” metric (Edelen, Stucky, & Chandra, 2015; Hansen et al., 2014); and
5. **Penalty function**: LASSO or MCP.

All simulation conditions were crossed, resulting in 32 unique conditions. Furthermore, all factors except for the penalty function are considered “between-subjects” factors because the levels do not vary across observations. In contrast, the penalty function factor is a “within-subjects” factor because each replication is tested for both levels of the factor. Three exogenous covariates, referred to as age (continuous), gender (dichotomous), and study (dichotomous), were also generated to affect both the latent variable distribution (impact) and a subset of the item parameters (DIF). The population parameters are shown in Table 4.
Table 4. Population Parameter Values for Monte Carlo Simulation Study

| Item | 6 Items | 12 Items | Intercept (Small DIF | Large DIF) | Loading (Small DIF | Large DIF) |
|------|---------|----------|----------------------|----------------|-------------------|
|      | DIF 1/3 | DIF 2/3  | Baseline  | Age | Gender | Study | Baseline  | Age | Gender | Study |
| 1    |         |          | - .5      |     |        |       | 1         |     |        |       |
| 2    | *       | *        | - .9      | .125 | .25   | .5    | - 1       | .5  | 1      |       |
| 3    | *       | *        | - 1.3     | .125 | - .25 | .5    |           | .5  | 1      |       |
| 4    | *       | *        | - 1.7     | .125 | .25   |       |           |     |        |       |
| 5    | *       | *        | - 2.1     |     | - .5  |       | .1       | .5  | 1      |       |
| 6    |         |          | - 2.5     |     |        |       |           |     |        |       |
| 7    |         |          | - .5      |     |        |       | 1         |     |        |       |
| 8    | *       |          | - .9      | .125 | .25   | .5    | - 1       | .5  | 1      |       |
| 9    | *       |          | - 1.3     | .125 | - .25 | .5    |           | .5  | 1      |       |
| 10   | *       |          | - 1.7     | .125 | .25   |       |           |     |        |       |
| 11   | *       |          | - 2.1     |     | - .5  |       | .1       | .5  | 1      |       |
| 12   |         |          | - 2.5     |     |        |       |           |     |        |       |

Note. Asterisks indicate items with DIF in the indicated condition. These data were originally simulated in Curran et al. (2016).
Second, I extended this simulation study by including two additional levels for the number of items condition, as well as a higher number of exogenous covariates. That is, in addition to the five study factors and their respective levels shown above, I included the following levels/conditions:

- **Number of items**: 48 or 96;
- **Number of exogenous covariates**: 3 or 6.

Unlike the simulation conditions generated by Curran et al., however, I did not cross these conditions with the other factors and their corresponding levels. This allowed me to limit the simulation study to a reasonable scope. Instead, I used a sample size of 2000, \( \frac{1}{3} \) proportion of DIF, large magnitude of DIF, and large mean / small variance impact for each of these additional levels. I chose this intersection of the other design factors when evaluating the new conditions because Reg-DIF was likely to perform well with such a large number of parameters in these types of conditions (Bauer et al., 2020; Belzak & Bauer, 2020). Moreover, since our previous approach of maximizing the observed data log-likelihood was unable to evaluate conditions with many item parameters (i.e., no more than 12 items with 3 DIF covariates), evaluating conditions with 48 and 96 items demonstrates the relative computational efficiency of the new algorithm in scenarios where regularization is particularly well suited – that is, identifying a sparse subset of parameters among many. I did, however, cross the additional number of items with the extra level for the number of exogenous covariates. Specifically, I evaluated 48 and 96 items with both 3 and 6 exogenous covariates. In the condition with 6 exogenous covariates, two were continuous and four were categorical, duplicating the pattern described previously with 3 exogenous covariates. The population parameters of this partial design followed the same block pattern as in Table 4. For instance, in the 48-item condition with 6 exogenous covariates, the first
24 items are specified block-wise with DIF (for $\frac{1}{3}$ proportion of DIF) and the first 3 exogenous covariates repeat twice with the same population parameter values on the DIF effects (for large magnitude of DIF). With 96 scale items and 6 exogenous covariates, for example, I estimated 96 items $\times$ 6 covariates $\times$ 2 intercept and slope DIF parameters, in addition to 6 covariates $\times$ 2 latent mean and variance impact parameters, with a total of 1164 parameters. This would be impossible if I were to directly maximize the observed marginal likelihood function, but it is in fact tractable using the penalized EM algorithm.

Finally, to explore the effect of $\gamma$ (i.e., the second tuning parameter in the MCP function) on the recovery of DIF effects, I evaluated three different values across all levels of sample size, number of items, proportion of DIF, and magnitude of DIF. Specifically, I compared $\gamma = 3$, $\gamma = 5$, and $\gamma = 7$. These values were chosen based on pilot testing. Recall that as the value of $\gamma$ increases, the closer MCP will mirror LASSO. In contrast, the closer the value of $\gamma$ gets to 1, the more the penalty will decrease in magnitude and the less biased the penalized estimate will become. Smaller values of $\gamma$ will also result in more local solutions due to the non-convexity of the MCP function. As such, I aimed to exclude values of $\gamma$ that would likely give many non-converged solutions. Given that the MCP function introduces non-convexity to the log-likelihood (whereas the LASSO function does not), I also evaluated different starting values for fitting Reg-DIF models with the MCP function. This will be discussed more in the Outcomes sub-section below and the Results section.

In summary, each simulated dataset (i.e., replication) included $J$ item responses for $N$ observations that were generated to have various amounts of DIF and impact from $K$ covariates (e.g., $K = 3$ covariates in the original data generation). For more information on the data generation process, I refer readers to Curran et al. (2016).
Procedure

I utilized my newly programmed regDIF R package to conduct the simulation study. First, for each replication I supplied a vector of 100 penalty values in descending order (e.g., 1.00, .99, ..., 0) to the estimation procedure, starting with a tuning parameter that penalized all DIF effects to zero. Note that this nicely handled the identifiability problem, as DIF effects entered the model until $\tau$ approached zero, rather than being removed from the model as they were in previous work. I also created an automatic stopping rule for the Reg-DIF procedure to ensure that there was at least one DIF parameter per covariate, per parameter, constrained to equal zero. Requiring that at least one DIF effect on both the intercept and slope of each item be constrained to zero is sufficient to produce a minimally identified model in the presence of a weak or zero penalty. Also note that the vector of tuning parameters had smaller increments near 0 to provide finer differentiation between DIF effects entering the model (see Belzak & Bauer, 2020, for a similar approach). Next, I ran all penalty values on 500 simulated replications per condition, saving information criteria and parameter estimates for each model. I used both BIC and AIC to select the optimal model. In particular, the model associated with the penalty value generating the minimum BIC or minimum AIC was selected for further analysis. Although previous research has shown that BIC outperforms AIC when using ad-hoc thresholding (Bauer et al., 2020), I aimed to re-examine this outcome using soft-thresholding. Finally, with the chosen model I used a variety of outcome metrics to evaluate the performance of the penalized EM algorithm, as described below.

Outcomes

To evaluate the quality of the penalized EM algorithm, I first recorded true positive (TP) and false positive (FP) rates for DIF effects that were identified with the best-BIC and best-AIC
models. A true positive refers to a non-zero DIF effect (intercept or slope) that exists within the population-generating model and also appears in the estimated model chosen for the sample. A false positive refers to a DIF effect that does not exist within the population-generating model but nevertheless appears in the estimated model chosen for the sample. I counted true and false positives at the item-level as opposed to the parameter-level. That is, if either the intercept or slope effect was identified with DIF, I counted this item with DIF. Although this method of counting yields both more true and false positives, it is consistent with the motivation of researchers who tend to care more about whether an item exhibits DIF rather than precisely how it exhibits DIF. We also used this method of counting TPs and FPs in Bauer et al. (2020); thus, the results can be readily compared between the two estimation approaches. As we also provided in Bauer et al., I present parameter-level true and false positive results in Appendix C to provide additional context to the item-level findings. In addition, prior work by our group evaluated TP and FP rates for DIF effects that not only remained in the final model but also were statistically significant when evaluated using naïve standard errors (Bauer et al., 2020; Belzak & Bauer, 2020). Although the naïve standard errors are incorrect due to their failure to incorporate uncertainty regarding model selection, using these significance tests nevertheless resulted in fewer Type I errors without much loss in power. In the present work, however, I counted effects that remained in the model regardless of statistical significance. This is a more conventional standard in regularization studies (Hastie et al., 2017).

Next, I examined mean squared error (MSE) of the parameter estimates from the chosen model and decompose the MSE into contributions due to bias versus sampling variance. In prior work, we computed MSE with and without the LASSO penalty after DIF was identified (Belzak & Bauer, 2020), and found that computing MSE without the LASSO penalty showed a decrease
in estimator bias but also an increase in variance, whereas computing MSE with the LASSO penalty showed an increase in bias but a concurrent decrease in variance. Notably, there were no considerable differences between the two approaches with respect to MSE. Therefore, in the current work I only computed MSE with the penalty function.

Lastly, for the Reg-DIF procedures that used MCP, I used two different sets of starting values aside from the default of “warm” values (i.e., estimates obtained from the previous tuning parameter value, which are then used to initialize the EM algorithm for the next tuning parameter value). The first (different) set of starting values were “cold” (i.e., all slopes set to 1 and all other parameters set to 0), which were used to re-estimate the best-fitting model according to BIC. After identifying the model with the minimum BIC, in other words, I re-estimated the model using the same level of penalization ($\tau$) and “cold” starting values. The second set of starting values were obtained from a LASSO model. That is, I randomly chose a tuning parameter value along the regularization path of the LASSO Reg-DIF procedure and re-fit the model using the MCP function and starting values obtained from the LASSO model at that value of $\tau$. For both of these different sets of starting values, I tallied the number of models in which the estimates from the “warm” starting values (obtained from the normal Reg-DIF procedure) were different than the “cold” starting values and the LASSO “random” starting values. I only checked two different conditions, however, to limit the scope of the simulation study. The conditions are (1) sample size of 500, number of items 12, proportion of DIF $\frac{2}{3}$, and large DIF magnitude, and (2) sample size of 2000, number of items 6, proportion of DIF $\frac{1}{3}$, and large DIF magnitude. Notably, these two conditions are at the extreme of what might occur in real data, that is, where MCP may be most (1) and least (2) likely to identify local maximums on the log-likelihood surface.
CHAPTER 4: RESULTS

I describe the simulation results next, beginning with false positive and true positive rates for core simulation conditions, then false and true positives for the extra conditions (many items and covariates), and finally mean squared error as a function of squared bias and variance. I also compare my findings with our group’s previous results in Bauer et al. (2020), where we analyzed many of the same simulation conditions with the observed marginal likelihood estimation approach. This provides a useful benchmark for the current results. Note that in comparing LASSO with MCP, I used \( \gamma = 3 \) for MCP because it showed a relatively nice balance between TPs and FPs. \( \gamma = 3 \) is also the most different from LASSO among the three values of \( \gamma \) when compared to LASSO; namely, it is closest to \( \gamma = 1 \), thus resulting in less estimator bias but also greater non-convexity in the likelihood.

**False Positive Rates**

Figure 1 shows false positive (FP) rates across the five core simulation factors while using BIC as the selection criterion. First, as expected, greater DIF in both magnitude and proportion resulted in larger FP rates overall. Larger sample sizes exacerbated these high FPs due to large amounts of DIF; in sample sizes of 2000, for instance, FP rates for the LASSO penalty reached as high as 78% when 66% of DIF effects were large in magnitude. MCP (\( \gamma = 3 \)) reached 56% in the same large DIF and large sample size condition. Second, also as expected, less DIF – both in magnitude and proportion – yielded smaller FP rates. Although the LASSO method exhibited more FPs in sample sizes of 2000 compared to 500, MCP showed smaller FP rates as sample sizes increased. In sample sizes of 2000 with 33% proportion of DIF, for
example, MCP never rose above 7% FPs. LASSO, on the other hand, was as high as 19% FPs in the same conditions. Finally, the number of items only had a marginal effect on FP rates, as FPs were slightly lower in the 12-item condition compared to the 6-item condition for both LASSO and MCP. The largest reduction occurred with the MCP approach in the largest sample size and the greatest amount of DIF condition, such that FPs decreased by 13% when increasing the number of items from 6 to 12.
Figure 1. LASSO and MCP False Positive Rates Using BIC

Figure 2 shows FP rates for the same conditions but using AIC instead. Not surprisingly, the more liberal AIC approach yielded considerably more FPs than using BIC. Namely, across all conditions, AIC showed far more FPs for both LASSO and MCP. It is worth noting, however,
that the MCP method showed fewer FPs than LASSO in all but one of the conditions. This was not observed with the BIC results, as LASSO appeared to have slightly fewer FPs in the smaller sample size condition. Larger sample sizes significantly exacerbated FP rates for LASSO, as was also observed to a lesser extent when using BIC to identify DIF. MCP FP rates were less affected by larger sample sizes. Nevertheless, FPs reached as high as 72% for MCP and as high as 96% for LASSO, both of which would likely be considered unacceptable by applied researchers.

Figure 2. LASSO and MCP False Positive Rates Using AIC

Note: MCP was implemented with $\gamma = 3$. 

43
Figure 3 shows FP rates for three different values of $\gamma$ (i.e., $\gamma = 3, 5, \text{or} 7$) using BIC for model selection. In most conditions, larger values of $\gamma$ showed higher rates of FPs. This was expected, given that LASSO showed higher FP rates on average and that MCP becomes more LASSO-like as $\gamma \to \infty$. Interestingly, in conditions with 66% of large magnitude DIF, larger values of $\gamma$ led to lower rates of FP, a reversal in pattern. This finding was surprising because LASSO showed the worst performance in these conditions; thus, I expected larger values of $\gamma$ (i.e., a more LASSO-like MCP function) to result in the highest FP rates for MCP in conditions with the most amount of DIF. Nevertheless, evaluating different values of $\gamma$ provides greater nuance to the importance of $\gamma$ when implementing Reg-DIF with MCP.
I now compare the current FP findings with previous results from the simulation study conducted by our research group in Bauer et al. (2020). In that estimation approach, we implemented Reg-DIF with a LASSO penalty using an observed marginal likelihood function and performed thresholding of the DIF effects outside of the optimization procedure. We also
compared LASSO with and without using naïve (and therefore under-estimated) standard errors, as well as with the IRT-LR-DIF method, which evaluates DIF one item at a time using likelihood ratio tests. In particular, we found that using LASSO with naïve standard errors resulted in the fewest number of FPs compared to using LASSO without standard errors and IRT-LR-DIF. Using LASSO without standard errors showed the next fewest FPs, and IRT-LR-DIF was the least discriminating method of the three.

The current findings provide some validation for our previous results. In Figure 4, I compared the penalized EM algorithm results with the marginal observed likelihood results. Note that this comparison did not include the use of standard errors for DIF identification. Specifically, when using BIC to select the best model, using LASSO (without standard errors) showed similar FP results across both estimation methods, particularly in smaller sample sizes and with less DIF. For instance, in sample sizes of 500, 6 items, and large DIF magnitude, the observed marginal likelihood approach showed 13% and 47% FP rates in the 33% and 66% proportion of DIF conditions respectively, whereas the penalized EM algorithm showed 10% and 46% FP rates in the same conditions. One evident deviation between the estimation methods occurred in the larger sample size condition with 66% proportion of DIF. The observed marginal likelihood approach exhibited FP rates of 40% and 48% in the small and large DIF magnitude conditions respectively, while the penalized EM algorithm showed 51% and 78% FP rates in the same conditions. The larger FP rates for the penalized EM algorithm could be due to the method of soft-thresholding, although it is not immediately clear why soft-thresholding would result in more FP rates. I consider this possibility more in the Discussion section. In addition to LASSO, the MCP method also showed mixed results in comparison to the observed marginal likelihood approach. Shown in Figure 4 as well, MCP exhibited fewer FPs in larger sample sizes, which
was also found when using LASSO with the penalized EM algorithm. On the other hand, MCP also exhibited higher FPs in conditions with small DIF magnitude when compared with the observed marginal likelihood approach. Thus, although MCP was mostly conservative when compared to the observed marginal likelihood approach, and even compared to LASSO, there were some mixed FP findings for MCP that complicate the overall conclusion. Finally, in contrast to using BIC for model selection, the AIC FP results showed relatively closer correspondence between the two estimation approaches. Although slightly more FPs were identified in sample sizes of 2000 when using the penalized EM algorithm with LASSO, roughly the same number of FPs were identified across both methods in sample sizes of 500.
Figure 4. Penalized EM (LASSO and MCP) and Observed Marginal Likelihood (LASSO) False Positive Rates Using BIC

In summary, the FP results here show that MCP exhibits fewer FP rates compared to LASSO as sample sizes increase to 2000. In other words, MCP appears to be more conservative.
than LASSO. Previous results using the observed marginal likelihood approach (with thresholding outside of the optimization procedure) also showed fewer FP rates than the penalized EM algorithm (with soft-thresholding within optimization) as sample sizes and the amount of scale DIF increased. In light of these findings, I turn next to describe the true positive results.

**True Positive Rates**

Figure 5 shows true positive (TP) rates across the core simulation factors while using BIC for model selection. In complement to the FP rates, and as I expected, TP rates were higher as sample sizes increased from 500 to 2000, reaching 100% recovery for the LASSO method and 99% for MCP. In contrast, LASSO exhibited its lowest TP rate at 39% in the small sample size condition with 33% of small DIF, and MCP exhibited its lowest TP rate at 53% in the small sample size condition with 66% of items showing small DIF. Although smaller proportions of DIF exhibited greater TPs on average – presumably because the scale is better defined by the latent variable without many DIF items – larger magnitudes were almost always more sensitive to recovering true DIF in the population. The LASSO method also exhibited higher TPs compared to MCP overall, which makes sense given that we observed higher FPs for LASSO as well. This was observed to a somewhat lesser extent in sample sizes of 500, where MCP sometimes exhibited higher TP rates, such as in the 33% proportion of small DIF conditions. Finally, as shown with FP rates, the number of items appeared to have minimal impact on TP rates, particularly as sample sizes increased.
Figure 6 shows TP rates while using AIC for model selection. As anticipated from the considerably higher FP rates, TP rates for AIC were also much higher relative to using BIC for model selection, particularly in the small sample size and small DIF magnitude conditions. The lowest TP rate was only 83% for LASSO and only 70% for MCP, both in the small sample size.

Note: MCP was implemented with $\gamma = 3$. 
condition with 66% of small DIF. In addition, LASSO exhibited uniformly higher TP rates compared to MCP, although this result was less pronounced with larger sample sizes. As before, the number of items had a marginal effect on DIF recovery when using AIC. The large number of FP rates for AIC selection, however, overshadow these high TPs.

Figure 6. LASSO and MCP true positive rates using AIC

Note: MCP was implemented with $\gamma = 3$. 

51
Figure 7 provides TP rates for the three different values of $\gamma$. In large part, these TP results mirror the FP results insofar that where larger values of $\gamma$ led to more FP rates (i.e., in conditions with less DIF), larger values of $\gamma$ also resulted in more TP rates. In contrast, where larger values of $\gamma$ led to the opposite effect of fewer FP rates (i.e., in conditions with more DIF), we found that larger values of $\gamma$ exhibited greater TP rates. This provides greater evidence that there is indeed a sweet spot for $\gamma$ which yields better rates for both TPs and FPs when there are large amounts of DIF.
In comparison to our research group’s previous results in Bauer et al. (2020), LASSO exhibited either slightly higher TP rates or approximately the same TP rates in the current study (using BIC for model selection). This is shown in Figure 8. For instance, the penalized EM LASSO achieved 39% TP rates in sample sizes of 500 and with a third of items showing small DIF, whereas the implementation of LASSO (without standard errors) in Bauer et al. reached
only 30% TP rates. This appears to account for some differences in FP rates between the two estimation approaches (i.e., higher FP rates for the penalized EM implementation of LASSO), particularly in conditions with smaller sample sizes and less DIF. In conditions with larger sample sizes and more DIF, however, higher TP rates could not account for higher FP rates. Namely, the penalized EM LASSO exhibited lower TP rates (and higher FP rates) when compared to the observed marginal likelihood LASSO in sample sizes of 2000 and with two thirds of items showing large DIF. In other words, the penalized EM algorithm performs worse overall in these high DIF conditions.
Figure 8. Penalized EM (LASSO and MCP) and Observed Marginal Likelihood (MCP) True Positive Rates Using BIC.

6 Items

12 Items

Note: P-EM stands for Penalized Expectation-Maximization, and O-ML stands for Observed Marginal Likelihood. LASSO O-ML results are reproduced from Bauer et al. (2020).

In summary, these core simulation findings showed that LASSO achieved higher TP rates compared to MCP, although this was overshadowed by higher FP rates as well. Notably, the
advantage of LASSO over MCP in TP rates appeared to be smaller than the advantage of MCP over LASSO in FP rates, suggesting that MCP may provide a better balance of TPs and FPs overall. In comparison to our LASSO findings in Bauer et al. (2020), the current TP results were somewhat different, showing more TPs and FPs using the penalized EM algorithm in conditions with smaller sample sizes and less DIF. That is, the penalized EM algorithm appeared to be slightly more liberal than the observed marginal likelihood approach in these conditions. On the other hand, the observed marginal likelihood approach appeared to outperform the penalized EM approach in large sample sizes and in scales with many DIF items. Despite these differences, however, I found more consistencies than discrepancies in comparing the results between the penalized EM algorithm and the observed likelihood approach, thus validating our previous results in Bauer et al.

**Many Items and Many Covariates**

Figure 9 shows TP and FP results for the four remaining conditions using BIC, namely, 48 and 96 items crossed with 3 and 6 covariates. First, the TP rates for both LASSO and MCP approached one, or perfect recovery of true DIF effects, across all combinations of many items and many covariates. MCP, however, exhibited slightly higher recovery of true DIF effects compared to LASSO. Parameter-specific findings in Figure 10 show that LASSO achieves much higher TP rates for slope DIF compared to MCP, albeit with some larger FP rates as well. On the other hand, MCP shows marginally better recovery of intercept DIF (i.e., higher true positive rates) while also maintaining better control of Type I error compared to LASSO. This is primarily how MCP showed higher *item*-level TPs compared to LASSO. Second, as for the FP results, MCP showed better control of FP rates compared to LASSO across all combinations of items and covariates. Specifically, MCP held FP rates near the nominal alpha rate of .05 for three
of the four conditions (ranging from .02 to .04), whereas LASSO had FP rates ranging from .16 to .17 for these three conditions. Interestingly, FP rates were higher for both MCP and LASSO in the condition with the greatest number of item parameters, that is, 96 items and 6 covariates. This could be due to a relatively small sample size of 2000 when compared with the total number of parameters present in the data-generating model.

Figure 9. LASSO and MCP True Positive and False Positive Rates for Conditions with Many Items and Many Covariates Using BIC.

![Graph showing TP and FP rates for LASSO and MCP with different conditions.](image)

Note: In this comparison, MCP was implemented with $\gamma = 3$, sample size was 2000, and 1/3 of the items were generated with large DIF.
Figure 10. LASSO and MCP True Positive and False Positive Rates for Intercept and Slope DIF Effects in Conditions with Many Items and Many Covariates using BIC.

<table>
<thead>
<tr>
<th></th>
<th>Intercept DIF</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Positive Rate</td>
<td>False Positive Rate</td>
<td>True Positive Rate</td>
<td>False Positive Rate</td>
<td></td>
</tr>
<tr>
<td>3 Covariates</td>
<td>95</td>
<td>0.00</td>
<td>95</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5 Covariates</td>
<td>97</td>
<td>0.00</td>
<td>96</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>6 Covariates</td>
<td>96</td>
<td>0.00</td>
<td>96</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>48 Items</td>
<td></td>
<td></td>
<td>96 Items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 Items</td>
<td></td>
<td></td>
<td>96 Items</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: In this comparison, MCP was implemented with $\gamma = 3$, sample size was 2000, and 1/3 of the items were generated with large DIF.

Figure 11 shows the same study results as Figure 9 but using AIC instead of BIC. In particular, TP rates for both LASSO and MCP across all four study conditions are perfect or equal to one. The cost of achieving perfect recovery of true DIF effects, however, is that there
were much higher FP rates for both penalization approaches. FP rates for LASSO hovered around .8, whereas they hovered around .5 for MCP. This difference between LASSO and MCP is consistent with the FP results we found when using BIC for model selection, such that MCP exhibited better control of FPs compared to LASSO. Higher FP rates overall for AIC is also consistent with our findings in the other study conditions. Namely, AIC is far too liberal in identifying DIF, leading to both higher TP rates and FP rates. In the conditions with many items and many covariates, the trade-off afforded by using AIC for model selection is negligible. That is, the modest gain in recovering true DIF does not outweigh the large loss in incorrectly identifying DIF where it does not exist. Therefore, using BIC appears to perform a much better job in model selection relative to AIC.

Figure 11. LASSO and MCP True Positive and False Positive Rates for Conditions with Many Items and Many Covariates Using AIC.

![Figure 11](image)

Note: MCP was implemented with $\gamma = 3$. Sample size was 2000. 1/3 of items had large DIF.
Mean Squared Error

In Figure 12, mean squared error (MSE) is shown as an aggregation of squared bias and variance for DIF effects that were present in the population. In other words, regardless of whether the effect was correctly identified with DIF or not in the final model, squared bias and variance is shown for those effects that were non-zero in the data-generating model. These results were computed using BIC for model selection and MCP with $\gamma = 3$. From a holistic perspective, MSE was expectedly smaller in larger sample sizes, with smaller magnitudes of DIF, and when fewer items had DIF, largely as a consequence of reduced variance. These results were collapsed (i.e., averaged) across the number of items because there were few differences between the 6-item and 12-item conditions. In conditions where 33% of the items had DIF, bias was approximately the same for both LASSO and MCP, but variance was notably higher for MCP. One explanation for this is that the more complex MCP penalty function may have led to greater estimator variation. For conditions where 66% of the items had DIF, both squared bias and variance were higher for MCP compared to LASSO. On one hand, this is not surprising given that MCP showed considerably worse performance in TP rates under the same conditions. On the other hand, it is somewhat unexpected that MCP showed greater bias on average than LASSO. Since MCP seeks to remove estimator bias induced through penalization, one might expect MCP to show less bias than LASSO. However, because MSE is computed in the process of identifying the correct level of model complexity (i.e., $\tau$), comparing bias between MCP and LASSO is less straightforward. If I were to use the correct model specification instead for both MCP and LASSO, for instance, then I would expect MCP to show less bias. In fact, when comparing bias only for DIF effects that were correctly specified (i.e., true positives) by either MCP or LASSO, MCP showed less bias than LASSO in larger sample sizes (i.e., $N = 2000$)
and larger numbers of items (i.e., \( J = 12 \)). This is evidenced in Figure 13, which breaks down MSE by different numbers of items, in addition to sample sizes. Notably, The advantage for MCP was also pronounced when correctly recovering intercept DIF over slope DIF. This makes sense given that MCP underperformed LASSO in recovering slope DIF (see Figure 4A). Despite the nuance in these results, LASSO appeared to yield lower MSE than MCP overall for DIF effects that were present in the population model (but not necessary for those that were present in the population and correctly recovered).

Figure 12. LASSO and MCP Mean Squared Error for DIF Effects that Were Present in the Population.

![Graph showing LASSO and MCP MSE](image)

Note: Model selection was accomplished using BIC. MCP was implemented with \( \gamma = 3 \).
Figure 13. LASSO and MCP Mean Squared Error for DIF effects that Were Present in the Population and Were Correctly Identified.

Note: Model selection was accomplished using BIC. MCP was implemented with $\gamma = 3$.

Figure 14 shows MSE as a function of squared bias and variance, but instead for DIF effects that were not present in the population. In other words, regardless of whether the effect
was correctly excluded or not from the final model, squared bias and variance is shown for DIF effects that were not included in the data-generating model. Again, MCP exhibits higher MSE, particularly in conditions where 66% of the items had DIF. The large MSE for MCP is also an unexpected result, not only because one might expect MCP to exhibit smaller bias overall, but also because MCP showed fewer FP rates than LASSO. On further inspection of the FP rates, LASSO showed a higher number of incorrectly identified DIF effects that were small in magnitude. This may be due to the fact that no statistical significance information was used to identify DIF with LASSO, as we did in Bauer et al. (2020). To the contrary, MCP showed a fewer number of incorrectly DIF effects but which were larger in magnitude. Thus, although LASSO had higher FPs due to a large number of small DIF effects, it still showed less bias and variance compared to MCP, particularly in conditions with more DIF.

**Different Starting Values**

It is important to determine whether the non-convexity of the MCP function will yield different model estimates based on different starting values, as this could have implications for model estimation/initiation when using MCP. Above (i.e., see Outcomes sub-section of Evaluation section), I have enumerated two different sets of starting values, including “cold” and LASSO “random” starting values, which were used to re-estimate models with the MCP function and compared with the default of using “warm” starting estimates. I also identified two conditions in which MCP would likely see the greatest (i.e., small sample size with more DIF) and least (i.e., large sample size with less DIF) number of differences in model estimates. These results are described next.

In the condition with small sample sizes and large amounts of DIF, I found that re-estimating the best-fitting model (according to BIC) using “cold” start values with the MCP
function resulted in 95% of models showing different estimates compared to using “warm” start values. In contrast, re-estimating models using LASSO “random” start values (at the same value of $\tau$) with the MCP function only resulted in 40% of models exhibiting different estimates. Note that I predicted there to be the greatest threat of identifying a local maximum in this simulated data condition. For the condition with large sample sizes and small amounts of DIF, however, there was expectedly less identification of different estimates when using different starting values. Specifically, using “cold” start values with the MCP function resulted in 46% of models showing different estimates compared to using “warm” start values, whereas using LASSO “random” start values resulted in 17% of models with different estimates. In sum, these results suggest that the risk of identifying local maxima with the MCP function could be high depending on which starting values are used and which data characteristics are present. I discuss the implications of these results later.
Figure 14. LASSO and MCP Mean Squared Error for DIF Effects that Were Not Present in the Population.

Note: Model selection was accomplished using BIC. MCP was implemented with $\gamma = 3$. 
CHAPTER 5: EMPIRICAL EXAMPLE

Having established the performance of the new penalized EM approach, I now provide an empirical demonstration with data measuring violent and non-violent delinquent behaviors. My goals are threefold with these data. First, I wish to demonstrate the relevance and usefulness of the new Reg-DIF estimation procedure for behavioral and social scientists. Second, this empirical example facilitates a tutorial for the regDIF R package, thereby providing researchers a straightforward guide for implementing Reg-DIF with the penalized EM algorithm. Lastly, these data were also analyzed by Bauer (2017) to illustrate the MNLFA model for DIF testing, wherein an iterative sequence of likelihood ratio tests was used to evaluate DIF across five predictors. Reanalyzing these data here allows me to evaluate the Reg-DIF results in light of Bauer’s findings, thus providing a head-to-head comparison between Reg-DIF (via the penalized EM algorithm) and the likelihood ratio test method⁴ (via the observed likelihood approach) in a real-world setting. It is important to note that the dimensions of the data (number of participants, items, and covariates) make the prior implementation of Reg-DIF via observed marginal likelihood computationally infeasible; hence, this approach will not be considered here. This further highlights the advantages of the new Reg-DIF procedure using penalized EM.

Before the analysis, I expected that fewer DIF effects would be identified using regularization, given that regularization favors sparse solutions. However, I also expected that similar items would exhibit DIF, in line with Bauer’s (2017) results. In addition, I expected that

---

⁴ An adapted likelihood ratio test method was used by Bauer (2017) which made multiple sweeps through the items, anchoring different sets of items at a time. Some pruning of non-significant DIF effects was also made at the final model fitting stage. See Bauer (2017) for more details on the adapted likelihood ratio test method.
MCP would provide more accurate results because the sample size was large. In the simulation results, that is, MCP showed better control of Type I errors compared to LASSO with larger sample sizes. These predictions reflected my findings from the Monte Carlo simulations.

Data and Items

The data were selected from the National Longitudinal Study of Adolescent to Adult Health (Add Health) and consist of a self-weighting subsample of 4,243 adolescents in the United States. Adolescents provided self-reports on two dimensions of delinquent behavior, with 10 items for violent behavior (V) and 8 items for non-violent delinquent behavior (NV). Although the items were presented with ordinal response options, Bauer (2017) dichotomized the responses (to yes/no) due to sparseness in the upper category levels. Table 5 provides the item descriptions and marginal endorsement rates as shown in Bauer (2017). Impact and DIF were evaluated as a function of age (ranging from 12 to 18, M_{age} = 14.9, SD_{age} = 1.7) and sex (47% male). In line with Bauer, five features were constructed from the two covariates: age, age^2, male, male \times age, and male \times age^2. In the original analysis, age was centered, and male was dummy-coded. For more details on these data, I refer readers to Bauer (2017).
Table 5. Items Evaluated for DIF in the Empirical Example

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item Descriptions</th>
<th>Marginal % Endorsement</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Non-Violent) Delinquency</td>
<td>D1. Paint graffiti/signs on someone else’s property or in a public space</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>D2. Deliberately damage property that did not belong to you</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>D3. Lie to parents/guardians about where you had been or whom with</td>
<td>53.2</td>
</tr>
<tr>
<td></td>
<td>D8. Drive a car without its owner’s permission</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>D9. Steal something worth more than $50</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>D10. Go into a house or building to steal something</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>D13. Steal something worth less than $50</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>D15. Were loud, rowdy, or unruly in a public place</td>
<td>48.2</td>
</tr>
<tr>
<td>Violent Behavior</td>
<td>V1. Saw someone shoot or stab another person</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>V2. Someone pulled a knife or gun on you</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>V3. Someone shot you</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>V4. Someone cut or stabbed you</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>V5. You got into a physical fight</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>V6. You were jumped</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>V7. You pulled a knife or gun on someone</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>V8. You shot or stabbed someone</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>D6. Hurt someone badly enough to need bandages or care from doctor/nurse</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td>D14. Take part in a fight where a group of your friends was against another group</td>
<td>17.8</td>
</tr>
</tbody>
</table>

**Previous Results**

Using the likelihood ratio test method, Bauer (2017) identified significant DIF in half of the items across both factors, as well as impact in all of the features. Specifically, for the non-violent factor, D2 (age and male), D3 (age, age^2, male, and male × age), D8 (age and age^2), and D15 (male) exhibited DIF. For the violent factor, DIF was identified in V1 (age, age^2, and male), V2 (age and age^2), V3 (age and age^2), D6 (age, male, and male × age), and D14 (age, male, and male × age). Note that lower-order terms were retained even if non-significant in the event that a higher-order term was significant. The non-violent factor exhibited mean impact for all features except the male × age^2 interaction, while it showed variance impact for age and male. Similarly, the violent factor showed mean impact for all features, as well as variance.
impact for age and male. The final model also allowed for the factor covariance to be a function of both age and male.

In addition to changing DIF detection procedures, the current application of Reg-DIF differed from Bauer’s (2017) prior analysis in three important ways. First, I evaluated DIF only in separate unidimensional models for each factor. In other words, I did not fit a two-factor model as the final model. Second, the automated Reg-DIF procedure does not currently make provisions for the retention of lower-order terms involved in higher-level effects. Third, the main effects (age and male) were standardized for purposes of using regDIF5, which were then used to compute the higher-level effects. This was not done in the original application when using likelihood ratio tests, meaning that the parameter estimates are not directly comparable between the prior and current work. Thus, the primary focus is on whether similar patterns of DIF were detected.

**Using regDIF R Package**

I provide a short tutorial here on using the regDIF R package here. To simplify the presentation, I focus on running LASSO on the non-violent factor.

First, I loaded the data into R using the following syntax:

```r
1. data <- read.table("~/nonviolent.dat")
```

where `nonviolent.dat` refers to the text file containing the non-violent delinquent behavior dataset. Next, I subset the `data` into two separate objects, one for the item responses and one for the exogenous covariates:

---

5 There is no consensus on whether standardizing interaction effects in regularization is advisable. Hastie et al. (2010) recommends standardizing all predictors in applications of regularization because it allows for all predictors to be penalized by the same degree. In the context of interactions, however, Bien, Taylor, and Tibshirani (2013) first standardize the main effects and then compute the higher-level effects. We use the latter approach because it is consistent with practice; that is, higher-level effects are not typically standardized themselves in typical regression analyses.
I assume here that the main effect predictors in pred_data have already been standardized, such that age and male have a mean and standard deviation of 0 and 1, and that the higher-level effects were computed using the standardized predictors. Finally, I used the main function in the regDIF R package to run the Reg-DIF procedure. This is shown as:

```r
4. fit <- regDIF(item.data = item_data,
5.     pred.data = pred_data,
6.     stdz = FALSE)
```

The first two arguments shown above must be specified by the user. Specifically, the item data is specified first as item.data, and the covariate data is specified second as pred.data. The third argument is optional and specifies whether to automatically standardize the predictors in the model. With interaction effects that have been created from standardized predictors, I set this argument to be FALSE so as not to standardize the higher-level effects themselves. There are other default arguments in regDIF that assume the item responses are distributed Bernoulli and that a 2-parameter logistic IRT model is specified to estimate the parameters. Finally, the regDIF function assumes that the penalty function is LASSO, but one may also specify MCP.

It should be noted that there are a number of other defaults used in the regDIF function. These defaults can be changed according to the user’s needs. By default, for instance, regDIF automatically computes 100 tuning parameter values, starting with a value that is large enough so that all DIF parameter estimates will be equal to zero. The tuning parameter values are also computed such that larger decrements occur at larger values of the tuning parameter and smaller decrements occur at smaller values of the tuning parameter. Other defaults include gamma = 3,
which controls the degree of “relaxation” of the MCP penalty function if it is used,\(^6\) and \(\text{stdz} = \text{TRUE}\), which standardizes the exogenous covariates. More information about the model defaults can be found at [https://github.com/wbelzak/regDIF](https://github.com/wbelzak/regDIF).

For the present application, I obtained the best-fitting model coefficients by using BIC for model selection. This is shown as

```
7. lasso_coefs <- coef(fit, tau = "min", method = "bic")
```

where "min" indicates that only the coefficients for the minimum value of the fit statistic should be returned, and "bic" indicates the fit statistic that should be used to select the best-fitting model. In addition to MCP, these LASSO model results are shown in the next section.

**Current Results**

The previous results from Bauer (2017) are now compared with the current results using Reg-DIF. In Table 6, parameter estimates for the non-violent factor are provided for both the LASSO and MCP methods, and in Table 7, parameter estimates for the violent factor are provided for both Reg-DIF methods. Furthermore, I use green, blue, and yellow color codes to indicate DIF effects that were (1) identified by both Reg-DIF (LASSO or MCP) and likelihood ratio tests (green), (2) identified by Reg-DIF but not by likelihood ratio tests (blue), and (3) identified by likelihood ratio tests but not by Reg-DIF (yellow).

Notably, 8 DIF effects were identified with both LASSO and likelihood ratio tests (see green highlights in Table 6), whereas 6 DIF effects were identified with both MCP and likelihood ratio tests. These consistent findings provide some measure of confidence in the identified effects. In particular, it appears that DIF occurs primarily in items D2 (property

---

\(^6\) Based on the simulation results, I use gamma = 3 as the default in regDIF because this value seems to provide the greatest difference between MCP and LASSO while not risking too much non-convexity in the log-likelihood function.
damage; DIF by age and sex), D3 (lying to parents; DIF by age and sex), D8 (taking a car without permission; DIF by age), and D15 (rowdy in public; DIF by sex). These results also seem to provide some instantiation of the simulation findings – namely, that LASSO is more sensitive (i.e., both higher true and false positives) for identifying DIF than MCP. In terms of DIF effects that were identified with Reg-DIF but not with Bauer’s (2017) method of using likelihood ratio tests (see blue highlights in Table 6), LASSO identified 3 additional DIF effects, whereas MCP identified 5 additional DIF effects. Two of these additional DIF effects were identified by both LASSO and MCP for the item D9 (steal > $50; DIF by age and sex). Given that LASSO and MCP showed high true positives when sample sizes were large, it is plausible that DIF in D9 exists in the population. Finally, for DIF effects that were identified by likelihood ratio tests but not by Reg-DIF (see yellow highlights in Table 6), Bauer found 2 additional quadratic age DIF effects on items D3 and D8, whereas LASSO and MCP did not. It is difficult to determine whether these DIF effects are true or false positives, however, because Bauer’s likelihood ratio test approach has not yet been examined using simulations.
Table 6. Reg-DIF Parameter Estimates for Non-Violent Delinquency Factor Using BIC

<table>
<thead>
<tr>
<th>Reference Parameter</th>
<th>LASSO</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MCP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Age</td>
<td>Age²</td>
<td>Male</td>
<td>Male × Age</td>
<td>Male × Age²</td>
<td>Baseline</td>
<td>Age</td>
<td>Age²</td>
<td>Male</td>
<td>Male × Age</td>
<td>Male × Age²</td>
</tr>
<tr>
<td>(Non-Violent) Delinquency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>−0.01</td>
<td>−0.29</td>
<td>0.06</td>
<td>0.03</td>
<td>0.04</td>
<td>0</td>
<td>−0.07</td>
<td>−0.29</td>
<td>0.13</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Variance</td>
<td>1</td>
<td>−0.17</td>
<td>0.23</td>
<td>0.32</td>
<td>0.07</td>
<td>−0.11</td>
<td>1</td>
<td>−0.14</td>
<td>0.23</td>
<td>0.17</td>
<td>0.03</td>
<td>−0.11</td>
</tr>
<tr>
<td>D1. Graffiti</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−2.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−2.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2. Property damage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−2.01</td>
<td>−0.05</td>
<td></td>
<td>0.14</td>
<td></td>
<td></td>
<td>−2.06</td>
<td></td>
<td></td>
<td>0.37</td>
<td>−0.01</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3. Lie parents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.51</td>
<td>0.22</td>
<td></td>
<td>−0.13</td>
<td>−0.06</td>
<td></td>
<td>0.55</td>
<td>0.40</td>
<td></td>
<td>−0.35</td>
<td>−0.19</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8. Car w/o permission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−2.55</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−2.56</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9. Steal &gt; $50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−3.74</td>
<td>0.08</td>
<td>−0.07</td>
<td></td>
<td></td>
<td></td>
<td>−3.79</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
<td>−0.20</td>
</tr>
<tr>
<td>Slope</td>
<td>1.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10. Steal from house</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−3.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−3.83</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D13. Steal &lt; $50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−1.75</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−1.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D15. Rowdy in public</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.24</td>
<td>−0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−0.24</td>
<td></td>
<td></td>
<td>−0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Green shading indicates parameters present in the final MNLFA model estimated by Bauer (2017), who used likelihood ratio tests with pruning of non-significant DIF effects), and also present in the best-fitting Reg-DIF model estimated here with the penalized EM algorithm. Yellow shading indicates parameters present only in the final MNLFA model estimated by Bauer (2017). Blue shading indicates parameters present only in the best-fitting Reg-DIF model estimated here. Finally, note that Reg-DIF with MCP was implemented with \( \gamma = 3 \).
Next, I compare Bauer’s (2017) likelihood ratio test findings with the current Reg-DIF findings for the violent behavior factor, shown in Table 7. Unlike before, I observed far less correspondence between the methods. In particular, only 2 out of 22 DIF effects were identified by both Reg-DIF (LASSO and MCP) and likelihood ratio tests (see green highlights in Table 7). From the simulation results in this dissertation, I would expect LASSO and MCP to show higher true positives with such a large sample size. This suggests that where Reg-DIF and Bauer identified DIF, namely in V1 and V2, is where DIF may likely exist in the population. For DIF effects identified by Reg-DIF (LASSO and MCP) but not by likelihood ratio tests (see blue highlight in Table 7), there was a single interaction effect between age and sex on V3, notably without its corresponding main effects. I reserve discussion on this odd finding for the paragraph below. Finally, there were 20 DIF effects across half of the violent behavior scale items that were identified by likelihood ratio tests but not by Reg-DIF (see yellow highlights in Table 7). If we expect Reg-DIF to have a high true positive rate with such a large sample size, one possible reason for this large discrepancy is that the majority of these DIF effects are false positives. Specifically, given that Bauer’s approach iteratively tests items using different sets of anchor items and prunes non-significant DIF effects in the final model-fitting stage, using likelihood ratio tests may lead to higher false positive rates. This has not been evaluated rigorously via simulations, however, so no firm conclusions can be made at this point. Another possible reason for this discrepancy is that the violent behavior scale may have different data characteristics than the simulation conditions examined here. This would imply that Reg-DIF may have a lower chance at recovering true DIF than what the simulation results would otherwise suggest for such a large sample size. For instance, there is greater sparsity among the item responses in the empirical data, which was not examined in the simulation study.
Table 7. Reg-DIF Parameter Estimates for Violent Behavior Factor Using BIC

<table>
<thead>
<tr>
<th>Reference Parameter</th>
<th>LASSO</th>
<th>MCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent Behavior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>-0.14</td>
<td>-0.15</td>
</tr>
<tr>
<td>Age(^2)</td>
<td>-0.27</td>
<td>-0.27</td>
</tr>
<tr>
<td>Male</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>Male × Age</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Male × Age(^2)</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>Age(^2)</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>Male</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Male × Age</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Male × Age(^2)</td>
<td>-0.08</td>
<td>-0.08</td>
</tr>
<tr>
<td>V1. Saw shoot/stab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.54</td>
<td>-2.54</td>
</tr>
<tr>
<td>Slope</td>
<td>1.35</td>
<td>1.37</td>
</tr>
<tr>
<td>V2. Pulled weapon on you</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.68</td>
<td>-2.70</td>
</tr>
<tr>
<td>Slope</td>
<td>1.70</td>
<td>1.71</td>
</tr>
<tr>
<td>V3. Shot you</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.43</td>
<td>-4.44</td>
</tr>
<tr>
<td>Slope</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>V4. Cut/stabbed you</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.98</td>
<td>-3.98</td>
</tr>
<tr>
<td>Slope</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>V5. Physical fight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.95</td>
<td>-0.95</td>
</tr>
<tr>
<td>Slope</td>
<td>2.08</td>
<td>2.07</td>
</tr>
<tr>
<td>V6. Were jumped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.87</td>
<td>-2.86</td>
</tr>
<tr>
<td>Slope</td>
<td>1.64</td>
<td>1.63</td>
</tr>
<tr>
<td>V7. You pulled weapon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.30</td>
<td>-4.30</td>
</tr>
<tr>
<td>Slope</td>
<td>1.85</td>
<td>1.84</td>
</tr>
<tr>
<td>V8. You shot/stabbed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.30</td>
<td>-5.30</td>
</tr>
<tr>
<td>Slope</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>D6. Hurt other badly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.07</td>
<td>-2.07</td>
</tr>
<tr>
<td>Slope</td>
<td>1.71</td>
<td>1.71</td>
</tr>
<tr>
<td>D14. Group fight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.80</td>
<td>-1.79</td>
</tr>
<tr>
<td>Slope</td>
<td>1.50</td>
<td>1.50</td>
</tr>
</tbody>
</table>
Note. Green shading indicates parameters present in the final MNLFA model estimated by Bauer (2017), who used likelihood ratio tests with pruning of non-significant DIF effects, and also present in the best-fitting Reg-DIF model estimated here with the penalized EM algorithm. Yellow shading indicates parameters present only in the final MNLFA model estimated by Bauer (2017). Blue shading indicates parameters present only in the best-fitting Reg-DIF model estimated here. Finally, note that Reg-DIF with MCP was implemented with $\gamma = 3$. 
One caveat to the empirical results is that some higher-level effects were included in the final model without their corresponding main effects for Reg-DIF. In the non-violent behavior factor (Table 6), for instance, the male × age interaction on D2 for MCP was included without the two main effects, and the male slope DIF effect (i.e., an interaction with the latent variable) was also included without the intercept DIF effect (i.e., main effect of male). In the violent behavior factor (Table 7), the male × age effect was included for both LASSO and MCP on V3 as well. One possible reason that these interactions and quadratic effects were included without their corresponding main effects was that they were not standardized themselves, implying that they could have had a larger or smaller weight in the penalization process. In turn, this could have contributed to a different pattern of results than if all of the predictors were standardized. Other penalty functions such as the hierarchical LASSO penalty prevent this from happening by only including the interaction effect when the main effects are also included. In the following discussion section, I consider this possible extension in more detail.

In sum, the empirical results show some degree of correspondence between Reg-DIF and Bauer’s (2017) likelihood ratio test approach. Nevertheless, there was also some clear inconsistencies in the findings. Namely, the non-violent delinquency factor exhibited more DIF effects when using Reg-DIF, whereas the violent factor exhibited more DIF effects when using Bauer’s likelihood ratio test approach. Because there has not been an extensive analysis of Bauer’s approach, few conclusions can be made about where DIF exists in either scale. Despite this limitation, the empirical example not only demonstrates a real-world application of the penalized EM algorithm, but also shows how to implement it in the regDIF R package. I turn now to discuss these results, as well as the simulation findings, in light of the current state of research.
CHAPTER 6: DISCUSSION

The main goals of this dissertation were to: (1) develop a more computationally efficient algorithm for evaluating DIF among many item responses and exogenous covariates; (2) incorporate soft-thresholding within parameter optimization, as opposed to ad hoc thresholding outside of the estimation process; (3) evaluate and compare the new penalized EM algorithm with previous implementations of regularized DIF and other penalty methods (i.e., firm-thresholding); and finally, (4) showcase the regDIF R package using an empirical example, thereby making the penalized EM algorithm accessible to other research scientists. Below, I summarize the progress made on each of these goals and discuss future directions for research. I end with a conclusion of the findings.

Developing an Efficient Algorithm

In previous work, our research group implemented a regularized approach for evaluating DIF which, despite outperforming other methods for identifying DIF, became computationally intractable when the number of items exceeded 12 and the number of covariates exceeded 3 (Bauer et al., 2020; Belzak & Bauer, 2020). A new estimation approach was thus sorely needed, particularly because applications of IRT models often involve scales with more than 12 items (e.g., achievement tests, clinical assessments, etc.). To address this computational limitation, I developed an algorithm that leveraged Bock and Aitken’s (1981) EM approach for estimating 2-parameter IRT models and expanded it to allow for multiple person-specific effects on the item responses and latent variable parameters. This novel, modified EM algorithm, although still computationally more expensive than Bock and Aitken’s EM approach due to the person-specific
covariates, is far more efficient for evaluating DIF among multiple covariates compared to our previous implementation of Reg-DIF via the observed marginal likelihood method. In the following, I describe how this approach was developed with respect to Bock and Aitken’s algorithm, primarily to emphasize the historical precedence for using an EM approach to improve estimation of Reg-DIF model parameters.

When initially estimating model parameters for Reg-DIF, our research group unknowingly encountered a problem that also vexed earlier developments of IRT estimation for 2-parameter models. In particular, Bock and Lieberman (1970) proposed a marginal likelihood approach to obtain slope and intercept parameters for a unidimensional factor model. Although this approach marked a significant step forward in IRT estimation, this method was largely impractical for more than a relatively small number of items (e.g., 5) because “the generation and inversion of a $2n \times 2n$ information matrix” (pp. 444) became too computationally slow, “even on the fastest computers, when $n$ exceeds 12” (where $n$ is the number of item responses). Our previous research observed the same problem when implementing regularization for DIF evaluation in the two-parameter IRT model (Bauer et al., 2020). Namely, having more than 12 items and 3 covariates on each item response resulted in waiting days for the model to converge. The run-time also increased exponentially when more items or covariates were included in the model. Therefore, for Reg-DIF to be practical for applied research use, there needed to be a more efficient approach for estimating many item parameters in 2-parameter models. Fortunately, a better approach had been developed for estimating 2-parameter IRT models which could be modified to implement Reg-DIF.

In particular, Bock and Aitken (1981) provided an elegant solution to the intractable 2-parameter estimation problem by reformulating the observed marginal likelihood function into a
missing data problem, where the latent variable parameters are treated as missing data and the item responses as observed data. The EM algorithm (Dempster, Laird, & Rubin, 1980) could then be used to simplify the estimation problem. That is, by taking the expected value of the complete data log-likelihood function (i.e., complete data includes both the observed and missing data) in the expectation step, one can get increasingly better estimates of the latent variable parameters in each iteration and, in effect, treat each item response as an independent regression problem in the maximization step. This reduced computation time greatly and allowed for many more items to be included in the analysis.

Following Bock and Aitken’s solution, I also use the EM algorithm to estimate Reg-DIF item parameters much more efficiently. Although the basic principle of taking the expected value of the complete data log-likelihood is the same, this modified EM algorithm includes exogenous predictors on both the latent variable scores (i.e., latent regression) and the item response probabilities (i.e., explanatory IRT). The main difference is that the log-likelihood function must first sum over observations, rather than response patterns (as is done with Bock and Aitken’s EM approach), before summing over quadrature points. Because the exogenous predictors vary randomly across observations, each observation has a unique likelihood value. In effect, there are more computations required to implement the penalized EM algorithm in comparison to Bock and Aitken’s EM algorithm. Despite this added computational burden, however, run times for the penalized EM algorithm are multiple times faster than the observed marginal likelihood approach. For example, in one of the simulated data sets that included 12 item responses, 3 exogenous covariates, and 2000 observations, running 100 values of tau (i.e., 100 fitted models) took approximately 141 hours using the observed marginal likelihood approach. This prevented us from rigorously evaluating Reg-DIF with these type of data (Bauer et al., 2020). In contrast,
this only took approximately 30 minutes using the penalized EM algorithm, or less than one half of 1% of the time that it took the observed likelihood approach.7

**Incorporating Soft-Thresholding for Penalization**

In addition to efficiently estimating Reg-DIF model parameters, I also aimed to integrate more principled thresholding into the new EM algorithm. In doing so, DIF parameters could then be penalized during optimization and, when necessary, contemporaneously removed from the model (i.e., set to zero) with a sufficiently large penalty. The second main goal of this dissertation was thus to implement soft-thresholding into the maximization step of the EM algorithm. This would not only align Reg-DIF with how regularization is typically implemented in the statistical sciences (Tibshirani, 1996), but would also allow for different penalty approaches (e.g., firm-thresholding) to be straightforwardly and properly included into the likelihood function. In the text below, I first summarize why soft-thresholding is important and how it, along with firm-thresholding, was implemented into the EM algorithm. I then discuss how soft- and firm-thresholding invariably fixed other issues that we encountered in previous work, pertaining especially to model under-identification.

First, to implement LASSO regularization, there needs to be some mechanism during the optimization of parameter estimates that can handle points of non-differentiability on the likelihood surface. These points of non-differentiability occur because there are absolute value operators in the LASSO penalty function, and by taking partial derivatives with respect to the

---

7 One caveat of this time comparison is that the observed marginal likelihood approach computes standard errors in the process of estimation, whereas the penalized EM algorithm does not with the regDIF R package. This implies that the computation time would take slightly longer with the penalized EM algorithm if standard errors were calculated using, for instance, the supplemental EM algorithm (Cai, 2008). I also used 21 fixed quadrature points for the penalized EM algorithm via regDIF in R, while the observed marginal likelihood approach via NLMIXED in SAS used 15 adaptive quadrature points. I found that using 21 fixed quadrature points with regDIF resulted in accurate recovery of parameter estimates to the second decimal place when compared to estimates obtained from SAS.
absolute-valued parameters, they become undefined when the estimates shrink to zero. As a result, this can impede maximum likelihood estimation if nothing is done to handle these points of non-differentiability. To overcome this issue in previous work, our research group developed a workaround which – despite being a non-standard approach for handling non-differentiability in the likelihood – appeared to perform well, insofar that regularization exhibited relatively good recovery of DIF (Bauer et al., 2020; Belzak & Bauer, 2020). Specifically, this nonconventional approach was necessary because of limitations in statistical software – namely, without access to the optimization code, we had to adapt a SAS estimation procedure (i.e., NLMIXED) to manually set penalized parameters (i.e., DIF effects) to zero when the corresponding estimates dropped below an arbitrarily small threshold value. Despite our empirical success with this approach, however, there is no theoretical support or precedent for performing regularization using this method. Therefore, in this dissertation I endeavored to use a more theoretically supported approach for handling non-differentiability with the newly modified EM algorithm.

Typically, handling non-differentiability with the LASSO penalty occurs via the soft-thresholding operator, which is simply a set of rules to determine the value of a parameter estimate in the presence of a tuning parameter. In the process of updating a parameter estimate using some root-finding approach (e.g., Newton-Raphson), that is, the soft-thresholding operator determines whether the tuning parameter (represented as $\tau$ in the current work) should be subtracted from the updated estimate (i.e., if the updated estimate is positive), whether it should be added to the updated estimate (i.e., if the updated estimate is negative), or whether the updated estimate should be set to zero (i.e., if $\tau$ is greater than the absolute value of the updated estimate). Therefore, in developing the modified EM algorithm for Reg-DIF, I also implemented the soft-thresholding operator with Reg-DIF to perform regularization in a more principled
manner. Furthermore, because I used coordinate descent optimization to update the parameter estimates, I also found that incorporating soft-thresholding into the M-step of the EM algorithm was a fairly straightforward process. In particular, I applied soft-thresholding operators to each individual DIF effect using the first and second partial derivatives of the likelihood function. It is worth noting that, had I updated parameter estimates in item-blocks as opposed to each individual parameter, soft-thresholding would need to be performed in tandem for each set of item estimates (i.e., all-at-once). It would still be relatively straightforward to implement soft-thresholding though, as the vector of updated parameter estimates would need to be compared to a single tuning parameter value only once. Regardless of whether the parameters are updated one-at-a-time or all-at-once for each item, the newly developed EM algorithm for Reg-DIF is well-suited for soft-thresholding.

In addition to soft-thresholding, I also implemented the minimax concave penalty (MCP) into Reg-DIF, that is, firm-thresholding. Like the soft-thresholding operator, firm-thresholding was straightforward to include with the EM algorithm mainly because it is just an extension of soft-thresholding. That is, the firm-thresholding operator scales the soft-thresholding operator using a second tuning parameter, represented as $\gamma$ in this work. More specifically, $\gamma$ controls how strong the penalty is in relation to the first tuning parameter $\tau$; for example, as $\tau$ increases, $\gamma$ reduces the overall effect of the penalty. In other words, there is a non-constant and non-linear amount of penalization of the (DIF) parameters that occurs when using firm-thresholding. With soft-thresholding, however, there is only a constant and linear amount of penalization that takes place. In essence, firm-thresholding typically yields less bias in the penalized parameter estimates compared to soft-thresholding, that is, assuming the model is properly specified.
Finally, by including both soft- and firm-thresholding into the penalized EM algorithm, I found model identification to no longer be a problem with Reg-DIF. In previous work, our research group had to use the Moore-Penrose inverse to estimate a saturated model, wherein all items were specified with DIF (Bauer et al., 2020). This avoided the model-identifying assumption that at least one item needed to be free of DIF. With soft- and firm-thresholding, however, we can now perform Reg-DIF in the “opposite direction” with $\tau$. This is to say that, rather than starting the Reg-DIF procedure with a fully-saturated model with $\tau = 0$, and then increasing $\tau$ until some DIF parameters are removed from the model, we can now specify a model with $\tau$ big enough so all DIF parameters are removed from the model, and then gradually reduce $\tau$ until DIF parameters enter the model. As the model nears full-saturation with a small enough $\tau$, the Reg-DIF procedure also automatically stops before the model becomes under-identified. Thus, because soft- and firm-thresholding are both implemented within the optimization process, there is far greater control at the beginning of the Reg-DIF procedure, which effectively solves the model (under)identification problem. This was not the case when using our previous ad hoc thresholding method with the observed marginal likelihood approach.

**Evaluating and Comparing Estimation Methods**

Beyond increasing the efficiency of parameter estimation in Reg-DIF, as well as penalizing DIF parameters according to sound statistical theory, I also developed the penalized EM algorithm with the notion that Reg-DIF would still outperform other iterative testing approaches in recovering DIF and thus perform similarly to the observed likelihood approach. In addition, it was important to examine whether the computational and theoretical improvements of the penalized EM algorithm actually translated into greater flexibility for Reg-DIF estimation. Therefore, the third main goal of this dissertation was to use Monte Carlo simulations to
determine (a) whether regularization remained effective for identifying DIF, and if so, whether the penalized EM algorithm could be a viable alternative to the observed marginal likelihood method, (b) whether there were any major differences between soft-thresholding and firm-thresholding in the recovery of DIF, and (c) whether the penalized EM algorithm could handle many more item responses and exogenous covariates given its computational advantages over the observed likelihood method. I highlight these findings next.

One of the first important findings of this dissertation was that the Monte Carlo simulation results here largely mirrored results in Bauer et al. (2020) in terms of outperforming likelihood ratio tests. As expected, I found that the penalized EM algorithm yielded fewer false positives (i.e., better control of Type I error) and fewer true positives (i.e., worse control of Type II error) on average relative to likelihood ratio tests. Reg-DIF’s superior performance occurred primarily in larger sample sizes and in scales with more DIF, as we also discovered in Bauer et al. This suggests that regularization remains effective in identifying DIF relative to iterative testing methods. It is important to point out, however, that we only compared Reg-DIF with one variant of using likelihood ratio tests in Bauer et al. – that is, we used a single sweep of likelihood ratio tests across the items, where each item was tested across the person-specific covariates only once while all other items were anchored. Therefore, it is possible that other DIF evaluation methods may perform better, such as iteratively using score tests or Wald tests, or using a different variant of likelihood ratio testing. As these iterative methods make untenable assumptions about the anchor items being free of DIF, however, it seems unlikely that they would outperform Reg-DIF in large sample sizes and greater amounts of scale DIF.

Furthermore, given that the results here were largely identical to our results in Bauer et al. (2020) in terms of iterative testing approaches, this also suggests that the penalized EM
approach may be able to replace the observed marginal likelihood approach in performing Reg-DIF without much loss of fidelity. In other words, researchers could take advantage of the more efficient penalized EM algorithm without losing the main benefits of Reg-DIF. On the other hand, there were some evident differences in the balance of true and false positives between the estimation approaches when performing Reg-DIF. In particular, the penalized EM algorithm tended to exhibit slightly more false positives than the observed likelihood approach, particularly in conditions where the amount of DIF was larger. Although this was offset by slightly higher true positives in some conditions, some concerning discrepancies in the FP results remained.

In an effort to explain the discrepancies between both estimation approaches, a few points must be considered first. For one, when the tuning parameter was large enough so that all DIF parameters were removed from the model, I showed that the penalized EM algorithm gave nearly identical parameter estimates as those given by the observed likelihood method. This implies that the differences in FP rates could not be due to faulty parameter estimation, at least when \( \tau \) is large and no DIF is included in the model. Instead, this points to differences in thresholding between the estimation methods. One possibility is that soft-thresholding was less specific (i.e., more likely to make Type I errors) to certain types of DIF compared to ad hoc thresholding. For instance, in examining the parameter-specific FP results more closely, I found that item intercepts were more prone to being misidentified as DIF (i.e., higher FPs) when using soft-thresholding as opposed to ad hoc thresholding,\(^8\) although this was offset with higher TPs. In contrast, soft-thresholding was less prone to misidentifying slope DIF effects (i.e., fewer FPs); this was similarly offset with fewer TPs. In consequence, these parameter-specific differences suggest that soft-thresholding may yield greater FPs overall when compared to ad hoc

---

\(^8\) This is shown in Figure 1A (Appendix C) and compared to results in Figure S1 from Bauer et al. (2020).
thresholding because intercept effects (i.e., main covariate effects) are often easier to identify with DIF, whether correctly or incorrectly, when compared to slope effects (i.e., interaction effects between covariates and the latent variable). Stated differently, it is plausible that the penalized EM algorithm showed more FPs at the item-level because it misidentified more intercept DIF than slope DIF, and therefore more DIF overall, compared to ad hoc thresholding. Of course, this is only one possible explanation for the discrepancies between the penalized EM algorithm and the observed likelihood approach.

A second important finding in this dissertation was that the MCP function (using firm-thresholding) appeared to be more conservative in the recovery of DIF compared to the LASSO penalty function (using soft-thresholding). Specifically, the MCP function exhibited fewer FPs and fewer TPs on average compared to LASSO. This was evident in examining the parameter-specific results (see Figures 1A-4A). That is, the MCP function showed greater specificity (i.e., fewer FPs), as well as less sensitivity (i.e., fewer TPs) in identifying intercept DIF compared to LASSO. Because intercept DIF is usually easier to identify (or misidentify), as discussed above, it is no surprise that MCP was more conservative overall. One caveat to this finding is that MCP showed greater estimator bias overall relative to LASSO when examining effects that were or were not present in the data-generating model. This was particularly the case in small sample sizes and scales with large amounts of DIF. As sample sizes and numbers of items increased, however, MCP showed less bias. Among effects that were correctly identified by both LASSO and MCP (i.e., true positives), MCP also showed less bias as sample sizes and number of items grew. This suggests that MCP may yield better results asymptotically. More specifically, MCP appeared to outperform LASSO uniformly as the number of items (and therefore parameters) grew larger. Whereas TP rates were roughly the same between LASSO and MCP in the 48- and
96-item conditions, FP rates hovered around the nominal rate for MCP only. In examining MCP more closely, I found that there was a “sweet spot” for specifying the second tuning parameter (i.e., $\gamma$), namely, where the greatest balance of true and false positives is achieved. At least from the Monte Carlo simulations conditions examined here, the “sweet spot” appears to be 3. One could also argue that $\gamma = 5$ performs well too. A more cautious conclusion which can be made from the MCP results is that researchers should decide what is more important for their DIF analysis – that is, having higher true positives or higher true negatives (i.e., lower false positives) – when choosing a $\gamma$ value to implement MCP with Reg-DIF.

Another important finding with the MCP function was that, in some instances, there were high rates of obtaining different model estimates when using different starting values. Certain data characteristics exacerbated this as well, such as having small sample sizes and large amounts of pervasive scale DIF. On one hand, this result is not particularly surprising given that the MCP function is known to be non-convex (Zhang, 2010). What is more surprising, on the other hand, is the magnitude of obtaining different results, reaching as high as 95% of models showing different estimates when comparing “cold” starting values from “warm” starting values. One major implication of this outcome is that different starting values will need to be provided when using the MCP function to ensure a global maximum is identified. If there are multiple maxima based on different starting values, one possible method for choosing the global maximum would be to compare the log-likelihood values from the different models and choose the model with the maximum value. In fact, I used this approach to determine whether the “warm” starting values or the alternative starting values (either “cold” or LASSO “random”) yielded better model fit when the estimates were different. Interestingly, this approach did not show that the “warm” starting values were a better fit than the “cold” or LASSO “random”
starting values in the majority of cases. To the contrary, in the small sample size and large DIF condition, nearly all of the “cold” starting values yielded a better model fit than the “warm” starting values (provided there was a divergence in final model estimates when using these two sets of starting values). Given these preliminary results, another possible approach could be that if different starting values yield different estimates, then the MCP method should not be trusted. I highlight some future directions related to the local solution problem below.

A final important finding here was that Reg-DIF can now be used to identify DIF when the number of items grows large, as well as when there are more background characteristics that may differentially affect the item responses and latent variable parameters. As mentioned throughout this dissertation, this was the main limitation of the observed marginal likelihood approach: that is, Reg-DIF was limited to approximately 12 items and 3 covariates due to computational inefficiencies. With the penalized EM algorithm, however, I have shown that Reg-DIF can now be performed with up to 96 items and 6 covariates, and probably with even more items and covariates as further improvements in computational efficiencies are made. In consequence, Reg-DIF can be reliably added to psychometric toolbox for both methodological and applied researchers. That is, methodological researchers can now evaluate Reg-DIF with many different data characteristics, and applied researchers can now utilize Reg-DIF across many different measurement scales. These possibilities are most evident by highlighting the regDIF R package that was concurrently developed here.

**Showcasing regDIF R package with Empirical Example**

The final main goal of this dissertation was to showcase the regDIF R package using a previously analyzed empirical example, while also using empirical data to directly compare Reg-DIF via the penalized EM algorithm against a variant of likelihood ratio testing via the observed
marginal likelihood approach. These data included two scales of adolescent delinquent behavior: one measuring non-violent delinquent behavior (8 Bernoulli-distributed items) and another measuring violent delinquent behavior (10 Bernoulli-distributed) items. Notably, our research group tried to evaluate Reg-DIF with these empirical data by using the observed marginal likelihood approach in previous work. It suffices to say that we were unsuccessful because the data included too many exogenous covariates (i.e., five features) and observations (i.e., $N > 4000$). In developing the more efficient penalized EM algorithm, however, I was able to evaluate Reg-DIF here and compare it to Bauer’s method of likelihood ratio testing. In the following lines, I thus highlight my empirical findings and summarize some of the features that are currently implemented in the regDIF R package.

The most noteworthy result from reanalyzing these data was that, in contrast to the simulation findings, I found more discrepancy than consistency in the identification of DIF when comparing the two approaches. For instance, both LASSO and MCP identified more items as exhibiting DIF in the non-violent behavior scale when compared to Bauer’s likelihood ratio test results, whereas LASSO and MCP identified fewer items as having DIF in the violent behavior scale. My expectation was that LASSO and MCP would identify fewer DIF effects across both scales given their superior control of Type I error over likelihood ratio tests. This obviously did not occur in the non-violent scale. Furthermore, although there were some items for which DIF was detected with both approaches – for instance, DIF was identified in Items D2, D3, D8, and D15 of the non-violent delinquency scale using both Reg-DIF and likelihood ratio tests – there were also many DIF effects detected by one and not the other approach. These discrepancies call for additional research, particularly on Bauer’s variant of likelihood ratio testing. That is, Bauer (2017) performed multiple sweeps of testing while progressively relaxing the anchor item set. He
also pruned DIF effects that were non-significant in the final model-fitting process. This was aimed at reducing Type I error rates that are common when using likelihood ratio tests to evaluate DIF. While our research group has evaluated a simpler method of using likelihood ratio tests to detect DIF in the MNLFA model (Bauer et al., 2020, i.e., a single sweep across the scale items), Bauer’s approach has not been rigorously studied with simulations yet. From this perspective, it is thus unclear whether Bauer’s approach outperformed or underperformed Reg-DIF. It should be emphasized, however, that the Monte Carlo results in this dissertation revealed that Reg-DIF (with firm-thresholding) had relatively good control of Type I error rates. This was especially pronounced in larger sample sizes, which is not much different than the empirical data analyzed here (i.e., $N > 4000$). It is thus possible that where Reg-DIF and the likelihood ratio tests differ in the results, Reg-DIF may be more likely to be correct. Of course, this conclusion is merely speculative, as we cannot know for sure where DIF occurs, or does not occur, with these empirical data.

In addition to directly comparing Reg-DIF with likelihood ratio testing using empirical data, I also provided readers a tutorial of the regDIF R package. Specifically, having developed regDIF to not only implement the penalized EM algorithm using coordinate descent and soft- and firm-thresholding, but also to run the simulations and empirical analyses, I also wanted to show how easy it is for other researchers to perform Reg-DIF with their own data. In particular, only two arguments need to be specified in the main regDIF R function: the `item.data` and the `pred.data` (i.e., predictor data). As part of this dissertation, I have also worked to get regDIF production ready so that the package is reliable to use and has the necessary features for most research applications. In particular, the regDIF R package features currently include the ability to model Bernoulli-distributed item responses, have continuous and categorical covariates, use
LASSO and MCP functions for penalization, specify 1-parameter and 2-parameter IRT models, automatically generate $\tau$ parameters, and automatically stop Reg-DIF to prevent under-identification (with small enough $\tau$), among other features.

Having described the progress made on each of the four main goals of this dissertation, I turn next to highlighting future directions for research on Reg-DIF and the penalized EM algorithm.

**Future Directions**

There are a variety of avenues that scientists can take to advance the development and evaluation of Reg-DIF, as well as the penalized EM algorithm more generally. I expand on some of the most important research directions below.

First, in light of the computational advantages of the penalized EM algorithm, even greater efficiencies could be made with future improvements of the algorithm. Because Reg-DIF includes not just covariates that are fixed across groups of observations but rather covariates that vary randomly over every observation, the EM computations must occur across large $N \times Q \times J$ arrays, where $N$ is the number observations, $Q$ is the number of quadrature points, and $J$ is the number of items. While there does not appear to be an easy way to reduce computations related to $N$ or $J$, the number of quadrature points $Q$ could be considerably reduced if a more precise calculation of the latent variable density is achieved. In this dissertation, I implemented fixed-point, rectangular quadrature and used 51 points to calculate the latent density. Although this approach was straightforward to implement, it is less effective when there are large differences in the latent means and variances between observations. In other words, when there is considerable impact on the latent variable parameters, many quadrature points are required to obtain accurate point estimates with a fixed-point quadrature approach. In contrast, our previous
implementation of Reg-DIF with the observed marginal likelihood approach only needed 15 quadrature points to get accurate point estimates because we used adaptive quadrature instead of rectangular quadrature (Bauer et al., 2020). Adaptive quadrature accounts for differences in the latent means and variances, and accordingly adjusts the quadrature points during model estimation. That is, it shifts the quadrature points right or left for different means, and expands or contracts the quadrature points for different variances. If adaptive quadrature can be implemented into the penalized EM algorithm, much greater efficiency could be easily achieved in estimating model parameters (Rabe-Hesketh, Skrondal, & Pickles, 2002). In fact, I am in the process of implementing adaptive quadrature into the regDIF R package.

Another computational improvement to the penalized EM algorithm could come in the form of calculating the partial derivatives more efficiently. The current approach uses coordinate descent and Newton-Raphson to update the parameter estimates, such that the first and second partial derivatives are calculated analytically for each parameter, and then each estimate is updated one-at-a-time (i.e., coordinate descent) by subtracting the previous estimate by the first partial derivative divided by the second partial derivative (i.e., Newton-Raphson). I implemented coordinate descent here because it has shown to perform well in logistic regression settings (Friedman et al., 2010), especially as the number of predictors in the model grows larger. One downside to coordinate descent, however, is that it needs to cycle through the parameters, again one-at-a-time, until the parameters converge to a stable solution. Additionally, coordinate descent requires re-calculation of the item response functions while optimizing each parameter. This is especially costly in the penalized EM algorithm because each observation requires a separate set of quadrature points due to the person-specific covariates; that is, the item response functions are $N \times Q$ matrices and must be calculated each time a parameter is updated. Given
that most applied researchers may not be evaluating large numbers of DIF predictors and that costly computations must be made when using coordinate descent, another optimization approach may be faster. An alternative would be to update the parameter estimates for each item all-at-once (i.e., block-wise) rather than one-at-a-time. Although this would require calculation of the cross-derivatives as well as an inversion of the Hessian, it could also lead to more efficient estimation compared to coordinate descent because computation of the gradient vector and Hessian matrix would need to occur only once for each item. Furthermore, the item response function would only need to be calculated once for every M-step. This would greatly cut down on the computations needed with the penalized EM algorithm.

Beyond computational advances, future research related to thresholding will also be important, particularly for including and evaluating other penalty functions with the Reg-DIF procedure. While this dissertation only examined the LASSO penalty and the MCP, it would be interesting for future work to examine whether other penalty functions have different tradeoffs in the recovery of DIF. For example, the elastic net combines the ridge and LASSO penalty functions using an additional parameter, typically denoted as $\alpha$. Increasing $\alpha$ can lead to a more LASSO-like penalty, whereas decreasing $\alpha$ can lead to a more ridge-like penalty. The main advantage of an elastic net penalty is that parameter selection (i.e., anchor item selection) can be achieved when there are highly correlated predictors (i.e., DIF covariates) in the model that may lead to indeterminate solutions. Future research thus could examine the benefits of using the elastic net penalty when the number of DIF covariates is large relative to the number of observations. Other penalty functions that were not examined here include the Smoothly Clipped Absolute Deviation (SCAD) penalty, which behaves similarly to MCP, and the hierarchical group LASSO function, which ensures that main effects are included in the model when
corresponding interaction effects are also included from the model due to regularization. In fact, this is one limitation of the current approach: namely, slope DIF could be included in the best-fitting DIF model without corresponding intercept DIF. This notably occurred in the empirical example. As such, incorporating the hierarchical group penalty will be critical to avoiding this problem. Another extension for Reg-DIF could be to apply different penalties to the DIF intercepts (i.e., main effects) and DIF slopes (i.e., interaction effects). Because slope DIF is typically more difficult to identify than intercept DIF, as interaction effects typically are, applying a smaller penalty to the slope DIF effects may result in better recovery. Of late, I am in the process of including the elastic net, SCAD, and hierarchical penalty functions into the regDIF R package, as well as different penalties to the intercepts and slopes.

In addition to examining different penalty approaches, it will also be critical for future research to examine the local solution problem when using the MCP function for DIF evaluation. As identified before, using different starting values can lead to the identification of local maxima when using the non-convex MCP function. Although I expected that using “warm” starting values, or estimates obtained from the previous tuning parameter value, would be the most reliable set of starting estimates for identifying a global maximum with the MCP function, I found the opposite result. Namely, in instances where different starting values led to different estimates, the “cold” starting values often yielded models with higher log-likelihoods and lower BICs (i.e., better model fit) when compared to models estimated with “warm” starting values. This perplexing result suggests that more empirical research is needed to determine the best approach for identifying the global maximum when using the MCP function. Other research in regression settings recommend a hybrid approach (Breheny & Huang, 2011), for instance, using model fit information and convexity diagnostic criteria to determine the global maximum.
Adapting this approach to IRT models and DIF evaluation in the presence of non-convex penalty functions like the MCP method will be a fruitful avenue for future research.

With respect to the selection of tuning parameters with Reg-DIF, future research could also examine alternative approaches to model selection. Following our research group’s previous work (Bauer et al., 2020; Belzak & Bauer, 2020), I used information criteria to identify the best-fitting model in this dissertation and found that BIC continues to work relatively well in identifying DIF. However, another approach for selecting the optimal degree of penalization would be to use $k$-fold cross-validation. In particular, other researchers have shown $k$-fold cross-validation to be a reliable method for model selection in DIF testing (Tutz & Schauerger, 2015). Model selection with the penalized EM algorithm could therefore be extended to cross-validation and then compared with using information criteria. One major downside to cross-validation, though, is that it is computationally intense to perform, whereas using information criteria is far less burdensome. In fact, this is the primary reason we have continued to use BIC and AIC in our evaluations of Reg-DIF. For cross-validation, the Reg-DIF procedure would need to be performed $k$ separate times when using $k$-fold cross validation. Although these analyses could be run independently, wherein researchers take advantage of multi-core computing resources, the penalized EM algorithm is already computationally heavy. Adding more layers of computational time may be unrealistic at this time.

In addition to the development of Reg-DIF procedures, there are a number of extensions that would allow researchers more opportunities to evaluate Reg-DIF. One straightforward extension would be to include different item response distributions into the Reg-DIF procedure, such as ordered categorical, continuous normal, and nominal item responses. This would require different partial derivatives to be calculated for each distribution, while also raising questions
about how DIF should be applied to the intercepts and slopes. With ordered categorical responses, for instance, DIF may be applied to \( C - 1 \) thresholds individually or to an overall intercept with \( C - 2 \) thresholds. Given that many item parameters already need to be estimated with Reg-DIF (i.e., intercept and slope effects for each covariate), the more plausible option would be to penalize the overall intercept of a categorical item response. This implies that all thresholds would exhibit the same amount of DIF in the same direction. This may or may not be a realistic assumption, but it would be a more parsimonious approach. Given that applications in psychology and education often utilize scales where the item responses are ordered categorical or continuous, it will be important for future research to evaluate the efficacy of identifying DIF using regularization with these different item responses. It is unknown whether we would see the same advantages of regularization over likelihood ratio tests when there are ordered categorical and/or continuous item responses. Note that I am currently implementing ordered categorical and Gaussian item responses into the regDIF R package. Thus, evaluating Reg-DIF with different item responses will soon be much easier to accomplish.

Another straightforward evaluation would be to compare Reg-DIF with other methods of identifying DIF, including Wald tests and score tests (i.e., modification indices). Given that modification indices have been used heavily in multiple-group confirmatory factor analysis models (MG-CFA; Steenkamp & Baumgartner, 1998) and have also seen some development for DIF testing in item response theory models (Wang, Strobl, Zeileis, & Merkle, 2018), it will be important to directly compare Reg-DIF with modification indices for DIF testing. This comparison may be better facilitated by including continuous item responses into the algorithm, as many applications of MG-CFA evaluate measurement invariance with continuous item responses. Beyond more classical methods like Wald tests and modification indices, other
approaches have been recently developed for evaluating DIF among multiple covariates and/or many groups. These methods include model-based recursive partitioning (Tutz & Berger, 2015; Strobl, Kopf, & Zeiles, 2015), latent class models (Cohen & Bolt, 2005), and alignment via optimization of a “similarity” function (Asparouhov & Muthen, 2014). As more researchers come to recognize the benefits of evaluating DIF among not just a single categorical covariate (or a few groups) but rather among multiple continuous and categorical covariates (or many groups), future research should compare these alternative approaches with Reg-DIF to determine both accuracy in the identification of DIF as well as computational efficiency.

In moving to future directions for the regDIF R package, an important development will be to include standard error calculations with the penalized EM algorithm. A possible reason why there were significant discrepancies between the penalized EM algorithm using LASSO and MCP and the observed likelihood approach using likelihood ratio tests is that we did not use standard error information to ultimately select DIF parameters in the best-fitting model with Reg-DIF. In examining the magnitude of the DIF effects with the empirical data more closely, Reg-DIF identified DIF where Bauer did not with effect sizes that were relatively small. This suggests that had standard error information been available with Reg-DIF, these effects may not have been significant. The main downside to using standard error information with the penalized EM algorithm is that getting standard errors are much more computationally difficult. This difficulty arises because the EM algorithm assumes more information about the estimation problem than is available, so one cannot simply use the diagonal of the Hessian matrix (evaluated at the maximum likelihood estimates) as can be done with, say, the observed marginal likelihood approach. Other standard error calculation approaches that consider the missing data are required, such as using a supplemental EM approach (Cai, 2008).
Lastly, an interesting direction for future development which does not necessarily concern the Reg-DIF procedure but rather the penalized EM algorithm more generally, is to develop other latent variable applications of the penalized EM algorithm. In particular, the penalized EM algorithm could be easily adapted to serve a number of model fitting and IRT testing purposes, such as applying a ridge or elastic net penalty function to the item intercepts and slopes of a latent variable measurement model when the number of item responses is larger than the number of observations (i.e., $N < P$). These data could arise in research contexts where it is far more costly to collect more observations than it is to collect more item responses. In these instances, estimating the model parameters would not otherwise be possible (due to under-identification) without a ridge-like penalty. It is worth mentioning that if the item responses and latent variable parameters were not regressed on exogenous covariates, penalized IRT models would be no more cumbersome to fit than any other IRT model that uses Bock and Aitken’s EM algorithm. Thus, the penalized EM algorithm could be adapted to other types of IRT models, such as shrinking the threshold-specific slope parameters to equality in Bock’s (1972) nominal response model, or shrinking cross-loading slopes in multidimensional IRT models to zero (see Sun et al., 2016). In other words, there are a variety of extensions for latent variable research in which the penalized EM algorithm may be beneficial for.

**Conclusion**

In summary, I have developed a new penalized EM algorithm in this dissertation to perform Reg-DIF (a) more efficiently – that is, with many more items and covariates – and (b) in a more principled manner – that is, with soft- and firm-thresholding. Through Monte Carlo simulation results and an empirical example, I have shown that Reg-DIF with the penalized EM algorithm remains an effective method for evaluating DIF compared to other iterative testing
approaches, especially when there are larger sample sizes and more DIF in a scale. In particular, the LASSO penalty function appears to yield more true positives in sacrifice of more false positives, whereas the MCP function is more conservative, exhibiting fewer false positives in sacrifice of fewer true positives. The simulation and empirical results also point to some discrepancies between the penalized EM approach and the observed marginal likelihood approach, both of which perform Reg-DIF. Namely, the penalized EM approach appears to be slightly more liberal (i.e., more FPs and TPs) than the observed likelihood method. Soft-thresholding with the EM algorithm may be a reason why. Thus, more research is required to determine the different tradeoffs between the estimation approaches, as well as between Reg-DIF and other DIF identification methods. Based on the results here, researchers should also determine whether false positives or false negatives (i.e., lower true positives) are more costly in the context of their research goals and choose the penalty function appropriately. The regDIF R package was also developed in parallel to this dissertation, making it far easier for applied researchers to implement Reg-DIF with their own measurement data. Although more computational development may need to be accomplished so that researchers can quickly fit many models with many tuning parameter values, it is now feasible to perform Reg-DIF when there are many more items and covariates. In contrast, there was no possibility that nearly 100 item responses and 6 covariates could be fit with the observed marginal likelihood approach. In short, the development of the penalized EM algorithm using coordinate descent optimization and soft-thresholding marks a critical step towards improving the evaluation of measurement bias among multiple covariates in latent variable measurement models.
APPENDIX A: CONCAVITY EVIDENCE

The log-likelihood function is defined as:

$$Q(y|y^{(t)}) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \log[\phi(Z_q|x_i; \xi^{(t)})]P(Z_q|y_i, x_i; y^{(t)})$$

$$+ \sum_{i=1}^{N} \sum_{q=1}^{Q} \sum_{j=1}^{J} \log[f(y_{ij}|Z_q, x_i; \alpha_j^{(t)})]P(Z_q|y_i, x_i; y^{(t)})$$

where \(\phi(Z_q|x_i; \xi^{(t)})\) is the normal density function for the latent variable that is governed by parameters \(\xi\), \(f(y_{ij}|Z_q, x_i; \alpha_j^{(t)})\) is the logistic likelihood function for item \(j\) that is governed by parameters \(\alpha_j\), and \(P(Z_q|y_i, x_i; y^{(t)})\) is the posterior distribution of the latent variable that is governed by parameters \(\gamma = \text{vec}(\xi, \alpha_j)\) for all \(J\) items. We want to maximize \(Q(y|y^{(t)})\) with respect to \(\gamma\).

To ensure the optimum \(\gamma^*\) is a global optimum rather than a local optimum, I need to establish that \(Q(y|y^{(t)})\) is concave for all values of \(\gamma\). It is important to note that if the additive components of a function are concave, then the function is also concave (Boyd & Vandenberghe, 2004). I may thus establish concavity separately for the normal density of the latent variable and for the item likelihoods.

I start with the latent variable component of \(Q(y|y^{(t)})\), written

$$Q_\theta(y_\theta|y_\theta^{(t)}) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \log[\phi(Z_q|x_i; \xi^{(t)})]P(Z_q|y_i, x_i; y^{(t)})$$

where \(Q_\theta(y_\theta|y_\theta^{(t)})\) represents the log-likelihood function with respect to the latent variable (\(\theta\)) parameters. One easy way to establish concavity is to check whether the second partial derivative of \(Q_\theta(y_\theta|y_\theta^{(t)})\) with respect to a parameter in \(y_\theta\) is equal to or less than 0 (non-positive) for all
values of the parameter in $\theta$ (Boyd & Vandenberghe, 2004). The second partial derivative with respect to a generic mean impact parameter is

$$\frac{\partial^2 Q(\theta | \theta^{(t)})}{\partial \zeta^2} = \sum_{i=1}^{N} \sum_{q=1}^{Q} - \frac{x_i^T x_i}{\psi} P(Z_q | \theta, x_i; \theta^{(t)}).$$

This second derivative is guaranteed to be non-positive for all values of $\zeta$ because $\frac{x_i^T x_i}{\psi} > 0$ (i.e., $\psi$ is always positive) and $P(Z_q | \theta, x_i; \theta^{(t)}) \geq 0$. Next, the second partial derivative with respect to a generic variance impact parameter is

$$\frac{\partial^2 Q(\theta | \theta^{(t)})}{\partial \zeta^2} = \sum_{i=1}^{N} \sum_{q=1}^{Q} - \frac{x_i^T x_i}{2\psi^2} P(Z_q | \theta, x_i; \theta^{(t)}).$$

This second derivative is also guaranteed to be non-positive for all values of $\zeta$ because $\frac{x_i^T x_i (Z_q - \alpha)^2}{2\psi^2} \geq 0$. Thus, I conclude that $Q(\theta | \theta^{(t)})$ is concave for all values of $\zeta$.

Next, the item likelihood component of $Q(\theta | \theta^{(t)})$ is written

$$Q_j(y_j | y_j^{(t)}) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \log[f(y_{ij} | Z_q, x_i; \theta_j^{(t)})] P(Z_q | \theta, x_i; y_j) P(Z_q | y_i, x_i; y_j^{(t)}),$$

where

$$f(y_{ij} | Z_q, x_i; \theta_j^{(t)}) = \mu_{ij}^{y_{ij}}[1 - \mu_{ij}]^{1-y_{ij}}$$

and

$$\mu_{ij} = \frac{1}{1 + \exp(-[(v_{0j} + x_i^T v_{1j}) + (\lambda_{0j} + x_i^T \lambda_{1j})\theta_i])}.$$
\[
\frac{\partial^2 Q_j(y_j, \zeta_j | y_j^{(t)}, \zeta_j^{(t)})}{\partial^2 \zeta_j} = \sum_{i=1}^{N} \sum_{q=1}^{Q} \left( \frac{\partial \eta_{ij}}{\partial \zeta_j} \right)^2 \mu_{ij} [1 - \mu_{ij}] P(Z_q | y_i, x_i; y_j^{(t)}).
\]

This second derivative is guaranteed to be non-positive because \( \mu_{ij} \geq 0 \) and \( P(Z_q | y_i, x_i; y_j^{(t)}) \geq 0 \). Thus, I conclude that \( Q_j(y_j | y_j^{(t)}) \) is concave for all values of \( \zeta_j \).

Finally, the LASSO penalty function is known to be convex in the penalized parameters (Tibshirani, 1996). Therefore, I have established that both \( Q(y | y^{(t)}) \) and \( Q(y | y^{(t)})_{\text{LASSO}} \) are concave for all values of \( y \). This is also supported by theoretical work from Bagnoli and Bergstrom (2005), who show that the log normal and log logistic PDFs and CDFs are concave in their parameters.
APPENDIX B: PARAMETER DERIVATIONS

To obtain the quadratic approximation of the expected value of the complete data log-likelihood function for item \( j \) (Equation 25), I provide analytical first- and second-order partial derivatives of the expected value of the complete data log-likelihood function for item \( j \) (Equation 19). I take the derivatives with respect to a generic item parameter \( \zeta_j \) holding all other parameters \( \mathbf{\Omega}_j \) at their current value. In the following, I first provide the mathematical steps taken to obtain the first and second partial derivatives, and then summarize these results by noting that my analytical derivatives are identical to numerically approximated derivatives.

Ignoring the latent distribution in Equation 19, the first partial derivative is written

\[
\frac{\partial Q_j(\mathbf{y}_j, \zeta_j | \mathbf{y}_j^{(t)}, \zeta_j^{(t)})}{\partial \zeta_j} = \frac{\partial}{\partial \zeta_j} \left[ \sum_{i=1}^{N} \sum_{q=1}^{Q} \log \left[ f(y_{ij} | Z_q, x_i; \mathbf{\Omega}_j^{(t)}) \right] P(Z_q | y_i, x_i; \mathbf{y}_j^{(t)}) \right],
\]

which, for Bernoulli-distributed item responses, expands to

\[
= \sum_{i=1}^{N} \sum_{q=1}^{Q} \left( y_{ij} \log \left[ P(y_{ij} = 1 | Z_q, x_i; \mathbf{\Omega}_j) \right] \right)
+ (1 - y_{ij}) \log \left[ \frac{1 - P(y_{ij} = 1 | Z_q, x_i; \mathbf{\Omega}_j)}{P(Z_q | y_i, x_i; \mathbf{y}_j^{(t)})} \right),
\]

where \( y_{ij} \) is an indicator variable for whether individual \( i \) endorsed item \( j \). Note that because the posterior probabilities \( P(Z_q | y_i, x_i; \mathbf{y}_j^{(t)}) \) are fixed and known from the E-step, the partial derivative of this density equals zero. Therefore, using the chain rule in Equation A3, I obtain
\[ \sum_{i=1}^{N} \sum_{q=1}^{Q} y_{ij} \frac{\partial \eta_{ij}}{\partial \zeta_j} \left[ 1 - P(y_{ij} = 1 | Z_q, x_i; \omega_j) \right] \]

\[ + \left( y_{ij} - 1 \right) \frac{\partial \eta_{ij}}{\partial \zeta_j} \left[ P(y_{ij} = 1 | Z_q, x_i; \omega_j) \right] P(Z_q | y_i, x_i; Y_j^{(t)}) , \]

\[ = \sum_{i=1}^{N} \sum_{q=1}^{Q} \frac{\partial \eta_{ij}}{\partial \zeta_j} \left[ y_{ij} - P(y_{ij} = 1 | Z_q, x_i; \omega_j) \right] P(Z_q | y_i, x_i; Y_j^{(t)}) . \] (A4)

In Equation A5, the partial derivative of the linear predictor \( \eta_{ij} = (v_{0j} + x_i v_{1j}) + (\lambda_{0j} + x_i \lambda_{1j}) \theta_i \) is easily computed. If \( \zeta_j = v_{0j} \), for instance, then \( \frac{\partial \eta_{ij}}{\partial \zeta_j} = 1 \). Equation A5 represents the solved first partial derivative of \( Q_j \) with respect to \( \zeta_j \).

The second partial derivative of \( Q_j (y_j, \zeta_j | Y_j^{(t)}, \zeta_j^{(t)}) \) with respect to \( \zeta_j \) is

\[ \frac{\partial^2 Q_j (y_j, \zeta_j | Y_j^{(t)}, \zeta_j^{(t)})}{\partial \zeta_j^2} \]

\[ = \sum_{i=1}^{N} \sum_{q=1}^{Q} \frac{\partial}{\partial \zeta_j} \left[ \frac{\partial \eta_{ij}}{\partial \zeta_j} \left[ y_{ij} - P(y_{ij} = 1 | Z_q, x_i; \omega_j) \right] \right] P(Z_q | y_i, x_i; Y_j^{(t)}) , \]

which is solved as

\[ = \sum_{i=1}^{N} \sum_{q=1}^{Q} \left[ \frac{\partial \eta_{ij}}{\partial \zeta_j} \right]^2 P(y_{ij} = 1 | Z_q, x_i; \omega_j) [1 - P(y_{ij} = 1 | Z_q, x_i; \omega_j)] \]

\[ \times P(Z_q | y_i, x_i; Y_j^{(t)}) . \] (A6)

I use Equations A5 (first derivative) and A7 (second derivative) to update each DIF parameter estimate at a time via soft-thresholding and each baseline parameter estimate at a time via Newton-Raphson.
APPENDIX C: PARAMETER-LEVEL RESULTS

The figures below include the true and false positives for the intercepts and slopes separately when using BIC for model selection. The “intercept only” findings are shown first, and the “slope only” findings next.
Figure 1A. LASSO and MCP False Positive Rates for Intercepts Only Using BIC

Note: MCP was implemented with $\gamma = 3$. 
Figure 2A. LASSO and MCP true positive rates for intercepts only using BIC.

Note: MCP was implemented with $\gamma = 3$. 
Figure 3A. LASSO and MCP false positive rates for slopes only using BIC.

Note: MCP was implemented with $\gamma = 3$. 
Figure 4A. LASSO and MCP true positive rates for slopes only using BIC.

Note: MCP was implemented with $\gamma = 3$. 
REFERENCES


