Comparison of machine learning methods with an underlying logistic regression model to detect differential item functioning in patient-reported outcome measures

by

Razieh Dorostimotlagh

A Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

of the University of Manitoba

in partial fulfilment of the requirements of the degree of

Master of Science

College of Community and Global Health

University of Manitoba

Winnipeg

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Dr. Lisa Lix, for her invaluable guidance, encouragement, and continuous support throughout my research and the writing of this thesis. Dr. Lix's expertise and insightful feedback greatly contributed to shaping this work.

I am also grateful to my committee members, Dr. Depeng Jiang and Dr. Yi Xiong, for their constructive comments and suggestions, which improved the quality of my research. I would also like to thank Dr. Ruth Ann Marrie from Dalhousie University for providing access to the data that made this research possible. I gratefully acknowledge the funding support from Dr. Lix.

Special thanks to Dr. Rob Balshaw and Barret Monchka for their support and guidance, which helped me better understand my research and broaden my perspective. I am also deeply appreciative of my colleagues and lab members at the George & Fay Yee Centre for Healthcare Innovation, whose collaboration and discussions enriched my academic experience.

Finally, I wish to thank my family and friends for their unwavering love, patience, and encouragement. Their support has been my source of strength throughout this journey.

Table of Contents

List of Tables

List of Figures

Abstract

Background: Patient-reported outcome measures (PROMs) are self-report instruments about health-related quality of life and well-being (i.e., physical and mental health). Assessing the validity of a PROM involves ensuring measurement invariance (MI), which confirms unbiased comparisons across groups. Differential item functioning (DIF), a form of measurement non-invariance, occurs when individuals with the same level of health status respond to an item differently. Many studies about DIF in PROMs focus on demographic characteristics (e.g., age), but other characteristics of individuals, such as the presence of comorbid health conditions, may also contribute to DIF. Machine learning (ML) methods may be advantageous to test for DIF across multiple covariates. The research purpose was to test for DIF using ML tree-based and penalized methods based on logistic regression (LR). The objectives were to 1) compare the performance of two ML methods to detect DIF on multiple covariates, and 2) test the association of demographic and clinical covariates with DIF.

Methods: DIF was tested using an item-focused tree (IFT) with an underlying LR model and a Least Absolute Shrinkage and Selection Operator (LASSO) regression model. For Objective 1, a simulation study was conducted in which data were generated under different analytical conditions by varying sample sizes, DIF effect magnitudes, and correlations among covariates. The performance of both the IFT and LASSO regression models was evaluated using Type I error and statistical power rates. For Objective 2, the association of the covariates with DIF was assessed in the 36-item Short Form Survey items completed by individuals diagnosed with immune-mediated inflammatory diseases. A hold-out bootstrap cross-validation technique was conducted to evaluate the performance of the IFT and LASSO regression models using the Brier score, accuracy, and mean square error (MSE) in test data.

Results: In the simulation study, the Type I error rate for the IFT regression model was below the nominal 5% level across simulation conditions. The LASSO regression model had lower Type I error rates than the IFT model when the correlation between covariates was strong. In conditions with small and moderate DIF effect sizes, the LASSO regression model had greater statistical power than the IFT regression model. In real-world data, the IFT regression model detected eight DIF items, and the LASSO regression model detected 16 DIF items. Sex and hypertension were the most frequent covariates associated with DIF. When both methods flagged

an item for DIF, they identified similar associated covariates. The estimated Brier score, accuracy, and MSE were similar for both methods.

Conclusions: Establishing MI in a PROM means that the latent construct is equivalent across population groups defined by socio-demographic and personal health characteristics. MI contributes to equity-focused PROM development. The IFT regression model is recommended to control the rate of false positives, where DIF is incorrectly detected. The LASSO regression model is recommended when DIF effects are small.

Chapter I: Introduction

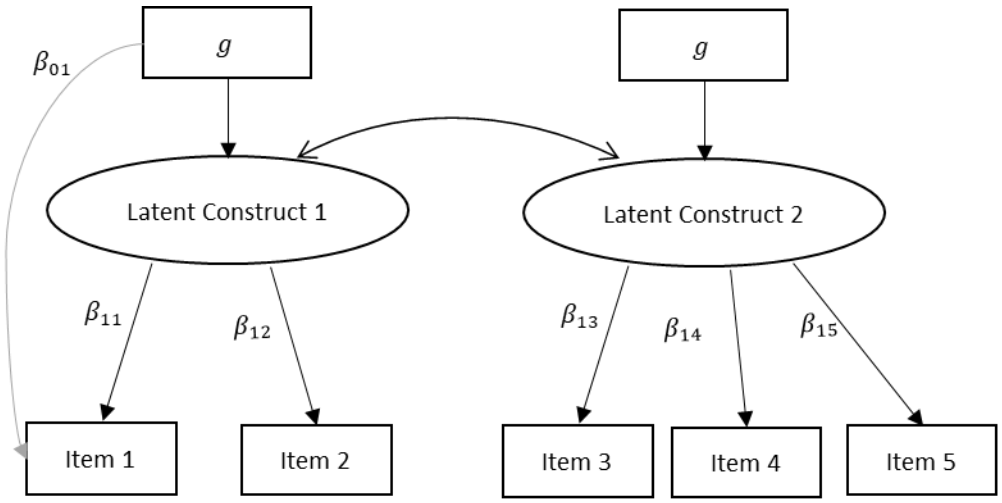
1.1. Background

Patient-reported outcome measures (PROMs) are self-report instruments that capture a patient's perspectives on their health-related quality of life (QoL) and well-being (1). PROMs measure latent (i.e., unobserved) constructs, such as physical, mental, and social health (2). Most PROMs contain multiple item(s) and are either disease-specific or generic instruments. The former assesses specific health conditions, while the latter are intended to be used in a variety of populations (3). The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), UCLA Prostate Cancer Index, and Health Assessment Questionnaire Disability Index are examples of disease-specific PROMs. The Short Form 36 Health Survey (the SF-36), EuroQol-5 Dimension (EQ-5D), and Hospital Anxiety and Depression Scale (HADS) are examples of generic PROMs (3).

Evaluation of the validity of a PROM can support its use in clinical or population-based studies (4). The validity of a PROM relies, in part, on the presence of measurement invariance (MI) to ensure fair comparisons across groups (5). A PROM exhibits MI if patients' interpretations and understanding of the items that comprise the PROM are equivalent (i.e., invariant) across patient subgroups (e.g., age groups) (4). If MI is a tenable assumption, item responses will be associated with the level of the measured latent construct. If MI is not a tenable assumption (i.e., measurement non-invariance), item responses will be influenced by other factor(s) in addition to the level of the latent construct (6). Differential item functioning (DIF) is a form of measurement non-invariance that occurs when patients with the same level of health status (i.e., the latent construct) respond differently to the PROM item(s) (7).

For example, suppose that a multi-item instrument comprised of binary items is represented by two latent constructs (i.e., unobserved domains). Furthermore, suppose items 1 and 2 are intended to measure one latent construct and items 3 to 5 are intended to measure the second latent construct.

Figure 1-1. Modeling measurement non-invariance.



In Figure 1-1, $\beta_{1i}, i \in \{1, \ldots, 5\}$ are item discrimination parameters indicating how well an item can distinguish between individuals with different levels of the latent construct. The parameter β_{01} represents the direct effect of subgroup g (e.g., males) on the item difficulty parameter for the first item. This parameter is used to quantify how difficult that subgroup finds choosing one response option over the other. A statistically significant value for β_{01} (i.e., rejecting H_0: $\beta_{01} = 0$) indicates a violation of the assumption of MI and results in DIF for item one caused by subgroup g. Note that in Figure 1-1, subgroup g is not associated with DIF in items two to five. Also, more than one subgroup (e.g., age and sex subgroups) can be associated with DIF in an item (8).

DIF analyses have been used in QoL studies to examine potential measurement biases in population subgroups defined by socio-demographic characteristics such as age, sex, income level, and race or ethnicity (9,10). DIF may also be affected by clinical characteristics, such as health status and the presence of various health conditions (i.e., comorbid health conditions) (11,12).

Methods to test for DIF, including methods based on logistic regression (LR), item response theory (IRT) and the Mantel-Haenszel (M-H) method, often test for DIF on one covariate at a time, which limits the ability to test for potential interactions and address confounding bias in observational studies (13–16). Incorporating multiple covariates in DIF analysis using these methods can be challenging. For example, IRT models that contain multiple

covariates require large sample sizes to ensure the stability of the large number of parameter estimates. These models are computationally expensive, and there is the potential for model non-convergence (17).

Consideration of multiple covariates in a study raises an interest in identifying the important covariates, improving the model accuracy, and making decisions about which covariates should be included in future studies (18). Note that DIF is specific to the covariate being tested. An item can exhibit a DIF effect for one covariate (e.g., sex) but not another one (e.g., age). By focusing on a single covariate at a time, assessing the importance of the covariates lies in assessing the estimated effect sizes (e.g. odds ratio) for individual covariates (13,14,19). Many of these methods require continuous covariates to be converted into categories; adopting arbitrary cutoff points may not necessarily align with the behavior of the covariates (13,20). For example, for a continuous covariate like age, researchers might split it into groups using the median age value to create two subgroups for the analysis, which might mask the presence of DIF effects (13). Moreover, the IRT and M-H methods are sensitive to deviations from the underlying model assumptions, which may not be tenable in applied analyses. For example, the IRT model assumes that the items in the dataset measure a single latent construct (i.e., unidimensionality), responses for each item are statistically independent of responses for all other items conditional on the latent variable (i.e., local independence), and the probability of choosing higher category of the item response increases as the latent construct increases (i.e., monotonicity) (21).

Including more than one covariate in the LR is possible. But most previous studies, such as the study by Crane et al. (22), used LR to test for DIF on one covariate at a time. When many covariates are tested for DIF, unstable estimates of DIF effects and inflated Type I error rates can occur. Covariate selection approaches, such as stepwise LR, can reduce the number of covariates (23). The LR model assumes a linear relationship between the log odds of the outcome and the values of the covariate (23). If the data do not satisfy derivational assumptions, this might result in biased estimates of model parameters and their confidence intervals (CIs) (24).

Machine learning (ML) methods such as regularized regression and decision trees have been proposed to test for DIF on multiple covariates and identify important covariates (13,25–

31). ML methods are a subset of artificial intelligence that automates model creation by learning from data, detecting patterns with minimal human involvement (16,32–34). Potential advantages of ML methods in DIF analysis include the ability to explore complex relationships (i.e., non-linear relationships), employ large sets of potentially correlated covariates, and rely less on statistical assumptions that may affect the validity of results (16,23,35). Moreover, ML methods can aid researchers in selecting important covariates from amongst a large set of covariates when testing for DIF. Thus, ML methods can help to reduce the number of covariates, resulting in a more parsimonious model that is less complicated to fit to a set of data (23).

Among ML methods, LR-based methods tend to be accurate in detecting DIF even with relatively small sample sizes compared to other methods, such as IRT-based methods that are sensitive to the sample size. This limitation of IRT-based methods becomes challenging when many covariates are involved, and the distribution of observations is highly unbalanced (23,24,36). Also, LR-based methods can incorporate different types of covariates, allowing continuous covariates to be included in the model without needing to convert them into categorical covariates (37–39). LR-based ML methods use the sum score obtained from a PROM's items as a realization of the latent construct (i.e., a proxy for measuring the level of health status) (31). In LR-based methods, the relative importance of covariates can be assessed based on likelihood-based statistics (e.g., the likelihood ratio test) to compare models with different sets of covariates and determine which model best fits the data. Once the best model is identified, the statistical significance of the covariates can be assessed based on test statistics. Another approach is bootstrapping from the original data and assessing the frequency of selected covariates or using quantile CIs (16,39–41).

1.2. Purpose and Objectives

The purpose of this research was to test for DIF using two LR-based ML methods, a penalized method and a tree-based method. The objectives were to:

1. Compare the performance of LR-based ML methods in detecting DIF on multiple covariates.
2. Test the association of demographic and clinical covariates with DIF.

The objectives addressed the following research questions:

1. Do LR-based ML methods perform similarly in controlling the Type I error rate and maximizing statistical power to detect DIF on multiple covariates in simulated data?
2. Are clinical and demographic covariates associated with DIF? If both types of covariates are associated with DIF, which covariates are most important?

1.3. Research Hypotheses

1. I hypothesize that the IFT regression model will control the Type I error rate to α.
2. I hypothesize that the LASSO regression model will have higher power than the IFT regression model in small sample sizes.
3. I hypothesize that demographic and clinical conditions will be associated with DIF, consistent with previous research.
4. I hypothesize that the LASSO regression model will identify more items and covariates associated with DIF compared to the IFT regression model when a large set of demographic and clinical conditions is tested for their association with DIF.

Chapter II: Literature Review

This literature review encompasses the following topics: 1) Prior research about DIF assessment, 2) type of covariates associated with DIF, and 3) Recent research about testing for DIF on multiple covariates in PROMs. The section concludes with a summary of the literature review and areas for future research.

2.1. Prior Research about DIF Assessment

Assessing DIF was considered first in educational settings in the 1960s, to study cultural differences in test performance between Black and Hispanic students and White students to detect test items biased against minority students (9,42). In mental health studies conducted in the 1960s and 1970s, identifying the reason for different prevalence rates among countries was of interest. Consequently, DIF detection methods were developed as a major quantitative assessment tool to measure whether the differences were due to the lack of MI (43).

In QoL research, assessing DIF has been considered a crucial step to confirm that the items comprising a multi-item PROM performed consistently across various subgroups of patients with the same level of health status (3,9). DIF effects can be categorized into two types: uniform and non-uniform DIF. Uniform DIF occurs when the likelihood of choosing a specific response category for an item is consistently different across different groups but follows a similar pattern for all response categories. Non-uniform DIF occurs when the effect of a group covariate on response probabilities varies depending on the response category, resulting in a non-parallel shift in response patterns amongst groups (7,9,43).

Early DIF analyses in QoL studies focused on testing for DIF on a single covariate at a time. Crane et al. (6) in their DIF assessment on the Cognitive Assessment Screening Instrument (CASI) recommended use of LR methods due to their ability to accommodate dichotomous, multinomial, and continuous covariates (i.e. age and years of education). They highlighted the limitations of M-H and IRT methods, particularly the need to categorize continuous covariates into two or more groups. Also, they pointed out that as a rule of thumb, IRT methods require a minimum of 250 individuals per group to obtain stable parameter estimates; this minimum sample size can be challenging to achieve for some covariates, such as ethnicity (6).

In the last decade, testing for DIF on multiple covariates became of interest in QoL research studies. A key challenge when several covariates are associated with DIF is identifying the most important covariate(s). Previously, researchers compared results from separate models for each covariate. One major issue was that multiple tests are conducted separately for individual covariates, which results in an inflated familywise Type I error rate (6,9), the overall probability of a Type I error amongst the set of tests. Bonferroni and Benjamini-Hochberg procedures were employed to adjust the observed p-values, addressing the issue of multiple comparisons; however, this approach becomes impractical when dealing with a large number of PROM items and covariates (6,32,33). Moreover, when covariates are correlated, comparing the results from separate models can lead to misleading conclusions (12,15,44).

Theoretically, IRT and LR models can incorporate multiple covariates. Stroble et al. (25,45) in their study about regression trees to test for DIF on multiple covariates, mentioned that applying IRT and LR methods may lead to sparse cells, requiring the exclusion of interaction effects (i.e., non-uniform DIF) from the models. Also, they stated that in psychological research, the underlying assumption of linearity in IRT and LR models is often violated. The authors recommended using ML methods such as decision tree models to overcome these limitations (25,45).

2.2. Type of Covariates Associated with DIF

Initial studies of DIF analyses in QoL often focused on testing for DIF on a single demographic characteristic only, such as sex/gender, age, ethnicity, education level, employment status, or type of job (7). But more recently, some studies have tested the association of clinical characteristics, such as comorbid health conditions, with DIF.

One of the studies that highlighted the association of health conditions with DIF was Dallmeijer et al. (46), who compared outcomes of the Functional Independence Measure (FIM^{TM}) in different patient groups with stroke, multiple sclerosis, and brain injury. Dallmeijer et al. (47) in another study found evidence of uniform DIF in the physical functioning domain of the SF-36 regarding stroke, amyotrophic lateral sclerosis, and multiple sclerosis. The evidence of uniform DIF in patients with neurological disorders, such as multiple sclerosis, demonstrated that adjustment for DIF may be required when comparing data of patients with neurological disorders

(46,47). According to Steultjens et al. (48), WOMAC items exhibited uniform and non-uniform DIF among patients with osteoarthritis, late-onset sequelae of poliomyelitis, and Parkinson's disease. Waller et al. (49) tested for DIF in the Beck Depression Inventory-II items and found that women with breast cancer and depression responded to the items differently. Smith et al. (50) in their study of testing for DIF on the EQ-5D items found that, in addition to sex and age, cancer type was associated with DIF. Pollard et al. (51) tested for DIF on the SF-36 items in an osteoarthritis population for socio-demographic characteristics (i.e., sex, age, social deprivation, social class, employment status), clinical and psychological characteristics (i.e., mood, Body Mass Index [BMI], the number of affected osteoarthritis joints and type of osteoarthritis). They found that age, employment status, social class, mood, type of osteoarthritis, social deprivation and BMI were associated with DIF.

2.3. Recent Research about Testing for DIF on Multiple Covariates in PROMs

Various methods have been developed to test for DIF in PROMs, focusing on multiple covariates and addressing the limitations of existing models, such as M-H and IRT models. These methods include structural equation modeling (SEM), which is a statistical method that models the relationship between the covariates and underlying latent constructs using a combination of Confirmatory Factor Analysis (CFA) and multiple LR models. Other methods include penalized regression methods such as Least Absolute Shrinkage and Selection Operator (LASSO)—which adds a penalty term to the absolute values of LR coefficients and shrinks the low effects to zero, Ridge— which adds a penalty term to the squared values of LR coefficients and shrinks all effects, and the Elastic Net—which combines LASSO and Ridge penalties. Also, tree-based methods such as item-focused trees (IFT) and Rasch trees (i.e., an IRT model that assumes that the probability of a correct response to an item is a logistic function of the difference between the person's latent construct and the item's difficulty) have been proposed to test for DIF in QoL research.

Lix et al. (44) tested for uniform DIF using the multiple indicators multiple causes (MIMIC) model, which is an SEM method, for the items in the mental health (MH) and physical functioning (PF) domains of the SF-36. They tested for DIF on the following covariates: sex, age group, body weight status (e.g., BMI), and self-perceived general health. They identified large

DIF effects for sex, age, and body weight status and recommended testing for DIF in population-based studies to confirm consistency of the SF-36 across subgroups. They did not investigate non-uniform DIF because interaction terms in the MIMIC model can lead to inflated Type I error rates. Their study only tested for DIF in two domains of the SF-36 because the MIMIC model is not appropriate to test for DIF on fewer than five items. They examined other domains using different methods, including LR. Although age and sex were the most common covariates in DIF analyses, they demonstrated that clinical covariates should be considered to improve the analysis.

Yadegari et al. (12) tested for uniform DIF using the MIMIC model for the items in the MH and PF domains of the SF-12 across multiple demographic and health status characteristics, including sex, age group, body weight status, and multi-morbidity (i.e., the presence of two or more chronic conditions) for patients having joint replacement surgery. They found that multi-morbidity had the largest contribution to the DIF model for the PF sub-scales in relative importance analysis, and older people reported more difficulties than younger individuals on the PF sub-scale. They highlighted the importance of the generalizability of their findings in terms of the measurement model fit to the data and occurrence of DIF, exploring multi-morbidity conditions in DIF analysis in other joint replacement populations.

Schauberger et al. (15) introduced a regularization approach named generalized partial credit model with LASSO (i.e., GPCMLASSO) approach in the presence of multiple, potentially correlated covariates. This approach first uses the Generalized Partial Credit Model (GPCM)—which is an IRT model that allows items to have different discrimination parameters—to model polytomous items and covariates. A LASSO penalty shrinks most of the covariate effects to zero. If at least one covariate is not shrunk to zero, the item is considered a DIF item, and the non-zero covariate(s) are associated with DIF in that item. The authors evaluated their method to test for uniform DIF, using simulated data sets with one to five covariates in conditions of no correlation and strong correlation among covariates. They suggested that GPCMLASSO is suitable in complex data situations such as multiple correlated covariates; they applied their method to the Child Depression Inventory (CDI) questionnaire, considering age, gender, race, and education. Their results showed that GPCMLASSO detected fewer DIF items compared to other methods, such as the simple LR model, due to its capability to avoid duplicated DIF effects from

correlated covariates (i.e., race and education). Their study incorporated both a simulation study and analysis of real-world data.

Ebrahimi et al. (14) used the Elastic Net regularized ordinal LR model as an alternative to the IRT model in small sample sizes. They focused on detecting uniform DIF in data from children with attention-deficit/hyperactivity disorder (ADHD) and their parents with respect to age and sex on the Persian version of the PedsQL™ 4.0 Generic Core Scales. They recommended using the Elastic Net regularized ordinal LR model when dealing with a small number of items (e.g., five or fewer) and a limited sample size (i.e., less than 150). However, in the application of the methods to real-world data, the results did not indicate a significant advantage over the non-regularized ordinal LR model. They highlighted that DIF analysis should be examined across several highly correlated continuous and categorical covariates.

Jafari et al. (52) explored uniform DIF at the item-level in a sample of children with and without ADHD, adjusting for age and sex. The authors assessed DIF in KINDLR- Quality of Life Questionnaire in Children and Adolescents along with their parents across the two groups using GPCMLASSO. They concluded that both children and parents have the same understanding of almost all items, indicating that the observed differences in KINDL scores between children and parents with and without ADHD were not likely to be due to measurement invariance and DIF. They recommended using GPCMLASSO in small sample sizes and with multicollinear covariates. They acknowledged that a limitation of their study was the lack of adjustment for comorbid conditions in their analysis, despite previous studies demonstrating the significant impact of these conditions.

Berrío et al. (20) identified DIF in the WHO Disability Assessment Schedule (WHODAS 2.0) amongst people with schizophrenia. Characteristics of individuals that were associated with DIF included socio-demographic characteristics, including gender, age, marital status, and education, as well as clinical covariates, including depression symptoms and symptoms of schizophrenia. They applied two procedures based on Rasch trees: TREE-PCM (i.e., Rasch tree for polytomous items) and PCM-IFT (i.e., item-focused tree for polytomous items). One limitation of the TREE-PCM procedure is that it does not identify the items responsible for DIF, only whether it is present in the data. Instead, its emphasis lies in identifying the global DIF

effect and the covariates that may contribute to DIF. PCM-IFT is a model-based method that also uses recursive partitioning to detect DIF. The authors confirmed the validity of the WHODAS 2.0 and indicated DIF was associated with age for just one of the 36 items in this PROM.

2.4. Summary of the Literature Review and Areas for Future Research

Over the past 60 years, multiple studies have aimed to detect DIF in PROMs. Many of these studies considered only one covariate at a time, focusing on socio-demographic characteristics such as sex or gender, age, ethnicity or race, and education level. Clinical covariates such as BMI and health conditions, such as cancer type, as well as comorbid conditions, found to be associated with DIF in PROMs in several studies (12,44,50).

Researchers have broadened the scope of DIF analyses to examine the importance of individual covariates when multiple covariates are associated with DIF. Fitting multiple models can lead to Type I error rate inflation due to multiple testing and may be challenging when the covariates are correlated due to multicollinearity effects (15). M-H and IRT-based models require categorizing continuous covariates before analysis, potentially masking DIF effects. Using IRT and LR methods to test for DIF on multiple covariates might be inefficient due to large sample requirements. As well, derivational assumptions, such as the assumption of linearity, must be satisfied for these statistical models if they are to produce valid conclusions about DIF; derivational assumptions may not always be satisfied in health-related studies (15). Statistical and ML methods such as SEM, regularized LR, and decision trees have been developed and may be able to address these challenges when testing for DIF in PROMs data on multiple covariates.

The literature highlights that testing for DIF requires considering both demographic and clinical covariates, and DIF analysis should be routinely conducted to ensure MI in PROMs before making clinical decisions. Previous studies that tested for DIF have used SEM, MIMIC models, tree-based methods, or regularized LR approaches. Therefore, comparing the performance of these methods would help to identify the optimal approach to test for DIF in practical settings. LR ML-based models were investigated in this study because they are straightforward to implement, accurate in detecting DIF in relatively small sample sizes, and provide parameter estimates that indicate the direction of DIF effects (and magnitude of DIF

effects in tree-based models). To compare methods within the LR framework, I compared a tree-based method (item-focused trees) with a regularized regression approach.

Among regularization methods, the LASSO regression model was selected because it performs both coefficient shrinkage and covariate selection by forcing some coefficients exactly to zero. This facilitates identification of covariates associated with DIF. Other regularization methods, such as ridge or Elastic Net, do not yield coefficients with zero values and require additional thresholding rules or tuning parameters for covariate selection, which can complicate their implementation.

Prior analyses on PROMs within chronic disease populations ensure a stable baseline for DIF that makes it an ideal candidate for methodological comparison. Using an already validated PROM allows the focus to be placed on comparing the performance of DIF detection methods, along with testing for DIF in a real-world setting.

Chapter III: Materials and Methods

This chapter describes the study design, data sources, and methods to address the research objectives. For Objective 1, I conducted a simulation study where the DIF items and the covariates associated with DIF were known. For Objective 2, DIF analyses were conducted using real-world data. I tested for uniform DIF and assessed the association of demographic and clinical covariates with DIF on the items comprising the SF-36. I used LASSO and IFT models to test for DIF on multiple covariates (27,28,31).

3.1. DIF Models

The underlying LR model for IFT and LASSO regression models to test for uniform DIF on item i, $i \in \{1, 2, \dots, I\}$ is:

$$Logit\left(P(Y_{pi} = 1)\right) = \log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \eta_{pi} = \beta_{0i} + \beta_{1i}x_p + \beta_{2i}g_p\,, \quad (1)$$

where $Y_{pi} \in \{0, 1\}$ is the response of person p for $p \in \{1, 2, \dots, P\}$, $Logit\left(P\left(Y_{pi} = 1\right)\right)$ is the logit of the probability of choosing an item response category with a value of one (i.e., the higher/healthier response option), β_{0i} is the item difficulty parameter for the i^{th} item, and β_{1i} is the item discrimination parameter for the i^{th} item, x_p is the sum of the item responses for person p, which is a proxy measure for the latent construct, β_{2i} is the DIF parameter, and g_p is a binary group membership covariate (31).

One approach to test for uniform DIF on the i^{th} item is to test the null hypothesis: $H_{0i}: \beta_{2i} = 0$ using a likelihood ratio test. The likelihood ratio test compares the change in the likelihood ratio for the model defined in equation 1, with a model without the group membership term (31). Rejection of H_{0i} indicates the item difficulty differs for the focal and reference groups (i.e., presence of DIF in item i). The g_p in equation 1 can be extended to a vector of categorical or continuous covariates. Therefore, β_{2i} is a vector of coefficients. Testing the statistical significance of the interaction of β_{1i} and β_{2i} assesses the presence of non-uniform DIF.

3.1.1. IFT Model

Constructing a decision tree involves iteratively identifying partitions of the covariate space. Each node within the tree represents a distinct subset of the covariate space. In the IFT model with an underlying LR model, a separate LR model with its own parameters is defined in each node. Uniform DIF can be tested by building a partition of the respondents with differing intercepts, given the covariates g_p. To structure a tree, all the DIF parameters are first estimated under H_{0i}. For each potential split on a covariate, the algorithm calculates a test statistic based on a likelihood ratio test, comparing models before and after the split. To assess whether this observed split is statistically significant, the algorithm constructs a null distribution by repeatedly permuting the covariate values (e.g., 1000 times) while keeping the item responses fixed to remove the association of covariates with the item responses (27). For each permutation, the likelihood ratio test statistic is recalculated. The p-value is then computed as the proportion of permuted test statistics that are greater than or equal to the observed one. This nonparametric procedure helps control the Type I error rate by evaluating how likely the observed split is under the assumption of H_{0i}. To avoid inflating the Type I error rate due to incorporating multiple items and covariates into the model, the model automatically does a Bonferroni correction, dividing the nominal significance level (α = 0.05) by the number of covariates (25,26).

The tree grows recursively using only statistically significant splits. The parent node A is divided into two subsets in the form of $A_1 = A \cap \{g_k \leq c\},\ A_2 = A \cap \{g_k > c\}$, based on the split point c on the covariate g_k. The first split for the k^{th} covariate and split point c_k yields:

$$\eta_{pi} = \beta_{1i} x_p + \left[\beta_{2il} I\left(g_{pj} \leq c_k\right) + \beta_{2ir} I\left(g_{pj} > c_k\right)\right], \qquad (2)$$

where $I(\cdot)$ denotes the indicator function with $I(a) = 1$ if a is true and $I(a) = 0$ otherwise. The parameter γ_{il} denotes the intercept in the left node $\left(g_{pk} \leq c_k\right)$ and γ_{ir} the intercept in the right node $\left(g_{pk} > c_k\right)$. The first split is determined by testing the H_{0i}: $\beta_{2il} = \beta_{2ir}$ using a likelihood ratio statistic. Among all rejected hypotheses, the first split is chosen based on the smallest p-value for any combination of covariates and split points across items using the likelihood ratio test statistic and then split the item into a left and a right node (27).

If the second split falls in the right node by s^{th} covariate at the split point c_s, two daughter nodes $I(g_{pk} > c_k)I(g_{ps} \leq c_s)$ and $I(g_{pk} > c_j)I(g_{ps} > c_s)$ will be obtained. In general, further splits are represented by:

$$node(\boldsymbol{g}_p) = \prod_{b=1}^{B} I(g_{pk_b} > c_{k_b})^{a_b} I(g_{pk_b} \leq c_{k_b})^{1-a_b}, \quad (3)$$

where B is the total number of indicator functions, c_{k_b} is the selected split point for the covariate k_b, and $a \in \{0, 1\}$ indicates either the below or above the split point indicator is involved. Therefore, the algorithm which yields a tree for each DIF item (e.g. item i) can be specified by:

$$\eta_{pi} = \beta_{1i} x_p + \sum_{t=1}^{T_i} \beta_{2it}\, node(\boldsymbol{g}_p), \quad (4)$$

where $t = 1, \dots, T_i$ identifies the terminal nodes, and $\sum_{t=1}^{T_i} \beta_{2it}\, node(\boldsymbol{g}_p)$ denotes the tree components, including subgroup-specific intercepts (i.e., item difficulty) denoted by the terminal nodes $node(\boldsymbol{g}_p)$. If an item does not result in any splits, the item is assumed to be free of DIF for all the investigated covariates (27).

3.1.2. LASSO Regression Model

If $\boldsymbol{\theta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i1}, \dots, \beta_{2iK})$ is the vector of regression coefficients from equation 1, where the regression model includes multiple covariates, where $k \in \{1, 2, \dots, K\}$ is the number of covariates, then $\widehat{\boldsymbol{\theta}}_i$ is the set of values that maximizes the log likelihood $l(\boldsymbol{\theta}_i)$ (31). LASSO regression is a regularization technique that adds a penalty term to $l(\boldsymbol{\theta}_i)$ to remove parameters that have little influence on the fit of the regression model (31,53). The penalty term is unique for each item and controlled by a tuning parameter (λ_i) multiplied by the sum of the absolute values of the regression coefficients (31,53). Therefore, the parameters are estimated by maximizing the penalized log likelihood:

$$\widehat{\boldsymbol{\theta}}_i(\lambda) = argmax\, l(\boldsymbol{\theta}_i) - \lambda_i \sum_{i=1}^{K} |\beta_{2ik}|. \quad (5)$$

As λ_i increases, the magnitude of the penalty increases and the absolute values of the DIF parameters shrink toward zero. If an item is free of DIF, the DIF parameters will be zero or close to zero (30,31,53).

Cross-validation (CV) is often used to select the optimal value of λ_i (31). CV involves dividing the dataset into F subsets or folds. The model is iteratively trained on $F-1$ folds, and the prediction error of the fitted model is calculated on the excluded fold. The optimal λ_i is the one that minimizes the overall prediction error across F folds (31).

LASSO may estimate a non-zero coefficient for a DIF parameter that could be shrunk to zero if the CV process is repeated and a different value of λ_i is chosen. To select statistically significant values of the estimated DIF parameters for item i, different approaches have been suggested, including a) nonparametric bootstrap quantile (NBQ) CI, b) variable inclusion probability (VIP), and c) selection inference (SI) (41,54–56). These methods can reduce the number of false positives (FPs) and control the familywise Type I error rate. Based on findings by Abram et.al (41) and Kammer et al. (57) in their comparison of different covariate selection approaches in LASSO regression, NBQ was shown to yield fewer false positives than VIP and SI (41,57).

For the NBQ and VIP approach, 500 bootstrap samples were generated by randomly sampling with replacement from each simulated dataset following the procedure described by Wang et al. (58), who found similar results regardless of the number of bootstrap samples (i.e., 200, 500 and 1000) in covariate selection for LASSO. Based on this finding and considering available computational resources, 500 bootstrap samples were selected. In each sample, CV was used to obtain the optimum value of λ_i, LASSO regression was then applied to each item using this optimal λ_i, and coefficient estimates were recorded.

For the NBQ approach for each covariate, the distribution of estimated coefficients across the 500 bootstrap samples was ranked from lowest to highest. Then a significant threshold γ—an empirical threshold—was chosen, and $\gamma/2$ and $1-\gamma/2$ centiles of the distribution were used to obtain the $(1-\gamma)\%$ CIs for each model coefficient. If the CI of the estimated coefficients did not include zero, the corresponding coefficients were considered as associated with DIF and the item was flagged as exhibiting DIF (41).

In the VIP approach, the frequency of non-zero coefficients for each covariate across the 500 bootstrap samples was calculated. The covariate was associated with DIF if the corresponding frequency was equal to or greater than $1 - \gamma$ (41,55).

The choice of γ to control the Type I error rate depends on study characteristics, such as sample size. Setting $\gamma = \alpha$ does not necessarily ensure Type I error control at the α level. A more reliable approach is to evaluate several γ values, estimate the corresponding Type I error and power rates, and then select the γ that achieves the desired error control and power. In practice, recommended values for γ range between 0.1 and 0.4 (41,54–56).

SI introduced by Lee et.al (59) is a statistical method to provide valid p-values and $(1 - \alpha)\%$ CIs for selected covariates in a LASSO model. The procedure provides CIs so that $P(\beta_{2ik,M} \in CI_{ik,M} | \widehat{M} = M) \geq 1 - \alpha, k \in M$, where M is the selected model by LASSO. This means that the inference is conditioned on the specific model that was selected from the data (57). Conditioning is done because traditional statistical inference (e.g., LR) assumes the model is fixed and independent of the data. In LASSO, the selected model depends on the observed data. Therefore, traditional p-values and CIs are invalid (57).

I considered NBQ with $\gamma = 0.1$ as the primary covariate selection approach for my study to compare with IFT. For comparison purposes among different post covariate selection approaches, NBQ with $\gamma = 0.2$, VIP with $\gamma = 0.2$ and $\gamma = 0.1$, and SI with α = 0.05 were considered (41,54,59).

3.2. Simulation Study

3.2.1. Study Design and Data Generation

An experimental study design was adopted in the simulation study where I generated datasets that varied in their sample size, number of items, percentage of DIF items, DIF effect size, and magnitude of correlation amongst the covariates. The characteristics of the simulation design are described in Table 3-1.

The simulation parameters and their values are similar to those selected by Bauer et al. (53) and Liang et al. (29) to test for DIF using IFT with a partial credit model (PCM) for ordinal

outcome measures and regularized structural equation modeling (i.e., Lasso, adaptive Lasso, and Elastic Net), respectively (29,53).

Table 3-1. Characteristics of simulation design

Simulation condition	Values
Sample size (n)	100, 200, 500, and 1000
Number of items (I)	• I = 6 when n = 100 and 200, • I = 12 when n = 500 and 1000

Number (I^*) of items with DIF	• n = 100 and 200
	  • I^* = 2 and 3
	• n = 500 and 1000
	  • I^* = 3 and 6

Number of covariates (k) and distributions	• $k = 7$
	  • $x_1,\ x_2, x_3 \sim N(0,1)^{\dagger}$
	  • $x_4, x_5 \sim B(0.5)^{\dagger}$
	  • $x_6, x_7 \sim B(0.3)$

DIF covariates (k^*)	• $k^* = 3$ • x_1, x_4, and x_6
DIF effect size	• No DIF (i.e., 0) • Small (i.e., 0.4) • Moderate (i.e., 0.8) • Large (i.e., 1.6)
Correlation between covariates (r)	• Uncorrelated ($r = 0$) • Weak ($r = 0.25$) • Strong ($r = 0.75$)

$^{\dagger}N(0,1)$ indicates normal distribution with mean 0 and variance 1. $B(p)$ indicates Bernoulli distribution with a probability of p

All covariates were sampled from a multivariate normal (N) distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$ of dimension k with diagonal values of 1 and off-diagonal values of r (i.e., the correlation was constant for all covariates). Then x_4 and x_5 were converted to binary covariates using the 0.5 quantile, and x_6 and x_7 were converted to binary covariates using the 0.7 quantile; since the 0.7 quantile of a standard normal distribution is approximately 0.52, values greater than or equal to 0.52 were coded as 1, and values below this threshold were coded as 0. This transformation resulted in a Bernoulli (B) distribution with a

probability of 0.3 (60). Dichotomizing a normally distributed covariate is known to reduce its correlation with other covariates (60).

Under H_{0i}, the probability of an item response for each person was simulated using the two-parameter logistic (2PL) IRT model shown in equation 6, where the item responses are a function of the latent construct (i.e., mental health). As the latent construct increases, the probability of choosing category 1 (e.g., the category that represents better health) for item i increases. The IRT model is defined as

$$Logit\left(P(Y_{pi}=1|\alpha_i,\theta_p,\beta_i)\right)=\alpha_i(\theta_p-\beta_i), \quad (6)$$

where α_i is the item discrimination parameter, β_i is the item difficulty parameter ($i=1,\dots,I$), and θ_p ($p=1,\dots,n$) the latent construct parameter. The latent construct parameter was randomly sampled from a $N(0,1)$ distribution. For item i, α_i was randomly sampled from a $U(1.5,2.5)$ distribution and β_i was an absolute value sampled from a $N(0,1)$ distribution following the simulation studies by Berger et al. (27) and Wang et al. (30).

To introduce uniform DIF in the ith item, $i \in \{1,\dots,I^*\}$ caused by x_1, x_4, and x_6 , and β_i in equation 6 was positively shifted by the DIF effect size in the focal group (i.e., where either $x_1>0$, $x_4=1$, or $x_6=1$):

$$Logit\left(P(Y_{pi}=1|\alpha_i,\theta_p,\beta_i)\right)=\alpha_i(\theta_p-\beta_i-DIF\ effect). \quad (7)$$

DIF effect sizes were set following Berger et al. (27) in their study about detecting DIF using IFT. The probabilities computed from the models in equation 6 for non-DIF items and equation 7 for DIF items were used as the probability of success in a Bernoulli distribution to generate binary item responses. The parameter θ_p in equation 6 is related to the total score (i.e., x_p) used in the LR model of equation 1, which was used as a proxy measure for the latent construct from the observed item responses. The parameter β_i in equations 6 is related to the DIF parameter (i.e., β_{2i}) in equation 1. However, these parameters arise from different modeling frameworks and are not mathematically equivalent. The 2PL IRT model was used for data generation to simulate item responses to allow direct control over the DIF effect magnitude by shifting the item difficulty parameter for the focal group. The LR model in equation 1 does not accommodate

a fixed DIF effect size because the combined regression coefficients determine the DIF effect size.

The strength of the correlation amongst the covariates was manipulated (i.e., no correlation, weak, and strong correlation among covariates); these conditions were consistent with those considered by Wang et al. (30) and Schauberger et al. (15).

The simulation study followed a nested experimental design. The datasets were generated for all combinations of simulation components and DIF effect sizes. The number of items and the number of DIF items were nested within sample size levels, resulting in 96 unique conditions. Each condition was repeated 1000 times, resulting in 96000 simulation replications. The number of replications was selected to increase precision (28,31,53). To ensure reproducibility of the simulations and to help ensure the simulation of unique datasets, seeds were selected randomly between 1 and 1,000,000 and recorded for each dataset.

To assess the validity of the simulated datasets, I generated datasets for a subset of simulation conditions with n = 2000 and 100 replications. These datasets were used to compute the average mean, variance, and 95% CIs for covariates, DIF parameters, correlation amongst covariates, and LR coefficient estimates over 2000 replications. I verified that the expected mean of each parameter fell within the CIs (61).

3.2.2. Covariates and Outcome

In both LASSO and IFT regression models, seven covariates (i.e., x_1 to x_7) were simulated and included in the models, and the outcome was the binary item response. Two-way interactions were not included in the models because I tested only for uniform DIF; interactions are required when non-uniform DIF is tested. In the LASSO regression model total score was included along with the covariates, where the total score was the sum of the item responses.

3.2.3. Model Performance Evaluation

Each item was characterized by a vector $\boldsymbol{\delta}_i^T = (\delta_{i1}, \ldots, \delta_{I7})$. The ground truth vector for DIF items was $\boldsymbol{\delta}_i^T = (1\ 0\ 0\ 1\ 0\ 1\ 0)$ and for non-DIF items, it was $\boldsymbol{\delta}_i^T = \mathbf{0}$. After applying the LASSO and IFT models, I obtained $\widehat{\boldsymbol{\delta}}_i^T$ vectors for each item. For item i, if H_{0i} was rejected,

then $\widehat{\boldsymbol{\delta}}_i^T = 1$. For covariate k in the item i, if H_{0ik} (i.e., covariate k is associated with DIF) was rejected, then $\widehat{\boldsymbol{\delta}}_{ik}^T = 1$. In the IFT model $\hat{\delta}_{ik} = 1$ if item i had split on covariate k and $\hat{\delta}_{ik} = 0$ otherwise. In the LASSO model, if the coefficient of covariate k in item i was significant, then $\hat{\delta}_{ik} = 1$, and $\hat{\delta}_{ik} = 0$ otherwise. At the item-level evaluation, item i exhibited DIF when at least one $\hat{\delta}_{ik} = 1$; Item i was considered a non-DIF item if $\widehat{\boldsymbol{\delta}}_i^T = \mathbf{0}$. Also, overall measures for evaluating each combination of items and covariates were calculated (27,28).

To test for DIF in item i at the item-level, H_{0i} indicated there is no evidence of DIF and H_{1i} indicated the presence of DIF regardless of which covariates were associated with it. Item-level Type I error and power were calculated as:

$$Item\ level\ Type\ I\ error = \frac{\sum_{i:\,\boldsymbol{\delta}_i=0} I(\widehat{\boldsymbol{\delta}}_i \neq \mathbf{0})}{\#\{i:\,\boldsymbol{\delta}_i = 0\}}, \tag{8}$$

$$Item\ level\ power = \frac{\sum_{i:\,\boldsymbol{\delta}_i\neq 0} I(\widehat{\boldsymbol{\delta}}_i \neq \mathbf{0})}{\#\{i:\,\boldsymbol{\delta}_i \neq 0\}}, \tag{9}$$

Furthermore, overall measures for evaluating the combination of items and covariates were calculated; H_{0ik} indicated no evidence that covariate k is associated with DIF, and H_{1ik} indicated covariate k is associated with DIF. Overall Type I error and power calculated as follows:

$$Overall\ Type\ I\ error = \frac{\sum_{i,k:\,\delta_{ik}=0} I(\hat{\delta}_{ik} \neq 0)}{\#\{i,k:\,\delta_{ik} = 0\}}, \tag{10}$$

$$Overall\ power = \frac{\sum_{i,k:\,\delta_{ik}\neq 0} I(\hat{\delta}_{ik} \neq 0)}{\#\{i,k:\,\delta_{ik} \neq 0\}}, \tag{11}$$

For each simulation condition, data were generated, and the item-level and overall Type I error rate, as well as item-level and overall power rates, were calculated. This process was repeated 1000 times, and the average rates were calculated across replications for each simulation condition (28,31,53).

3.2.4. Random Sampling Errors for Type I Error Rates

To quantify the precision of the estimated Type I error rates from the simulation study, 95% CIs were computed for each condition. At $\alpha = 0.05$, it is expected that 5% of null hypotheses that are true are erroneously rejected due to random sampling error. The Type I error rate can be treated as a Bernoulli trial. Therefore, Type I error rates follow a binomial distribution with n = 1000 and 0.05 success probability. Accordingly, the 95% CI for the Type I error rate is:

$$95\%\ CI = \ p \pm 1.96\sqrt{\frac{p(1-p)}{N}}\ . \qquad (12)$$

The 95% CI for $\alpha = \ 0.05$ is

$$95\%\ CI = \ 0.05 \pm 1.96\sqrt{\frac{0.05 \times 0.95}{1000}} = (0.036, 0.064) \qquad (12)$$

If the calculated Type I error rate from simulated data falls outside (3.6%, 6.4%), it indicates that the calculated rate is significantly different from the expected rate (62).

3.3. Real-World Study

3.3.1. Study Design and Data Source

A secondary dataset was analyzed for a Manitoba cohort of individuals with immune-mediated inflammatory disease (IMID) from multiple clinics and community sources. Study data were from a prospective 3-year longitudinal study conducted from November 2014 to July 2016. Only data from the first year of measurement were used. All study participants were 18 years of age or older and were required to have adequate knowledge of the English language to complete the study questionnaires (44).

This dataset was selected for several reasons. First, it contained a substantial number of covariates, encompassing both demographic (e.g., sex, age, marital status) and clinical characteristics (e.g., comorbid conditions, smoking status), that may be associated with DIF.

Second, the dataset had been previously assessed for DIF and DIF items were found across various types of PROMs (44,63,64). Last but not least, the dataset had few missing values, which reduced the likelihood of selection bias associated with missing data (44,65).

3.3.2. Covariates and Outcome

The dataset contained demographic covariates along with clinical covariates and comorbid conditions. Table 3-2 shows the covariates and their type.

Table 3-2. Characteristics of the dataset

Demographic covariates	Clinical covariates	Comorbid conditions‡	
Age*	IMID type†:	Anxiety	Lung cancer
Sex‡	Multiple sclerosis	Bipolar	Lung disease
Race†	Inflammatory bowel disease	Breast cancer	Lupus
Education†	Rheumatoid arthritis	Cholesterol	Migraine
Income†	Smoking status‡	Colon cancer	Osteoarthritis
Marital status†		Depression	Fibromyalgia
Occupation†		Diabetes	Other cancers
		Epilepsy	Peptic Ulcer
		Heart disease	Peripheral vascular disease
		Hypertension	Schizophrenia
		Irritable bowel syndrome	Skin cancer
		Kidney disease	Thyroid
		Liver disease	

*Continuous covariates
†Categorical covariates
‡Binary covariates

The outcomes were the items for the SF-36 domains. The SF-36 is a generic PROM comprising 36 items measuring eight health domains including mental health (MH), physical functioning (PF), general health (GH), role limitation due to physical health (RP), bodily pain (BP), social functioning (SF), role limitation due to emotional health (RE), and vitality (VT) (66). Previous research examining DIF, mostly on MH and PF of the SF-36 domains, suggested that these two domains are more likely to exhibit DIF (44,63,64,67).

For the MH domain, the covariates Depression and Anxiety were excluded from the fitted models because MH domain items are intended to measure depression and anxiety. Although these health conditions are important predictors of MH, including them in DIF models where the outcome is the response for each item in the MH domain—rather than the latent construct (i.e.,

MH)—would lead to their association with DIF, and this might affect other associated covariates.

3.3.3. Data Preparation

Descriptive analyses were conducted to assess the data distribution, frequency of response categories for the PROMs, and prevalence of missing values. Demographic covariates, including race, education, marital status, and occupation, were defined as categorical covariates. Epilepsy and lupus were dropped from the study due to low prevalence (i.e., fewer than 10 observations in a category). Schizophrenia was combined with bipolar disorder because of their clinical similarity. I excluded participants who had missing observations on the covariates or item responses and those who selected "I do not wish to answer" for income. The number of comorbid conditions (NCC) was derived from the comorbid conditions listed in Table 3-2.

Correlations among covariates were examined using Spearman's correlation coefficient. Item responses were dichotomized by grouping categories into positive and negative responses (68,69). This methodology aligned with one employed by Grassi et al. (69), who demonstrated that dichotomizing SF-36 items can often preserve the instrument's validity and reliability.

To assess the sensitivity of the analysis to the choice of response cut point, I applied two different dichotomization strategies. In both approaches, the original polytomous responses collapsed into binary categories, but the cut-points differed. The first approach dichotomized responses at the midpoint of the scale, and the second approach contrasted the maximum response option (e.g., "not at all") against all other categories. Details of the dichotomization strategies, including the cut-points used to map the original polytomous categories into binary responses, are presented in Table 3-3. I considered the first approach as the primary dichotomization approach, since the results were consistent with previous research and provided greater comparability with established findings (12,44). The SF-36 items are designed so that higher sum scores on the domain show a better level of health status (23); in the process of scoring the SF-36, some item responses needed to be reversed to be consistent with other items.

Table 3-3. Dichotomization strategies for the SF-36 domain items.

Domain	Approach I		Approach II	
	Category I	Category II	Category I	Category II
Mental Health Vitality	All of the time Most of the time A good bit of the time	Some of the time A Little of the time None of the time	All of the time Most of the time A good bit of the time Some of the time A Little of the time None of the time	None of the time
Physical Functioning	Yes, limited a lot Yes, limited a little	Not limited at all	Yes, limited a lot	Yes, limited a little Not limited at all
General Health, Item 1	Fair Poor Good	Very good Excellent	Fair Poor Good Very good	Excellent
General Health, Items 2 to 5	Definitely true Mostly true Don't know	Mostly false Definitely false	Definitely true Mostly true Don't know Mostly false	Definitely false
Bodily Pain	Extremely Quite a bit	Moderately Slightly Not at all	Extremely Quite a bit Moderately Slightly	Not at all
Social Functioning, Item 1	Extremely Quite a bit	Moderately Slightly Not at all	Extremely Quite a bit Moderately Slightly	Not at all
Social Functioning, Item 2	All of the time Most of the time	Some of the time A Little of the time None of the time	All of the time Most of the time Some of the time A Little of the time None of the time	None of the time

Note: Role limitation due to physical health and role limitation due to emotional health domains were originally binary. Item 1 from general health and the two items comprising the social functioning domain had different response formats.

The total score for each domain after the dichotomization process was calculated for inclusion in the LASSO model. Although dichotomization reduces variation in the underlying latent construct and may introduce bias, this approach was adopted to ensure methodological consistency with the IFT regression model. In the R package implementation, the IFT regression model requires total scores from dichotomized item responses.

3.3.4. Checking Model Assumptions

I assessed the linearity of continuous covariates, including age, NCC, and total score, with the logit of the item responses in the SF-36 domains. For each item and covariate pair, I first fit a logistic regression model with the covariate as a linear predictor. To visualize the relationship, the covariate was divided into 10 equally sized bins, and within each bin, I computed the mean covariate value and the observed probability of endorsing the item. The logit of the observed probabilities was then plotted alongside the predicted logits from the fitted logistic regression model. Deviations of the observed logits from the predicted line indicate violations of the linearity assumption between the logit of the response and the continuous covariate.

The unidimensionality of the items that comprised each SF-36 domain was assessed to ensure that all items measured a single construct. Confirmatory factor analysis (CFA) was conducted using a single-factor model. Model fit indices include χ^2 and the corresponding p-value, comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Since χ^2 is sensitive to the sample size, the corresponding p-values may result in rejection of the H_0 (i.e., the items are unidimensional). The recommended cutoff point for indicating a good fit for CFI is values greater than 0.90. Values greater than 0.95 are considered an excellent fit (70). RMSEA values below 0.05 are considered a close fit and below 0.08 as a reasonable fit (70). A good fit for SRMR is values less than 0.08 (70). If these criteria were met, the item set was considered unidimensional (70). CFA was limited to domains with at least four items to estimate reliable latent constructs. Domains with fewer than four items provide limited degrees of freedom, which can result in unstable models and parameter estimates, and poor model fit (70).

3.3.5. Strength and Direction of Association

In the IFT regression model, if more than one split was observed, the importance of covariates was assessed by the order in which they appeared in the tree. The direction of DIF effect (i.e., comparing the probability of choosing category one) between the splits (e.g., focal group and reference group in a binary covariate) was assessed by comparing the estimated sub-group intercepts (i.e., γ_{il} and γ_{ir}) in the terminal nodes. If $\gamma_{il} I(g_{pj} \leq c_k) > \gamma_{ir} I(g_{pj} > c_k)$,

$g_{pj} \leq c_k$ indicated a higher probability of choosing the category of one (i.e., the higher/healthier response option) in item i.

In the LASSO regression model, covariate importance was assessed by the absolute values of the estimated coefficients, with larger values indicating stronger associations with DIF. The direction of the DIF effect was evaluated by interpreting the corresponding odds ratios: For categorical covariates, values greater than one indicated a higher probability of choosing the category of one for the focal group in item i compared to the reference group, and values less than one indicated a lower probability. For continuous covariates, higher positive values of the covariate were associated with an increased probability of choosing category one, whereas negative coefficients indicated a decreased probability.

3.3.6. Model Performance Evaluation

I used the hold-out CV technique combined with bootstrap resampling to evaluate the model performance in the absence of ground truth. In the hold-out bootstrap CV approach, I resampled from the data 500 times with replacement, with the size of the real-world data, consistent with my simulation study, where I used the bootstrap approach for covariate selection in LASSO for each domain. I divided each resampled data set into training (80%) and test (20%) datasets. This split point was suggested by Bami et al. (71) in their study of a train-test split algorithm for CV and hold-out iteration and used by Jafari et al. (52) in their study of DIF using penalization approaches.

I applied LASSO and IFT regression models to the training datasets and derived the fitted regression models. These models were used to estimate item response probabilities on the test datasets. These probabilities were then dichotomized according to the prevalence of category 1 in the dichotomized SF-36 item responses. Performance metrics, including the Brier score, mean squared error (MSE), and accuracy, were calculated for the test dataset for each bootstrap sample. The average values of each performance metric across the 500 bootstrap samples were computed (1,4,28):

Brier score: The Brier score is a measure of calibration that describes the accuracy of predictions of binary outcomes based on the squared differences between actual item responses and the probability of predicted item responses by the model (72):

$$Brier\ score\ for\ item\ i = \frac{1}{N}\sum_{p=1}^{N}\left(f_p - O_p\right)^2 , \tag{14}$$

where f_p is the predicted probability of person p, O_p is the actual item response of the person p, and N is the number of individuals in each domain. The Brier score can range in value from zero to one, with zero indicating a perfect model with 100% accuracy (72).

Accuracy: Accuracy is the proportion of correctly predicted item responses for item i relative to the total number of responses:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{p} I(y_p = \hat{y}_p) . \tag{15}$$

Mean square error (MSE):

$$MSE = \frac{1}{N}\sum_{p=1}^{n}\left(y_p - \hat{y}_p\right)^2 . \tag{16}$$

where N is the number of observations, y_p, $\hat{y}_p$ are the actual item responses and predicted item responses to item i, respectively (26,28).

3.4. Software and Packages

Analyses were conducted with R software version 4.2.2. I used the "DIFtree" package to apply the IFT regression model. This package can be used to test for uniform and non-uniform DIF and can be applied to dichotomous and polytomous item responses (27). To model IFT using the "DIFtree" package, I set α = 0.05 and the number of permutations to 1000 in the model (27). To model LASSO regression, the optimum value for each λ_i were chosen by a 5-fold CV with the "cv.glmnet" package, and then the selected λ_i was used on each item with the "glmnet" package (31,73). The "selectiveInference" package was used for the SI approach. Unidimensionality through CFA was assessed with the "lavaan" package. To predict the item

response probabilities on the test datasets in the CV-bootstrap process, I employed the "predict" function for the LASSO regression model and "IFTpredictor" for the IFT regression model.

Due to the computational intensity of the simulation study analysis, I implemented parallel processing using the "future" and "purrr" packages in R.

Chapter IV: Results

This section is organized into two main parts. First, results from the simulation study are presented. These include Type I error rates at the item-level (i.e., the item was incorrectly identified as a DIF item) and overall (i.e., the covariates were incorrectly identified as associated with DIF) and power rates at the item-level (i.e., the DIF items were correctly identified) and overall (i.e., DIF covariates were correctly identified). Simulation results are reported under varying sample sizes, correlation among covariates, and DIF effect magnitudes. Second, findings from the real-world analysis of the SF-36 item responses are described. This part begins with descriptive statistics of the study population and item responses, followed by results of testing for DIF from the IFT and LASSO models.

4.1. Simulation Study

4.1.1. Validity of Data Generation

Table 4-1 shows the average of the mean and variance estimates with 95% CIs for the distribution of simulated covariates over 100 replications for the condition where the DIF effect = 1.6, $r = 0.75$, $I = 12$, and $I^* = 3$, given $n = 2000$. All the estimated means and variances were close to the corresponding population values, and their 95% CIs contained the population value from where the covariates were generated.

Table 4-1. Average of the mean and variance estimates (95% CIs) for simulated covariate distributions.

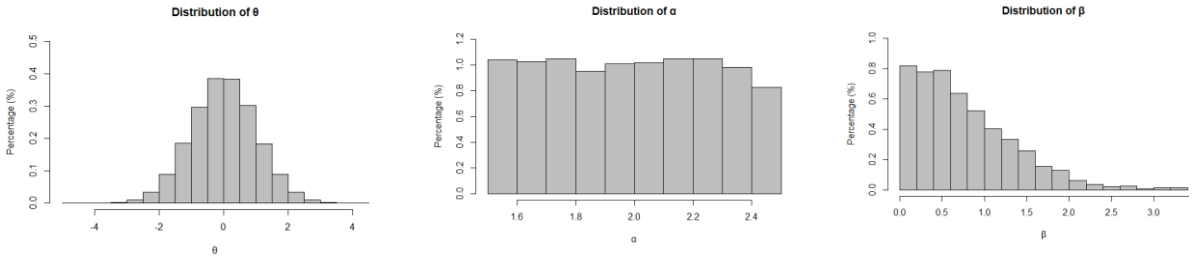
Covariate & Distribution	μ	$\overline{\hat{\bar{x}}}$ (95% CI)	σ^2	$\overline{\widehat{\delta^2}}$ (95% CI)
$x_1 \sim N(0,1)$	0	0.00 (-0.04, 0.03)	1	1.00 (0.95, 1.06)
$x_2 \sim N(0,1)$	0	0.00 (-0.04, 0.03)	1	0.98 (0.93, 1.03)
$x_3 \sim N(0,1)$	0	0.00 (-0.02, 0.05)	1	1.09 (1.04, 1.15)
$x_4 \sim B(0.5)$	0.5	0.49 (0.47, 0.51)	0.25	0.25 (0.24, 0.26)
$x_5 \sim B(0.5)$	0.5	0.49 (0.48, 0.51)	0.25	0.25 (0.24, 0.26)
$x_6 \sim B(0.3)$	0.3	0.30 (0.28, 0.32)	0.21	0.20 (0.19, 0.21)
$x_7 \sim B(0.3)$	0.3	0.29 (0.28, 0.31)	0.21	0.20 (0.19, 0.21)

Figure 4-1 shows the distributions of the estimates for the latent construct (θ), item discrimination (α), and item difficulty (β) parameters for generating the item responses in the

logistic function over 100 replications. The histograms indicate that the generated parameters align with their expected distributions.

Figure 4-1. Histograms for the distribution of estimated parameters for simulating the item responses.

Panel A. $\theta \sim N(0,1)$ Panel B. $\alpha \sim U(1.5, 2.5)$ Panel C. $\beta \sim abs(N(0,1))$



θ is a latent construct, α is item difficulty, and β is item discrimination

Table 4-2 presents the estimated correlations among covariates along with their 95% CIs over 100 replications. Converting continuous covariates to binary has introduced some bias, resulting in lower estimated correlations for the binary covariates (i.e., x_4 to x_7) compared to the expected values (i.e., 0.75). The estimated 95% CIs for the correlations among continuous covariates contained the population values (i.e., 0.75).

Table 4-2. Estimated correlation matrix for simulated covariates (95% CIs).

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1	0.73 (0.73, 0.75)	0.73 (0.73, 0.75)	0.62 (0.61, 0.62)	0.62 (0.61, 0.62)	0.58 (0.57, 0.58)	0.58 (0.57, 0.58)
x_2		1	0.74 (0.73, 0.75)	0.62 (0.62, 0.62)	0.62 (0.61, 0.62)	0.58 (0.58, 0.58)	0.58 (0.57, 0.58)
x_3			1	0.62 (0.61, 0.62)	0.62 (0.61, 0.62)	0.58 (0.57, 0.58)	0.58 (0.57, 0.58)
x_4				1	0.54 (0.54, 0.54)	0.49 (0.49, 0.5)	0.49 (0.49, 0.5)
x_5					1	0.49 (0.49, 0.5)	0.49 (0.49, 0.49)
x_6						1	0.53 (0.52, 0.53)
x_7							1

Table 4-3 presents the estimated coefficients from simple logistic regression models assessing associations between simulated covariates and item responses over 100 replications. The coefficients for x_1, x_4, and x_6 were statistically significant (i.e., the 95% CI did not include

zero) for the DIF items, reflecting the intended simulation design. Coefficients for non-DIF items were not statistically significant (i.e., the 95% CIs included zero), supporting the accuracy of the data generation process to simulate DIF and non-DIF items.

Table 4-3. Average coefficient estimates and 95% confidence intervals for the simple LR model.

Covariate	DIF items	Non-DIF items
x_1	-0.79 (-0.99, -0.6)	0.00 (-0.10,0.09)
x_2	-0.01 (-0.19, 0.18)	0.00 (-0.10, 0.10)
x_3	0.00 (-0.19,0.18)	0.00 (-0.1, 0.09)
x_4	-5.12 (-40.50, -3.98)	0.00 (-0.19,0.19)
x_5	0.01 (-0.37, 0.37)	0.00 (-0.18,0.19)
x_6	-5.00 (-61.65, -2.65)	0.01 (-0.20,0.21)
x_7	-0.03 (-0.51, 0.41)	0.01 (-0.20,0.21)

4.1.2. Model Performance Evaluation

4.1.2.1. Type I Error Rates

Table 4-4 presents the item-level and overall Type I error rate estimates for IFT and LASSO regression models. The results are stratified by sample sizes and correlations among covariates.

Generally, IFT showed item-level Type I error rate estimates (i.e., the item was incorrectly identified as a DIF item) lower than the nominal 5% level, ranging from 2.7% to 4.4%. The observed Type I error rates fell within the interval of (3.6%, 6.4%)—the random sampling error bounds, except for two simulation conditions ($n = 500$ and 1000, $r = 0.75$) where the estimates were lower than the lower random sampling error bounds (i.e., 3.6%). LASSO showed higher item-level Type I error rates than the nominal level in most conditions and reached 9.7% when $n = 1000$ and $r = 0$, which exceeded the upper random sampling error bounds. However, when $r = 0.75$, the Type I error rate for LASSO was close to the nominal level of 5% and ranged from 4.5% to 5.6%, falling in the random sampling error bounds. The largest observed Type I error rate difference between the LASSO and IFT methods was 6.2% when $n = 1000$ and $r = 0$. As the correlation among covariates increased, item-level Type I error rates tended to decrease, particularly for the LASSO regression model. When covariates are strongly correlated, LASSO tends to randomly keep only one covariate and shrink the other(s) to

zero. This reduces the total number of selected covariates, which lowers the chance of false positives and therefore Type I error rates (74).

The overall Type I error rate (i.e., the covariates were incorrectly identified as associated with DIF) for both methods remained below 5% across all conditions, ranging from 0.4% to 0.7% for IFT and 0.7% to 1.5% for LASSO. LASSO consistently showed slightly higher overall Type I error rates than IFT, with the largest difference being 1%.

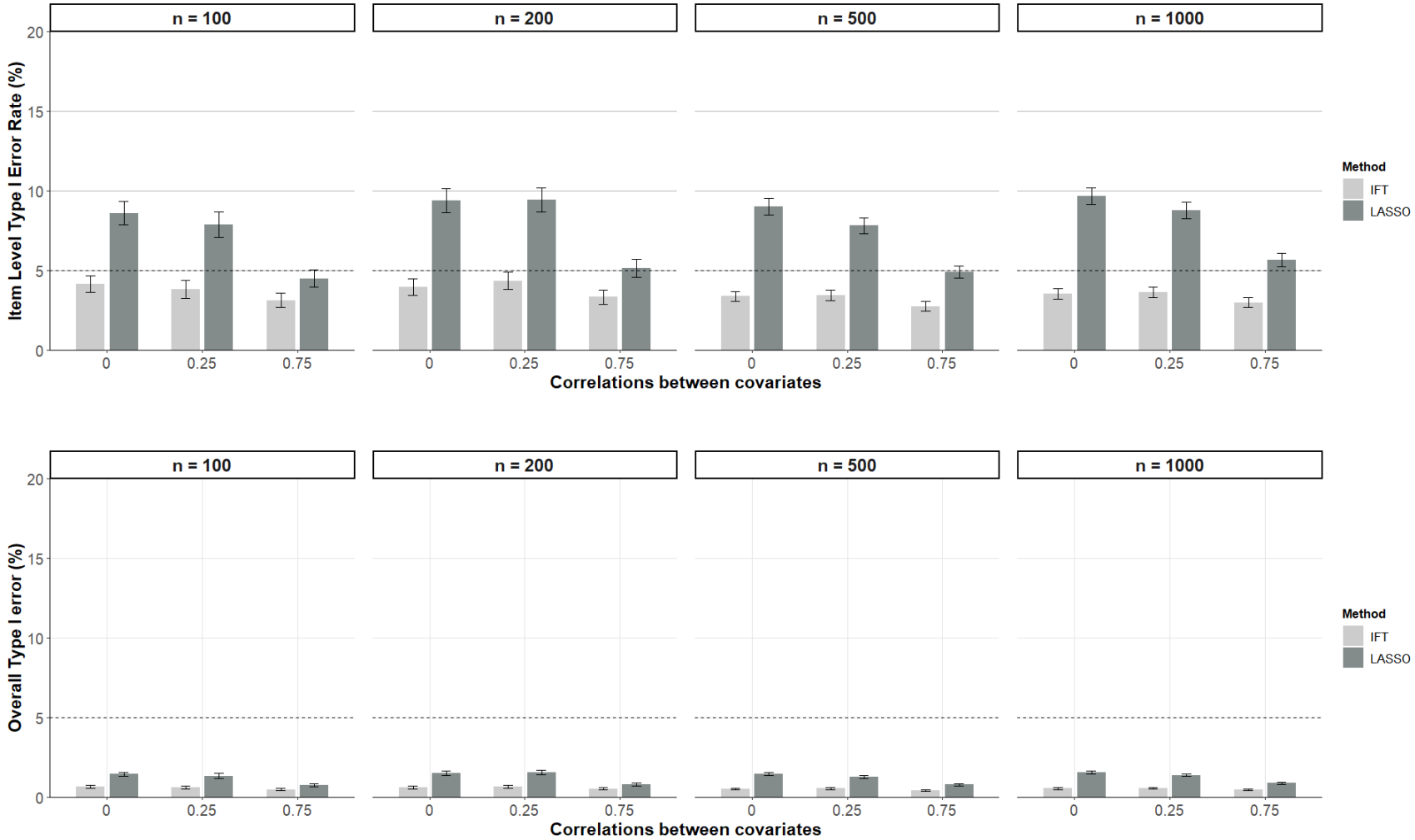
The overall Type I error rates were lower than the item-level Type I error rates. This difference is expected, given how the two metrics are defined in equations 8 and 10. Item-level Type I error is the ratio at which an item is incorrectly identified as showing DIF (i.e., false positive) over the number of non-DIF items. For example, in the condition with six non-DIF items, if x_1 is associated with DIF for one non-DIF item, the item-level Type I error rate for that simulated data is 1/6 or 0.17. But for the overall Type I error rate, the numerator remains one false positive, and the denominator is the number of covariates across all non-DIF items, which is 42 (i.e., each item has seven covariates), yielding an overall Type I error rate of 1/42 or 0.02 in this example. As a result, overall Type I error rates were more conservative and typically lower than item-level rates, consistent with Berger et al. (27) and Bollmann et al. (28). Since all the overall Type I error rates were much lower than the random sampling error bounds, 95% CIs for them were not reported.

Table 4-4. Type I error rates for LASSO and IFT regression models stratified by sample size and correlation among covariates.

Conditions		Item-level (95%CI)		Overall	
n	**r**	**IFT**	**LASSO**	**IFT**	**LASSO**
100	0	4.1 (3.6, 4.7)	8.6 (7.9, 9.3)	0.7	1.4
	0.25	3.8 (3.3, 4.4)	7.9 (7.1, 8.7)	0.6	1.3
	0.75	3.1 (2.7, 3.6)	4.5 (3.9, 5.0)	0.5	0.7
200	0	3.9 (3.4, 4.4)	9.4 (8.6, 10.1)	0.6	1.5
	0.25	4.4 (3.8, 4.9)	9.5 (8.7, 10.2)	0.7	1.5
	0.75	3.3 (2.9, 3.8)	5.1 (4.5, 5.7)	0.5	0.8
500	0	3.4 (3.0, 3.7)	9.0 (8.5, 9.5)	0.5	1.4
	0.25	3.4 (3.1, 3.7)	7.8 (7.3, 8.3)	0.5	1.3
	0.75	2.7 (2.4, 3.0)	4.9 (4.5, 5.3)	0.4	0.8
1000	0	3.5 (3.2, 3.9)	9.7 (9.1, 10.2)	0.5	1.5
	0.25	3.6 (3.3, 3.9)	8.8 (8.3, 9.3)	0.5	1.4
	0.75	3.0 (2.7, 3.3)	5.6 (5.2, 6.1)	0.4	0.9

Figure 4-2 displays the item-level and overall Type I error rate estimates in percentages for IFT and LASSO regression models stratified by sample size and correlations among covariates.

Figure 4-2. Item-level and overall Type I error rate estimates (%) for IFT and LASSO regression models stratified by sample size and correlations among covariates.



4.1.2.2. Power Rates

DIF items = 25%: Item-level power (i.e., the DIF items were correctly identified) and overall power (i.e., DIF covariates were correctly identified in DIF items) results are presented in Table 4-5, when 50% of items were simulated to exhibit DIF. Item-level and overall power increased as the sample size and DIF effect increased and for both methods. Compared to IFT, LASSO showed higher item-level and overall power regardless of sample size. For small DIF effects (i.e., DIF effect = 0.4) with no and weak correlation (i.e., $r = 0$ to 0.25) among covariates, the item-level power across sample sizes ranged from 3.8% to 18.8% for IFT, and from 9.2% to 28.4% for LASSO. However, under conditions of strong correlation among covariates (i.e., $r = 0.75$), item-level power for the LASSO model declined and fell below that of IFT. In the presence of strong correlation among covariates, the item-level power increased to 47.8% for IFT and decreased to 24.0% for LASSO. The largest item-level power difference observed between the methods was 29.8% where IFT showed higher item-level power than LASSO when the sample size was 500, the DIF effect was moderate (i.e., DIF effect = 0.8), and there was a strong correlation among covariates.

The overall power was less than the item-level power for corresponding conditions. This difference is expected, given how the two metrics are defined in equations 9 and 11. Item-level power is the ratio is the number of items correctly identified as DIF (i.e., true positive) to the number of DIF items. For example, in the condition with six DIF items–two non-DIF items and four DIF items, if x_1 is associated with DIF for one DIF item, the item-level power for that simulated data is 1/4 or 0.25. But for the overall power, the numerator is the number of true positives, and the denominator is the number of covariates associated with DIF, which is 12, given that each item has three associated covariates. This yields an overall power of 1/12 or 0.08 in this example. As a result, overall power rates were more conservative and lower than item-level rates, consistent with Berger et al. (27) and Bollmann et al. (28).

The overall power of LASSO to identify the covariates associated with DIF was higher than IFT across sample sizes, except when the covariates were strongly correlated. The largest overall power difference observed between the methods was 14.2% where IFT showed higher overall power than LASSO in the condition with a sample size of 1000, a large DIF effect (i.e., DIF effect = 1.6), and strong correlation among covariates.

Table 4-5. Item-level and overall power (%) estimates for IFT and LASSO regression models stratified by DIF effects magnitude, correlations among covariates, and sample size, DIF items = 25%.

Conditions		Item-level		Overall	
DIF effect	**r**	**IFT**	**LASSO**	**IFT**	**LASSO**
$\boldsymbol{n = 100, I = 6}$					
0.4	0	3.8	9.2	0.7	1.9
	0.25	5.0	9.4	0.8	1.6
	0.75	6.0	5.2	1.0	1.0
0.8	0	5.5	14.2	1.0	3.5
	0.25	7.6	15.0	1.6	3.7
	0.75	14.8	11.3	2.5	2.5
1.6	0	10.7	29.8	3.5	10.5
	0.25	20.3	35.2	6.8	12.5
	0.75	49.2	28.9	11.6	9.0
$\boldsymbol{n = 200, I = 6}$					
0.4	0	5.0	11.6	0.9	2.1
	0.25	5.1	10.5	1.0	2.3
	0.75	9.2	8.2	1.6	1.7
0.8	0	8.7	19.8	2.1	5.6
	0.25	12.1	23.8	3.7	6.9
	0.75	29.8	18.7	6.3	4.9
1.6	0	26.8	51.4	12.8	23.4
	0.25	46.5	61.3	20.7	26.7
	0.75	75.8	56.0	24.4	19.3
$\boldsymbol{n = 500, I = 12}$					
0.4	0	6.8	15.8	1.6	4.2
	0.25	10.1	18.6	2.6	4.9
	0.75	25.3	14.8	5.4	3.8
0.8	0	18.2	41.9	7.3	16.2
	0.25	38.0	50.9	15.2	19.7
	0.75	76.3	46.5	22.2	15.3
1.6	0	72.4	90.1	58.8	63.6
	0.25	88.2	94.1	69.2	64.2
	0.75	95.8	91.0	53.3	43.1
$\boldsymbol{n = 1000, I = 12}$					
0.4	0	10.1	26.0	3.0	7.8
	0.25	18.8	28.4	5.7	9.1
	0.75	47.8	24.0	10.4	6.8
0.8	0	40.9	69.3	24.9	34.6
	0.25	69.0	78.8	41.6	39.2
	0.75	93.9	74.8	39.4	28.3
1.6	0	92.3	98.0	86.0	86.2
	0.25	96.2	98.7	88.0	83.2
	0.75	98.5	97.5	70.2	56.0

Figures 4-3 and 4-4 show the corresponding item-level and overall power estimates under small, moderate, and large DIF effects. As shown in the plots, item-level and overall power

estimates increased with sample size and DIF effect magnitude. Notably, LASSO exhibited higher power in no and weak correlation conditions, whereas IFT outperformed LASSO when covariates are strongly correlated.

Figure 4-3. Item-level power rates for IFT and LASSO regression models stratified by DIF effects magnitude, correlations among covariates, and sample size, DIF items = 25%.

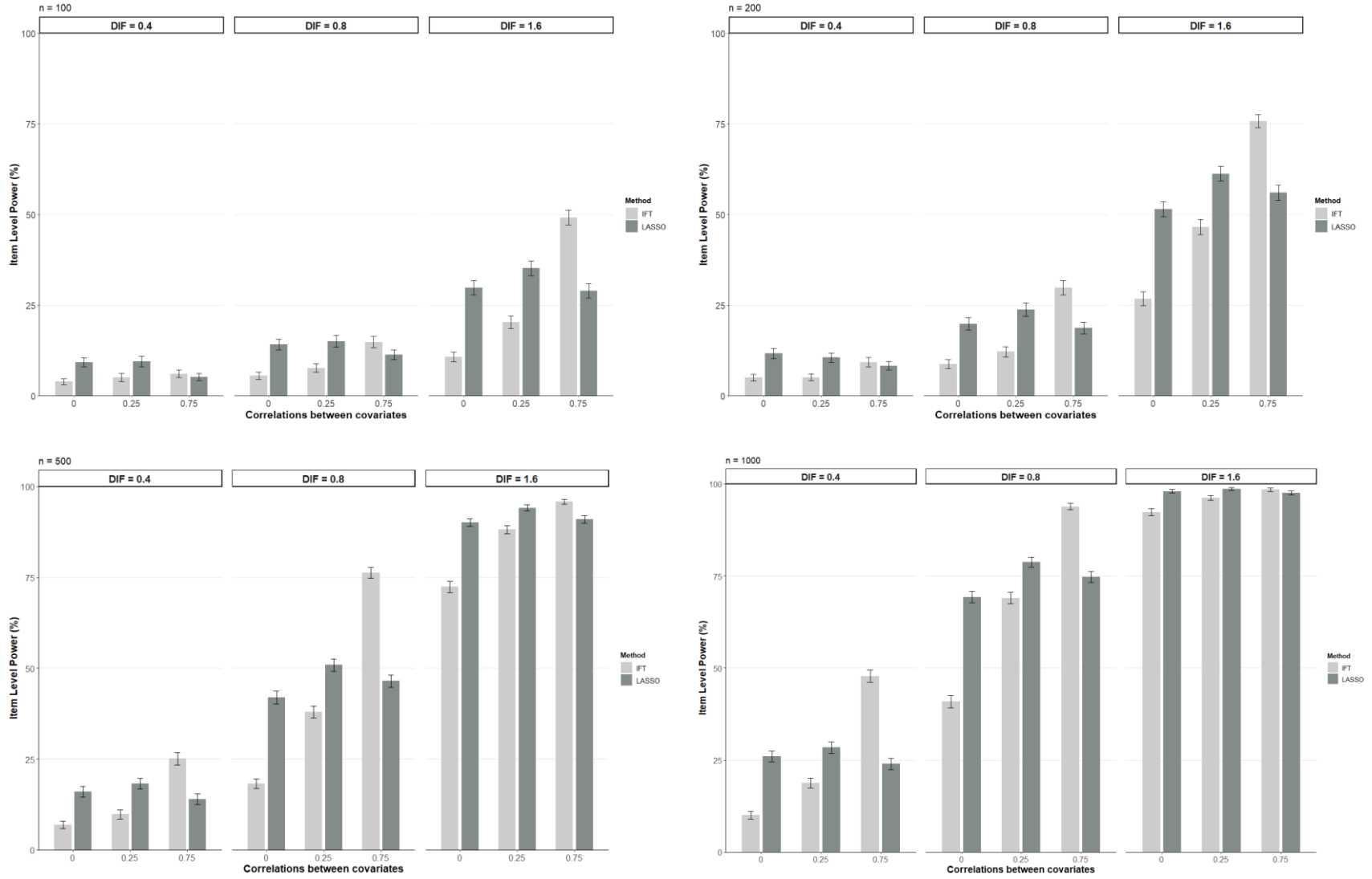
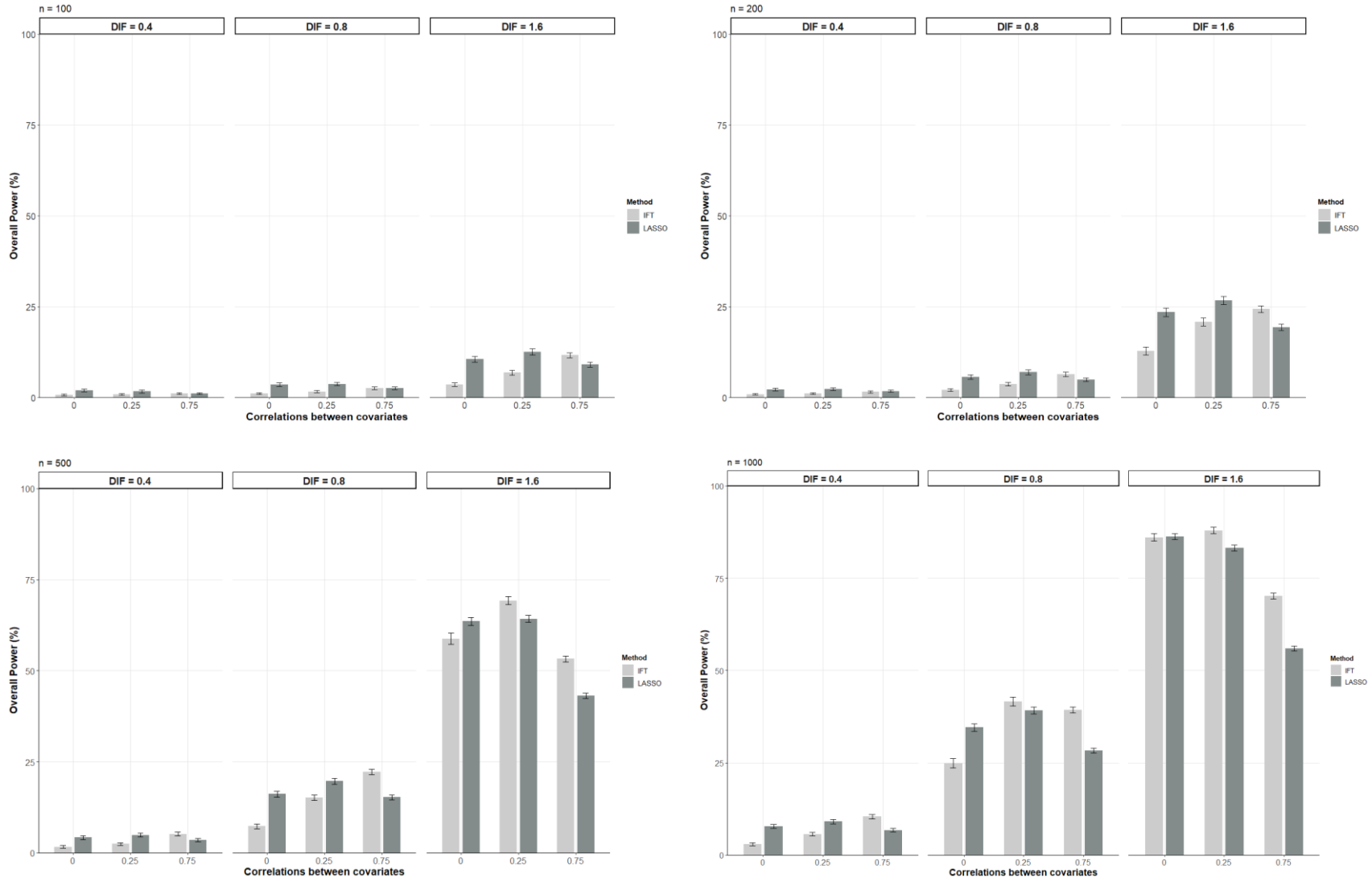


Figure 4-4. Overall power estimates for IFT and LASSO regression models stratified by DIF effects magnitude, correlations among covariates, and sample size, DIF items = 25%.



DIF items = 50%: Table 4-6 presents the item-level and overall power estimates for IFT and LASSO regression models stratified by DIF effects magnitude, correlations among covariates, and sample size when 50% of items were simulated to exhibit DIF.

The item-level and overall power for 50% DIF items were slightly lower than 25% DIF items in their corresponding conditions. A similar pattern to 25% DIF items condition in the item-level and overall power was observed when 50% of the items were simulated to exhibit DIF. LASSO continued to show higher power than IFT in conditions with uncorrelated and weakly correlated covariates, with the largest difference being 29.6%; IFT outperformed LASSO in conditions with strong covariate correlations, with the largest difference being 23.8%.

Table 4-6. Item-level and overall power estimates for IFT and LASSO regression models stratified by DIF effects magnitude, correlations among covariates, and sample size, DIF items = 50%.

Conditions		Item-level		Overall	
DIF effect	**r**	**IFT**	**LASSO**	**IFT**	**LASSO**
$\boldsymbol{n = 100, I = 6}$					
0.4	0	4.8	9.9	0.8	1.8
	0.25	4.4	8.3	0.6	1.3
	0.75	4.3	5.1	0.6	0.9
0.8	0	4.7	12.2	0.8	2.6
	0.25	5.8	13.1	1.2	3.2
	0.75	9.3	7.8	1.5	1.7
1.6	0	7.0	20.6	1.9	6.3
	0.25	11.5	24.5	3.2	7.4
	0.75	28.1	21.1	6.1	5.6
$\boldsymbol{n = 200, I = 6}$					
0.4	0	4.5	10.8	0.7	2.0
	0.25	5.3	10.8	0.8	2.1
	0.75	7.1	6.3	1.1	1.2
0.8	0	5.9	15.4	1.2	3.7
	0.25	9.4	17.1	2.1	4.2
	0.75	17.0	12.5	3.4	3.2
1.6	0	14.3	35.2	5.3	13.1
	0.25	25.6	41.9	9.1	15.7
	0.75	53.2	36.4	13.7	11.6
$\boldsymbol{n = 500, I = 12}$					
0.4	0	4.2	12.3	0.8	2.7
	0.25	6.6	12.6	1.3	2.8
	0.75	10.7	8.5	1.8	1.8
0.8	0	8.5	23.2	2.5	7.4
	0.25	16.0	27.8	4.9	9.0
	0.75	40.3	25.1	9.0	7.2
1.6	0	33.3	62.9	18.0	31.1
	0.25	55.3	70.5	29.0	34.5
	0.75	81.2	66.7	30.8	25.1
$\boldsymbol{n = 1000, I = 12}$					
0.4	0	6.1	17.0	1.5	4.3
	0.25	9.1	18.6	2.3	4.8
	0.75	20.8	15.3	4.2	3.8
0.8	0	17.0	41.0	6.5	15.8
	0.25	32.7	48.0	12.9	18.0
	0.75	69.3	45.5	19.6	14.6
1.6	0	63.2	84.4	48.6	56.1
	0.25	80.9	88.4	58.6	56.4
	0.75	91.9	86.3	49.5	39.5

Figures 4-5 and 4-6 display the corresponding item-level and overall power estimates under small, moderate, and large DIF effects. Similar to the 25% DIF items condition, item-level and overall power estimates increased with sample size and DIF effect magnitude. LASSO showed higher power in no and weak correlation conditions, and IFT outperformed LASSO when covariates are strongly correlated.

Figure S-1 and S-2 in the supplementary material section show item-level Type I rate and power for different post-covariate selection approaches in LASSO, including SI, VIP and NBQ with $\gamma = 0.2$ and 0.1. Among these methods, the item-level Type I error rate of NBQ with $\gamma = 0.1$ was less than other approaches across simulation conditions.

Figure 4-5. Item-level power rate estimates for LASSO and IFT regression models stratified by DIF effects magnitude, correlations between covariates, and sample size, DIF items = 50%.

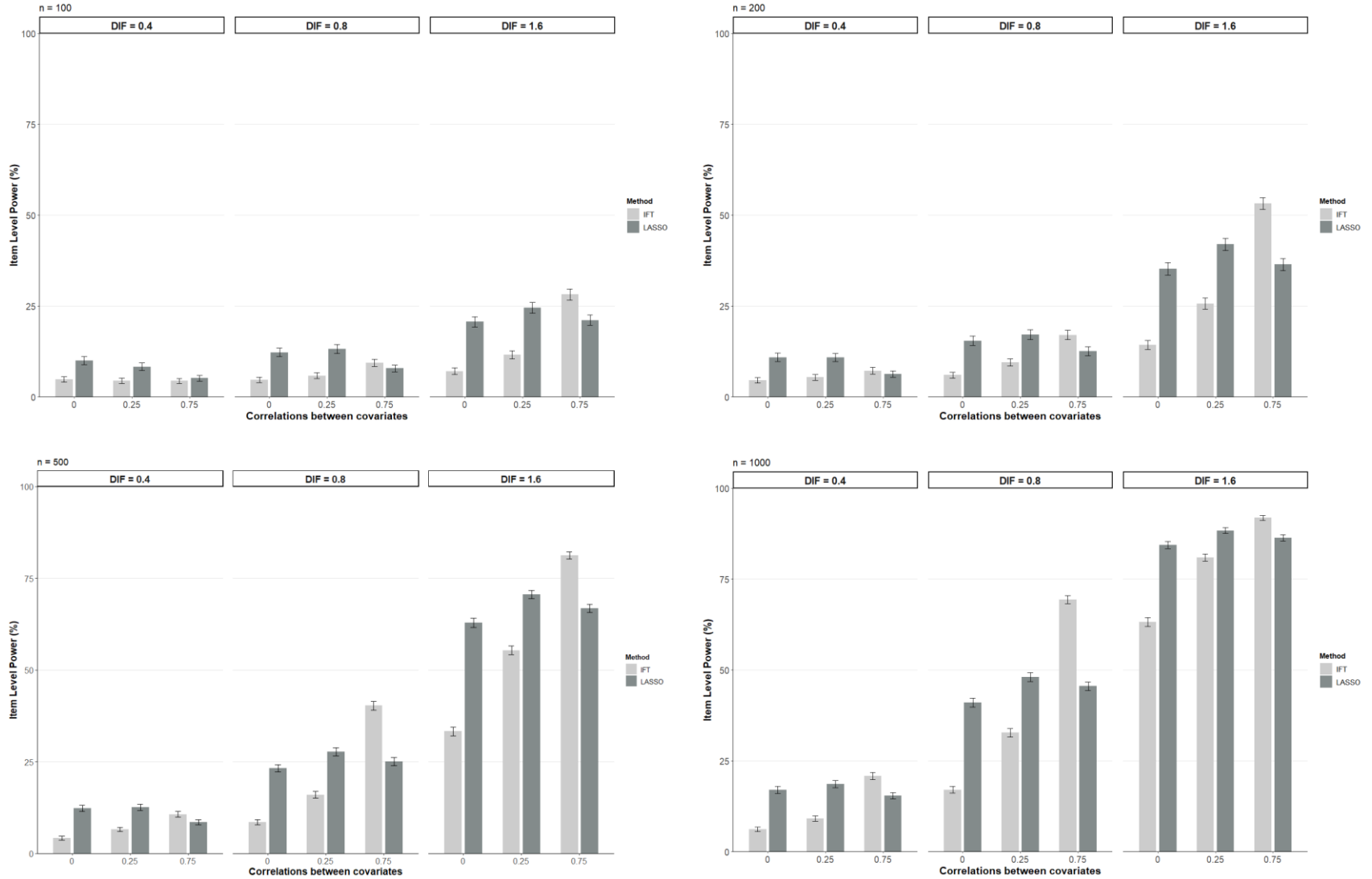
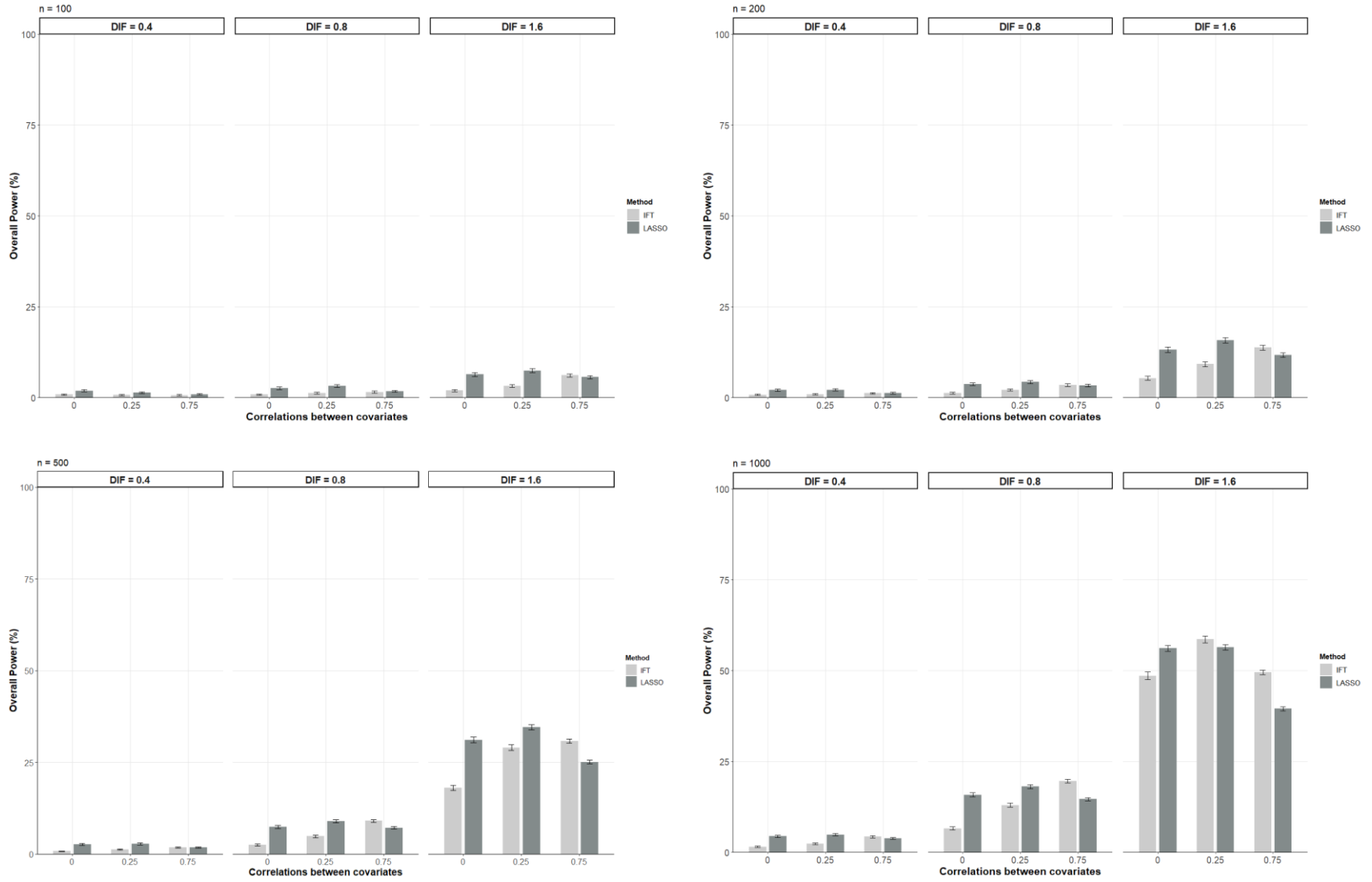


Figure 4-6. Overall power estimates for LASSO and IFT regression models stratified by DIF effects magnitude, correlations between covariates, and sample size, DIF items = 50%.

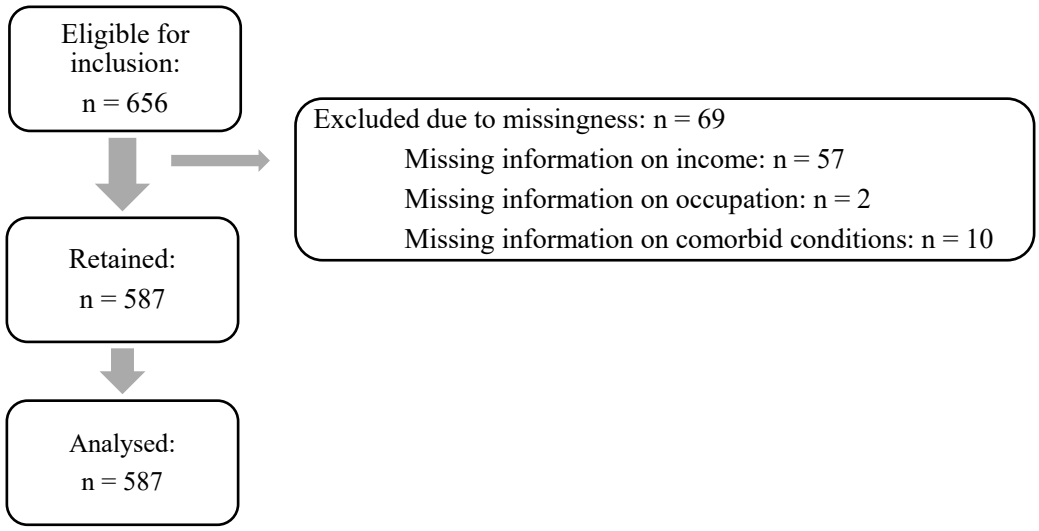


4.2. Real-World Study

4.2.1. Descriptive Analysis

The study cohort comprised 656 individuals. After the data preparation process (Figure 4-7), 587 individuals remained for the analysis. Of this number, 74.6% were female, the majority of study participants were white (91.0%), the mean age was 50.5 years (SD = 13.7), and 64.9% reported an annual income over $50,000.

Figure 4-7. Flow diagram of data preparation



The cohort included 221 (37.6%) individuals with multiple sclerosis (MS), 224 (38.2%) with inflammatory bowel disease (IBD), and 142 (24.2%) with rheumatoid arthritis (RA). The median number of comorbid conditions (NCC) was two (range: 1 – 9). 58.8% of participants were smokers. Depression was the most prevalent comorbid condition (33.7%), followed by hypertension (25.7%). The frequencies of participant characteristics are summarized in Table 4-7.

Table 4-7. Characteristics of the study cohort, n = 587

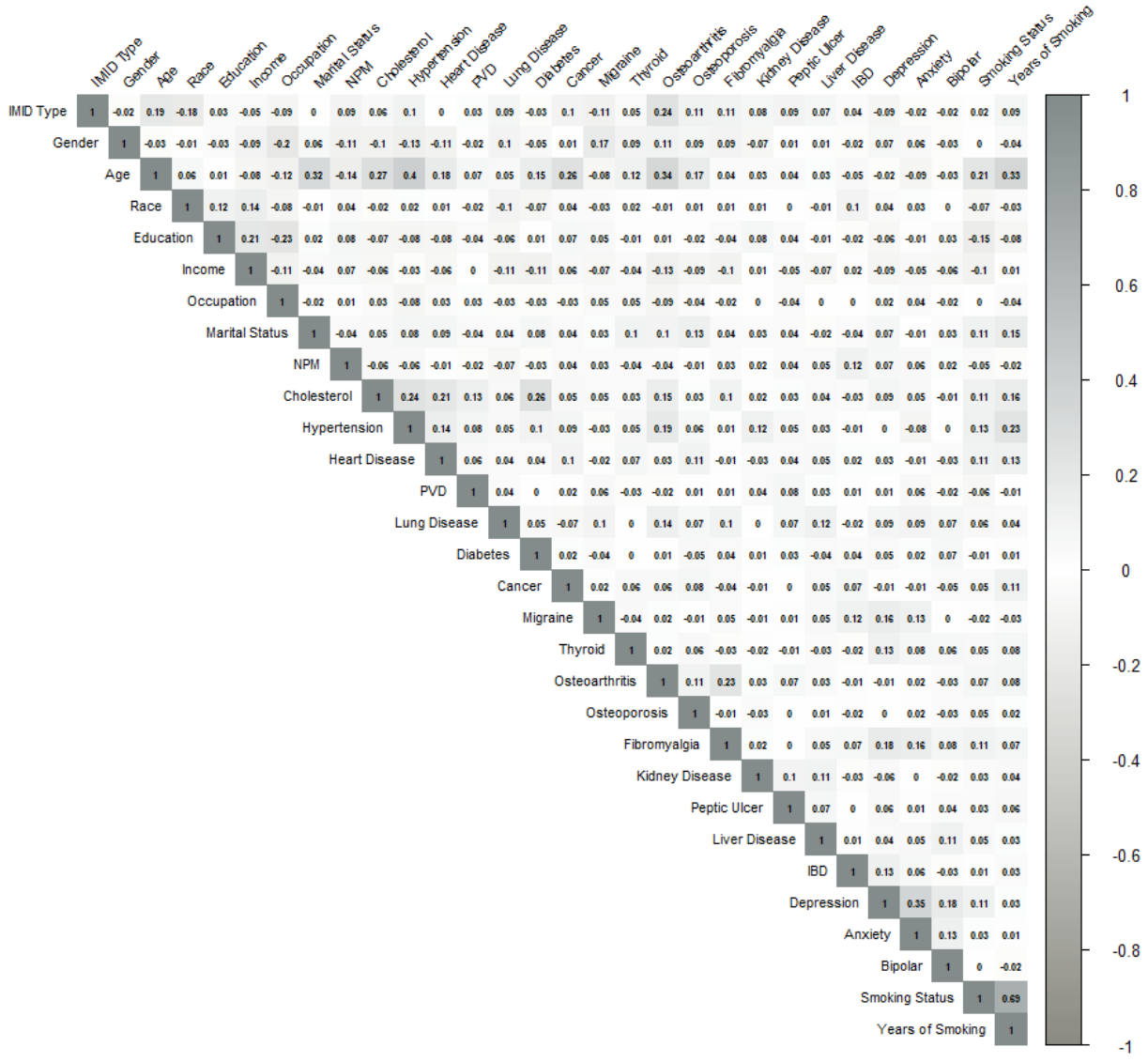
Covariate	N (%)
Demographic Covariates	
Female	438 (74.6)
Age*	50.5 ± 13.7
Race	
White	534 (91.0)
Non-white	53 (9.0)
Education	
Less than high school, High school diploma, GED	172 (29.3)
College, Technical, Trade, University bachelor's degree	367 (62.5)
University master's degree, University doctorate	48 (8.2)
Income	
Less than $15,000	52 (8.9)
$15,000 - $29,999	56 (9.5)
$30,000 - $49,999	98 (16.7)
$50,000 - $100,000	245 (41.7)
Over $100,000	136 (23.2)
Marital Status	
Single (never married)	91 (15.5)
Married/Common law	388 (66.1)
Divorced/Widowed/Separated	108 (18.4)
Occupation	
Professional	353 (60.1)
Other	234 (39.9)
Clinical Covariates	
Immune-mediated inflammatory disease	
Multiple sclerosis	221 (37.6)
Inflammatory bowel disease	224 (38.2)
Rheumatoid arthritis	142 (24.2)
Smoker	345 (58.8)
Number of prescribed medications	
0	233 (39.7)
1	297 (50.6)
2	57 (9.7)
Number of comorbid conditions[†]	2 (1 – 9)
Presence of Health Condition	
Anxiety	97 (16.5)
Bipolar	12 (2.0)
Cancer	53 (9.0)
Cholesterol	106 (18.1)
Depression	198 (33.7)
Diabetes	35 (6.0)
Fibromyalgia	28 (4.8)
Heart disease	28 (4.8)
Hypertension	151 (25.7)
Irritable Bowel Syndrome	86 (14.7)

Kidney disease	13 (2.2)
Liver disease	18 (3.1)
Lung disease	90 (15.3)
Migraine	100 (17.0)
Osteoarthritis	94 (16.0)
Osteoporosis	26 (4.4)
Peptic ulcer	21 (3.6)
Peripheral vascular disease	17 (2.9)
Thyroid	65 (11.1)

*Values reported are the mean and standard deviation
†Values reported are median and range

The Spearman correlation coefficients between covariates are presented in Figure 4-7. Age showed a positive correlation with hypertension ($r = 0.40$). Other correlations were low and negligible ($r < 0.39$), suggesting limited multicollinearity among the covariates (75). NCC was excluded from the correlation matrix because it is obviously positively correlated with the covariates from which it is derived.

Figure 4-8. Spearman correlation heatmap between covariates.



IMID type: Immune-mediated inflammatory diseases type; NPM: Number of prescribed medications; NCC: number of comorbid conditions; PVD: Peripheral vascular disease; IBS: Irritable bowel syndrome.

Table 4-8 summarizes the frequency distribution of responses to the SF-36 domain items. The proportion of participants reporting positive responses to the items was high. For example, 77.1% reported being “not limited at all” in response to the item “Limited in bathing or dressing yourself”.

Table 4-8. Frequency of responses for the SF-36 domain items.

Item	Response options N (%)					
Mental health (n = 586)						
	All of the time	**Most of the time**	**A good bit of the time**	**Some of the time**	**A little of the time**	**None of the time**
Been very nervous	7 (1.2)	17 (2.9)	43 (7.3)	81 (13.8)	190 (32.4)	248 (42.3)
Felt down in the dumps	6 (1.0)	14 (2.4)	26 (4.4)	52 (8.9)	95 (16.2)	393 (67.1)
Felt calm and peaceful	34 (5.8)	188 (32.1)	102 (17.4)	118 (20.1)	112 (19.1)	32 (5.5)
Felt downhearted and blue	7 (1.2)	24 (4.1)	53 (9.0)	92 (15.7)	187 (31.9)	223 (38.1)
Been happy	58 (9.9)	255 (43.5)	86 (14.7)	105 (17.9)	71 (12.1)	11 (1.9)
Physical functioning (n = 582)						
	Yes, limited a lot	**Yes, limited a little**	**Not limited at all**			
Vigorous activities	323 (55.5)	155 (26.6)	104 (17.9)			
Moderate activities	112 (19.2)	198 (34.0)	272 (46.7)			
Lifting or carrying groceries	78 (13.4)	177 (30.4)	327 (56.2)			
Climbing several flights of stairs	166 (28.5)	199 (34.2)	217 (37.3)			
Climbing one flight of stairs	75 (12.9)	128 (22.0)	379 (65.1)			
Bending, kneeling, or stooping	143 (24.6)	210 (36.1)	229 (39.3)			
Walking more than one mile	192 (33.0)	145 (24.9)	245 (42.1)			
Walking several blocks	140 (24.1)	136 (23.4)	306 (52.6)			
Walking one block	57 (9.8)	113 (19.4)	412 (70.8)			
Bathing or dressing	24 (4.1)	109 (18.7)	449 (77.1)			
General health (n = 584)						
	Excellent	**Very good**	**Good**	**Fair**	**Poor**	
Health in general	32 (5.5)	154 (26.4)	237 (40.6)	123 (21.1)	38 (6.5)	
	Definitely true	**Mostly true**	**Don't know**	**Mostly false**	**Definitely false**	
Get sick more easily than others	51 (8.7)	83 (14.2)	95 (16.3)	164 (28.1)	191 (32.7)	
As healthy as anybody I know	81 (13.9)	166 (28.4)	91 (15.6)	136 (23.3)	110 (18.8)	
Expect health to get worse	75 (12.8)	135 (23.1)	213 (36.5)	90 (15.4)	71 (12.2)	
My health is excellent	40 (6.8)	216 (37.0)	49 (8.4)	134 (22.9)	145 (24.8)	
Role limitation due to physical health (n = 586)						
	Yes	**No**				
Cut down time on work/activities	249 (42.5)	337 (57.5)				
Accomplished less	347 (59.2)	239 (40.8)				
Limited in work/activities	294 (50.2)	292 (49.8)				

Difficulty performing the work/activities	310 (52.9)	276 (47.1)				
Bodily pain (n = 583)						
	Not at all	**Slightly**	**Moderately**	**Quite a bit**	**Extremely**	
Intensity of bodily pain	84 (14.4)	220 (37.7)	147 (25.2)	119 (20.4)	13 (2.2)	
Extent pain interfere normal work	227 (38.9)	157 (26.9)	105 (18.0)	82 (14.1)	12 (2.1)	
Social functioning (n = 584)						
	Not at all	**Slightly**	**Moderately**	**Quite a bit**	**Extremely**	
Interferes with social activities	20 (3.4)	64 (11.0)	141 (24.1)	152 (26.0)	207 (35.4)	
	All of the time	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**	
Time interferes with social activities	208 (35.6)	202 (34.6)	97 (16.6)	67 (11.5)	10 (1.7)	
Role limitation due to emotional health (n = 587)						
	Yes	**No**				
Cut down time work/ activities	183 (31.2)	404 (68.8)				
Accomplished less	266 (45.3)	321 (54.7)				
Not as careful in work/activities	185 (31.5)	402 (68.5)				
Vitality (n = 584)						
	All of the time	**Most of the time**	**A good bit of the time**	**Some of the time**	**A little of the time**	**None of the time**
Felt full of pep	9 (1.5)	122 (20.9)	98 (16.8)	153 (26.2)	135 (23.1)	67 (11.5)
Had a lot of energy	11 (1.9)	103 (17.6)	109 (18.7)	150 (25.7)	132 (22.6)	79 (13.5)
Felt worn out	27 (4.6)	81 (13.9)	116 (19.9)	151 (25.9)	144 (24.7)	65 (11.1)
Felt tired	71 (12.2)	120 (20.5)	128 (21.9)	141 (24.1)	100 (17.1)	24 (4.1)

4.2.2. Checking Model Assumptions

Figures S-3 to S-10 show the relationship between continuous covariates (i.e., age, NCC, and sum score) and the logit of binary responses for the SF-36 domain items. The corresponding plots are provided in the Supplementary Material section.

The one-dimensional latent trait structure models obtained from CFA on the SF-36 domains demonstrated an acceptable fit to the data. The model fit statistics for each domain are presented in Table 4-8. GH demonstrated excellent fit (CFI = 0.99, RMSEA = 0.06, SRMR = 0.02). RP showed acceptable fit in two criteria (CFI = 0.96, RMSEA = 0.20, SRMR = 0.03). In contrast, MH, PF, and VT exhibited suboptimal fit, with CFI values below 0.90 and RMSEA values exceeding 0.15. CFA did not converge for the BP, SF, and RE domains due to the limited number of items for model identification. Therefore, the results should be interpreted with caution if evidence of DIF is found in these domains.

Table 4-9. Confirmatory factor analysis for assessing unidimensionality of the SF-36 domains.

SF-36 Domain	Model fit statistics			
	$\chi^2(df, p)$	CFI	RMSEA	SRMR
Mental health	161.6 (5, <0.001)	0.89	0.23	0.05
Physical functioning	696.3 (355, <0.001)	0.87	0.18	0.05
General health	18.6 (5, 0.002)	0.99	0.06	0.02
Role limitation due to physical health	1161.3 (6, <0.001)	0.96	0.20	0.03
Bodily pain	NA	NA	NA	NA
Social functioning	NA	NA	NA	NA
Role limitation due to emotional health	NA	NA	NA	NA
Vitality	206.6 (2, <0.001)	0.87	0.42	0.06

4.2.3. Covariates Associated with DIF

Table 4-9 summarizes the covariates associated with DIF in the SF-36 items, as identified by LASSO and IFT regression models. At the item-level, IFT identified nine items with at least one covariate associated with DIF, and LASSO identified 16 items. Although LASSO detected more overall associations (i.e., covariates associated with DIF), both methods identified similar covariates within each item. Among comorbid conditions, hypertension was the most frequently observed covariate, associated with three items. Sex was the most frequent demographic characteristic associated with three items.

Table S-1 in the supplementary material presents the covariates associated with DIF under the alternative dichotomization approach. IFT detected four items with DIF: Two items in the PF domain and two items in the SF domain. LASSO detected nine items with DIF: one item in the MH domain, seven items in the PF domain, and one item in the GH domain. Less agreement for identifying the DIF items and associated covariates between IFT and LASSO was observed. Marital status and IMID type were the most frequent covariates associated with DIF. Table S-2, in the supplementary material, shows the covariates associated with DIF identified by other post-covariate selection methods (i.e., SI, VIP, and NBQ) for LASSO regression.

Table 4-10. Associated covariate(s) with DIF in the SF-36 items identified by LASSO and IFT regression models.

Item	Associated Covariate(s) with DIF	
	IFT	LASSO
Mental health		
Been very nervous	–*	Migraine
Felt down in the dumps	–	–
Felt calm and peaceful	Hypertension	Hypertension
Felt downhearted and blue	Hypertension	–
Been happy	–	–
Physical functioning		
Vigorous activities	–	Osteoporosis
Moderate activities	–	–
Lifting or carrying groceries	–	–
Climbing several flights of stairs	Sex	Sex
Climbing one flight of stairs	–	Peptic
Bending, kneeling, or stooping	Age	Age
Walking more than one mile	–	
Walking several blocks	Osteoarthritis Thyroid	Osteoarthritis Income
Walking one block	–	–
Bathing or dressing	–	Sex Income
General health		
Health in general	–	Diabetes Peptic
Get sick more easily than others	Age	Age Number of prescribed medications
As healthy as anybody I know	–	–
Expect health to get worse	–	–
My health is excellent	–	–
Role limitation due to physical health		
Cut down time on work/activities	Income	Income
Accomplished less	Cancer	Cancer
Limited in work/activities	–	Sex Depression
Difficulty performing the work/activities	–	Hypertension
Bodily pain		
Intensity of bodily pain	–	–
Extent pain interfere normal work	–	–
Social functioning		
Interferes with social activities	–	–
Time interferes with social activities	–	Number of prescribed medications
Role limitation due to emotional health		
Cut down time work/ activities	–	–
Accomplished less	–	Education
Not as careful in work/activities	–	–

Vitality		
Felt full of pep	–	–
Had a lot of energy	–	–
Felt worn out	–	–
Felt tired	–	–

*No association was detected.

Table 4-11 presents the model summary results from the IFT regression model, showing the regression models (i.e., equation 4) for each node for items exhibiting DIF.

Table 4-11. Model summaries for DIF items in IFT

Item	Sub-group item difficulties
Mental health	
Felt calm and peaceful	$\eta_p = \beta_1 x_p - 19.3\, I(\text{Hypertension} = \text{no}) - 18.0\, I(\text{Hypertension} = \text{yes})$
Felt downhearted and blue	$\eta_p = \beta_1 x_p - 4.5\, I(\text{Hypertension} = \text{no}) - 6.5\, I(\text{Hypertension} = \text{yes})$
Physical functioning	
Climbing several flights of stairs	$\eta_p = \beta_1 x_p - 6.4\, I(\text{Sex} = \text{male}) - 7.6\, I(\text{Sex} = \text{female})$
Bending, kneeling, or stooping	$\eta_p = \beta_1 x_p - 3.4\, I(\text{Age} \leq 39) - 4.9\, I(\text{Age} > 39)$
Walking several blocks	$\eta_p = \beta_1 x_p - 6.9\, I(\text{Osteoarthritis} = \text{no}) - 5.0\, I(\text{Osteoarthritis} = \text{yes}) I(\text{Thyroid} = \text{no}) - 22.1(\text{Osteoarthritis} = \text{yes}) I(\text{Thyroid} = \text{yes})$
General health	
Get sick more easily than others	$\eta_p = \beta_1 x_p - 2.6\, I(\text{Age} \leq 30) - 1.2\, I(\text{Age} > 30)$
Role limitations due to physical health	
Cut down time on work/activities	$\eta_p = \beta_1 x_p - 2.8\, I(\text{Income} \leq \$100k) - 4.6\, I(\text{Income} > \$100k)$
Accomplished less	$\eta_p = \beta_1 x_p - 5.5\, I(\text{Cancer} = \text{no}) - 7.3\, I(\text{Cancer} = \text{yes})$

η_p is the logit of the item response for person p, β_1 is item discrimination, x_p is the total score for person p, $I(\cdot)$ is an indicator function

Table 4-11 summarizes the results from the LASSO regression model, including the estimated odds ratios for selected covariates, their corresponding NBQ with $\gamma = 0.1$, and the optimal value of λ_i for each item identified with DIF. According to Tables 4-10 and 4-11, the presence of hypertension in answering the item "felt calm and peaceful" decreased the probability of choosing positive responses. Also, women reported more difficulty in climbing several flights of stairs than men.

Table 4-12. Model summaries for LASSO

Item	Optimum lambda	Covariate (F/R): Odds ratio (90% NBQ CI)
Mental Health		
Been very nervous	0.01	Migraine (yes/no): 0.7 (0.2 – 0.9)
Felt calm and peaceful	0.01	Hypertension (yes/no): 2.0 (1.2 – 6.7)
Physical functioning		
Vigorous activities	0.02	Osteoporosis (yes/no): 0.2 (0.1 – 0.9)
Climbing several flights of stairs	0.01	Sex (female/male): 0.4 (0.2 – 0.7)
Climbing one flight of stairs	0.01	Peptic (yes/no): 0.3 (0.1 – 0.7)
Bending, kneeling, or stooping	0.01	Age (one year unit): 0.9 (0.8 – 0.9)
Walking several blocks	0.01	Osteoarthritis (yes/no): 1.9 (1.3 – 8.5) Income ($30k – $49k/less than $15k): 0.7 (0.1 – 0.9)
Bathing or dressing	0.01	Sex (female/male): 1.9 (1.1 – 4.6) Income ($15k – $29K/less than $15k): 0.5 (0.2 – 0.9)
General health		
Health in general	0.02	Diabetes (yes/no): 0.4 (0.1 – 0.9) Peptic (yes/no): 0.5 (0.1 – 0.8)
Get sick more easily than others	0.02	Age (one year unit): 1.1 (1.1 – 1.2) Number of prescribed medications (one unit increase): 0.4 (0.2 – 0.7)
Role limitation due to physical health		
Cut down time on work/activities	0.01	Income (over $100k/less than $15k): 0.4 (0.1 – 0.9)
Accomplished less	0.02	Cancer (yes/no): 0.4 (0.1 – 0.6)
Limited in work/activities	0.01	Sex (female/male): 0.5 (0.2 – 0.7) Depression (yes/no): 1.9 (1.2 – 6.6)
Difficulty performing the work/activities	0.02	Hypertension (yes/no): 0.6 (0.2 – 0.9)
Social functioning		
Time interferes with social activities	0.01	Number of prescribed medications (2/0): 0.4 (0.1 – 0.9)
Role limitation due to emotional health		
Accomplished less	0.01	Education (College, Technical, Trade, University bachelor's degree/Less than high school, High school diploma, GED): 0.5 (0.1 – 0.9)

4.2.4. Model Performance Evaluation

Table 4-10 presents the predictive performance of the IFT and LASSO models, evaluated by hold-out bootstrap CV across 500 replications. Overall, both methods demonstrated similar predictive accuracy on the test data.

Brier scores were below 0.12 with two exceptions: Item “Walking one block” in PF domain, where the Brier score was 0.28 (95% CI = 0.28-0.28) for IFT and 0.27 (95% CI = 0.27, 0.28) for LASSO; and item “Expect health to get worse” in the GH domain with Brier scores of 0.16 (95% CI = 0.16-0.16) for IFT and 0.15 (95% CI = 0.15-0.16) for LASSO. Accuracy estimates across items ranged from approximately 70% to over 95%. The lowest accuracy was observed for the item “Bathing or dressing” in the PH domain, with 73.01% (95% CI = 73.01%-73.02%) for IFT and 70.69% (95% CI = 70.68%- 70.69%) for LASSO. This item also showed the largest MSE with 0.27 (95% CI = 0.27-0.27) for IFT and 0.29 (95% CI = 0.29, 0.30) for LASSO.

Table 4-13. Hold-out bootstrap cross-validation results of the predictive performance of the IFT and LASSO models in the SF-36 items.

Item	IFT			LASSO		
	Brier score (95% CI)	Accuracy % (95% CI)	MSE (95% CI)	Brier score (95% CI)	Accuracy % (95% CI)	MSE (95% CI)
Mental Health						
Been very nervous	0.06 (0.06, 0.06)	87.34 (87.34, 87.35)	0.13 (0.12, 0.13)	0.06 (0.05, 0.06)	86.41 (86.41, 86.41)	0.14 (0.13, 0.14)
Felt down in the dumps	0.03 (0.03, 0.03)	92.05 (92.05, 92.05)	0.08 (0.08, 0.08)	0.03 (0.02, 0.03)	92.09 (92.09, 92.09)	0.08 (0.08, 0.08)
Felt calm and peaceful	0.05 (0.04, 0.05)	93.50 (93.49, 93.50)	0.07 (0.06, 0.07)	0.04 (0.04, 0.05)	93.74 (93.73, 93.74)	0.06 (0.06, 0.06)
Felt downhearted and blue	0.05 (0.04, 0.05)	91.74 (91.74, 91.74)	0.08 (0.08, 0.08)	0.05 (0.05, 0.05)	89.02 (89.02, 89.03)	0.11 (0.11, 0.11)
Been happy	0.06 (0.06, 0.06)	92.62 (92.62, 92.62)	0.07 (0.07, 0.08)	0.06 (0.06, 0.06)	90.74 (90.74, 90.74)	0.09 (0.09, 0.09)
Physical Functioning						
Vigorous activities	0.04 (0.03, 0.04)	94.75 (94.75, 94.75)	0.05 (0.05, 0.05)	0.04 (0.04, 0.04)	94.75 (94.75, 94.75)	0.05 (0.05, 0.05)
Moderate activities	0.08 (0.08, 0.08)	86.24 (86.24, 86.24)	0.14 (0.14, 0.14)	0.07 (0.07, 0.08)	86.41 (86.41, 86.41)	0.14 (0.13, 0.14)
Lifting or carrying groceries	0.08 (0.08, 0.08)	83.83 (83.83, 83.84)	0.16 (0.16, 0.16)	0.08 (0.08, 0.08)	82.65 (82.65, 82.65)	0.17 (0.17, 0.18)
Climbing several flights of stairs	0.08 (0.07, 0.08)	89.52 (89.51, 89.52)	0.10 (0.10, 0.11)	0.07 (0.07, 0.07)	90.40 (90.40, 90.41)	0.10 (0.09, 0.10)
Climbing one flight of stairs	0.06 (0.06, 0.06)	87.18 (87.18, 87.18)	0.13 (0.13, 0.13)	0.06 (0.06, 0.06)	86.42 (86.42, 86.43)	0.14 (0.13, 0.14)
Bending, kneeling, or stooping	0.16 (0.16, 0.16)	84.18 (84.17, 84.18)	0.16 (0.16, 0.16)	0.16 (0.16, 0.16)	83.89 (83.88, 83.89)	0.16 (0.16, 0.16)
Walking more than one mile	0.08 (0.08, 0.08)	88.62 (88.62, 88.62)	0.11 (0.11, 0.12)	0.09 (0.09, 0.09)	89.64 (89.63, 89.64)	0.10 (0.10, 0.11)
Walking several blocks	0.09 (0.09, 0.09)	90.46 (90.46, 90.47)	0.10 (0.09, 0.10)	0.10 (0.10, 0.10)	90.23 (90.22, 90.23)	0.10 (0.10, 0.10)
Walking one block	0.28 (0.28, 0.28)	85.48 (85.47, 85.48)	0.15 (0.14, 0.15)	0.27 (0.27, 0.28)	84.01 (84.01, 84.01)	0.16 (0.16, 0.16)
Bathing or dressing	0.09 (0.09, 0.10)	73.01 (73.01, 73.02)	0.27 (0.27, 0.27)	0.11 (0.11, 0.11)	70.69 (70.68, 70.69)	0.29 (0.29, 0.30)
General Physical Health						
Health in general	0.09 (0.09, 0.09)	85.32 (85.32, 85.32)	0.15 (0.14, 0.15)	0.08 (0.08, 0.08)	86.29 (86.28, 86.29)	0.14 (0.13, 0.14)
Get sick more easily than others	0.12 (0.12, 0.12)	75.47 (75.47, 75.47)	0.25 (0.24, 0.25)	0.11 (0.11, 0.12)	77.59 (77.59, 77.59)	0.22 (0.22, 0.23)
As healthy as anybody I know	0.10 (0.10, 0.11)	84.48 (84.47, 84.48)	0.16 (0.15, 0.16)	0.10 (0.10, 0.10)	85.84 (85.84, 85.84)	0.14 (0.14, 0.14)
Expect health to get worse	0.16 (0.16, 0.16)	73.58 (73.58, 73.59)	0.26 (0.26, 0.27)	0.15 (0.15, 0.16)	74.65 (74.65, 74.66)	0.25 (0.25, 0.26)
My health is excellent	0.07 (0.07, 0.07)	90.93 (90.92, 90.93)	0.09 (0.09, 0.09)	0.07 (0.07, 0.07)	90.83 (90.83, 90.83)	0.09 (0.09, 0.09)

Role Limitation Due to Physical Health						
Cut down time on work/activities	0.07 (0.07, 0.07)	88.37 (88.37, 88.37)	0.12 (0.11, 0.12)	0.06 (0.06, 0.06)	90.33 (90.33, 90.33)	0.10 (0.09, 0.10)
Accomplished less	0.07 (0.07, 0.07)	89.42 (89.42, 89.42)	0.11 (0.10, 0.11)	0.06 (0.06, 0.06)	90.64 (90.64, 90.64)	0.09 (0.09, 0.10)
Limited in work/activities	0.06 (0.06, 0.06)	91.47 (91.47, 91.47)	0.09 (0.08, 0.09)	0.05 (0.05, 0.05)	92.84 (92.84, 92.84)	0.07 (0.07, 0.07)
Difficulty performing the work/activities	0.07 (0.06, 0.07)	90.58 (90.58, 90.59)	0.09 (0.09, 0.10)	0.06 (0.06, 0.06)	91.44 (91.44, 91.45)	0.09 (0.08, 0.09)
Bodily Pain						
Intensity of bodily pain	0.01 (0.00, 0.02)	99.14 (99.13, 99.15)	0.01 (0.00, 0.02)	0.00 (0.00, 0.01)	99.14 (99.13, 99.15)	0.01 (0.00, 0.02)
Extent pain interfere normal work	0.01 (0.00, 0.02)	99.14 (99.13, 99.15)	0.01 (0.00, 0.02)	0.00 (0.00, 0.00)	100.00 (100.00, 100.00)	0.00 (0.00, 0.00)
Social Functioning						
Interferes with social activities	0.03 (0.03, 0.03)	93.45 (93.45, 93.45)	0.07 (0.06, 0.07)	0.03 (0.03, 0.03)	91.81 (91.81, 91.82)	0.08 (0.08, 0.08)
Time interferes with social activities	0.03 (0.03, 0.03)	94.66 (94.65, 94.66)	0.05 (0.05, 0.06)	0.03 (0.02, 0.03)	94.31 (94.31, 94.31)	0.06 (0.06, 0.06)
Role Limitation Due to Emotional Health						
Cut down time work/ activities	0.04 (0.04, 0.04)	93.46 (93.46, 93.46)	0.07 (0.06, 0.07)	0.04 (0.04, 0.04)	93.73 (93.73, 93.74)	0.06 (0.06, 0.06)
Accomplished less	0.03 (0.03, 0.03)	95.65 (95.65, 95.65)	0.04 (0.04, 0.04)	0.02 (0.02, 0.02)	96.77 (96.76, 96.77)	0.03 (0.03, 0.03)
Not as careful in work/activities	0.07 (0.07, 0.07)	88.31 (88.31, 88.31)	0.12 (0.11, 0.12)	0.06 (0.06, 0.06)	89.43 (89.42, 89.43)	0.11 (0.10, 0.11)
Vitality						
Felt full of pep	0.07 (0.07, 0.07)	91.39 (91.39, 91.39)	0.09 (0.08, 0.09)	0.07 (0.06, 0.07)	91.44 (91.44, 91.44)	0.09 (0.08, 0.09)
Had a lot of energy	0.05 (0.05, 0.06)	93.39 (93.39, 93.40)	0.07 (0.06, 0.07)	0.05 (0.05, 0.05)	92.97 (92.97, 92.97)	0.07 (0.07, 0.07)
Felt worn out	0.06 (0.06, 0.06)	90.77 (90.77, 90.77)	0.09 (0.09, 0.09)	0.06 (0.06, 0.06)	91.80 (91.80, 91.80)	0.08 (0.08, 0.08)
Felt tired	0.09 (0.09, 0.10)	86.37 (86.37, 86.37)	0.14 (0.13, 0.14)	0.09 (0.09, 0.09)	86.55 (86.55, 86.56)	0.13 (0.13, 0.14)

Chapter V: Discussion

5.1. Summary of Key Findings

This study investigated two ML LR-based methods to test for DIF on multiple covariates. A simulation study was conducted to compare Type I error rates and statistical power rates for DIF detection across different conditions of sample size, magnitude of correlation among covariates, and DIF effect sizes. A real-world data set comprised of individuals' responses to a general-purpose PROM, the SF-36, was analyzed to compare their consistency in detecting DIF on multiple sociodemographic and clinical covariates.

The simulation results highlighted the performance characteristics of each method. IFT consistently maintained item-level Type I error rates below the nominal α across all simulation conditions, while LASSO tended to show inflated error rates, particularly in conditions with no or weak covariate correlations. In conditions with strong correlation among covariates, LASSO controlled the item-level Type I error rate at the nominal α. Zou et al. (74) in their study about using Elastic Net for covariate selection indicated that LASSO arbitrarily selects one covariate from the set of correlated covariates and shrinks the others to zero, reducing model complexity (73). This behavior decreases item-level Type I error rates under the condition of strong correlation among covariates, but at the cost of reduced item-level power (74). The overall Type I error rate for both methods was below 1% in most conditions, showing that both methods were conservative in identifying the associated covariates to DIF.

The results of Type I error rates for the IFT regression model in the simulation study were consistent with Berger et al. (27), and Tutz et al. (26). The item-level Type I error rate reported by Berger et al. (27)—who tested for uniform DIF under conditions of 400 and 800 sample size, three covariates, and no correlation among covariates—ranged from 0.49 to 0.51, and the overall Type I error rate was below 0.01. Tutz et al. (26)—who tested for uniform DIF under conditions of a 500-sample size, four covariates, and no correlation among covariates—reported item-level Type I error rate ranging from 0.02 to 0.03 and the overall Type I error rate below 0.01. Although simulation conditions in the current study involved more complex conditions, such as a sample size of 100, seven covariates, and different correlation strengths

among covariates, item-level Type I error rates remained close to the nominal α, and overall Type I errors remained below 0.01.

Item-level Type I error rate inflation for LASSO in simulation studies has been highlighted in several studies. Bauer et al. (53)—who used LASSO to test for uniform DIF under conditions of 500 and 1000 sample sizes, and three correlated covariates—reported item-level Type I error rates as high as 0.24 in most conditions. The authors considered nonzero coefficients as evidence of DIF and evaluated a post-selection approach in which the selected model was re-estimated without penalty. P-values obtained from the refitted models were "biased" because they do not account for the data-driven variable selection performed by LASSO. Once LASSO identifies nonzero coefficients, re-estimating the model without penalty and computing p-values as if the predictors had been specified a priori leads to overly optimistic significance levels. They used the biased p-values from the refitted model to determine the significance of the nonzero coefficients. Belzak et al. (76), who simulated data with sample sizes of 250, 500, and 1000 and one covariate, reported item-level Type I error rate below the nominal α, except for one condition of 500 sample size (i.e., 0.11). They selected the optimum LASSO model using the minimum Bayesian Information Criteria (i.e., BIC) and followed Belzak et al. (76) to obtain p-values. This approach appeared to control the Type I error rate at the nominal α. However, their simulation results were based on one covariate in the models; they expanded the results of the simulation study to real-world data, applying the method for models with multiple covariates. These studies highlight the sensitivity of the LASSO regression in detecting DIF to the simulation conditions and post-covariate selection approaches.

As expected, by increasing the sample size and DIF effect magnitude, item-level and overall power for both methods increased, consistent with previous studies (26,27,53,76). Item-level and overall power for LASSO were slightly higher than IFT under weak (i.e., 0.4) and moderate (i.e., 0.8) DIF effects. Overall power for both methods was less than item-level power in their corresponding simulation conditions. When there was a strong correlation amongst covariates (i.e., $r = 0.75$), IFT showed greater item-level and overall power than LASSO, outperforming in detecting DIF items and associated covariates. Item-level and overall power for both methods decreased slightly when the percentage of DIF items increased from 25% to 50%.

However, the comparative pattern and relative advantage of IFT and LASSO remained consistent across conditions.

Low item-level power rates for IFT under weak and moderate DIF effects in the simulation study were consistent with those found by Berger et al. (27) and Tutz et al. (26). Overall power for IFT was less than item-level power in their corresponding simulation conditions, consistent with Tutz et al. (26). However, Berger et al. (27) found overall power similar to item-level power in all corresponding simulation conditions. This inconsistency may reflect the greater complexity of the simulation design that I considered, which included seven covariates and two binary covariates sampled from a $B(0.3)$ distribution. Consistent with Berger et al. (27) and Tutz et al. (26), item-level and overall power rates for IFT decreased slightly when the percentage of DIF items increased from 25% to 50%.

The item-level power results for LASSO decreased slightly when the percentage of DIF items increased from 25% to 50%, consistent with Bauer et al. (53) and Belzak et al. (76). No study reported overall power rates; therefore, it was not possible to compare this measure with those used in previous research. Other patterns, such as the effect of correlation among covariates, cannot be compared, as the authors of previous studies did not consider this condition. The simulation study results from the current study and relevant studies highlight the effect of simulation conditions, such as the number of DIF items, the magnitude of DIF effects, correlation among covariates, and post-covariate selection approaches, on the item-level power rates for LASSO regression.

In the analysis of real-world data, LASSO flagged more items with evidence of DIF than IFT. This was consistent with the expectations based on the simulation results, where LASSO showed greater power rates in identifying small and moderate DIF effects. Where both methods identified an item with DIF, there was an agreement in terms of the associated covariate and direction of DIF. However, compared to IFT, LASSO identified more than one covariate associated with DIF. Sex and hypertension were the most frequent covariates associated with DIF across all items. The majority of the items in the MH, PF and RP domains showed evidence of DIF across more than one demographic covariate and comorbid condition covariate in this data. Some of the DIF effects observed for the PF domain items were consistent with findings in

previous studies by Lix et al. (44) and Fan et al. (77). Women tended to have more difficulties in climbing several flights of stairs and fewer problems in bathing or dressing than men (44,77). This study revealed that comorbid conditions, including hypertension, osteoporosis, peptic ulcer, cancer, depression, migraine, and thyroid, were associated with DIF. The results of this study and previous research about the SF-36 suggest that the SF-36 domain scores should account for potential DIF effects across chronic disease populations. Moreover, the association of comorbid conditions with DIF should be investigated when testing for DIF, especially in populations with diverse health statuses.

5.2. Study Strengths and Limitations

The main strength of this study was using LR-based ML methods to test for DIF on multiple covariates. The methods were advantageous because they can be used to investigate DIF on continuous covariates, as they do not require specifying a cut-point a priori (5,13,14,52). These methods enabled the capture of complex relationships without being strict on statistical assumptions, such as the assumption of a linear relationship between the log-odds of the outcome and covariates (16,35). The design of the simulation study allowed for controlled experimentation and manipulation of different types of covariates, including continuous and categorical, with different sample sizes, strength of correlations, and DIF effects; these features enhanced the relevance of the findings of the simulation study to the real-world study (31,76). Moreover, the real-world dataset contained a large set of demographic and clinical characteristics, such as comorbid health conditions (12,14,44,52). All domains of the SF-36 were assessed for the presence of DIF, whereas in the literature, similar studies have often focused only on the MH and PF domains of the SF-36 (14,31).

This study had some limitations. I did not test for non-uniform DIF (i.e. testing the interaction between each covariate and the total score). Theoretically, the methods can be extended to test for non-uniform DIF. Previous research tested for uniform DIF in simulation studies (27,76). Also, evidence of uniform DIF has been found in real-world studies, but there is less evidence of non-uniform DIF (12,44,77). The simulated data were generated in controlled settings; for example, I assumed the presence of uniform DIF but not non-uniform DIF. The number of covariates and their correlations were also controlled. PROM data may exhibit more complex DIF patterns, including nonlinearity. Therefore, extrapolating simulation findings to

real-world data and interpretations should be done cautiously. While dichotomizing ordinal outcome measures when applying methods for detecting DIF simplifies the analysis and facilitates interpretation, it comes at the cost of potentially reducing power and increasing vulnerability of the results to the influence of arbitrary cut points (78,79). By computing total scores after dichotomizing items, variation in the underlying latent construct was reduced, which may potentially introduce bias in the models. The CFA results indicated that several SF-36 domains exhibited suboptimal model fit. Moreover, CFA cannot be applied to domains with three items or fewer than three items (e.g., BP) because of model identification issues. Therefore, interpreting DIF findings for these domains should be done with caution. The study's ability to detect significant effects or differences might have been constrained by the sample size, limiting confidence in the study's conclusions and generalizability of the findings to other cohorts (14,24). Also, BMI was not captured in this dataset; obesity has been found in other studies to be associated with DIF in the SF-36 items (44). In this study, Spearman correlation coefficients were used as a simplified measure of correlation across all covariate types. While this approach provides a practical, nonparametric summary of the relationships, it does not fully account for differences in covariate types. Although tetrachoric correlation coefficients are well-known for assessing correlation for binary covariates, Khamis et. al (79) suggested Kendall's and phi coefficients for mixtures of nominal and binary covariates (80).

5.3. Future Research

The scope of previous research has primarily involved comparing ML approaches—such as tree-based methods (e.g., IFT) or regularization methods (e.g., LASSO)—with their underlying theoretical framework (i.e., IRT or LR models) (14,26,76). Future research could broaden this comparison by investigating tree-based methods and regularization approaches to test for DIF in polytomous items. For example, Partial Credit Models, which are ML IRT-based models, have been applied in tree-based DIF detection by Bollmann et al. (28) and in regularization approaches by Schauberger et al. (15), both showing Type I error rates close to the nominal α, and improvement in power rates. These models can be compared with LR-based models for polytomous items, such as the Elastic Net regularized ordinal LR approach used by Ebrahimi et al. (14).

The performance of the IFT model under correlated covariates could be investigated. According to the simulation results in the current study, IFT performs well under correlated covariates. Previous research has primarily relied on regularization approaches to address the challenges that may arise from correlated covariates (15).

The methods in previous and current studies analyzed each item independently. In practice, items belonging to the same domain might not be independent. Most PROMs, including the SF-36, contain item clusters within each domain that share similar content. The presence of correlation amongst the items could result in increased Type I error and reduce the power of the methods to detect DIF. Huang et al. (81) in a simulation study about LASSO regression to test for DIF on dependent items found that this model could control Type I error rates and had higher power to detect DIF than the LR model. Future research should compare the performance of the IFT and LASSO regression models in the presence of correlated items (26,30,81).

Primary DIF detection methods were first developed to test for uniform DIF; extensions to test for non-uniform DIF have been less frequently investigated because they require larger sample sizes (24). Future research could extend IFT and LASSO regression models to test for non-uniform DIF. These methods are capable of exploring complex relationships amongst covariates.

The current study used the sum score as a proxy measure of the latent construct. Sum scores assume all items contribute equally to the latent construct. They are observed indicators subject to measurement error and may not fully capture the latent construct. If the SF-36 domain is multidimensional (i.e., does not measure a single latent construct) or has a small set of items, the sum score may not fully control for true differences in the latent construct in the regression model, leading to measurement error that could mask the presence of DIF (9,82). Scott et al. (9) suggested using IRT-based scoring methods to estimate the latent construct score for use in the LR model. Future studies could compare models that use the sum score to measure the latent construct with a latent variable model to estimate the latent construct score (9,82).

In the real-world dataset, some covariates (e.g., bipolar disorder) had very small numbers of observations, which could potentially affect the stability and precision of the DIF effect estimates. In future studies on cohorts containing a large number of covariates, IFT and LASSO regression models could be used to select the most relevant covariates and reduce model

complexity before testing for DIF. Sanchez-Pinto et al. (23) in their comparison of different covariate selection methods, suggested using tree-based methods when the overall sample size is large, and using regression-based methods for smaller cohorts.

Missing values may be related to health status, disease type, age, or education rather than being completely at random. Therefore, removing missing values from the analysis can bias the results and affect the generalizability of the study findings. However, I found that there were a few missing values in the real-world dataset; exclusion of these values would have a negligible impact on the study findings.

5.4. Study Significance

In conclusion, this research aimed to guide on selection of DIF detection methods for PROMs before their use in practical settings. The presence of MI ensures that PROMs reflect the true level of latent construct across groups rather than being biased, improving equity-focused PROM development.

Based on the findings of this study, IFT is recommended when controlling false positives is of interest, as it maintains false positive rates close to the nominal α, consistent with Berger et al. (27). However, this came at the expense of reduced power, particularly for weak and moderate DIF effects. LASSO is recommended when the small DIF effects may affect the decisions, and some increase in false positives is acceptable, as it demonstrated higher power in those conditions, but at the cost of inflated Type I error rates, in line with Bauer et al. (53) and Belzak et al. (76).

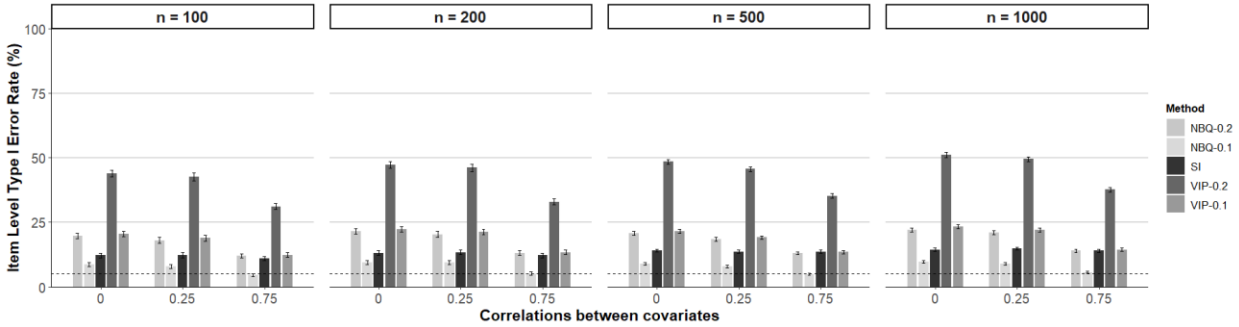
In practice, IFT regression might be preferable to LASSO regression due to its interpretability and ease of implementation. LASSO, on the other hand, is sensitive to the choice of model selection criteria (e.g., BIC or k-fold CV), post-covariate selection approaches, and the arbitrary thresholding used in NBQ and VIP approaches.

One notable strength of DIF analysis is that by identifying specific items within a PROM instrument that exhibit DIF, clinicians gain information about potential disparities in how individuals from different groups respond to those items. This knowledge may empower them to consider modifying or tailoring the content or meaning of items, ensuring that PROMs meet the

needs of all individuals, ultimately improving the quality of care provided (26–28,31). However, removing an item with DIF within a PROM is not a preferred strategy since it might affect the comparability of domain scores across different populations and the content validity of well-established and well-known PROMs such as the SF-36. In practice, it is useful to compare PROM scores with and without DIF adjustment as a sensitivity analysis. DIF-adjusted scores can be generated by recalculating domain scores after removing or down-weighting items that show DIF. The impact of DIF can then be evaluated by comparing unadjusted and adjusted scores. If the two scores are similar, DIF may have limited practical impact on the interpretation of study findings; if differences are large, DIF-adjusted scoring may be warranted (12,44).

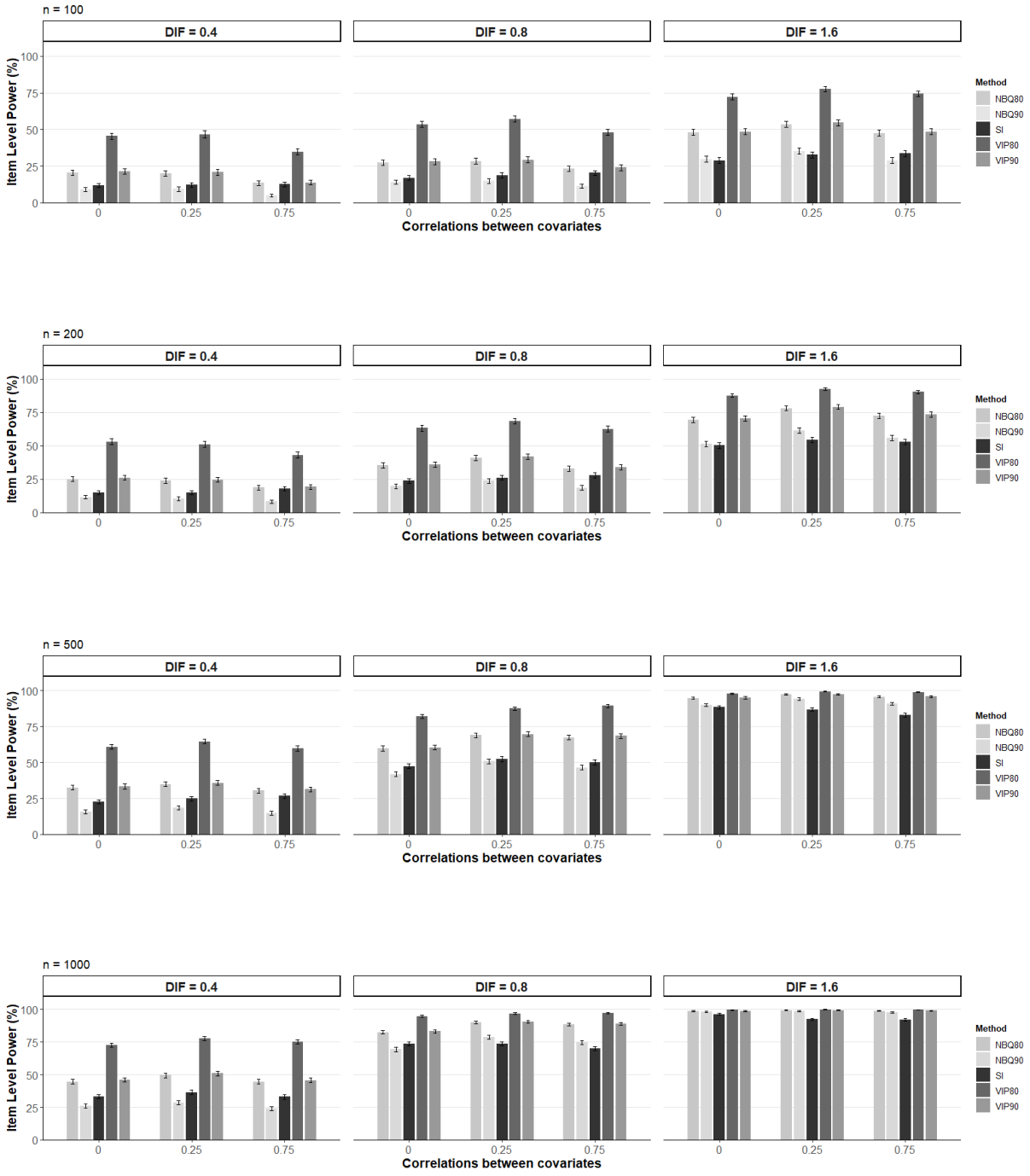
Supplementary Material

Figure S-1. Item-level Type I error rate estimates (%) for post covariate selection approaches, stratified by sample size and correlations between covariates.



SI= Selective inference, VIP-0.2= Variable inclusion probability with 0.2 threshold, VIP-0.1= Variable inclusion probability with 0.1 threshold, NBQ-0.2= non-parametric bootstrap quantiles with 0.2 threshold, NBQ-0.1= non-parametric bootstrap quantiles with 0.1 threshold.

Figure S-2. Item-level power rate estimates (%) for post covariate selection approaches, DIF items = 25%, stratified by sample size and correlations between covariates



SI= Selective inference, VIP-0.2= Variable inclusion probability with 0.2 threshold, VIP-0.1= Variable inclusion probability with 0.1 threshold, NBQ-0.2= non-parametric bootstrap quantiles with 0.2 threshold, NBQ-0.1= non-parametric bootstrap quantiles with 0.1 threshold.

Figure S-3. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 mental health items.

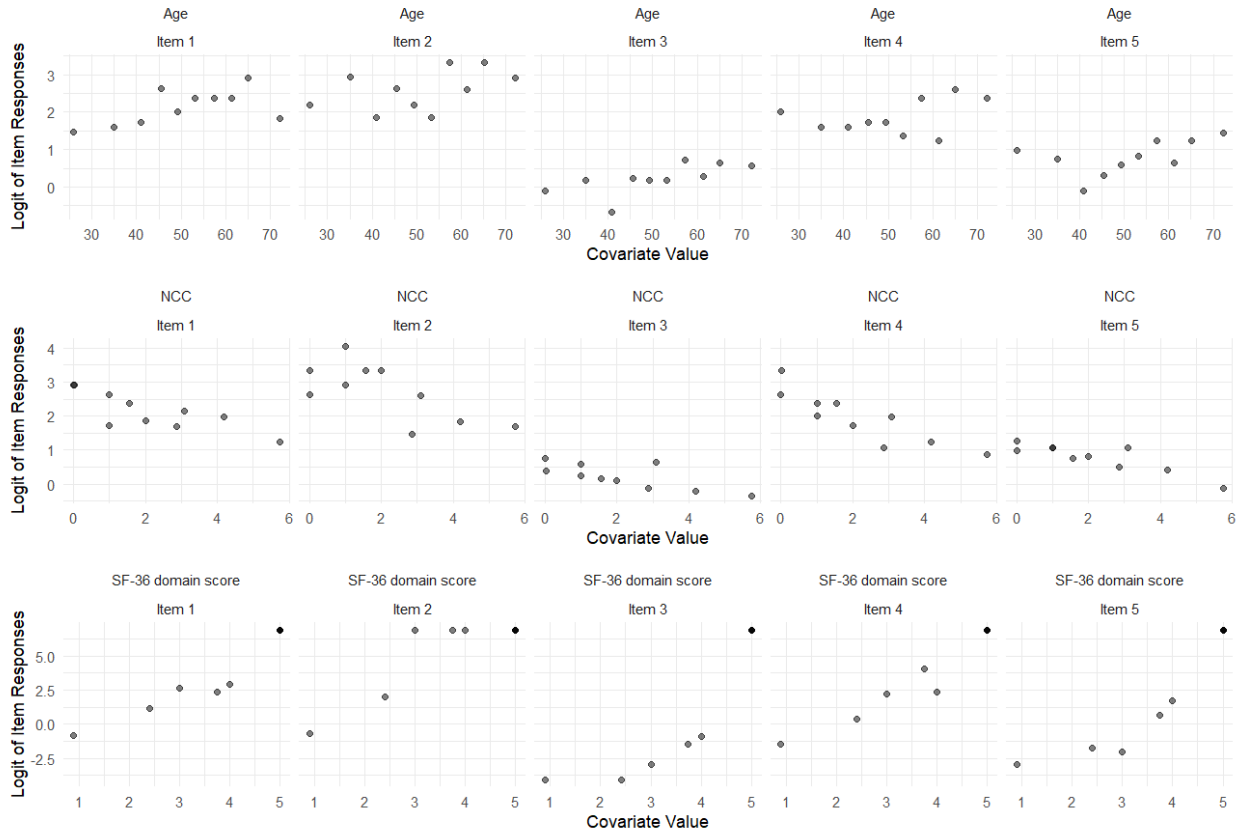


Figure S-4. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 physical functioning items.

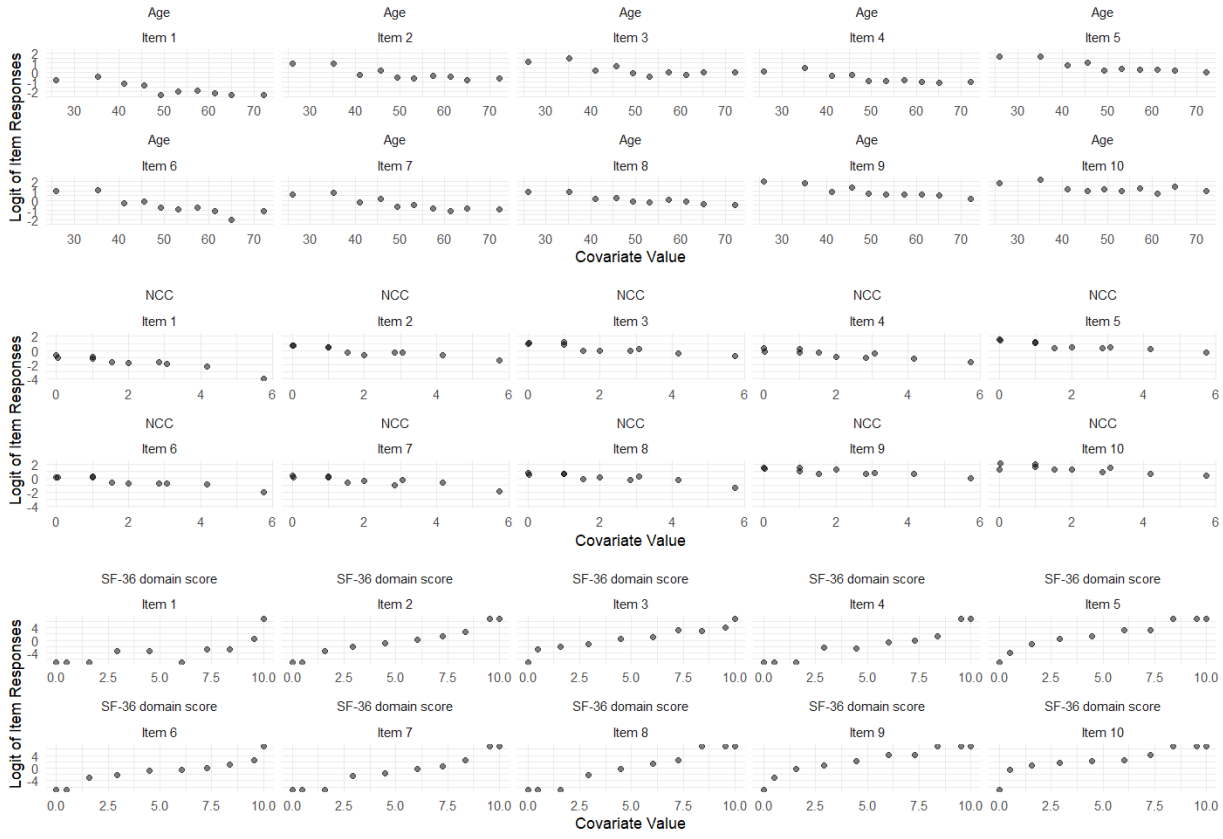


Figure S-5. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 general health items.

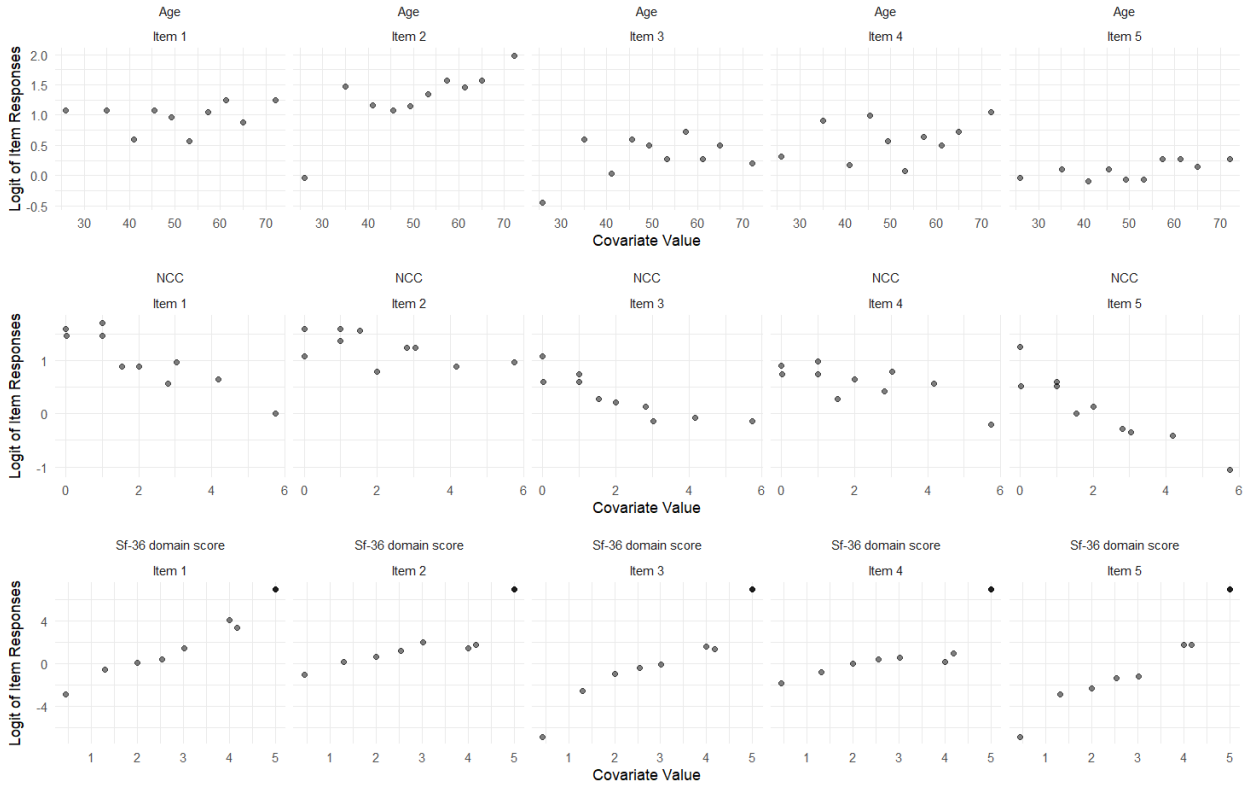


Figure S-6. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 role limitation due to physical health items.

Figure S-7. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 bodily pain items.

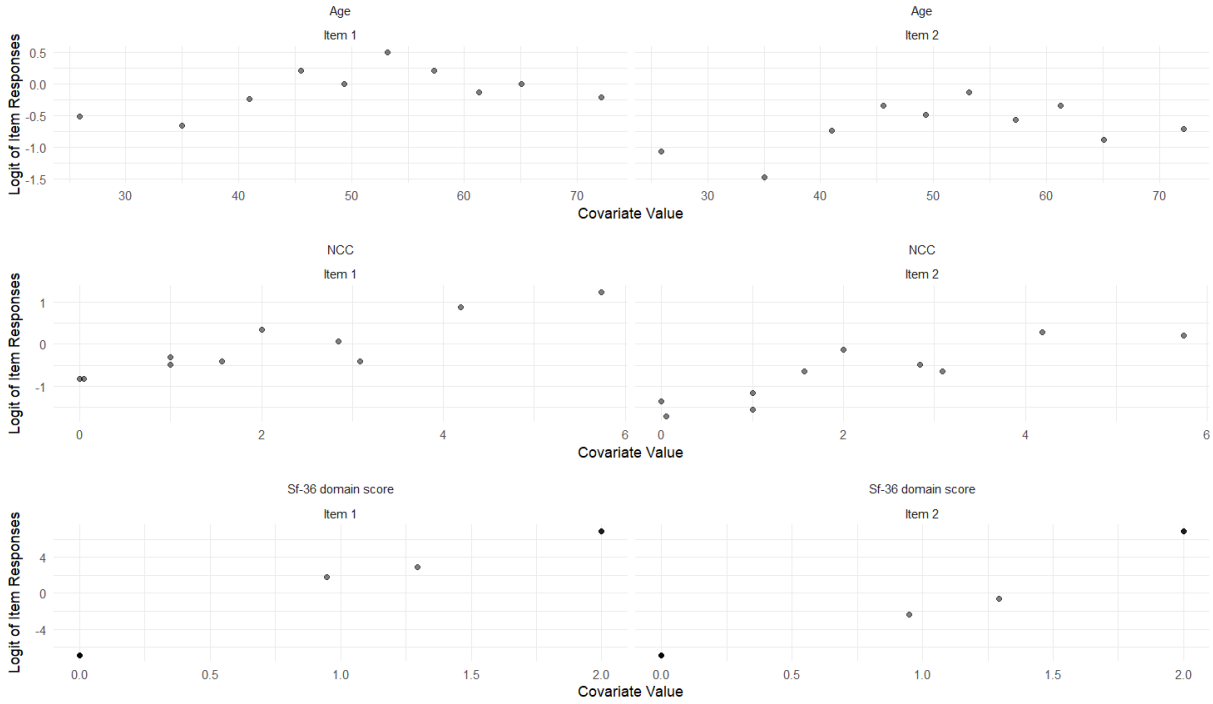


Figure S-8. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 social functioning items.

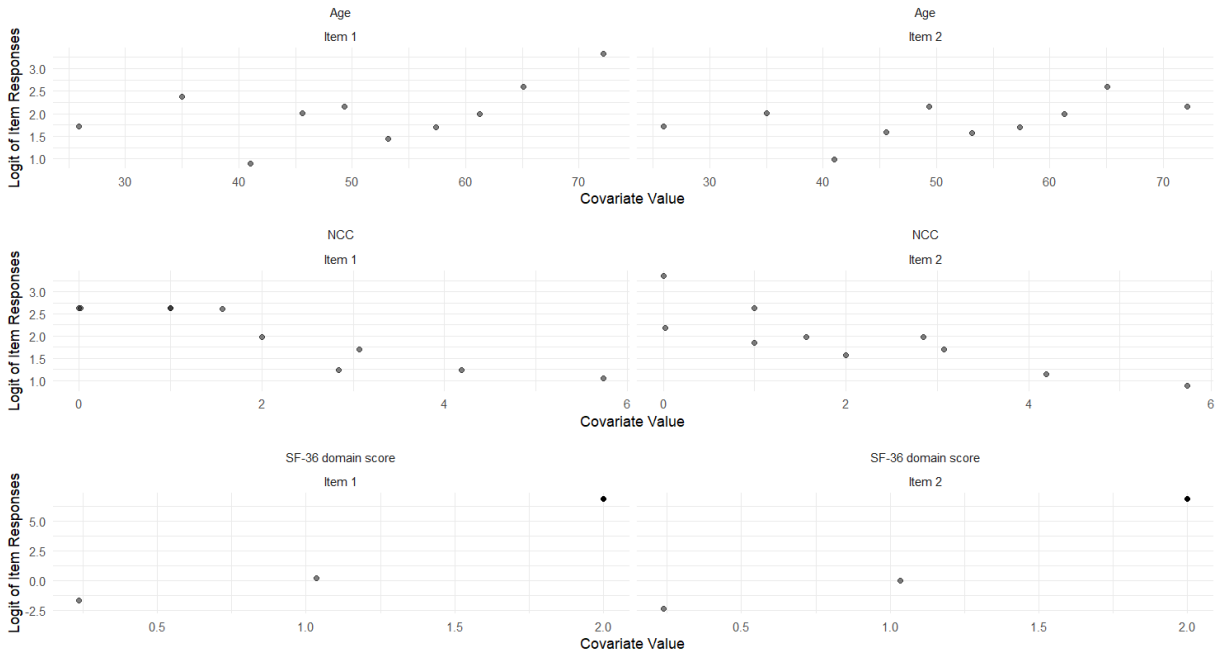


Figure S-9. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 role limitation due to emotional health items.

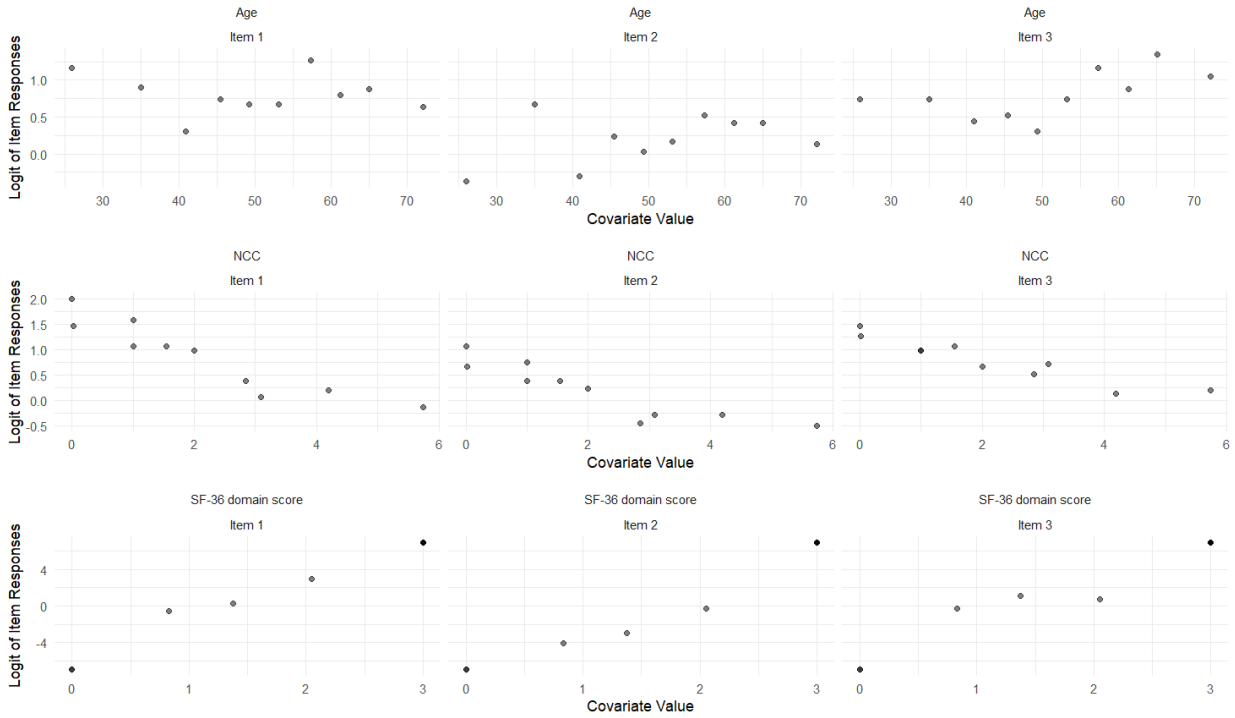


Figure S-10. Checking for linearity assumption between age, number of comorbid conditions (NCC), and SF-36 total score and the logit of binary item responses for the SF-36 vitality items.

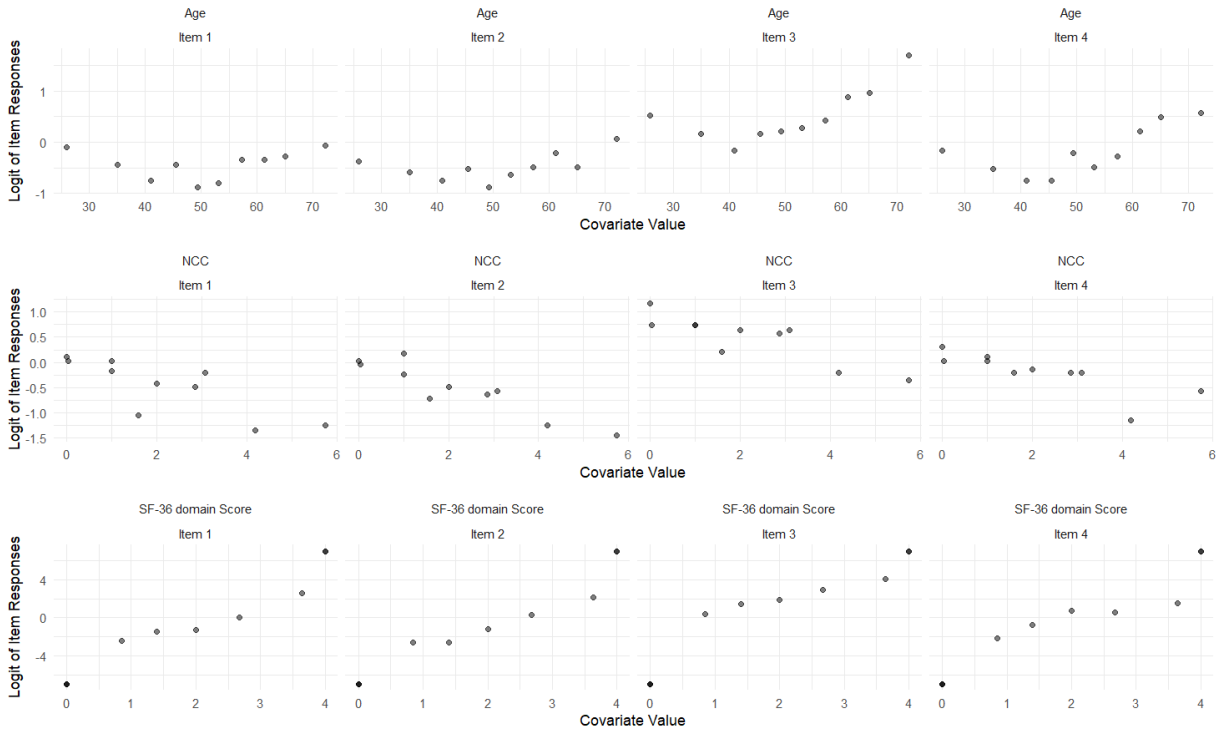


Table S-1. Covariates associated with DIF as identified by LASSO and IFT regression models based on an alternative approach to dichotomize SF-36 item responses

Item	IFT	LASSO
Mental Health		
Been very nervous	–	–
Felt down in the dumps	–	Cholesterol
Felt calm and peaceful	–	–
Felt downhearted and blue	–	–
Been happy	–	–
Physical Functioning		
Vigorous activities	–	Age
Moderate activities	–	Lung disease
Lifting or carrying groceries	–	Osteoporosis
Climbing several flights of stairs	IMID type	IMID type
Climbing one flight of stairs	–	–
Bending, kneeling, or stooping	IMID type	IMID type
Walking more than one mile	–	IMID type
Walking several blocks	–	IMID type
Walking one block	–	–
Bathing or dressing	–	–
General Health		
Health in general	–	–
Get sick more easily than others	–	–
As healthy as anybody I know	–	–
Expect health to get worse	–	Age
My health is excellent	–	–
Bodily Pain		
Intensity of bodily pain	–	–
Extent pain interfere normal work	–	–
Social Functioning		
Interferes with social activities	Marital status	–
Time interferes with social activities	Marital status Bipolar	–
Vitality		
Felt full of pep	–	–
Had a lot of energy	–	–
Felt worn out	–	–
Felt tired	–	–

Note: Role limitation due to physical health and Role limitation due to emotional health were already binary

Table S-2. Covariates associated with DIF as identified by different LASSO post covariate selection approaches.

Item	SI	VIP ($\gamma = 0.2$)	VIP ($\gamma = 0.1$)	NBQ ($\gamma = 0.2$)	NBQ ($\gamma = 0.1$)
Mental Health					
Been very nervous	Migraine	Income Migraine Osteoarthritis	Migraine	Migraine	Migraine
Felt down in the dumps	–	Thyroid	–	–	
Felt calm and peaceful	Hypertension	Hypertension Income Irritable bowel syndrome	Hypertension	Hypertension	Hypertension
Felt downhearted and blue		Hypertension Osteoarthritis	Hypertension	Hypertension	
Been happy	Thyroid	Thyroid Heart disease Income Race	Thyroid	Thyroid	
Physical Functioning					
Vigorous activities	–	–	–	–	
Moderate activities	Osteoporosis	Osteoporosis Cholesterol Marital status	Osteoporosis	Osteoporosis	
Lifting or carrying groceries	–	Osteoporosis Race	–	–	
Climbing several flights of stairs	–	IMID type Sex Irritable bowel syndrome	IMID type	IMID type	IMID type
Climbing one flight of stairs	Sex	Osteoporosis Sex Income Number of prescribed medications	Osteoporosis Sex Income	Osteoporosis Sex Income	Sex
Bending, kneeling, or stooping	Peptic	Peptic Osteoporosis Income IMID type	Peptic Osteoporosis Income	Peptic Osteoporosis Income	Peptic
Walking more than one mile	Age	Sex Lung disease IMID type Smoking status Age Osteoarthritis	Age	Age	Age
Walking several blocks	Income	Income Irritable bowel syndrome Sex IMID type	Income	Income	

Walking one block	–	Osteoarthritis Income Osteoporosis	Income Osteoarthritis	Osteoarthritis Income	Osteoarthritis Income
Bathing or dressing	–	Irritable bowel syndrome Number of prescribed medications	–	–	
Vigorous activities	Income Sex	Income Sex Race IMID type Migraine Age	Income Sex IMID type	Income Sex IMID type	Income Sex
General Health					
Health in general	Diabetes Peptic	Diabetes Peptic Income Osteoporosis Number of prescribed medications Marital status	Diabetes Peptic	Diabetes Peptic	Diabetes Peptic
Get sick more easily than others	Number of prescribed medications Cholesterol IMID type Peptic Age	Number of prescribed medications Cholesterol IMID type Peptic Lung disease Age	Number of prescribed medications IMID type Lung disease Age	Number of prescribed medications IMID type Lung disease Age	Number of prescribed medications Age
As healthy as anybody I know	–	Income Heart disease	Income	Income	
Expect health to get worse	–	IMID type Osteoarthritis Irritable bowel syndrome Smoking status	IMID type	IMID type	
My health is excellent	–	Depression	Depression	Depression	
Role Limitation Due to Physical Health					
Cut down time on work/activities	Income IMID type	Peptic Income IMID type Education Age	Peptic Income IMID type Age	Peptic Income IMID type Education Age	Income
Accomplished less	Cancer	Cancer Depression	Cancer	Cancer	Cancer
Limited in work/activities	Sex Depression	Sex Depression Osteoarthritis Fibromyalgia	Sex Depression	Sex Depression	Sex Depression

		Education Age			
Difficulty performing the work/activities	IMID type	IMID type Income Hypertension Fibromyalgia	IMID type Hypertension	IMID type Hypertension	Hypertension
Bodily Pain					
Intensity of bodily pain	–	Diabetes Bipolar	–	–	
Extent pain interfere normal work	–	Diabetes Bipolar	–	–	
Social Functioning					
Interferes with social activities	–	Peripheral vascular disease Number of prescribed medications Thyroid	Number of prescribed medications	Number of prescribed medications	
Time interferes with social activities	–	Peripheral vascular disease Number of prescribed medications Thyroid	Number of prescribed medications	Number of prescribed medications	Number of prescribed medications
Role Limitation Due to Emotional Health					
Cut down time work/ activities	Hypertension Income NCC	Hypertension	–	–	
Accomplished less	–	Education Occupation Marital status Cancer Lung disease Smoker	Education Occupation Cancer	Education Occupation Cancer	Education
Not as careful in work/activities	Hypertension	Hypertension Occupation Cancer	Occupation	Occupation	
Vitality					
Felt full of pep	–	Age	Age	Age	
Had a lot of energy	–	Osteoporosis		–	
Felt worn out	Hypertension	Osteoporosis Hypertension Cholesterol Occupation Age	Osteoporosis Hypertension Age	Osteoporosis Hypertension Age	
Felt tired	–	Age	–	–	

References

1. Weldring T, Smith SMS. Article commentary: Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights. 2013;6:61–8.
2. Fayers PM, Machin D. Quality of life : assessment, analysis, and interpretation. John Wiley; 2000. 404 p.
3. Parker DJ, Werth PM, Christensen DD, Jevsevar DS. Differential item functioning to validate setting of delivery compatibility in PROMIS-global health. Qual Life Res. 2022;31(7):2189–200.
4. Kim M, Fong J, Pusic AL, Fischer JP, Mehrara BJ, Nelson JA. Editorial: maintaining the integrity of PROMs in research and practice. Ann Surg Oncol. 2023;30(7):3879–81.
5. Komboz B, Strobl C, Zeileis A. Tree-based global model tests for polytomous Rasch models. Educ Psychol Meas. 2018;78(1):128–66.
6. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Stat Med. 2004 Jan 30;23(2):241–56.
7. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res. 2007;16:33–42.
8. El Y, Editor M. Patient Reported Outcome Measures in Rheumatic Diseases. Springer International Publishing Switzerland; 2016. 449 p.
9. Scott NW, Fayers PM, Aaronson NK, Bottomley A, De Graeff A, Groenvold M, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes. 2010;8:81.
10. Lai JS, Teresi J, Gershon R. Procedures for the analysis of differential item functioning (DIF) for small sample sizes. Eval Health Prof. 2005;28(3):283–94.
11. Bellavia A, Rotem RS, Dickerson AS, Hansen J, Gredal O, Weisskopf MG. The use of logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis. Epidemiol Methods. 2020;9(1).
12. Yadegari I, Bohm E, Ayilara OF, Zhang L, Sawatzky R, Sajobi TT, et al. Differential item functioning of the SF-12 in a population-based regional joint replacement registry. Health Qual Life Outcomes. 2019;17(1).
13. Guo F, Min H, Jex S, Choi Y. Old enough to perceive things differently? Detecting measurement invariance across age groups using item-focused tree. Work Aging Retire. 2023;9(1):59–70.
14. Ebrahimi V, Bagheri Z, Shayan Z, Jafari P. A machine learning approach to assess differential item functioning in psychometric questionnaires using the Elastic Net

regularized ordinal logistic regression in small sample size groups. Biomed Res Int. 2021;2021.

15. Schauberger G, Mair P. A regularization approach for the detection of differential item functioning in generalized partial credit models. Behav Res Methods. 2020;52(1):279–94.

16. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14:91–118.

17. French AW, Miller TR. Logistic regression and its use in detecting differential item functioning in polytomous items national council on measurement in education. J Educ Meas. 1996;33(3):315–32.

18. Rivera AJ, Muñoz JC, Pérez-Goody MD, de San Pedro BS, Charte F, Elizondo D, et al. XAIRE: An ensemble-based methodology for determining the relative importance of variables in regression tasks. Application to a hospital emergency department. Artif Intell Med. 2023;137:102494.

19. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12–8.

20. Berrío ÁI, Gómez-Benito J, Guilera G. Differential item functioning in the WHODAS 2.0 scale in schizophrenia: an application of the Rasch trees method based on demographic and clinical covariates. Assessment. 2022;29(8):1858–68.

21. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Care. 2000;38(9):28–42.

22. Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential Item Functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Med Care. 2006;44(11):115–23.

23. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform. 2018;116:10–7.

24. Lee S. Detecting differential item functioning using the logistic regression procedure in small samples. Appl Psychol Meas. 2017;41(1):30–43.

25. Strobl C, Kopf J, Zeileis A. Rasch trees: a new method for detecting differential item functioning in the Rasch model. Psychometrika. 2015;80(2):289–316.

26. Tutz G, Berger M. Item focused trees for the identification of items in differential item functioning. Psychometrika. 2015;81:727–50.

27. Berger M, Tutz G. Detection of uniform and nonuniform differential item functioning by item-focused trees. Journal of Educational and Behavioral Statistics. 2016;41(6):559–92.

28. Bollmann S, Berger M, Tutz G. Item-focused trees for the detection of differential item functioning in partial credit models. Educ Psychol Meas. 2018;78(5):781–804.

29. Liang X, Jacobucci R. Regularized structural equation modeling to detect measurement bias: evaluation of Lasso, adaptive Lasso, and Elastic Net. Structural Equation Modeling. 2020;27(5):722–34.

30. Wang C, Zhu R, Xu G. Using Lasso and adaptive Lasso to identify DIF in multidimensional 2PL models. Multivariate Behav Res. 2023;58(2):387–407.

31. Magis D, Tuerlinckx F, De Boeck P. Detection of differential item functioning using the Lasso approach. J Educ Behav Stat. 2015;40(2):111–35.

32. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.

33. Singal AG, Mukherjee A, Joseph Elmunzer B, Higgins PDR, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108(11):1723–30.

34. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artif Intell Med. 2017;90:1–14.

35. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol. 2018;18(1):179.

36. Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14(1).

37. Cuevas M, Cervantes VH. Differential item functioning detection with logistic regression. Mathématiques et sciences humaines. 2012;(199):45–59.

38. Van Den Noortgate W, De Boeck P. Assessing and explaining differential item functioning using logistic mixed models. J Educ Behav Stat. 2005;30(4):443–64.

39. Tonidandel S, LeBreton JM. Determining the relative importance of predictors in logistic regression: an extension of relative weight analysis. Organ Res Methods. 2010;13(4):767–81.

40. Cortina JM, Landis RS. When small effect sizes tell a big story, and when large effect sizes don’t. Statistical and methodological myths and urban legends; 2009. 287–308 p.

41. Abram S V., Helwig NE, Moodie CA, DeYoung CG, MacDonald AW, Waller NG. Bootstrap enhanced penalized regression for variable selection with neuroimaging data. Front Neurosci. 2016;10(1):344.

42. Crano WD, Brewer MB. Principles and methods of social research. Lawrence Erlbaum Associates; 2002. 416 p.

43. Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Q. 2008;50(4):538.

44. Lix LM, Wu X, Hopman W, Mayo N, Sajobi TT, Liu J, et al. Differential item functioning in the SF-36 physical functioning and mental health sub scales: A population-based investigation in the Canadian Multicentre Osteoporosis Study. PLoS One. 2016;11(3):e0151519.

45. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14(4):323–48.

46. Dallmeijer A, Dekker J, Roorda L, Knol D, van Baalen B, de Groot V, et al. Differential item functioning of the functional independence measure in higher performing neurological patients. J Rehabil Med. 2005;37(6):346–52.

47. Dallmeijer AJ, de Groot V, Roorda LD, Schepers VPM, Lindeman E, van den Berg LH, et al. Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: A study using Rasch analysis. J Rehabil Med. 2007 Mar;39(2):163–9.

48. Steultjens MPM, Stolwijk-Swüste J, Roorda LD, Dallmeijer AJ, Van Dijk GM, Post B, et al. WOMAC-pf as a measure of physical function in patients with Parkinson's disease and late-onset sequels of poliomyelitis: Unidimensionality and item behaviour. Disabil Rehabil. 2012;34(17):1423–30.

49. Waller NG, Compas BE, Hollon SD, Beckjord E. Measurement of depressive symptoms in women with breast cancer and women with clinical depression: A differential item functioning analysis. J Clin Psychol Med Settings. 2005 Jun;12(2):127–41.

50. Smith AB, Cocks K, Parry D, Taylor M. A Differential Item Functioning Analysis of the EQ-5D in Cancer. Value in Health. 2016 Dec 1;19(8):1063–7.

51. Pollard B, Johnston M, Dixon D. Exploring differential item functioning in the SF-36 by demographic, clinical, psychological and social factors in an osteoarthritis population. BMC Musculoskelet Disord. 2013;(14):346.

52. Jafari P, Mehrabani-Zeinabad K, Javadi S, Ghanizadeh A, Bagheri Z. A machine learning approach to assess differential item functioning of the KINDL quality of life questionnaire across children with and without ADHD. Child Psychiatry Hum Dev. 2022;53(5):980–91.

53. Bauer DJ, Belzak WCM, Cole VT. Simplifying the assessment of measurement invariance over multiple background variables: using regularized moderated nonlinear factor analysis to detect differential item functioning. Structural Equation Modeling. 2020;27(1):43–55.

54. Taylor J, Tibshirani R. Post-selection inference for ℓ1-penalized likelihood models. Canadian Journal of Statistics. 2018 Mar 1;46(1):41–61.

55. Meinshausen N, Bühlmann P. Stability selection. J R Statist Soc B. 2010;(72):417–73.

56. Zhang Z, He Z, Qin Y, Shen Y, Shia BC, Li Y. Variable selection with scalable bootstrapping in generalized linear model for massive data. Journal of Data Science. 2023 Jan 1;21(1):87–105.

57. Kammer M, Dunkler D, Michiels S, Heinze G. Evaluating methods for Lasso selective inference in biomedical research: a comparative simulation study. BMC Med Res Methodol. 2022 Dec 1;22(1).

58. Wang S, Nan B, Rosset S, Zhu J. Random lasso. Annals of Applied Statistics. 2011;5(1):468–85.

59. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. Ann Stat. 2016 Jun 1;44(3):907–27.

60. Chen M. Generating nonnegatively correlated binary random variates. Stata J. 2015;15(1):301–8.

61. Sargent RG. Verification and validation of simulation models. Journal of Simulation. 2013;7(1):12–24.

62. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006 Dec 30;25(24):4279–92.

63. Bagheri Z, Jafari P, Mahmoodi M, Hossein M. Testing whether patients with diabetes and healthy people perceive the meaning of the items in the Persian version of the SF-36 questionnaire similarly: a differential item functioning analysis. Qual Life Res. 2017;26(4):835–45.

64. Fan Yu Y, Peng Yu A, Ahn J. Investigating differential item functioning by chronic diseases in the SF-36 health survey a latent trait analysis using MIMIC models. Med Care. 2007;45:851–9.

65. Tennenhouse LG, Marrie RA, Bernstein CN, Lix LM. Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease. J Psychosom Res. 2020;134:110126.

66. Ware JE. SF-36 health survey update. Spine (Phila Pa 1976). 2000;25(24):3130–9.

67. Perkins AJ, Stump TE, Monahan PO, Mchorney CA. Assessment of Differential Item Functioning for demographic comparisons in the MOS SF-36 health survey. Research. 2006;15(3):331–48.

68. Joelson A, Strömqvist F, Sigmundsson FG, Karlsson J. Single item self-rated general health: SF-36 based observations from 16,910 spine surgery procedures. Quality of Life Research. 2022 Jun 1;31(6):1819–28.

69. Grassi M, Nucera A, Zanolin E, Omenaas E, Anto JM, Leynaert B. Performance comparison of Likert and binary formats of SF-36 version 1.6 across ECRHS II adults populations. Value in Health. 2007;10(6):478–88.

70. Montoya AK, Edwards MC. The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. Educ Psychol Meas. 2021;81(3):413–40.

71. Bami -Zahra, Bami Bamiz Z. A new flexible train-test split algorithm, an approach for choosing among the hold-out, k-fold cross-validation, and hold-out iteration. ArXiv. 2025;(2501):06492.

72. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.

73. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. JSS Journal of Statistical Software [Internet]. 2010;33(1):1–22. Available from: http://www.jstatsoft.org/

74. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Source: Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67(2):301–20.

75. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal. 2012;24(3):69–71.

76. Belzak WCM, Bauer DJ. Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. Psychol Methods. 2020;25(6):673–90.

77. Fan Y, Peng Yu A, Ahn J. Investigating differential item functioning by chronic diseases in the SF-36 health survey: A latent trait analysis using MIMIC models. Care. 2007;45(9):851–9.

78. Sankey SS, Weissfeld LA. A study of the effect of dichotomizing ordinal data upon modeling. Communications in Statistics Part B: Simulation and Computation. 1998;27(4):871–87.

79. Irwin JR, Mcclelland GH. Negative consequences of dichotomizing continuous predictor variables. J Mark Res. 2003;40(3):366–71.

80. Grønneberg S, Moss J, Foldnes N. Partial Identification of Latent Correlations with Binary Data. Psychometrika. 2020 Dec 1;85(4):1028–51.

81. Huang J, Miller MD, Huggins-Manley AC, Leite WL, Knopf HT, Ritzhaupt AD. Evaluating the performance of a regularized Differential Item Functioning method for testlet-based polytomous Items. Educ Psychol Meas. 2025;85(6):1180–99.

82. Widaman KF, Revelle W. Thinking about sum scores yet again, baybe the last time, we don’t know, oh no..1: A comment on McNeish (2023). Educ Psychol Meas. 2024;84(4):637–59.