Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 18

4-14-2020

Investigating the Performance of Propensity Score Approaches for Differential Item Functioning Analysis

Yan Liu
University of British Columbia, yan.liu@ubc.ca

Chanmin Kim
Boston University, chanmink@bu.edu

Amrey D. Wu
University of British Columbia, amery.wu@ubc.ca

Paul Gustafson
University of British Columbia, gustaf@stat.ubc.ca

Edward Kroc
University of British Columbia, ekroc@stat.ubc.ca

See next page for additional authors

Recommended Citation

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., Kroc, E., & Zumbo, B. D. (2019). Investigating the performance of propensity score approaches for differential item functioning analysis. Journal of Modern Applied Statistical Methods, 18(1), eP2744. doi: 10.22237/jmasm/1556669280

Investigating the Performance of Propensity Score Approaches for Differential Item Functioning Analysis

Cover Page Footnote

This research was partly funded by the UBC Paragon Research Initiative.

Authors

Yan Liu, Chanmin Kim, Amrey D. Wu, Paul Gustafson, Edward Kroc, and Bruno D. Zumbo

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2744.
doi: 10.22237/jmasm/1556669280

בס"ד

ISSN 1538 − 9472

Investigating the Performance of Propensity Score Approaches for Differential Item Functioning Analysis

Yan Liu
University of British Columbia
Vancouver, British Columbia

Chanmin Kim
Boston University
Boston, MA

Amery D. Wu
University of British Columbia
Vancouver, British Columbia

Paul Gustafson
University of British Columbia
Vancouver, British Columbia

Edward Kroc
University of British Columbia
Vancouver, British Columbia

Bruno D. Zumbo
University of British Columbia
Vancouver, British Columbia

To evaluate the performance of propensity score approaches for differential item functioning analysis, this simulation study was conducted to assess bias, mean square error, Type I error, and power under different levels of effect size and a variety of model misspecification conditions, including different types and missing patterns of covariates.

Keywords: Propensity score, differential item functioning (DIF), collapsibility, model misspecification, logistic regression, conditional logistic regression

Introduction

The major advantage of randomized experimental study designs over quasi-experimental or observational designs is that random assignment tends to make treatment groups comparable, i.e., balanced over both observed and unobserved covariates. However, randomized experiments are not always feasible or ethical in many fields, and so quasi-experimental or observational designs are widely used instead. In order to approximate causal inferences, propensity score matching has been recommended and applied in medical, epidemiological and economic research, and these methods have lately been extended to social, psychological and educational research (e.g., Austin, 2008; Guo & Fraser, 2010; Hong & Raudenbush, 2005; Thoemmes & Kim, 2011).

doi: 10.22237/jmasm/1556669280 | Accepted: August 19, 2018; Published: April 14, 2020.
Correspondence: Bruno D. Zumbo, bruno.zumbo@ubc.ca

The popularity of propensity score methods has given rise to the application of propensity score in differential item functioning (DIF) analysis. There are a few studies that have recommended and demonstrated the application of the propensity score approach in DIF analysis (Joldersma & Bowen, 2010; Bowen, 2011; Lee & Geisinger, 2014; Liu, et al., 2016). However, none of these studies have systematically investigated under what conditions and to what degree propensity score DIF methods perform better than conventional DIF methods. This paper aims to address this current gap in the literature.

The purpose of this study is to compare the performance of DIF analysis methods based on propensity score approaches with that of conventional logistic regression DIF analysis under different levels of effect size and in the presence of different selections of covariates and a variety of model misspecification conditions. In addition, logistic regression DIF analysis with covariance adjustment is also included for comparison as it is an alternative method to matching. Covariance adjustment regression analysis allows one to include confounders and, hence, to adjust for the confounding effects in DIF analysis. However, this method may not be able to give a reliable adjustment for the differences in the observed covariates when there are substantial differences in the distribution of the covariates between the two groups (Cochran, 1957; Rubin, 2001). A detailed description can be found in Liu et al. (2016). The paper is organized as follows: (i) a review of propensity score matching and two important issues related to its application, (ii) a review of previous studies on the application of propensity score in DIF analysis, (iii) a brief description of logistic regression DIF analysis, (iv) a description of a Monte Carlo simulation study comparing propensity score DIF analysis methods with the conventional logistic regression DIF analysis, and (v) conclusion and discussion.

Review on Propensity Score and Two Important Issues

Propensity Score

Propensity score matching was first proposed by Rosenbaum and Rubin (1983). The propensity score is defined as the conditional probability of assigning an individual to the treatment condition given a set of observed covariates. Symbolically, this is

$$\mathrm{e}\left(\mathbf{X}_i\right)=\mathrm{P}\left(Z_i=1|\,\mathbf{X}_i\right),$$

where $\mathbf{X}_i$ is a vector of scores on the observed covariates, e($\mathbf{X}_i$) denotes the propensity score for each individual i; Z_i is an indicator for grouping variable/treatment conditions, and $Z_i = 1$ refers to participants belonging to the treatment group or the focal group in the DIF context, whereas $Z_i = 0$ refers to participants belonging to the control group or the reference group in the DIF context (Rosenbaum & Rubin, 1983). The propensity scores are usually estimated via logistic regression

$$\mathrm{P}\left(Z_i = 1 \mid \mathbf{X}\right) = \frac{\mathrm{e}^{\beta_0 + \boldsymbol{\beta}(\mathbf{X})}}{1 + \mathrm{e}^{\beta_0 + \boldsymbol{\beta}(\mathbf{X})}}, \tag{1}$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is a vector of coefficients on the covariates, and $\mathbf{X}$ is a vector of scores on the observed covariates (e.g., Rosenbaum, 2010, p. 167; Rosenbaum & Rubin, 1983).

Propensity score matching is used to approximate a randomized experimental study by reducing the pre-existing group differences in the data collected from quasi-experimental or observational studies. Propensity score methods can help to balance the characteristics of non-equivalent groups, so that two subgroups with the same propensity score values have the same distribution on observed covariates (e.g., Rosenbaum, 1995, 2002, 2010; Rosenbaum & Rubin, 1983, 1985; Rubin, 2001; Schafer & Kang, 2008). In order to solve the sparseness problem raised by exact matching methods, the propensity score method creates a single composite score from all observed covariates and hence observations from two groups can be matched on one single score alone.

A variety of propensity score methods have been developed. In the present study, optimal pair and full matching as well as stratification methods were chosen for the DIF analyses since they are commonly used in practice and are readily available for implementation in R packages, such as MatchIt (Ho et al., 2011). More detailed information about these methods can be found in the books by Guo and Fraser (2014) and Pan and Bai (2015), as well as the paper by Liu et al. (2016).

Two Important Issues

Strong Ignorability of Treatment Assignment. Propensity score matching is a widely used matching method, possibly even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010, p. 114). However, propensity score approaches have a crucial assumption, strong ignorability of treatment assignment, which is hardly, if ever, fully met in practice. This

assumption states that the treatment assignment and observed outcome(s) are conditionally independent after controlling for the effects of a collection of observed covariates that determine the assignment mechanism (Rosenbaum & Rubin, 1983).

In order to make a causal claim, the strong ignorability of treatment assignment assumption has to be met. This can only be achieved when the treatment assignment mechanism is fully explained by the observed covariates; under this condition, selection bias can be completely removed. In practice, researchers rarely know whether the observed covariates satisfy this assumption. Hence, model misspecification is always a potential issue for propensity score methods, understood as either when some covariates are missing from the analysis, or when the functional form describing the relationship of the covariates to the treatment assignment is misspecified.

Covariate Selection. The selection of covariates is a crucial step in any observational design as the selection has a major impact on how well propensity scores uncover the unknown mechanism of self-selection into groups. In practice, however, it is rare that researchers know the selection process exactly; more often they are confronted with decisions regarding which covariates to include from a huge pool of candidates (e.g., Steiner et al., 2010). One recommendation is to simply include every available covariate for the propensity score estimation (e.g., D'Agostino, 1998; Zigler & Dominici, 2014). It is not known if these variables are really unrelated to the treatment assignment. The redundant or irrelevant covariates may cause some modeling problems, such as multicollinearity which can result in an inflation of standard errors of regression coefficients. However, the biased standard errors do not affect propensity score estimation, so the inclusion of redundant or irrelevant covariates should not be a matter of concern (e.g., Blackstone, 2002), and including all available covariates in the model is better than omitting some covariates.

Some researchers, however, have started to challenge this recommendation and have shown that including all available covariates into a model may create problems for propensity score estimation and bias the final conclusions (e.g., Brookhart et al., 2006; Cuong, 2013; Zhao, 2008; Zigler & Dominici, 2014). For example, based on their simulation studies Brookhart et al. (2006) suggested that the inclusion of variables in the propensity score estimation, which are related to the exposure but not to the outcome, will increase the variance of the estimated exposure effect. Zhao (2008) found that over-parameterization can bias the parameter estimates in the final analysis. Cuong (2013) showed that the inclusion

of all covariates that were related to outcome or both outcome and grouping (assignment) variables improved the efficiency of the parameter estimate of grouping variable, but the inclusion of covariates that were only related to a grouping variable tended to increase the mean square error of that parameter estimate. Currently, exactly what kind of covariates should be included in the propensity score estimation phase is still a controversial issue.

Application of Propensity Score Methods in DIF

Whether a test is fair to all test takers in the target population is an essential issue in achievement, licensure, and credentialing examinations. For example, when developing or adapting a test to another language or cultural group, it is important to make sure that a comparison of test scores is meaningful. Various DIF methods have been developed to address this issue (e.g., Angoff, 1972, 1993; Holland & Thayer, 1988; Shepard, 1982; Swaminathan & Rogers, 1990; Zumbo, 1999, 2007). An item displays DIF when individuals from different groups do not have the same probability of getting the item right after matching on their ability or attribute of interest. After an item is identified as DIF, test developers or researchers need to decide whether the items should be removed from the test; this would be the case if the item indeed put one group at a disadvantage due to certain extraneous characteristics other than the test taker's ability or attribute (e.g., Ellis, 1989; Hambleton & Patsula, 1998; Wu & Ercikan, 2009).

The problem with conventional DIF analyses is that they can only detect DIF but cannot disentangle the sources of DIF since many confounders may covary with the outcome variable. Unlike randomized experimental studies, DIF studies are based on observational data and typically do not have equivalent groups before testing. For instance, researchers would not know if the DIF of an item were due to translation problems or other factors, such as students' learning motivation, self-confidence, parents' education, and social economic status. This is very common in educational or psychological settings because a lot of confounders covary with outcome variables. Hence, a typical DIF analysis cannot help test developers decide whether they should throw away an item flagged as DIF due to, for instance, problems in language translation.

Dorans and Holland (1993) suggested that propensity score matching might be a good solution instead of matching directly on multiple observed covariates. Joldersma and Bowen (2010) applied the propensity score approach to examine translation effects (English vs. Spanish) using Mantel-Haenszel DIF analysis. Bowen (2011) conducted a simulation study to compare the conventional Mantel-

Haenszel to Mantel-Haenszel DIF analyses based on the propensity score matched data and found that propensity score DIF analyses exhibited lower Type I error rates, but higher Type II error rates. However, this simulation study contained only one replication, manipulated only one matching factor (i.e., ability distribution differences), and matched only on one covariate (i.e., total test scores).

Lee and Geisinger (2014) compared the conventional DIF analyses with the propensity score approach for examining gender DIF using an empirical data set. Their study showed that the Mantel-Haenszel and logistic regression methods based on propensity scores detected a fewer number of gender DIF items than did the conventional Mantel-Haenszel and logistic regression methods. Liu et al. (2016) demonstrated the application of propensity score optimal matching and stratification in logistic regression DIF analyses using data from the Trends in International Mathematics and Science Study (TIMSS), and suggested that propensity score matching is a promising approach for studying causal DIF if the key covariates are collected and pre-test differences between groups can be balanced to a condition akin to a random assignment. However, most previous studies focus on either applications or demonstrations. Only one of them conducted a simulation study (Bowen, 2011). However, this study only simulated one replication per experimental condition and included only one matching variable. None of these studies systematically investigated under what conditions propensity scores DIF methods would perform better than conventional DIF methods in terms of both uniform and non-uniform DIF.

Logistic Regression DIF Analysis

Logistic regression as a test of DIF was proposed by Swaminathan and Rogers (1990) and has been highly recommended due to its flexibility as it can test both uniform and non-uniform DIF (Zumbo, 1999, 2007). Of course, other analytical methods can also be used for detecting DIF, but they are not generally as flexible as logistic regression DIF analysis. For example, alternatives to logistic regression such as item response theory and multiple indicators and multiple causes (MIMIC) model based on structural equation model framework usually require larger sample sizes; furthermore, the Mantel-Haenszel method is designed to detect uniform DIF. The more general logistic regression approach is adopted in this DIF simulation study.

In a conventional logistic regression DIF analysis, group, ability, and an interaction between group and ability are used to predict the probability of a correct

answer to an item (or endorsing that item) on a given sample. The model is specified as follows:

$$\ln \frac{\mathrm{p}\left(Y=1 \mid \mathrm{tot}, G\right)}{1-\mathrm{p}\left(Y=1 \mid \mathrm{tot}, G\right)}=b_0+b_1 \mathrm{tot}+b_2 G+b_3\left(\mathrm{tot} * G\right), \quad (2)$$

where p is the estimated probability for a participant to answer a particular item correctly or endorse that item; tot indicates the total test score for each participant, which is used as a proxy for ability; *G* is the dummy coded grouping variable (0 = reference group, 1 = focal group); and tot***G* indicates the interaction between these two. The coefficient b_1 indicates the relation between a person's total test score and the score on the item; b_2 captures the mean score difference between the two groups on the item; and b_3 displays the interaction between the person's total test score and group membership. If b_2 is statistically significant, it suggests that the probability of answering the item correctly is different between these two groups after controlling for the ability (uniform DIF). If b_3 is significant, it suggests that there is an interaction effect between group membership and total test scores (non-uniform DIF).

In the following Monte Carlo simulation study, conditional logistic regression was used for analyzing the matched data obtained from optimal pair and full matching methods. Conditional logistic regression differs from the conventional logistic regression in that the parameters of the conditional logistic regression are estimated using paired or clustered samples. The conditional logistic regression is used to take care of data dependency due to pairs or clusters and is widely used for matched case-control studies. The detailed description of conditional logistic regression can be found in Hosmer et al. (2013, pp. 227-267), and Breslow and Day (1980). The description of conditional logistic regression in a DIF analysis context can be found in Liu et al. (2016).

Monte Carlo Simulations

As indicated in the literature, covariate selection is essential in the application of propensity score methods, as it can dramatically affect the conclusions made from propensity score analysis. Hence, the types of covariates and different missing patterns of covariates were investigated in the context of propensity score DIF analysis in the present study. Here, covariate selection refers to the types of relationship between covariates and the outcome/grouping variable, or simply the

types of covariates (i.e., covariates related to the outcome only, related to the grouping variable only, or related to both the outcome and grouping variables).

Monte Carlo simulations were utilized to investigate how propensity score DIF methods perform in the presence of a variety of model misspecification conditions and under different levels of effect size. The magnitude of bias, mean square error (MSE), Type I error, and power were examined. In addition, model performance is examined.

Consider three propensity score methods: optimal pair matching, optimal full matching, and stratification. For propensity score DIF methods, the simulated data were matched first and then analyzed by conditional logistic regression, or first stratified and then analyzed by the regular logistic regression. The R package MatchIt (Ho et al., 2011) was used for optimal pair and full matching. The R package Epi was used for the conditional logistic regression DIF analyses (Carstensen et al., 2016). The DIF results obtained from propensity score methods were compared with those obtained from the conventional logistic regression as well as covariance adjustment logistic regression. For information in the implementation of logistic regression DIF analysis based on propensity score approach, see the step-by-step demonstration as well as R code in Liu et al. (2016).

Simulation Design

A detailed description of simulation models, testing models, and hypotheses is provided as follows. Equation (2) can be used for a basic understanding of DIF concepts: b_2 is the regression coefficient for the grouping variable (G) and b_3 is the coefficient of the interaction of grouping variable and total scores (tot*G).

Simulation Models. In the simulation models, a variety of conditions were simulated by systematically varying two factors, effect size and the types of covariates. Each condition had 1000 replications. These conditions are:

a) Three levels of effect size of regression coefficients, i.e., no effect, moderate effect, and strong effect (0, 1, 2), for both grouping variable (G) and the interaction (tot*G) in three DIF scenarios
   - No-DIF [$b_2 = b_3 = 0$ in equation (2)];
   - Uniform DIF [$b_2 = 1$ or 2, but $b_3 = 0$ in equation (2)];
   - Non-uniform DIF [$b_3 = 1$ or 2, regardless $b_2 = 0$, 1 or 2 in equation (2)]; and

b) Three types of covariates

- Only related to the outcome (*Y*) with three levels of regression coefficients, i.e., weak, medium, and strong (0.5, 1, 2);
- Only related to the grouping variable (*G*) at a medium level;
- Related to both the outcome (*Y*) and the grouping variable (*G*).

In order to focus on the two factors of interest, other factors often manipulated for examining DIF methods were fixed. The sample size was fixed to be 1000 and the ratio of sample sizes is 3:7 (focal vs. reference groups); that is, around 30% of the sample was from the focal group (treatment group). Multicollinearity concerns was minimized by setting the correlations to zero among covariates and fixing the correlations between covariates and total scores as well as the correlation between the grouping variable and total scores to zero. The covariates were generated using the mnormt R package (Azzalini & Genz, 2016).

The simulations included three sets of separate models: propensity score estimation [equation (3)], outcome variable generation [equation (4)], and corrected true value generation [equation (6)]. The details of the corrected true value generation model are provided in the description of "Collapsibility & Corrected True Values". The propensity score simulation model is defined as follows:

$$\mathrm{p}_i\left(G=1|\,\mathbf{W}\right)=\frac{\exp\left(-1+X_2+X_3+X_5-X_7\right)}{1+\exp\left(-1+X_2+X_3+X_5-X_7\right)}, \tag{3}$$

where $\mathrm{p}_i(G = 1 \mid \mathbf{W})$ refers to the estimated propensity scores, G denotes the grouping variable, and $\mathbf{W}$ represents a vector of covariates (X_1, X_3, X_5, X_7) used in the simulation model for propensity score estimation. The outcome variable simulation model is defined as

$$\begin{aligned}&\ln\frac{\mathrm{p}\left(Y=1|\,\mathbf{X},\mathrm{tot},G,G*\mathrm{tot}\right)}{1-\mathrm{p}\left(Y=1|\,\mathbf{X},\mathrm{tot},G,G*\mathrm{tot}\right)}\\&=-0.5-X_1-X_2-0.5X_3-0.5X_4-2X_5-2X_6+2\,\mathrm{tot}+b_2G+b_3\,\mathrm{tot}*G\end{aligned} \tag{4}$$

where $\mathbf{X}$ on the left side of the equation represents a vector of covariates (X_1, X_2, X_3, X_4, X_5, X_6), tot denotes total test scores, G denotes the grouping variable, and tot*G denotes the interaction between the grouping variable and total scores. All regression coefficients of covariates and total test scores in equation (4) were fixed except the regression coefficients for G and tot*G, which are manipulated to vary at three levels (0, 1, and 2).

Table 1. Testing models for model misspecifications and different types of covariates

		Related to *G*	
		No	**Yes**
Related to *Y*	No	--	X_7 (Model #6)
	Yes	X_2 (Model #1) $X_2 + X_6$ (Model #2)	$X_1 + X_5$ (Model #3); $X_1 + X_5 + X_7$ (Model #4); $X_1 + X_2 + X_5 + X_6$ (Model #5); $X_1 + X_2 + X_5 + X_6 + X_7$ (Model #7)

Note: *Y* denotes the outcome variable; *G* denotes the grouping variable in DIF analysis; X_1 and X_5 are related to both outcome and grouping variables; X_7 is only related to grouping variable; X_2 and X_6 are only related to outcome variable

To simulate three types of covariates, three covariates (X_1, X_3, X_5) were included, which are related to the outcome and grouping variables, three covariates (X_2, X_4, X_6), which are only related to the outcome, and one covariate (X_7), which is only related to the grouping variable. Three other variables were also included in the outcome variable simulation model, i.e., total scores, the grouping variable, and the interaction between them. These three variables should not be construed as covariates because they are variables of interest in the final DIF analyses.

Testing Models. The main purpose of the present study is to test how propensity score DIF methods perform in the presence of different scenarios of covariate selection and a variety of model misspecification conditions. Seven models were tested (see Table 1). In this contingency table, the rows represent the status of the associations between covariates and the outcome variable (*Y*), i.e., not related or related, and similarly the columns display the associations between covariates and the grouping variable (*G*).

Hypotheses of the Present Study. In the literature, some studies showed that covariates related to the outcome variable or related to both outcome and grouping variables increased the precision of estimates, but covariates only related to grouping variable introduced bias into the results. Based on previous studies, it was hypothesized propensity score DIF methods performed better than the conventional logistic regression DIF method (referred to as the conventional method hereafter), but the covariance adjustment logistic regression DIF method (referred to as the covariance adjustment method hereafter), might exhibit similar performance as propensity score DIF methods. In addition, there are three hypotheses for the comparisons of the seven testing models:

- Models #3, #4, #5, and #7 should perform better than Models #1, #2, and #6 because they included covariates that are related moderately and/or strongly to both the group variable and the outcome;
- Model #6 would show the poorest performance because the model included only one covariate that was related only to the grouping variable;
- Model #3 may perform better than model #4, and model #5 better than model #7 because the inclusion of X_7, the covariate only related to the grouping variable, may introduce some errors in the parameter estimates.

Collapsibility & Corrected True Values

The collapsibility issue has been discussed and presented in different terms in the literature. (e.g., Yule, 1903; Cohen & Nagel, 1934; Greenland et al., 1999; Greenland & Pearl, 2011). In a linear regression context, the same relation between Y and G (grouping variable) is seen whether confounding by **X** is dealt with by regression adjustment, or by creating covariate balance (either physically by randomization or via propensity score methods). That is, collapsability means that dealing with **X** by regression adjustment or dealing with **X** by comparing balanced groups leads to the same thing. However, a binary logistic regression is known to suffer from non-collapsibility because of a non-linear link function. In a causal inference context, the relation of the outcome and the treatment condition remains the same when covariates are included or omitted under randomization experimental designs. Thus, in a binary logistic regression model for DIF analysis, the model is collapsible over all covariates only if all covariates are not statistically significant under the randomization experiment setting. However, most DIF analyses are based on observational studies so that regression coefficients b_1^*, b_2^*, and b_3^* in equation (6) are not equal to b_1, b_2, and b_3 in equation (5). Equations (5) and (6) are defined as follows:

$$\ln \frac{\mathrm{p}\left(Y=1 \mid \mathbf{X}, \mathrm{tot}, G, G*\mathrm{tot}\right)}{1-\mathrm{p}\left(Y=1 \mid \mathbf{X}, \mathrm{tot}, G, G*\mathrm{tot}\right)} = b_0 + b_1\mathrm{tot} + b_2 G + b_3\left(\mathrm{tot}*G\right) + \mathbf{X\beta}, \qquad (5)$$

where **X** is a vector of covariates and **β** is a vector of regression coefficients of **X**, b_1 is the regression coefficient of tot, and all other notations are the same as equation (4);

$$\ln \frac{\mathrm{p}\left(Y=1 \mid \mathbf{X}, \mathrm{tot}, G, G*\mathrm{tot}\right)}{1-\mathrm{p}\left(Y=1 \mid \mathbf{X}, \mathrm{tot}, G, G*\mathrm{tot}\right)} = b_0^* + b_1^*\mathrm{tot} + b_2^* G + b_3^* \left(\mathrm{tot}*G\right), \qquad (6)$$

where b_0^* denotes the intercept b_1^*, b_2^*, and b_3^* denote the regression coefficients for tot, *G*, and *G**tot, respectively.

Equation (6) represents the ideal situation if the data were obtained from a randomized experimental study. Using the propensity score DIF approach, the attempt is made to approximate the random assignment mechanism and reduce the pre-existing group differences in the observational data. In the conventional DIF analysis [equation (2)], this non-collapsibility issue is simply ignored. Thus, the DIF simulations mimicked the non-collapsibility scenarios in which the relationship between *G* and *Y* is partly dependent on **X**. Correspondingly, b_1, b_2, and b_3 in equation (5) were not the true values obtained from a randomized experiment. Therefore, define b_1^*, b_2^*, and b_3^* in equation (6) as the corrected true values, and these were used for examining bias, MSE, model performance, but not for type I error and power. The corrected true values of b_2^* and b_3^* are provided in each graph in the results section.

Outcome Variables of the Simulation Study

We compared DIF results obtained from three propensity score methods, optimal pair matching, optimal full matching and stratification, to those obtained from conventional and covariance adjustment methods in terms of five indices: bias, MSE, type I error rate, power, as well as model performance. These five indices are the outcome variables of this simulation study and they were assessed via regression coefficient estimates of both *G* and *G**tot in three scenarios: no DIF, uniform DIF, and non-uniform DIF.

In the present study, bias is used to examine the magnitude of inflation or deflation of the estimates of regression coefficients, defined as $\mathrm{E}\left(\hat{\theta}-\theta\right)$, where θ denotes the population parameters, $\hat{\theta}$ denotes an estimate of the regression coefficient based on one simulated data set, and $\mathrm{E}\left(\hat{\theta}-\theta\right)$ is the average value of regression coefficient estimates computed from the 1000 replications in the simulation. MSE incorporates the information about the variance of the estimator in addition to bias, defined as $\mathrm{MSE} = \mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right]$.

Type I error is the incorrect rejection of a true null hypothesis, H_0: $\theta = 0$. In the DIF context, when the Type I error rate is high there is greater risk of concluding the existence of DIF when it actually does not exist. Power was used to examine how often DIF results were correctly identified.

The model performance of propensity score methods was also examined for each simulation condition and in each testing model. Model performance used in the present study refers to the situation $H_0 : \theta = \hat{\theta}$ where θ denotes the population parameters and $\hat{\theta}$ denotes the average value of 1000 regression coefficient estimates from the simulations. Theoretically, model performance is actually Type I error for testing a null hypothesis that is defined at a particular nonzero value. For example, $H_0 : b_2^* = 0.7$ instead of $H_0 : b_2^* = 0$ in order to assess the model performance in the uniform DIF scenario. Hence, model performance was used to distinguish it from the conventional Type I error.

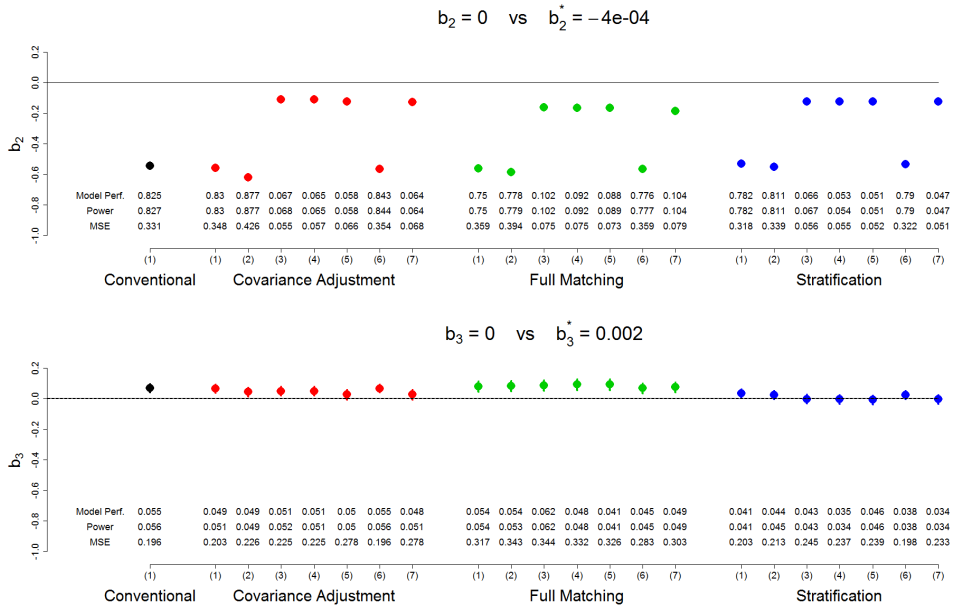

Figure 1. Comparisons of conventional and covariance adjustment logistic regression DIF methods and propensity score DIF methods (optimal full matching and stratification) for no-DIF scenario ($b_2 = 0$ and $b_3 = 0$); Note the performance of bias, MSE, type I error, and model performance was demonstrated in this figure; the seven models are listed as follows: Model #1 (X_2); Model #2 ($X_2 + X_6$); Model #3 ($X_1 + X_5$); Model #4 ($X_1 + X_5 + X_7$); Model #5 ($X_1 + X_2 + X_5 + X_6$); Model #6 (X_7); Model #7 ($X_1 + X_2 + X_5 + X_6 + X_7$)

Simulation Results

The simulation results are reported in the following three scenarios: no DIF, uniform DIF, and non-uniform DIF. Results from optimal pair matching were not used, because they were similar to those obtained from optimal full matching.

No-DIF Scenario. In Figure 1, the solid reference lines represent the corrected true values $\left(b_2^* = -0.0004 \text{ and } b_3^* = 0.0022\right)$, collapsed over the covariates, while the dashed reference lines represent the original values used in outcome simulation model ($b_2 = 0$ and $b_3 = 0$), ignoring the collapsibility issue.

Bias. The dots in the graph showed the magnitude of bias for each testing model and for each DIF analysis method. Downward biases were found for the G $\left(b_2^*\right)$ across all testing models and methods. The first black dot shows the average bias for the conventional method, -0.544 for b_2^* and 0.07 for b_3^*. Aligned with these hypotheses, propensity score DIF methods and covariance adjustment DIF methods performed better than the conventional method when the models included covariates correlated to both Y and G. The magnitude of bias of b_2^*, i.e., the distance between a dot and the solid reference line, for models #3, #4, #5 and #7 (i.e., the models with covariates related to both Y and G) was much smaller than that of models #1, #2, and #6 (i.e., the models omitting some important covariates). However, contrary to the hypothesis, models #3 and #5 did not perform better than models #4 and #7, a result which was found across all three DIF scenarios. The magnitude of bias for b_3^* was very small across all DIF methods in the no-DIF scenario.

MSE. Across all figures, bars for 95% confidence intervals of the parameter estimates were plotted together with the point estimates, but they are invisible because variances are too small relative to the magnitude of bias. In the no-DIF scenario, MSE values were driven by the bias term, and thus the findings only recapture the bias results discussed above.

Model Performance. Model performance is the Type I error attached to the null hypothesis that the population parameter is actually the corrected true value. Again, the results of model performance echoed the findings for bias.

Type I error. The findings were almost identical to those obtained from model performance. This is unsurprising, as the corrected true values $\left(b_2^* = -0.0004 \text{ and } b_3^* = 0.0022\right)$ are very close to zero, and thus the results for model performance and Type I error should be similar.

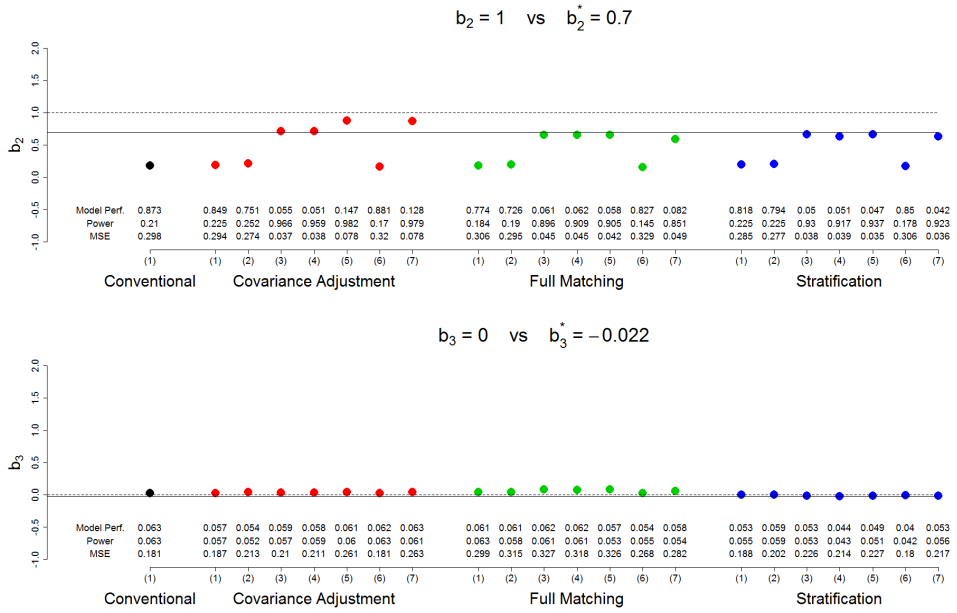

Figure 2. Comparisons of conventional and covariance adjustment logistic regression DIF methods and propensity score DIF methods (optimal full matching and stratification) for uniform DIF scenario $b_2 = 1$ and $b_3 = 0$; Note the performance of bias, MSE, type I error, and model performance was demonstrated in this figure; the seven models are listed as follows: Model #1 (X_2); Model #2 ($X_2 + X_6$); Model #3 ($X_1 + X_5$); Model #4 ($X_1 + X_5 + X_7$); Model #5 ($X_1 + X_2 + X_5 + X_6$); Model #6 (X_7); Model #7 ($X_1 + X_2 + X_5 + X_6 + X_7$)

Uniform DIF Scenarios. Figures 2 and 3 showed the results obtained from uniform scenarios. The solid reference lines represent the corrected true values, $b_2^* = 0.7$ and $b_3^* = -0.022$, in Figure 2 and $b_2^* = 1.385$ and $b_3^* = -0.00008$ in Figure 3; the dashed reference lines indicate the original values used in outcome simulation models ($b_2 = 1$ and $b_3 = 0$) in Figure 2 and ($b_2 = 2$ and $b_3 = 0$) in Figure 3.

Bias. Aligned with these hypotheses, covariance adjustment and propensity score methods performed better than the conventional method when the models included covariates correlated to both Y and G. With the exception of models #5 and #7 for the covariance adjustment method, the magnitude of bias for b_2^* was smaller for models #3, #4, #5 and #7, ranging from -0.11 to 0.02 $\left(b_2^* = 0.7\right)$ and from -0.005 to 0.145 $\left(b_2^* = 1.385\right)$, than that of models #1, #2 and #6, ranging from -0.7 to -0.49 $\left(b_2^* = 0.7\right)$ and from -0.515 to -0.335 $\left(b_2^* = 1.385\right)$. However,

models #5 and #7 in the covariance adjustment method tended to recover the original values (b_2 = 1 or 2) used in the outcome simulation model and had an even larger magnitude of bias (around 0.5) when the effect size was increased to 2. Similar to the no-DIF scenario, the bias for b_3^* was small across all models and all methods.

MSE. Similar to the results of the no-DIF scenario, the findings echoed those obtained for bias because variances of estimates were small. However, the variances of estimates were relatively large for b_3^*, so that MSE values for b_3^* became larger than those for b_2^* in general.

Model Performance. The results for b_2^* are similar to those obtained from bias. For b_3^*, all values of model performance were small, falling in an acceptable range, less than 0.065.

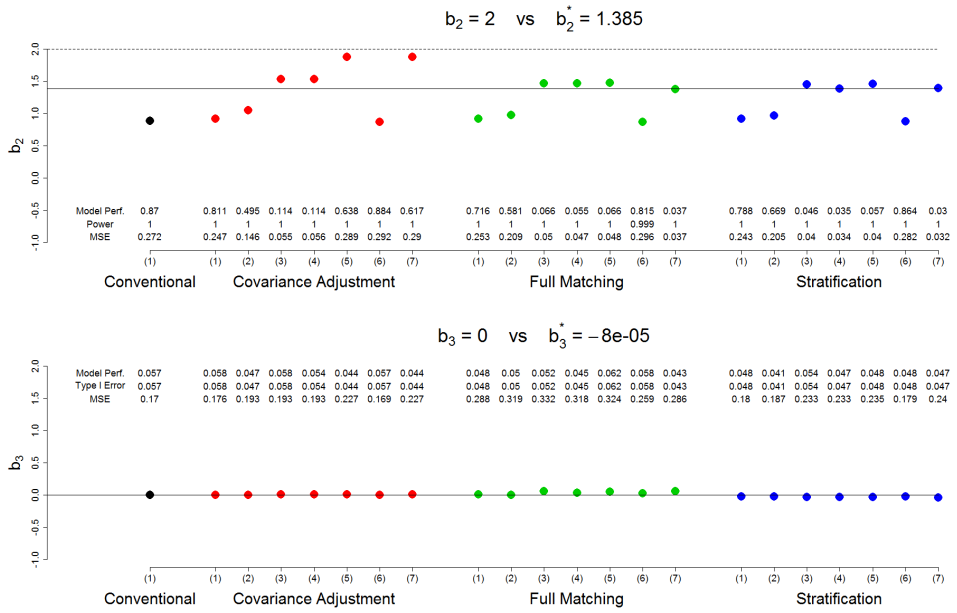

Figure 3. Comparisons of conventional and covariance adjustment logistic regression DIF methods and propensity score DIF methods (optimal full matching and stratification) for uniform DIF scenario b_2 = 2 and b_3 = 0; Note the performance of bias, MSE, type I error, and model performance was demonstrated in this figure; the seven models are listed as follows: Model #1 (X_2); Model #2 (X_2 + X_6); Model #3 (X_1 + X_5); Model #4 (X_1 + X_5 + X_7); Model #5 (X_1 + X_2 + X_5 + X_6); Model #6 (X_7); Model #7 (X_1 + X_2 + X_5 + X_6 + X_7)

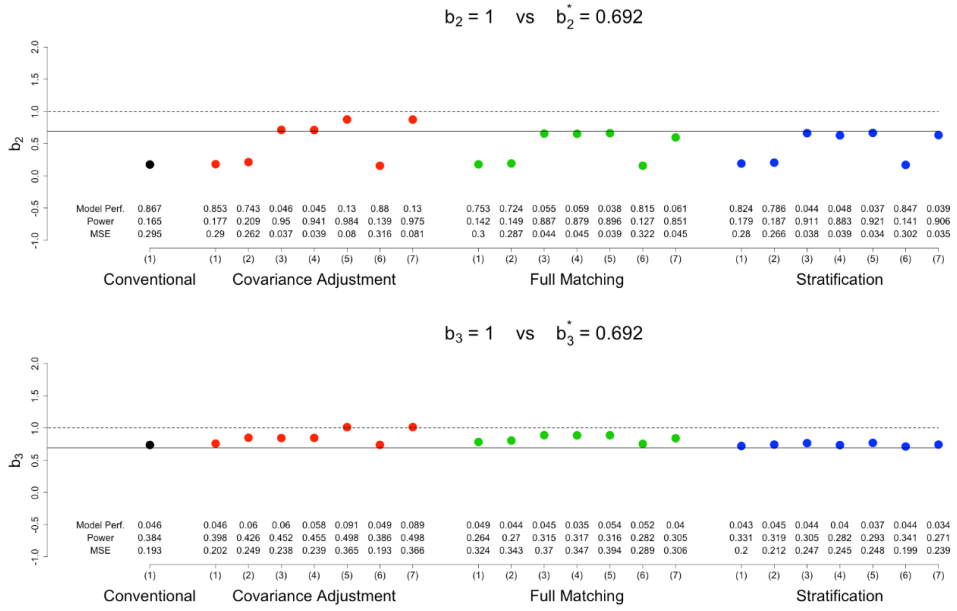

Figure 4. Comparisons of conventional and covariance adjustment logistic regression DIF methods and propensity score DIF methods (optimal full matching and stratification) for non-uniform DIF scenario $b_2 = 1$ and $b_3 = 1$; Note the performance of bias, MSE, type I error, and model performance was demonstrated in this figure; the seven models are listed as follows: Model #1 (X_2); Model #2 ($X_2 + X_6$); Model #3 ($X_1 + X_5$); Model #4 ($X_1 + X_5 + X_7$); Model #5 ($X_1 + X_2 + X_5 + X_6$); Model #6 (X_7); Model #7 ($X_1 + X_2 + X_5 + X_6 + X_7$)

Type I error. The Type I error rates for b_3^* were small in both Figures 2 and 3, falling in an acceptable range, less than .065.

Power. The conventional method had a low power (0.21) when $b_2^* = 0.7$, but increased to a power of one when $b_2^* = 1.385$. Again, aligned with these hypotheses, the covariance adjustment and propensity score methods performed better than the conventional method when $b_2^* = 0.7$: models #3, #4, #5 and #7 (0.896-0.982) outperformed models #1, #2, and #6 (0.133 0.252). However, power was increased to almost one across all methods when $b_2^* = 1.385$.

Non-Uniform DIF Scenarios. Figures 4 and 5 showed the non-uniform scenarios. Again, the solid reference lines represent the corrected true values $\left(b_2^* = 0.692 \text{ and } b_3^* = 0.692\right)$ in Figure 4 and $\left(b_2^* = 1.396 \text{ and } b_3^* = 1.443\right)$ in Figure 5; the dashed reference lines represent the original values used in outcome

simulation models ($b_2 = 1$ and $b_3 = 1$) in Figure 4 and ($b_2 = 2$ and $b_3 = 2$) in Figure 5.

For the non-uniform DIF scenario, the results for b_2^* were not reported, because they were similar to those obtained from the uniform DIF scenario and also because the interpretation of G becomes less important when the interaction is found to be statistically significant. Hence, focus on the results for the interaction term $\left(b_3^*\right)$ for this scenario. Most of the findings on the interaction term did not align with these hypotheses.

Bias. Contrary to these hypotheses, the magnitude of bias for the conventional method (0.05 when $b_3^* = 0.692$; 0.04 when $b_3^* = 1.443$) was smaller than that of most other models across all other methods. In addition, models #1 and #6 display small bias compared to other models. Although the conventional method outperformed other methods under most conditions, the overall magnitude of bias

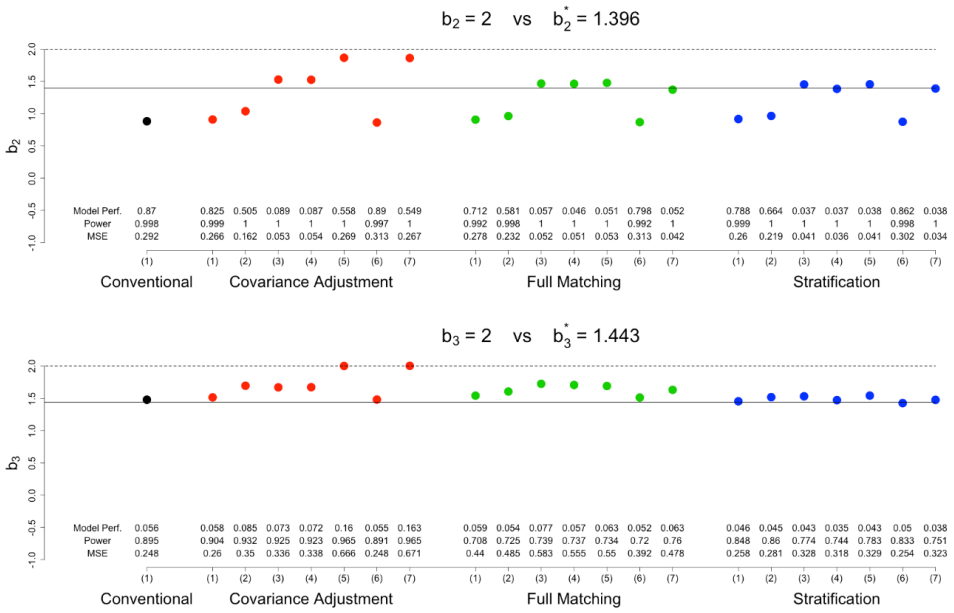

Figure 5. Comparisons of conventional and covariance adjustment logistic regression DIF methods and propensity score DIF methods (optimal full matching and stratification) for non-uniform DIF scenario $b_2 = 2$ and $b_3 = 2$; Note the performance of bias, MSE, type I error, and model performance was demonstrated in this figure; the seven models are listed as follows: Model #1 (X_2); Model #2 ($X_2 + X_6$); Model #3 ($X_1 + X_5$); Model #4 ($X_1 + X_5 + X_7$); Model #5 ($X_1 + X_2 + X_5 + X_6$); Model #6 (X_7); Model #7 ($X_1 + X_2 + X_5 + X_6 + X_7$)

under these other methods was not large, ranging from 0.007 to 0.26, much smaller than the bias magnitude for *G* in the no-DIF and uniform DIF scenarios. Again, models #5 and #7 under covariance adjustment tended to recover the original values used in the outcome simulation models and had the largest magnitude of bias (0.318 when $b_3 = 1$; 0.567 when $b_3 = 2$) using the corrected true value as reference.

MSE. The conventional method and models #1 and #6 based on covariance adjustment and stratification methods had smaller MSE values (0.19-0.2 when $b_3 = 1$; 0.24-0.26 when $b_3 = 2$) than others. The stratification method performed better than optimal full matching and covariance adjustment in general in the non-uniform DIF scenario.

Model Performance. The conventional method and all propensity score methods showed acceptable model performance (0.035-0.075) whereas the covariance adjustment method had a larger magnitude of model performance (indicating poorer performance). Again, models #5 and #7 under covariance adjustment method tended to recover the original values used in the outcome simulation model (0.091-0.089 when $b_3 = 1$; 0.16; 0.163 when $b_3 = 2$).

Power. The magnitude of power for $b_3 = 1$ was found to be low across all methods, ranging from 0.27 to 0.498. When b_3 was increased from 1 to 2, power increased substantially across all methods: the conventional method had an average power of 0.895; the covariance adjustment method had the highest power (0.891-0.965); the optimal full matching method had relatively low power (0.708-0.76); and the stratification method was slightly better than optimal full matching (0.744-0.86). Contrary to these hypotheses, the conventional and covariance adjustment methods seemed to generate greater power than propensity score methods.

Conclusion and Discussion

Aligned with these hypotheses, models with covariates moderately or strongly related to both *G* and *Y* exhibited substantially lower bias, lower MSE, better model performance, and smaller type I error than did models with covariates related to *Y* only or to *G* only in the no-DIF and uniform DIF scenarios. These models also produced higher power in the uniform DIF scenario. However, models #3 and #5 were found to perform no better than models #4 and #7, which suggests that the inclusion of a covariate correlated only to the grouping variable did not affect the conclusions about DIF if the model already included covariates at least moderately correlated with both *Y* and *G*.

Contrary to these hypotheses, the results showed different patterns in the non-uniform DIF scenario. The conventional method in fact induced less bias and larger

power for the interaction term (tot*G) than most models based on propensity score approaches. However, the magnitude of bias for all methods was small in the non-uniform DIF scenario compared to the no-DIF and uniform DIF scenarios.

It is also interesting to note that the levels of effect size of regression coefficient estimates greatly affected both the power of G in the uniform scenario and that of tot*G in the non-uniform DIF scenario. Power was dramatically increased across all models and methods when the effect size was raised from one to two. This finding indicates that the magnitude of population effect sizes plays an important role in identifying DIF, a result that is consistent with the established theory of statistical power.

Our findings suggest that propensity score methods work better than the conventional method in the no-DIF and uniform DIF scenarios when including covariates moderately or strongly correlated to both outcome and grouping variables, but that these methods do not perform well when including covariates solely correlated to either the outcome or grouping variable. However, the results obtained from the non-uniform DIF scenarios were more complex than the no-DIF and uniform DIF scenarios. These results suggest that propensity score methods do not perform better – and sometimes perform even worse – than the conventional method when aiming to identify non-uniform DIF.

Because the results from optimal pair matching were similar to those from optimal full matching, only the results obtained from optimal full matching were included. There were some interesting differences between these two optimal matching methods. The optimal pair matching always produced larger variances than other methods, which might have been caused by the smaller sample size after matching (dropping from 1000 to around 600). Correspondingly, the optimal pair matching method exhibited larger MSE and less power than other methods.

Aligned with these hypotheses, the covariance adjustment method performed similarly to propensity score methods under many conditions, but there were some notable differences between them. The results revealed that propensity score methods in general produced better model performance in uniform and non-uniform DIF scenarios than the covariance adjustment method, but that the covariance adjustment method had higher power than propensity score methods for tot*G in the non-uniform DIF scenario. More specifically, propensity score methods tended to approximate the randomization mechanism, whereas the covariance adjustment method tended to recover the original values used in the outcome simulation model. This finding suggests that the differences in the algorithms used to fit the covariance adjustment and propensity score models may lead to different conclusions under some conditions.

Another issue is the collapsibility issue when using propensity score methods. The ultimate goal of using matching methods is to balance the pre-group differences and to approximate the random assignment mechanism. Hence, in simulations, researchers need to consider the use of the corrected true values, which mimic randomization, instead of using the original values adopted in the outcome simulation models. This is an important issue as the conclusions from a simulation study would be different if one used a different reference.

The important messages to practitioners and psychometricians are: (a) it is crucial to include key covariates that are moderately or strongly related to both *G* and *Y* in propensity estimation models; (b) the conventional method produces high type I error rates and correspondingly flags more DIF items incorrectly, while propensity score DIF methods can provide more accurate results on identifying DIF items for the no-DIF and uniform DIF scenarios; and (c) propensity score methods have relatively higher Type II error rates than the conventional method in the presence of non-uniform DIF. In addition, researchers must be careful when using the covariance adjustment method for DIF analysis. It may produce misleading results under certain conditions when researchers use it as an alternative to matching methods and aim to approximate the random assignment mechanism in their data analyses. A thorough description of the problem of using covariance adjustment as an alternative to matching can be found in Liu et al. (2016).

The conclusions can be affected by the signs of covariates in the simulation models. This issue has not been discussed in the propensity score simulation literature. A simple scenario was chosen and adopted the same signs for all regression coefficients of covariates in the outcome variable simulation model [equation (4)]. The DIF results may change if these signs are mixed due to cancelation effects. There is a need for more studies to investigate this issue.

More studies are needed to investigate the use of propensity score methods for examining non-uniform DIF. In addition, the present study did not consider the effects of correlated covariates (multicollinearity) and non-linear functional forms of covariates; thus, it may be interesting for future studies to consider these factors.

References

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27*(12), 2037-2049. doi: 10.1002/sim.3150

Azzalini, A., & Genz, A. (2016). The R package 'mnormt': The multivariate normal and '*t*' distributions (version 1.5-5). Retrieved from http://azzalini.stat.unipd.it/SW/Pkg-mnormt

Blackstone, E. H. (2002). Comparing apples and oranges. *The Journal of Thoracic and Cardiovascular Surgery, 123*(1), 8-15. doi: 10.1067/mtc.2002.120329

Bowen, D. F. (2011). *The effects of controlling for distributional differences on the Mantel-Haenszel procedure* [Master's thesis]. University of North Carolina Chapel Hill.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: The analysis of case-control studies* (Vol. 1). Lyons, France: International Agency for Research on Cancer.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156. doi: 10.1093/aje/kwj149

Carstensen, B., Plummer, M., Laara, E., & Hills, M. (2016). Epi: A package for statistical analysis in epidemiology [R package version 2.0]. Retrieved from https://CRAN.R-project.org/package=Epi

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*(3), 261-281. doi: 10.2307/2527916

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and the scientific method*. New York: Harcourt Brace.

Cuong, N. V. (2013). Which covariates should be controlled in propensity score matching? Evidence from a simulation study. *Statistica Neerlandica, 67*(2), 169-180. doi: 10.1111/stan.12000

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265-2281. doi: 10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

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

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*(6), 912-921. doi: 10.1037/0021-9010.74.6.912

Greenland, S., & Pearl, J. (2011). Adjustments and their consequences–Collapsibility analysis using graphic methods. *International Statistics Reviews, 79*(3), 401-426. doi: 10.1111/j.1751-5823.2011.00158.x

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 14*(1), 29-46. doi: 10.1214/ss/1009211805

Guo, S., & Fraser, W. M. (2014). *Propensity score analysis: Statistical methods and applications* (2nd edition). Thousand Oaks, CA: Sage Publications, Inc.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*, 153-171. doi: 10.1023/a:1006941729637

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. Retrieved from: https://r.iq.harvard.edu/docs/matchit/2.4-20/

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

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205-224. doi: 10.3102/01623737027003205

Hosmer, D. W., Lemeshow, S., Jr., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd edition). Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781118548387

Joldersma, K., & Bowen, D. (2010). *Application of propensity models in DIF studies to compensate for unequal ability distribution*. Paper presented at the annual meeting of National Council on Measurement in Education, Denver, CO.

Lee, H., & Geisinger, K. F. (2014). The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF. International Journal of Testing, 14(4), 313-338. doi: 10.1080/15305058.2014.922567

Liu, Y., Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A.D. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research, and Evaluation, 21*, 13. doi: 10.7275/ewqz-n963

Pan, W., & Bai, H. (2015). *Propensity score analysis: Fundamentals and developments*. New York: The Guilford Press.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology, 40*(1), 75-149. doi: 10.1111/j.1467-9531.2010.01228.x

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer. doi: 10.1007/978-1-4757-2443-1

Rosenbaum, P. R. (2002). *Observational studies* (2nd edition). New York: Springer. doi: 10.1007/978-1-4757-3692-2

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer. doi: 10.1007/978-1-4419-1213-8

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. doi: 10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling that incorporates the propensity score. *The American Statistician, 39*(1), 33-38. doi: 10.1080/00031305.1985.10479383

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188. doi: 10.1023/a:1020363010465

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*(4), 279-313. doi: 10.1037/a0014268

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

Steiner, P. M., Cook, T. D., Shadish, W. R. S., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267. doi: 10.1037/a0018719

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118. doi: 10.1080/00273171.2011.540475

Wu, A. D., & Ercikan, K. (2009). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287-300. doi: 10.1207/s15327574ijt0603_5

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2*(2), 121-134. doi: 10.1093/biomet/2.2.121

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters, 98*(3), 309-319. doi: 10.1016/j.econlet.2007.05.010

Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association, 109*(505), 95-107. doi: 10.1080/01621459.2013.869498

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-types (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. doi: 10.1080/15434300701375832