PROPENSITY SCORE ADJUSTMENT IN MEASUREMENT INVARIANCE

A Dissertation

by

QIAN CAO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Myongsun Yoon
Committee Members,	Oi-Man Kwok
	Wen Luo
	Lei-Shih Chen
Head of Department,	Shanna Hagan-Burke

December 2016

Major Subject: Educational Psychology

ABSTRACT

Measurement invariance testing is prerequisite if meaningful comparisons of latent construct across groups are important to the study in social science. If measurement invariance is rejected, the result of non-invariance might be from unbalanced covariates across groups. Propensity score is one approach to correct unbalanced covariates in the data when these unbalanced covariates are the source of measurement non-invariance.

The main purpose of this dissertation is to evaluate propensity score adjustment in testing measurement invariance in both empirical data and Monte Carlo simulation study. The traditional logistic regression and machine learning estimation method (i.e., random forest) were applied to obtain accurate propensity score.

In empirical study, when propensity score was applied as a new covariate to adjust unbalanced covariates across groups, measurement invariance was improved from metric invariance to scalar invariance. Weighting by odds method with random forest estimation improved the metric invariance to scalar invariance, but weighting with logistic regression did not.

The results of a simulation study indicated a substantial Type I error rate inflation if ignoring the unbalanced covariates among groups and using multiple group CFA to conduct the measurement invariance test. Type I error rate inflation was also observed if logistic regression was applied to adjust measurement invariance. On the other hand,

using random forest estimation method to balance covariates across groups gave accurate measurement invariance test conclusion.

DEDICATION

I dedicated this dissertation to my parents and my husband who give me unconditional love and support and encourage me to pursue my dream.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Yoon, for her thoughtful comments and critical insights throughout my dissertation. I would like to thank my co-chair Dr. Kwok, and my committee members, Dr. Luo and Dr. Chen, for their guidance and support and encouragement throughout the course of this research.

I also would like to thank Dr. Hughes for her support and care for my study and kindness to share the data with me.

Thanks also go to my friends and colleagues, Brandie, Huan, Jenny, Maria, Mark, Mirim, Yuhong, Shuqiong, Siqi, and Yu-Yu, for making my time at Research, Measurement and Statistics program a wonderful experience.

Finally, thanks to my mother and father for their encouragement and to my husband for his patience and love.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Measurement invariance has become a common practice before using a measurement in social science. We often study latent constructs measured by multiple observed items. In order to compare people from different groups (e.g., male vs. female), the relationship between the observed items and latent construct should be same across groups (Yoon & Millsap, 2007). For example, in order to compare males and females on the depression scale, if males and females are identical on the latent depression structure, they should have the same distribution of observed depression items. In this case, measurement invariance holds for males and females. If measurement invariance is violated at group level, persons with same latent depression construct but from different gender groups, will receive different scores on the observed items depending on group membership. In this situation, the difference on observed depression scores might not represent the true difference on latent depression construct, thus the use of test scores on the measurement is inappropriate, and it will not be valid to compare test scores among groups. Therefore, measurement invariance is a critical condition before comparing group differences on the observed items.

In practice, sometimes measurement invariance across groups is not well established in research (e.g., Hox, De Leeuw, & Zijlmans, 2015; Cham, Hughes, West, & Im, 2015). If the comparison of group means is important to the study, failure to establish measurement invariance is problematic. Therefore, dealing with measurement non-invariance becomes an important topic. Measurement non-invariance might come

from some non-invariance items across all items, or other unbalanced covariates across groups (Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015). If the non-invariance is due to the effect of other covariates that are not balanced across groups, measurement non-invariance can be explained and corrected by the unbalanced covariates through propensity score method (Hox et al., 2015). Propensity score can be used to balance groups on the observed covariates. However, propensity score is rarely used in testing measurement invariance.

The purposes of this dissertation are to review different aspects in measurement invariance and propensity scores, and to apply propensity score adjustment in measurement invariance test in both empirical data and Monte Carlo simulation study. The current literature review consists of two parts: measurement invariance and propensity score. The first part consists of four sections to introduce measurement invariance: definition of measurement invariance, multiple group CFA under SEM framework, and hierarchical procedure of measurement invariance. The second part includes the framework of propensity score as a method to deal with the situation when violations of measurement invariance exist. The second part consists of four sections: logistic regression, machine learning techniques, group equating, and balance check.

Measurement Invariance

Measurement invariance is established when the observed score's probability in a test given the identical ability is equivalent across different groups (Mellenbergh, 1989; Meredith & Millsap, 1992; Yoon & Millsap, 2007). The formal definition of measurement invariance is expressed as follows.

$$P(X \mid \xi, G) = P(X \mid \xi) \quad , \qquad (1)$$

X is the observed score, ξ is latent construct underlying X , and G is the grouping variable. From this formula, the conditional probability of X given ξ is independent of grouping variable. In other words, measurement invariance holds when individuals with same latent construct score have the same probability distribution of observed scores regardless of grouping variable.

The commonly used grouping variables are subgroups of population, different time points, and different test forms (Meade & Bauer, 2007). Subgroups of population may include demographics such as gender, ethnicity and country. In longitudinal studies, researchers repeatedly conduct a measurement across time, measurement invariance across time should be tested (Millsap, 2010). Measurement invariance can be tested across different test forms such as face-to-face interview, online survey, or telephone survey (Hox et al., 2015).

Factorial Invariance

Measurement invariance in a factor model is defined as factorial invariance, and factorial invariance is a special case of measurement invariance (Yoon, 2008). Measurement invariance concerns the entire distribution of scores, while factorial invariance only considers its means and covariance in factor structures. Therefore, measurement invariance has a broader scope than factorial invariance.

The most common way to conduct measurement invariance is multiple group confirmatory factor analysis (MGCFA). Under confirmatory factor analysis framework, factorial invariance is examined for the equivalence of parameters specified in the model

across groups. In a single unidimensional factor model, the relationship between the latent factor ξ and the continuous observed scores X in the CFA model are represented as:

$$X_{ij} = \tau_j + \lambda_j \xi_j + \delta_{ij} \text{ , (2)}$$

Where X_{ij} is an observed score of an individual i on an item j ; τ_j and λ_j are intercept and factor loading on an item j; ξ_j is the latent factor for an individual j; δ_{ij} is the unique factor score. For the multiple groups, the corresponding measurement model is,

$$X_g = \tau_g + \lambda_g \xi_g + \delta_g \text{ , (3)}$$

where g indicates group membership. Given the assumption that ξ factor score and δ unique factor scores are uncorrelated with each other (i.e., $\text{cov}(\xi, \delta) = 0$) in each group, the covariance structure of X in group is:

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g \text{ , (4)}$$

where Σ_g is a population covariance matrix of X in group g, Φ is a variance covariance matrix for factors in group g, Θ_g is a variance covariance matrix for the unique factors in group g. The expectation of X in each group is

$$E(X_g) = \tau_g + \lambda_g k_g \text{ , (5)}$$

where k_g is the factor mean in group g . If the highest level of measurement invariance (i.e., strict factorial invariance) holds, it follows that same intercepts (i.e., $\tau_g = \tau$), same factor loadings (i.e., $\lambda_g = \lambda$) and same unique factors (i.e., $\delta_g = \delta$) across groups. The formula (4) and (5) can be simplified as:

$$\Sigma_g = \Lambda \Phi_g \Lambda' + \Theta, \quad (6)$$

$$E(X_g) = \tau + \lambda k_g, \quad (7)$$

The equation 6 reveals that the group difference in covariance structure of observed scores (X) is due to the difference in covariance structure of latent factors (Φ_g). Similarly, in equation 7, the group differences in means of X are due to factor means (k_g).

Hierarchical Procedure of Measurement Invariance

The procedures to test measurement invariance are hierarchically: configural invariance, metric invariance, scalar invariance, and strict invariance (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Yoon & Millsap, 2007). Depending on which set of parameters are tested for group equality, different levels of factorial invariance are established sequentially, and it is easy to locate at which level of invariance is violated. Each level is described in details as follows.

Configural Invariance

The first level is the least restrictive model, which merely assuming same patterns in factor loadings, same numbers of latent factors in each group, and same locations of the zero and nonzero loadings in each group. No other invariance constraints are placed on this level. This is the baseline model, after establishing the configural invariance we can further test subsequent higher level of factorial invariance. If the configural invariance is violated, the subsequent factorial invariance testing is not meaningful. The violation of configural invariance indicates lack of configural

invariance across groups, meaning either the number of factors varies across groups or the factors are defined by different variables in each group.

Metric or Weak Invariance

After establishing configural invariance, the second level the factor loading are constrained to be equal across all groups ($\lambda_g = \lambda$), but allow differences in intercepts and unique factor variances. Metric invariance is essential for most purposes, since factor loadings affect the means, variance, and correlations among the measured variables. If the metric invariance is violated, the linear relation between factor score and observed score is not equal across groups, and one unit change in latent factor will lead to different unit change in the observed score in different group.

Scalar or Strong Invariance

After establishing metric invariance, further test for the invariance of intercepts is conducted. The scalar invariance is defined as the equivalence of intercepts across groups (i.e., $\tau_g = \tau$) in addition to metric invariance. The intercepts will not affect the variance or correlations among the measured variables, but they do affect the means. If the invariance for intercept is violated across groups, the observed score in one group will systematically higher or lower than in another group given same latent factor. Therefore, to makes group mean comparisons meaningful, scalar invariance is a required step, as seen in equation 7.

Strict Invariance

The most restrict invariance level require the unique factors are invariant across groups (i.e., $\delta_g = \delta$), after the previous conditions of equal factor loading, intercept are

established. With strict invariance, observed group differences on means or covariance are from true group difference on latent factors. Violations of strict invariance do not affect the means, but do affect correlations and covariance among observed variables. It may indicate that the reliabilities of at least some measured variables differ across groups. This is not a necessary step to compare the latent factor means across groups (Widaman & Reise, 1997). In reality, it is difficult to achieve strict invariance.

Propensity Score

Propensity score methods are originally proposed by Rosenbaum and Rubin (1983), experiencing tremendous increase of interest in many scientific areas including the social science. Propensity score is frequently interpreted in the context of causal effect. To estimate causal effect, researchers try to equate the treatment and control group prior to any treatment (i.e., baseline covariates) by randomization in practice (West et al., 2014). When sample size is large enough, randomization guarantees that the means of the treatment and control groups are equal on all possible baseline covariates, whether measured or unmeasured. In this situation, the average causal effect is calculated by average treatment effect, $ATE = \overline{Y_t} - \overline{Y_c}$. In social science, it is not always practical or ethical to randomly assign participants in control or treatment group, and therefore it is difficult to make a strong causal effect conclusion.

The main purpose of propensity scores is to balance the treatment and comparison groups on observed baseline covariates, and therefore propensity score can increase researcher's ability to draw causal inferences. Propensity score e(X) is defined

as a conditional probability that an individual is to be assigned to a treatment group given a set of observed covariates X at baseline (Rosenbaum and Rubin, 1983):

$$e(X) = \Pr(T = 1 \mid X) \quad , (8)$$

A propensity score reduces the selection bias through balancing groups based on the observed covariates. In order to provide causal inference for the outcome, assumptions for propensity score analysis (i.e., ignorable treatment assignment) involve (Rosenbaum & Rubin, 1983; Cham, 2013): (1) given a set of observed covariates X, the potential outcomes of a participant in the treatment and control groups are conditionally independent of the treatment assignment; (2) given the covariates X, the participant has non-zero probabilities of being assigned to either the treatment or control group. When the strong ignorability assumption holds for the covariates, the assumption also holds for the propensity score. Balance on participants' propensity scores e(x) between treatment conditions provides unbiased average treatment effect estimate in the non-randomized study.

Covariates Selection

Before the propensity score is estimated, the selection of a composite set of covariates (X) at baseline is the most critical issue in propensity score analysis. Since propensity scores are only estimated from the observed covariates, covariates selection could seriously affect the accuracy and precision for propensity score (West et al., 2014). Only a rich set of covariates can meet the strongly ignorable assumption, and it is critical giving detailed information about collecting the covariates (Thoemmes & Kim, 2011). West et al. (2014) suggested to use all possible variables that might be related to both

treatment (grouping variables) and outcome, and include as many as possible in the covariates measured at baseline. That is, any potential confounder that might bias the treatment effect must be collected at the baseline in the propensity score estimation.

There are three types of relationship among the covariates X, the treatment condition T, and the outcome variable Y: (1) X is a confounder that causes both T and Y; (2) X only causes Y but has no relation with T; (3) X only causes T but has no relation to Y. In scenario (1), X acts as a confounder (i.e., influence both the grouping variables and the outcome) and must be controlled to achieve an unbiased estimate of the causal effect. In scenario (2) and (3), X does not confounder the causal effect, and therefore X does not need to be controlled for an unbiased estimate of the causal effect. Assuming all potential confounders are collected and successfully balanced, the consistent estimates for the average causal effect will be established.

Sensitivity Analysis

It should be noted that researchers' attempt to employ all important covariates cannot be empirically tested. Sensitivity analysis is a necessary step to test how the current results might be affected if there were one or more unmeasured confounders (Rosenbaum, 1986). The unmeasured variables are hidden variables, and they act as hidden bias. The adjustment treatment effect is estimated as (West et al., 2014),

$$d^* = d - \gamma(smd^*), \quad (9)$$

where d is the treatment effect after controlling observed covariates, γ is the correlation between unobserved covariate with the outcome, smd^* is the rescaled by

using $\frac{smd}{\sqrt{2}}$, $\gamma(smd^*)$ together is the hidden bias from unobserved covariates. Thus the adjustment treatment effect is calculated by removing the hidden bias due to unmeasured covariates from estimated treatment effect. In practice, γ and smd^* are unknown, we need to assume their values either from the observed data or from theory and literature. For example, Hong (2004) suggested to use the largest correlation between observed covariates and outcome as γ, and the largest absolute smd^* value among observed covariates as smd^* to represent the worst scenario. After obtaining the adjustment treatment effect, we can conclude whether this effect remains statistically significant. If the adjustment treatment effect remains statistically significant under the worst scenario, the hidden bias due to unobserved covariates is ignorable, and the results will not be affected by the unobserved covariates.

After a complex set of covariates are collected, propensity score can be estimated. Any statistical model estimating the probability of group membership is propensity score. There are two different traditions in propensity score estimation. One is classical statistical modeling, assuming the data are generated by a given data model and there are nature functions associate the predictors with the outcomes (e.g., logistic regression, or discriminant analysis). The other is machine learning algorithms techniques, it treats the data mechanism as unknown, and need to use an algorithm to find the relation between predictors and outcomes (Breiman, 2001) (e.g., classification and regression trees, random forest).

Classical Methods

Logistic Regression

Logistic regression is the typical method to estimate propensity scores. It is estimated with a logistic regression model using treatment group as the dependent variable and using all covariates as the independent variables. For example, if we have 65 covariates without considering their interactions, the equation will be

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + ...b_{65}X_{65}, \quad (10)$$

$$\text{or } p = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + ...b_{65}X_{65})}{1+\exp(b_0 + b_1X_1 + b_2X_2 + ...b_{65}X_{65})}$$

where p is probability of being in the treatment group, given the 65 covariates, the b_0 is the intercept, b_1 to b_{65} are the slope for corresponding 65 covariates (West et al., 2014). Logistic regression is a familiar and well-understood tool of researchers, and it is easy to be conducted in most statistical software (e.g., SPSS, SAS, STATA, R) (Westreich, Lessler, & Funk, 2010). Logistic regression is the most commonly used method in social science to estimate propensity score (Thoemmes & Kim, 2011). However, when the covariates are a large set (e.g., more than 10 covariates), and the relations among these covariates are complex (e.g., involving interactions, nonlinear quadratic relation, the estimation will lead to a great bias (Cham et al., 2015; Lee, Lessler, & Stuart, 2010). In this situation, covariance balance may not be achieved by conditioning on the impropriate estimated propensity score, and therefore lead to biased effect estimate (Cham, 2013).

Discrimination Analysis

Discriminant analysis is mainly used when more than two groups membership to estimate (Hox et al., 2015). Several discriminant functions will be produced, and the number of functions is the number of group minus 1 or equal to the number of predictor variables, whichever is smaller (Tabachnick & Fidell, 2013). The first discriminant function provides the maximum discrimination among groups. The second discriminant function also maximally separates groups, but on the basis that the second function is uncorrelated with the first. This procedure will continue until all possible function is built. Typically, only the first or two functions is used, remaining functions will provide no additional information about group membership. After the discriminant function is selected, the probability in each group membership will be computed as propensity score.

Machine Learning Techniques

Contrary to strong assumptions in logistic model, machine learning techniques try to extract the relations between the outcome and predictors by a learning algorithm without a priori data model (Breiman, 2001). Machine learning techniques can outperform classical statistical techniques (e.g., logistic regression), especially when dealing with high-dimension data with large amount of covariates (Breiman, 2001), and they can give better predictive accuracy than data models, and provide better information about the underlying mechanism. Classification and regression trees (CART), bagged CART, and random forests are commonly used as machine learning techniques.

Lee et al. (2010) compared different machine learning techniques (i.e., CART methods, random forests) to logistic regression for propensity score estimation, and found random forests provided consistently superior performance than logistic model. More and more researches have applied machine learning techniques to estimate propensity score (Cham et al., 2015; Lee et al., 2010). Although machine learning techniques are sometimes criticized for lack of easy etiologic interpretation in the output of the machine learning classifiers, the "black box" nature of these techniques does not rule out them as potentially useful tools for propensity score analysis (Westreich, Lessler, & Funk, 2010).

Classification and Regression Tree (CART)

CART is also called as decision trees, it is a method to partition a data set into regions such that each region is as homogeneous as possible. The purpose of this method is Decision trees are referred to as classification trees if the predicted outcome is categorical, or as regression trees if the predicted outcome is continuous. In this study, since grouping membership is categorical, we refer CART as classification tree. It is the earliest machine learning method, and provides foundation for other new methods (e.g., bagged CART, random forests). CART is very useful in dealing with large amount of covariates, and can be used to identify important variables and interactions (Sutton, 2005). CART has been widely used among data mining community, and it also can be applied in imputation of missing values (Harrell, 2001; Cham & West, 2016).

Steps

In Classification Tree model, it will estimate the tree model recursively from top to bottom. At each step, a covariate is selected with a split value, and then participants are classified into two nodes at the next level.

To illustrate these steps, an example is shown in Figure 1. We use 5 covariates (i.e., age, income, gender, education level and GPA) to estimate propensity score. The first covariate (i.e., age) is selected with its cuts-off value (i.e., 30), and participants are classified into two nodes (age > 30 and age < 30). In the next level, participants in the left node (age > 30) are classified into two groups on income (income > 5000 and income < 5000), and participants in the right node (age < 30) are classified into two groups on gender. This process will continue to the terminal node, and the propensity score will be the percentage of participants in the treatment group for each terminal node.

Covariate and Cut-off Value Selection

The most important procedure is to select the covariate and its cut-off value to maximize the reduction of the impurity measure after splitting. To achieve this purpose, researches frequently use Gini index for each node:

$$G = 1 - \sum_{i=1}^{levels} \left(p_i \right)^2 \quad , (11)$$

where p is the proportion of participants in group i (Hastie, Tibshirani, & Friedman, 2009). For example, in Figure 1, to obtain the Gini for Age,

$G\_parent = 1 - 50\%^2 - 50\%^2 = 0.5$ for the starting point,

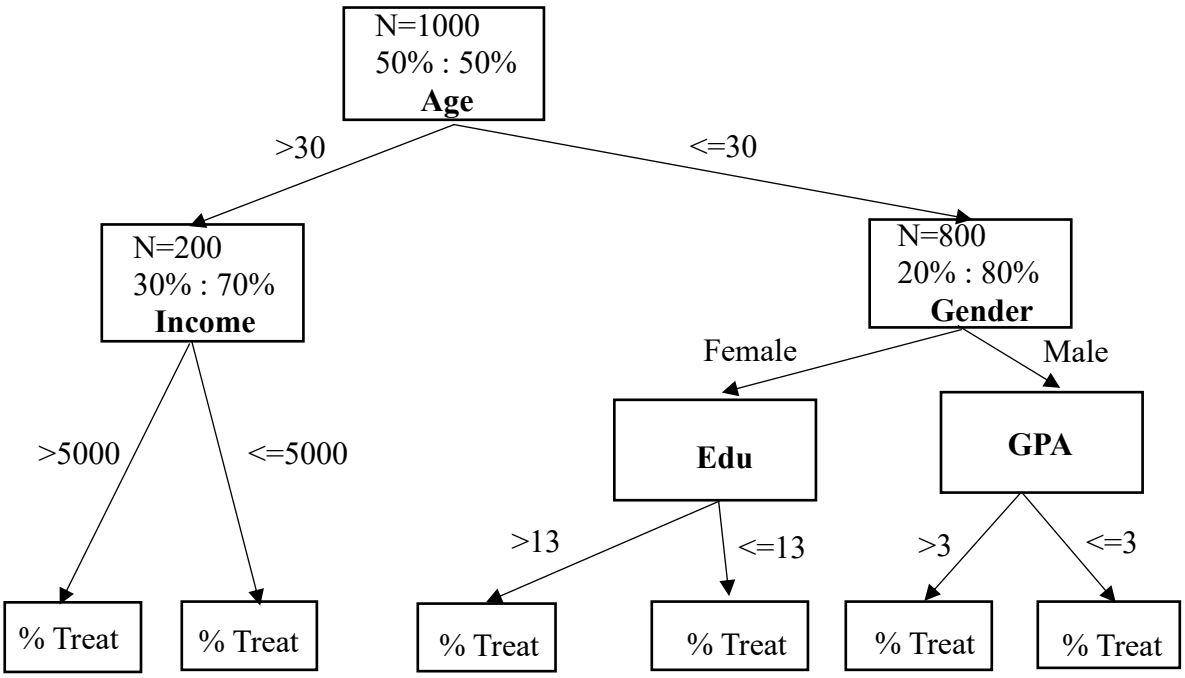

Figure 1. Classification Tree model. 50%:50% is percent of the participants in treatment vs. control group. Propensity score is the percent of participants in treatment across each final splitting group.

$G\_left = 1 - 30\%^2 - 70\%^2 = 0.42$ for the next left branch,

$G\_right = 1 - 20\%^2 - 80\%^2 = 0.32$ for the right branch.

Then we can estimate the Gini worth,

$$G\_worth = G\_parent - \left(\frac{n_1}{N}\right) G\_left - \left(\frac{n_2}{N}\right) G\_right \quad , (12)$$

where n1, n2 and N are the total number of participants in left, right branch and the starting point, respectively. In this example,

$$G\_worth = 0.5 - \left(\frac{200}{1000}\right) 0.42 - \left(\frac{800}{1000}\right) 0.32 = 0.16 .$$

After we obtain all Gini worth index for each covariate, we pick the covariate with maximum Gini worth. However, this method tends to select continuous or multi-nominal covariates, and ignores binary covariates (Berk, 2008; Hastie et al., 2009). An alternative approach is the conditional significance test for each covariate. This test is recommended since it reduces the covariate selection bias (Cham & West, 2016; Hothorn, Hornik, & Zeileis, 2006). After testing whether each covariate is associated with the treatment group given all other covariates, the covariate with smallest p value will be selected as the first node.

Tree Size

Tree may turn out to be of very high complexity with hundreds of levels, therefore it need to be optimized before applying to a new data. Optimization by minimum number of points in each node method is widely used. The splitting will stop when the number of observations in the node is less than a required minimum number.

In practice this required number is usually set to 10% of the sample size. This method works fast and easy to use.

Bagged CART

By incorporating bootstrapping subsamples of the observations into traditional CART, and then averaging over subsamples, scientists developed an advanced method - bootstrap CART or bagged CART. Bagged CART is a technique combining many classification trees to reduce the variance associated with predictions and improve prediction process. The first step is to draw many bootstrap samples from the original data. Some studies recommend use 25 to 50 bootstrap samples, and over 25 bootstrap samples will not lead to much additional improvement (Sutton, 2005). Others suggest some additional improvement may occur when bootstrap samples increase from 50 to 100 (Hastie et al., 2009). Observations not drawn in the bootstrap samples are called "out-of-bag" observations, we will discuss it later in random forests. The second step is to produce a tree from each bootstrap sample, for example, if we draw 25 bootstrap samples, we will build 25 different trees. These trees may differ from each other dramatically, so interpretations based on one single tree might be risky (Sutton, 2005). In the third, assign each observation to a group membership based on the probability over different trees. In other word, if one observation is classified in treatment group during 51% of the time over different trees, then it is assigned to the treatment group.

By averaging over the results from a large number of bootstrap samples, bagging can reduce the variance of unstable procedures and without increasing the bias, leading to improved performance. Bagging CART can improve both the stability and estimates

of the class probabilities (Hastie et al., 2009). However, interpreting the classification tree would be difficult, since bagging procedure will build more than a single tree.

Random Forests

Random forests method is a substantial modification of bagged CART that builds different trees on bootstrapping observations and then average them (Breiman, 2001). Random forests method uses similar bootstrap method to select subsamples of observations (similar to bagged CART), but it also selects a random sample of predictors/covariates before splitting each node. Random forests are implemented using the R package *randomForest*, and this method performs most accurate and interpretable results in estimating group membership (Hastie et al., 2009; Liaw & Wiener, 2002).

Steps

Random forests use similar way in CART to build each classification tree model, but it incorporates more steps in random sample of observations and covariates. Four steps in random forest are proposed to estimate propensity score.

The first step is to draw multiple random sub-samples from the data to build trees. Three issues are considered during this step. (1) *Out-of-bag samples*. An important feature of random forests is the use of out-of-bag samples. The sample is selected only from the not used observations in the classification trees model, rather than from all participations in original data. By this method, the estimation is less biased in propensity score variability and ATE estimation (Berk, 2008; Strobl, Malley, & Tutz, 2009; Cham, 2013). (2) *The number of sub-samples*. In practice, 500 sub-samples is a good option, and 500 trees will be built for each sub-sample later. (3) *Sub-sample size*. For each tree,

different subsample size is suggested to minimize the covariate selection bias, for instance, 63.2% (Strobl, Boulesteix, Zeileis, & Hothorn, 2007), 50% (Friedman & Hall, 2007) of the original sample size.

The second step is to build a Classification Tree model for each bootstrap sample. Two important features are unique to random forests. (1) *The number of covariates*. A random number of covariates (number of covariates=m) is drawn from all original covariates (number of covariates = k), which can stabilize the propensity score estimation across repeated sampling. Researchers proposed different suggestions on sub-sample of covariates (m), for instance, $m = \sqrt{k}$ (Strobl, Boulesteix, Kneib, Augustin, Zeileis, 2008), or m=k (Cham et al., 2015). The default value is set as $m = \sqrt{k}$ in the package. The optimal choice may depend on the original sample size, and further study need to clarify this option. (2) *The node size*. The number of observations in terminal node of each tree could be very small, and the node size can be very large, so the tree will be of high complexity to make the tree as less bias as possible. In the final step, propensity scores will be estimated from all classification trees, and the average scores across all trees are the final propensity scores.

Random forest is an advanced version of bagged CART by randomly draw both samples and covariates. It can reduce the variance, since it can substantially improve stability by averaging over trees. In addition, it can reduce bias, since a very large number of covariates can be considered. Due to the splitting rule, a few covariates will more likely to be selected, whereas many other competitive covariates (perform a little bit worse than those selected covariates) are rarely selected as splitting covariate. With

random forests, each covariate will have opportunities to be selected as splitting covariate, and thus the bias will be reduced. By randomly draw covariates at each possible split, the fitted values across trees are more independent, and benefit from averaging many trees will be more dramatic (Sutton, 2005).

Equating Propensity Scores among Groups

After propensity scores are estimated, next step is to equate the estimated propensity score distributions between the treatment and the control groups. Various equating methods are frequently applied, such as matching, weighting, Analysis of Covariance (Thoemmes & Kim, 2011; West et al., 2014).

Matching

Matching is frequently used and has complex procedures. Matching can be distinguished by several different dimensions: (1) Proportions between treated units match to control units (e.g., 1:1 or 1: many), it can be achieved by matching each participant to a fixed or variable number of participants in the other group. (2) Matching algorithms (exact or approximate): exact matching requires identical propensity score among groups and it is hard to achieve in practice, whereas approximate matching involves nearest neighbor matching. A caliper of one quarter of a standard deviation of the logit of the propensity score is suggested to avoid bad matches (Rosenbaum & Rubin, 1985). (3) Whether match to minimize average absolute distance on all sample (optimal matching) or whether a single match is formed with the best available unit one at a time (greedy matching). In practice, researches can combine different options in each dimensions and find a most appropriate matching method for their study.

Limitations include that a large portion of participants may be deleted during matching procedure.

Weighting

This weighting procedure reflects survey sampling weighting procedures to estimate parameters and their associated standard errors (Asparouhov, 2005). The advantages of this method includes: (1) it can utilize full sample of participants; (2) This procedure weights both the treatment group and control group, and it can obtain both the average treatment effect (ATE) over the population and average treatment effect in treatment group (ATT). Two weighting schemes have been applied to weight groups (treatment vs control) in researches. One is inverse weighting method: participants in treatment group is weighted by of $\frac{1}{\hat{\pi}}$, whereas control group is weighted by of $\frac{1}{1-\hat{\pi}}$, where $\hat{\pi}$ refers to an individual's estimated propensity score. An alternative is odds method: participants in treatment group are weighted by 1, whereas control group is weighted by of $\frac{\hat{\pi}}{1-\hat{\pi}}$. With this procedure, participants in control group but are more similar to the treatment group, will have a large propensity score (close to 1) and large weights. Similarly, participants in treatment group but are more similar to the control group, will have a small propensity score (close to 0) and small weights. The limitation includes that the weights may be highly influenced by the propensity score close to 0 or 1 (Kang & Schafer, 2007).

ANCOVA

Analysis of covariance (ANCOVA) can be conducted using the propensity score as a new covariate in the treatment effect model (or regression adjustment). This method

replaces the complex set of covariates with estimated propensity score as the only covariate in the model. The ANCOVA method is used when the propensity score and the outcome is in a linear relation, and no interaction exists between propensity score by the grouping variable (Themmes & Kim, 2011; West et al., 2014). This method is applied in some empirical studies (Hox et al., 2015; Weitzen, Lapane, Toledano, Hume, & Mor, 2004).

Covariates Balance Checks

After equating, we need to check the performance of equated propensity score by assessing whether the distribution of propensity score between treatment group and control group overlap perfectly (e.g., kernel density plots or boxplots) (West et al., 2014). Further check on the balance each covariate's distribution between treatment and control group is even more important. Two indexes are used to check the balance: absolute standardized mean difference (SMD) and variance ratio (VR). SMD is closely related to Cohen's d, except using the denominator as treatment group SD, which can standardize the equated and unequated mean difference (Stuart, 2010). A SMD of 0 indicates perfect balance, and any covariate with large SMD (i.e., SMD>.25) indicates a substantial lack of balance (Ho, Imai, King, and Stuart, 2007), it usually ranges from 0 to 1 in empirical study (Cham et al., 2015). The ratio of the variance in the treatment group and control group is also used to check the balance. A value of 1 indicates perfect balance and large VR indicates lack of balance (e.g., VR> 2.0, Rubin, 2001), it usually ranges from 1 to 2.

Research Purpose

Although failure to achieve measurement invariance happens in practice, few studies focus on how to deal with measurement non-invariance. Measurement non-invariance might come from two sources: some non-invariance items across all items, or other unbalanced covariates across groups (Van De Schoot et al., 2015). If non-invariance is due to some non-invariance items, Bayesian restricted latent factor analysis (RFA) method is applied to detect those non-invariance items (Barendse, Albers, Oort, & Timmerman, 2014). If the non-invariance is due to the artificial effect of other covariates that are not balanced across groups, propensity score can be applied to adjust the unbalanced covariates. For example, Hox et al. (2015) demonstrate how the measurement non-invariance across different mode groups (e.g., web survey, telephone survey, face-to-face interview) can be explained and corrected by other unbalanced covariates such as demographics and baseline scales. They found measurement invariance can be improved from metric invariance to scalar invariance after adding the propensity score as a covariate into the multiple group CFA. Therefore, once the potential unbalanced covariates are balanced through propensity scores, and factorial invariance test will be more accurate.

In previous studies, logistic regression is the most frequently used method in estimating propensity score, however, this method will give increasing bias especially in dealing with a composite set of covariates (Lee et al., 2010). This study will incorporate a relatively new propensity score estimation method (i.e., random forests) to test

measurement invariance, and provide a practical guide to researchers in dealing with measurement non-invariance.

The purpose of this study is to achieve accuracy in measurement invariance test by applying propensity score to adjust the potential unbalanced covariates across groups. This study will address several research questions. First, this study will demonstrate how propensity score adjustment is applied to factorial invariance test using empirical data. Second, this study will investigate the effects of propensity score adjustment on factor invariance tests. Third, this study will compare logistic regression with random forest for propensity score estimation, since both methods are commonly applied in education area. Fourth, this study will compare ANCOVA with constraining across groups for equating groups to check which method will give more accurate conclusion about measurement invariance.

CHAPTER II

STUDY ONE: EMPIRICAL STUDY

This dissertation will include two studies. First study will use an empirical data to demonstrate how propensity score adjustment is applied in measurement invariance. Second study will use Monte Carlo simulation method to examine the performance of different propensity score estimation, different equating method in measurement invariance.

Method

Data Source

To demonstrate the proposed propensity score analysis for testing measurement invariance, we analyzed the data from a previously published study (Cham et al., 2015). The study examined the effect of retention status in elementary school on grade 9 motivation for educational attainment. The retained and promoted students were equated on 67 covariates observed at baseline (i.e., grade 1). Participants were 561 students (54.37% boys), who recruited in the fall of 2000 or 2001 into a larger longitudinal study (N=784) when they were in grade 1. The ethnic composition of the participants was 35.29% Caucasian, 24.24% African American, 36.36% Hispanic, and 4.11% other. The students were from one of three school districts (one urban and two small city districts) in Texas. The criterion of selecting participants was that their scores were below the median on a district-administered test of literacy in the spring of kindergarten or the fall of grade 1. They also conducted measurement invariance test between the retained and promoted students in the bifactor model, and Chi-square difference test showed

configural invariance. However, using propensity score in measurement invariance test was not discussed in their study.

Measures

Retention Status

Students were considered retained in a given grade if they were in the same grade for two consecutive years. In this study, 177 students (31.55%) retained, and 384 students (68.45%) continuously promoted in elementary school.

Teacher Educational Expectations

The Teacher Educational Expectations subscale (5 items) is from a 32-item measuring adolescents' motivation for educational attainment (Cham, Hughes, West, & Im, 2014). The students at grade 9 responded the 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). Example items include "My teachers expect that I will do well in the future", "I am one of the students teachers believe will be successful" and "My teachers believe that I will graduate from high school". The reliability coefficient for this subscale was reported satisfactory (ω=0.8) (Reise, 2012; Cham et al., 2015).

Covariates for Propensity Score Analysis

To estimate propensity scores of the retained and promoted students, a total of 67 covariates (potential confounders) were measured at baseline (i.e., grade 1). These 67 covariates were selected based on the associations with grade retention or academic achievement (Cham et al., 2015).

Analysis

Hox et al. (2015) mentioned two ways to conduct propensity score adjustment in measurement invariance: 1) regress the propensity score on items, 2) and use propensity score as a weighting variable. To compare how the propensity score analysis can adjust measurement invariance test, we test measurement invariance in three different models: baseline model, regression model, and weighting model.

In baseline model, no propensity score was considered in the multiple group CFA analysis. In Model 1, the propensity score loading on the corresponding five items (i.e., λ_1 to λ_5) were constrained same across two groups (i.e., retained and promoted groups). In model 2, the weighting by the odds method was applied to equate the estimated propensity score distributions between two groups. Logistic regression and random forest method were both used to estimate propensity scores. Maximum likelihood MLMV estimation and DIFFTEST option in Mplus were applied to compute the χ^2 test, and χ^2 difference test (Muthén & Muthén, 1998-2014). This method can correct for both the sampling weights and the non-normal distributions of the items (Bentler & Dudgeon, 1996).

Results

As discussed in Cham et al. (2015), the propensity score balanced the retained and promoted group students on the set of observed 67 covariates at baseline.

Table 1 shows the results of measurement invariance in three different models with considering two propensity score estimation (i.e., logistic regression and random forest). This table listed the values of χ^2 test statistics, RMSEA, CFI and SRMR. The χ^2

Table 1. Empirical results for measurement invariance in baseline model and propensity scores adjustment model

			χ^2	df	*p*	RMSEA	CFI	SRMR	$\Delta\chi^{2\,a}$	df	*p*	Invariance Test
Without PS												
Baseline	1	configural	17.18	10	0.07	0.05	0.99	0.03				
	2	metric	21.98	14	0.08	0.05	0.99	0.04	5.33	4	0.25	
	3	scalar	31.23	18	0.03	0.06	0.98	0.05	12.41	4	0.01	Metric invariance
Panel A. PS with Logistic Regression												
Model 1	1	configural	21.70	15	0.12	0.04	0.99	0.04				
	2	metric	26.33	19	0.12	0.04	0.99	0.05	5.23	4	0.26	
	3	scalar	32.77	23	0.09	0.04	0.98	0.05	6.99	4	0.14	Scalar invariance
Model 2	1	configural	17.23	10	0.07	0.06	0.94	0.04				
	2	metric	30.12	14	0.01	0.07	0.87	0.11	17.00	4	<.01	
	3	scalar	50.08	18	0.00	0.09	0.74	0.16	39.04	4	<.01	Configural invariance
Panel B. PS with Random Forest												
Model 1	1	configural	22.10	15	0.11	0.04	0.99	0.03				
	2	metric	27.59	19	0.09	0.04	0.99	0.05	6.23	4	0.18	
	3	scalar	35.36	23	0.05	0.05	0.98	0.05	9.35	4	0.05	Scalar invariance
Model 2	1	configural	13.22	10	0.21	0.04	0.96	0.05				
	2	metric	19.11	14	0.16	0.04	0.94	0.09	8.40	4	0.08	
	3	scalar	23.23	18	0.18	0.04	0.94	0.10	5.88	4	0.21	Scalar invariance

Note: PS is propensity score. In baseline model, no PS adjustment is applied. In Model 1, PS is constrained same loading across groups; In Model 2, the weighting by the odds method is applied. [a] DIFFTEST option in M*plus* is applied to conduct chi-square difference test.

difference test of the models that investigate same factor pattern, factor loading and latent intercepts in a sequential order (i.e., configural vs. metric, metric vs. scalar).

Baseline model. The null hypothesis of the χ^2 difference test is that the more restricted invariance model fits the data equally well as the less restricted invariance model. Before considering propensity score adjustment, the model comparison test (i.e., configural vs. metric, metric vs. scalar in Model 1) showed metric invariance, χ^2 (4) =12.41, $p < .05$, CFI = .99, RMSEA = .05, SRMR = .04.

Logistic regression. Panel A is the logistic regression estimation for propensity score. After constraining same loading across groups (Model 1), the model comparison test (i.e., metric vs. scalar) showed scalar invariance, χ^2 (4) =6.99, $p = .14$, CFI = .98, RMSEA = .04, SRMR = .05. The regression adjustment improved metric invariance to scalar invariance. In Model 2, after applying weighting by odds method across groups, the model comparison test showed configural invariance, χ^2 (4) = 17.00, $p < .01$, CFI = .94, RMSEA = .06, SRMR = .04. The weighting by odds method did not improve measurement invariance test under logistic regression estimation. The measurement invariance decreased from metric invariance to configural invariance.

Random Forest. Panel B shows the random forest estimation for propensity score. After constraining same loading across groups (Model 1), the model comparison test (i.e., metric vs. scalar) showed scalar invariance, χ^2 (4) = 9.35, $p = .053$, CFI = .98, RMSEA = .05, SRMR = .05. The regression adjustment did improve metric invariance to scalar invariance. In Model 2, after applying weighting by odds method across groups, the model comparison test (metric vs. scalar) showed scalar invariance, χ^2 (4) = 5.88, p =

.21, CFI = .94, RMSEA = .04, SRMR = .10. Chi-square difference test showed that weighting by odds method improved measurement invariance test under random forest estimation, but the fit index (i.e., SRMR) did not show good fit for scalar invariance.

CHAPTER III

STUDY TWO: MONTE CARLO SIMULATION STUDY

Method

Data Generation

Following simulation structure in published papers (Lee et al., 2010; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008), we generate 10 covariates as standard normal random variables (see Figure 2). Six among these 10 covariates are binary, and others are continuous variables. Four covariates are associated with both grouping variable and outcome (i.e., confounders, W1 to W4), three covariates are associated with grouping variable only (i.e., exposure predictors, W5 through W7), and four covariates are associated with outcome only (i.e., outcome predictors, W8 through W10). The binary grouping variable was generated from confounders and exposure predictors, and the continuous outcome was generated from grouping variable, four confounding variables and three outcome predictors (Lee et al., 2010; Setoguchi et al., 2008). Equal size of two groups were generated.

After outcome and the grouping variable was generated from the 10 covariates, we use the outcome variable as the latent factor in the MGCFA model. A single factor model with six observed variables (i.e., X1 to X6) was used to generate data for each simulation condition. The population model we generated was scalar invariance with the propensity score adjustment. For the population parameters, we referred previous simulation studies on measurement invariance to generate scalar invariance model (Yoon& Kim, 2014; Yoon & Millsap, 2007). Propensity score loadings were constrained

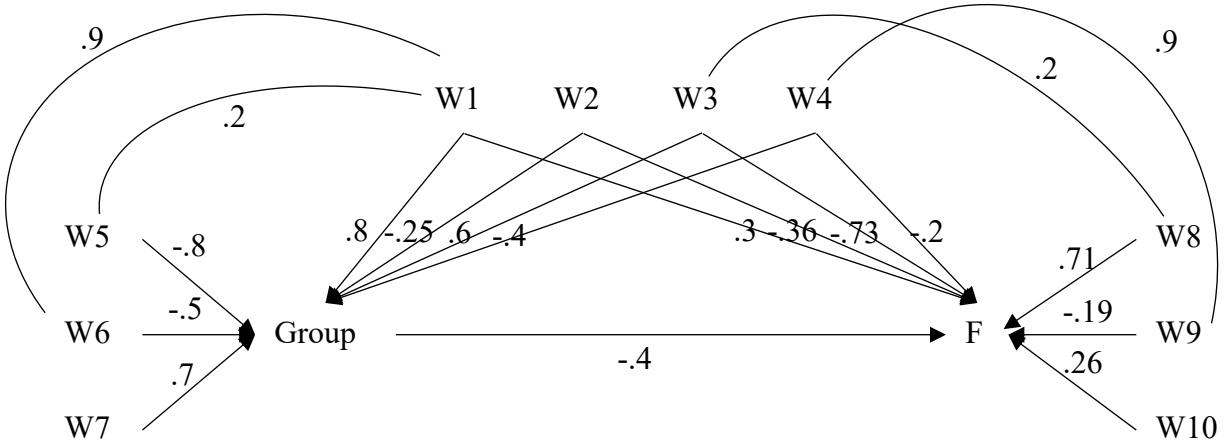

Figure 2. Covariates relations in simulation data structure (Reprint from Lee, Lessler, & Stuart, 2009).

same between group 1 and group 2, ranged from -0.4 through 0.2, which were based on empirical study propensity score loadings. Factor loadings were constrained same across two groups, ranged from 0.6 through 0.9 (see Figure 3). The intercepts were set same across two groups, ranged from -0.25 through 0.25. The residual variance of X1 through X6 were all set to 0.3 for group 1 and group 2.

Simulation Conditions

We conducted a Monte Carlo simulation study using R package and M*plus* for data generation and analysis. Two simulation conditions were investigated: (1) degree of non-linearity with grouping variable and interaction among covariates; (2) Sample size.

Degree of Non-linearity and Interaction among Covariates

Three scenarios differed in the degrees of non-linearity (i.e., quadratic) and interaction among covariates are generated in the true propensity score model (Lee et al., 2010): (1) linearity (main effects only); (2) mild interactions and non-linearity (three two-way interaction terms and one quadratic term); (3) and moderate interactions and non-linearity (10 two-way interaction terms and three quadratic terms).

Sample Size

Sample sizes were simulated as 500 (small), 1000 (medium) and 2000 (large). The selected levels of sample size are frequently used in propensity score simulation studies (e.g., Lee et al., 2010; Setoguchi et al., 2008), and measurement invariance studies (e.g., Yoon & Kim, 2014; Yoon & Millsap, 2007).

In sum, the scalar invariance model was generated for each of 9 conditions: 3 (degree of non-linearity and interaction among covariates) × 3 (sample size). Since 1000

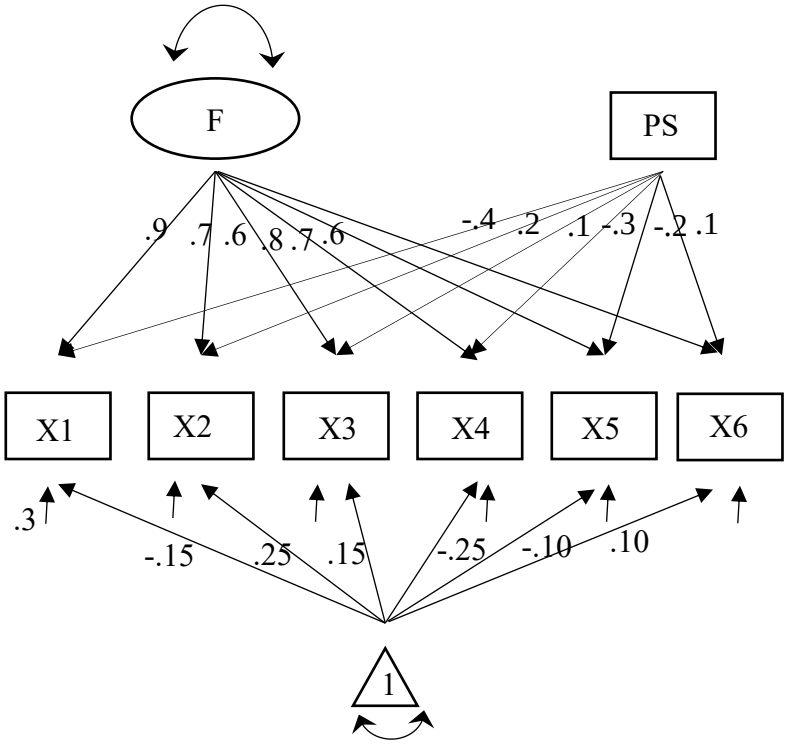

Figure 3. Multiple group CFA for group 1 with population parameters. The factor loadings, propensity score loadings and intercepts in group 2 were constrained same as group 1.

replications were frequently used in previous propensity score studies, this study will generate 1000 replications for each condition using R software.

Analysis

After generating multiple group data, we used multiple-group CFA to conduct the analysis without the propensity score adjustment and with propensity score adjustment. Maximum Likelihood parameter estimation was used for model estimation. For all analyses, cases that had any improper solutions such as non-convergence model or negative unique variance were dropped from results.

Propensity Score Estimation

First, propensity score estimated from logistic regression with main effect only for each covariate. Second, propensity score estimated by Random forests, using the *randomForest* package with default parameters (i.e., ntree = 1000, mtry = 2). (3) No propensity score was considered in the model.

Model Evaluation

Several fit statistics were examined at various simulation conditions. Chi-square statistics were used to check rejection rates of correctly specified or misspecified models. Measurement invariance under the SEM framework is typically tested through the chi-square difference test for comparing two competing nested models (i.e., the less invariant model with sequentially the more invariant model) at the significance level of $\alpha \leq 0.05$. That is, all competing models: configural vs. metric, metric vs. scalar, were tested sequentially.

In addition to Chi-square difference test, the comparative fit index (CFI), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) were used as alternative fit index, they were evaluated by looking at means. Larger value of CFI indicated a better fit, and the value of > .95 is considered as a good fit. Smaller value of RMSEA indicates a better fit, and a value of < .06 is considered as good fit. Smaller value of SRMR indicates a better fit, and a value of < .08 is considered as good fit (Hu & Bentler, 1999).

First, admissible solutions were checked. Inadmissible models include the cases in which parameter estimates and standard errors were not provided or not within a plausible range (e.g., negative variance).

Second, Type I error occurs when the hypothesis of scalar invariance is rejected, resulting in incorrect identification of configural or metric invariance. In current study, Type I error rate was defined as the proportion of replications in which scalar invariance is incorrectly rejected. In other words, any model did not have fair fit (i.e., CFI < .95, RMSEA > .06, or SRMR > .08) was Type I error. Type I error rates were considered acceptable when they did not exceed the sampling error rates (i.e., 5%).

Finally, the examination of Type I error rate, analysis of variance (ANOVA) was conducted to entangle the factors influencing Type I error in factorial invariance testing. Eta-square of two design factors effects (i.e., sample size, degree of non-linearity) on Type I error was computed.

Results

In current study, no replication had any improper solutions such as non-convergence model or negative unique variance. All models converged successfully across different conditions, and therefore all models had full admissible solutions.

Multiple Group CFA without Propensity Score Adjustment

Type I error rates referred to the proportion of the replications in which the null hypothesis (i.e., scalar invariance) was incorrectly detected as non-invariance (i.e., CFI < .95, RMSEA > .06, SRMR > .08 for scalar invariance model). As shown in Table 2, when the propensity score was completely ignored in multiple group CFA, Type I error rates became seriously inflated, ranging from 82.9% to 92.4%. In other words, the null hypothesis was overly rejected when propensity score was not applied to adjust unbalanced covariates. Therefore, the null hypothesis of scalar invariance was incorrectly rejected when the propensity score adjustment was completely ignored.

The one-way ANOVA test showed that sample size had a statistical significant effect on Type I error rates ($\eta^2 = 64.60\%$), but not significant for nonlinearity. As sample size increased, Type I error rates increased as well. Nonlinearity did not affect Type I error rate as much as sample size did ($\eta^2 = 26.65\%$).

CFI had means from 0.958 to 0.966. For N=500 or 1000, 8.9% to 29.2% of the replications correctly identified the model as inadequate good fit, using the cutoff value of .95 (Hu & Bentler, 1999). However, as sample size increased to 2000, small percent of the replications (0.8% to 2.3%) failed the cutoff of .95, incorrectly identified the model as adequate good fit.

Table 2. Simulation results without propensity score adjustment

			CFI		RMSEA		SRMR		$\Delta\chi^2$	
Nonlinearity	N	Error	Mean	% < .95[a]	Mean	% > .06[a]	Mean	% > .08[a]	Mean	% sig[b]
Low	500	82.9	0.959	25.6	0.074	82.7	0.067	15.5	44.170	100.0
	1000	86.0	0.963	8.9	0.071	86.0	0.056	0.6	75.990	100.0
	2000	85.9	0.966	0.8	0.067	85.9	0.049	0.0	133.420	100.0
Medium	500	84.1	0.958	28.0	0.075	84.0	0.067	14.6	45.046	100.0
	1000	88.9	0.960	11.6	0.072	88.9	0.056	0.4	79.140	100.0
	2000	92.4	0.964	2.3	0.069	92.4	0.048	0.0	140.357	100.0
High	500	84.0	0.958	29.2	0.074	83.8	0.066	13.9	44.570	100.0
	1000	88.6	0.962	12.2	0.072	88.6	0.055	0.4	78.439	100.0
	2000	89.9	0.965	0.9	0.069	89.9	0.047	0.0	137.467	100.0

[a] percent of replications with the fit index smaller or larger than the specified value.
[b] percent of replications with $\Delta\chi^2$ being statistically significant at $p < .05$.

RMSEA had means from 0.067 to 0.075. Among all conditions, 82.7% to 92.4% of the replications failed the cutoff of .06 (Hu & Bentler, 1999), correctly identified the model as inadequate good fit among all conditions.

SRMR had means from .047 to .067. For N=500, 13.9% to 15.5% of the replications failed the cutoff of .08 (Hu & Bentler, 1999), correctly identified the model as inadequate good fit. As sample size increase, SRMR did not identify inadequate good fit model (ranged from 0% to 0.6%). SRMR was more sensitive to non-invariance when sample size was small.

Chi square difference test ($\Delta\chi^2$) detected all the non-invariance, it flagged 100% of the replications as non-invariance of intercepts.

Multiple Group CFA with Random Forest Adjustment

As shown in Table 3, after applying random forest propensity score adjustment, Type I error rates reasonably ranged from 0 to 3.7%, which were within the acceptable range of Type I error rate (i.e., 0-5%). That is, after applying random forest propensity score to adjust unbalanced covariates, multiple group CFA measurement invariance became acceptable under the null hypothesis conditions.

Non-linearity or sample size did not have statistical significant effect on Type I error rates based on ANOVA test.

CFI had means from 0.994 to 0.999. None of the replications detect the model as inadequate good fit, passing the cutoff value of .95.

RMSEA had means from 0.007 to 0.020. Small percent of replications (ranged from 0% to 1%) identified the model as inadequate good fit among all conditions, using

Table 3. Simulation results for random forest

			CFI		RMSEA		SRMR		$\Delta\chi^2$	
Nonlinearity	N	Error	Mean	% < .95[a]	Mean	% > .06[a]	Mean	% > .08[a]	Mean	% sig[b]
Low	500	0.4	0.996	0.0	0.013	0.4	0.046	0.0	4.960	6.3
	1000	0.0	0.998	0.0	0.010	0.0	0.033	0.0	5.130	5.1
	2000	0.0	0.999	0.0	0.007	0.0	0.024	0.0	4.993	5.8
Medium	500	1.6	0.995	0.0	0.016	0.6	0.049	1.0	4.962	4.6
	1000	0.0	0.997	0.0	0.013	0.0	0.037	0.0	5.036	4.9
	2000	0.0	0.998	0.0	0.011	0.0	0.029	0.0	4.797	3.6
High	500	3.7	0.994	0.0	0.020	1.0	0.054	3.3	5.026	4.7
	1000	0.3	0.995	0.0	0.019	0.1	0.044	0.3	5.115	5.6
	2000	0.0	0.996	0.0	0.020	0.0	0.038	0.0	5.073	6.2

[a] percent of replications with the fit index smaller or larger than the specified value.
[b] percent of replications with $\Delta\chi^2$ being statistically significant at $p < .05$.

the cutoff value of 0.06.

SRMR had means from .024 to .054. Small percent of replications (ranged from 0% to 3.3%) identified the model as inadequate good fit, using the cutoff value of .08.

$\Delta\chi^2$ detected small percent of replications, it flagged 3.6% to 6.3% of the replications as non-invariance.

Multiple Group CFA with Logistic Regression Adjustment

As shown in Table 4, after logistic regression adjustment, Type I error rates ranged from 20.6% to 35.9% when N = 500, which were off the acceptable range of Type I error rate (i.e., 0% to 5%). As sample size increased, Type I error rates decreased. Type I error rates were inflated especially under high degree of non-linearity condition.

Sample size had a statistical significant effect on Type I error rates, but not significant for degree of nonlinearity. The factor related to Type I error rate was mainly sample size (η^2 = 83.68%). Nonlinearity did not affect the Type I error rate as much as sample size did (η^2 = 11.67%).

CFI had means from 0.973 to 0.985. Small percent of the replications (ranged from 0% to 5.4%) failed to pass the cutoff value of .95, identifying the model as inadequate good fit.

RMSEA had means from 0.041 to 0.054. For N=500, 20.2% to 35.1% of the replications correctly identified the model as inadequate good fit. However, as sample size increased to 2000, small percent of the replications (0% to 1.3%) identified the model as inadequate good fit.

Table 4. Simulation results for logistic regression

			CFI		RMSEA		SRMR		$\Delta\chi 2$	
Nonlinearity	N	Error	Mean	% < .95[a]	Mean	% > .06[a]	Mean	% > .08[a]	Mean	% sig[b]
Low	500	20.6	0.978	2.3	0.047	20.2	0.054	1.1	26.110	97.1
	1000	4.9	0.982	0	0.045	4.9	0.042	0	40.800	100.0
	2000	0.0	0.985	0.0	0.041	0.0	0.033	0.0	62.268	100.0
Medium	500	23.4	0.978	1.8	0.048	23.2	0.055	1.8	25.670	96.2
	1000	5.4	0.981	0.0	0.046	5.4	0.043	0.0	40.269	100.0
	2000	0.1	0.985	0.0	0.041	0.1	0.034	0.0	60.965	100.0
High	500	35.9	0.973	5.4	0.054	35.1	0.060	5.3	28.890	97.7
	1000	15.8	0.976	0.0	0.052	15.8	0.050	0.3	47.365	99.9
	2000	1.3	0.980	0.0	0.048	1.3	0.042	0.0	73.710	100.0

[a] percent of replications with the fit index smaller or larger than the specified value.
[b] percent of replications with $\Delta\chi^2$ being statistically significant at $p < .05$.

SRMR had means from .033 to .060. Small percent of replications (ranged from 0% to 5.3%) identified the model as inadequate good fit (i.e., fail to pass the cutoff value of .08).

Chi-square difference test ($\Delta\chi^2$) detected almost all replications (ranged from 96.2% to 100.0%) as non-invariance.

CHAPTER IV

DISCUSSION AND CONCLUSIONS

Discussion

In educational psychology field, latent constructs were measured by multiple observed items. The relation between observed items and latent construct should be same across groups. Measurement invariance is critical before comparing group difference on the observed items. In practice, measurement invariance across groups might not be well established (i.e., measurement non-invariance), which is problematic if comparison of group means is of research interest. Propensity score is one approach to correct unbalanced covariates across groups if these unbalanced covariates are the source of measurement non-invariance (Hox et al., 2015).

The main purpose of this dissertation is to evaluate propensity score adjustment in testing measurement invariance in both empirical data and Monte Carlo simulation study. Empirical study demonstrated how to conduct propensity score adjustment in measurement invariance test. Monte Carlo simulation considered different conditions, including sample size, degree of non-linearity and propensity score estimation methods. Specifically, Type I error rates were defined as the proportion of the replications in which the null hypothesis of scalar invariance was incorrectly detected as non-invariance (i.e., CFI < .95, RMSEA > .06, or SRMR > .08). Chi-square difference test were also presented in the study. Multiple group confirmatory factor analysis was conducted to test measurement invariance in both empirical data and Monte Carlo simulation study.

In empirical study, when propensity score was estimated by logistic regression or random forest, and applied as a new covariate to adjust unbalanced covariates across groups, measurement invariance was improved from metric invariance to scalar invariance. The improvement after applying propensity score adjustment was consistent with previous study (Hox et al., 2015).Weighting by odds method with random forest estimation improved the metric invariance to scalar invariance, but weighting with logistic regression did not. One possible reason is that logistic regression estimation is not as accurate as random forest estimation, especially when the empirical data included 67 observed covariates to estimate propensity score. In this situation, weighting the propensity score may enlarge the estimation bias.

However, the empirical study did not tell which estimation method was better, and how the result may differ if inappropriate method was applied. The Monte Carlo simulation study can answer this question and give guidelines for using propensity score estimation method under different conditions.

In the Monte Carlo simulation study, one of the most salient findings is that substantial Type I error rate inflation occurred when propensity score adjustment in unbalanced covariates was ignored and multiple group CFA was applied in measurement invariance test. That is, the invariant model is more likely to be rejected and misleading concluded to be non-invariant when the unbalanced covariates are not taken into account for the analysis. Therefore, multiple group CFA without propensity score adjustment is not recommended for measurement invariance test if unbalanced covariates exist in the data given the considerable inflation in the Type I error rates.

In addition, Type I error rate was within the acceptable range when random forest propensity score adjustment was employed in measurement invariance test. In other words, the invariant model is more likely to fail to reject and correctly conclude to be invariant after the unbalanced covariates are taken into account in the measurement invariance test. Therefore, propensity score adjustment with random forest estimation is recommend in measurement invariance test if unbalanced covariates exist in the data.

Finally, substantial Type I error rate inflation was also observed if logistic regression was applied to adjust the unbalanced covariates. This inflation was not as serious as completely ignoring the propensity score adjustment. As degree of non-linearity became high (including high level of interactions and quadratic terms), the inflation became more serious. This indicates that logistic regression could not successfully estimate propensity score when the covariates have complex interaction. As sample size was large (i.e., 2000), Type I error rate inflation was minimized.

Limitation and Future Research

In this study, only a limited number of conditions were considered, for example, one single factor and two groups were included in simulation study. In empirical cases, measurement invariance often involves more than one single factor. Current study aimed to examine how different propensity score estimation methods would perform differently in the simplest cases. In future research, more conditions can be considered such as multiple groups and factors.

Another limitation of current study is that weighting method was not used to simulate the data since simulation in M*plus* or R is not suitable for this specific data.

Weighting method is another promising method to balance group (Hox et al., 2015). In future study, more flexible software can facilitate simulation data and will give interesting results.

Conclusions

In conclusion, when ignoring the unbalanced covariates among groups and using multiple group CFA to conduct the measurement invariance test, large Type I error rate inflation was observed. Therefore, the invariant models were overly rejected and concluded to be non-invariant when unbalanced covariates were not adjusted by propensity score in measurement invariance test. Therefore, when unbalanced covariates exist in the data, an appropriate method that can balance covariates is required before conducting measurement invariance test. The current study combined the empirical data and Monte Carlo simulation study to evaluate propensity score adjustment in testing measurement invariance. The latest machine learning estimation method (i.e., random forest) was applied to obtain accurate propensity score.

REFERENCES

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434.

Barendse, M.T., Albers, C.J., Oort, F.J., and Timmerman, M.E. (2014). Measurement bias detection through Bayesian factor analysis. *Front. Psychol.* 5:1087. doi:10.3389/fpsyg.2014.01087

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*(1), 563-592.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Cham, H. (2013). Propensity score estimation with random forests (Doctoral dissertation). Arizona State University. Retrieved from ProQuest Dissertations and Theses. (Accession Order No. UMI 3567836)

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21*(3), 427-445. doi:10.1037/met0000076

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2014). Assessment of adolescents' motivation for educational attainment. *Psychological Assessment, 26*(2), 642.

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2015). Effect of retention in elementary grades on grade 9 motivation for educational attainment. *Journal of School Psychology, 53*(1), 7-24.

Friedman, J. H., & Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference, 137*(3), 669-683.

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

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199-236.

Hong, G. (2004). Causal-inference for multi-level observational data with application to kindergarten retention (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674.

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, *6*. doi: 10.3389/psyg.2015 00087

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.doi:10.1080/10705519909540118

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523-539.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18-22.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*(4), 611-635.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57(*2), 289-311.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5-9.

Muthén, L. K., & Muthén, B. O. (1998-2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224.doi:10.2307/1165073

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.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41,* 103–116. http://dx.doi.org/10.2307/2530647

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546-555.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-107.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 1.

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

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics*, *24*, 303-329.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson/Allyn & Bacon.

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.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, *13*(12), 841-853.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*(5), 906.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Yoon, M. (2008). Statistical power in testing factorial invariance with ordinal measures. *Dissertation Abstracts International*, *68*(11), 7705B-7854B.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435-463.

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, *46*(4), 1199-1206.