Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 21, Number 13, December 2016 ISSN 1531-7714

Investigating Causal DIF via Propensity Score Methods

Yan Liu, Bruno D. Zumbo, Paul Gustafson, Yi Huang, Edward Kroc, Amery D. Wu,
University of British Columbia

A variety of differential item functioning (DIF) methods have been proposed and used for ensuring that a test is fair to all test takers in a target population in the situations of, for example, a test being translated to other languages. However, once a method flags an item as DIF, it is difficult to conclude that the grouping variable (e.g., test language) is responsible for the DIF result because there may exist many confounding variables that lead to the DIF result. The present study aims to (i) demonstrate the application of propensity score methods in psychometric research on DIF for day-to-day researchers, and (ii) describe conditional logistic regression for matched data in a DIF context. Propensity score methods can help to achieve the comparability between different populations or groups with respect to participants' pre-test differences, which can assist in examining the validity of making a causal claim with regard to DIF.

In the development of educational, psychological, or licensure tests, or in the adaptation of tests to another language, an essential issue is to make sure that the test is fair to all test takers in the target population and the comparison of test scores is meaningful. For example, in recent years more than 60 countries have participated the Third International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA). Many researchers have used the results of these international tests to inform their educational policy and school practice. It should be noted that these tests are often developed in one language first and then adapted to other languages for participants from different countries (e.g., Johansone & Malak, 2008). Therefore, an important question that has been raised is, "Do the test items in different languages measure the same abilities?" A similar question has been raised regarding the development of computerized tests, "Do the same items function the same in different test administration modes (e.g., paper-and-pencil vs. computerized tests) for all test takers?"

Various differential item functioning (DIF) methods have been introduced to address these kinds of issues (e.g., Angoff, 1972, 1993; Cardall & Coffman, 1964; 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 has been flagged as DIF, test developers often proceed to examine whether it indeed puts one group at disadvantage and favors the other due to some extraneous sources other than the ability or attribute, such as the translation or administration mode. The researchers then make a decision whether the items should be removed from the test.

The present study has two purposes. The first purpose is to demonstrate the application of propensity score methods in assessment and testing research on DIF for day-to-day researchers. Propensity score matching methods can help to achieve the comparability between different populations or groups with respect to participants' pre-test differences, which can assist in examining *the claim of DIF being the cause of item bias*. The second purpose is to introduce the use of *conditional logistic regression* for data analysis based on matched data to the fields of assessment and testing

literature. These propensity score DIF methods are also compared to other conventional DIF methods. More specifically, the present study demonstrates how to apply propensity score methods with logistic regression analysis when examining DIF due to the effect of translation from English to French using Trends in International Mathematics and Science Study (*TIMSS*) 2007 mathematics test data.

At the outset, it is important to point out that conditional logistic regression differs from the conventional logistic regression such that the conditional maximum likelihood function is specified only for the discordant pairs/clusters[1] Conditional logistic regression is widely used for case-control studies in epidemiology and biostatistics research, but has been largely neglected in assessment research (e.g., Breslow, Day, Halvorsen, Prentice, & Sabai, 1978; Langholz & Goldstein, 2001; Le & Lindgren, 1988; Lienhardt, et al., 2005). It will be explained in more detail in the section of *Description of Conditional Logistic Regression DIF Analysis* as well as the demonstration section

This paper is organized into the following six sections: (i) group non-equivalence: a description of the fundamental problem at hand, (ii) a review of logistic regression DIF analysis, (iii) a description of propensity score matching methods, (iv) a description of conditional logistic regression, (v) a demonstration of conditional logistic regression DIF analysis using propensity score optimal matching methods, and (vi) a general discussion.

Group Non-Equivalence: A Description of the Fundamental Problem At Hand

One major challenge for all conventional DIF analyses is that they can only detect DIF, but cannot disentangle, for example, the effect of translation or administration mode from other confounders, personal or contextual factors (Zumbo, 2007). For instance, researchers would not know if the DIF of an item were due to translation problem or other factors when they found existent differences in students' learning motivation, parents' education, and social economic status. This is more likely the case in educational settings because a lot of confounders covary with outcome variables. Hence, a typical DIF analysis cannot help test developers to decide on whether they should throw away an item flagged as DIF due to, for instance, translation problems. Unlike randomized experimental studies, DIF studies are based on observational data. Randomized experimental design can create equivalent groups and balance out the confounders by the randomization process (i.e., random assignment). However, observational studies, such as DIF studies, typically do not have equivalent groups before the testing.

The most common attempts to approximate group equivalence are matching and covariance adjustment. In the context of DIF, matching is a method of selecting units from the reference group who are similar to those in the focal group with respect to the observable covariates that are related to group membership mechanism. Herein, the reference group is equivalent to the control group and the focal group is equivalent to the treatment group in an experimental design. However, exact matching becomes onerous or even impossible when matching on a large number of covariates, especially when several continuous covariates are involved. This will result in the sparse data problem, that is, some units from the treatment group do not have matched units from the control group. Rosenbaum and Rubin (1983) described this problem and indicated the needs to find approximate matching methods instead of exact matching.

Stratification is an alternative to matching. Using this method, groups are classified into several strata and in each stratum units from the focal group are comparable to the units from the reference group (Rosenbaum, 2002). While easier to implement than the exact matching methods, stratification methods may still produce extremely unbalanced groups within certain strata. However, stratification may also run into the sparse data problem as exact matching methods. Cochran (1965) pointed out that the number of strata (combinations of different values/categories of the covariates) grows exponentially when the number of covariates increases, even for binary covariates. For example, when we have ten binary categorical covariates, there will be 1024 strata (210). With so many strata, some of strata may only include units from the focal group, but not from the reference group or vice versa. Thus, it is impossible to directly

[1]A discordant pair is a pair of participants matched on the propensity scores, one from the focus group and the other from the reference group, whose outcome scores on an item are different. Similarly, in a discordant set the score from one participant of the focus group is different from the scores obtained from the matched participants of the reference group or vice versa.

compare the two groups when one group contains no units within a stratum.

Another common strategy is the covariance adjustment, such as ANCOVA or regression analysis, of which conventional DIF methods allowing for the adjustment of confounders are an example (Zumbo, 2007, 2008). While familiar to most researchers, these methods may not be able to give a reliable adjustment on the differences in the observed covariates when there are substantial differences in the distribution of these covariates between the two groups (Cochran, 1957; Rubin, 2001). We will use a hypothetical example to illustrate this problem. To appropriately compare student mathematics performance in private versus public schools, we want to adjust for parent annual income. However, parent income in public schools ranged from $10,000 to $30,000 while in private schools it ranged from $40,000 to $60,000. Hence, the distributions of parent annual income for public and private schools do not overlap at all.

Figure 1 illustrates how covariance adjustment does not work well for comparing the mathematics performance among students using the above hypothetical example. The two lines represent the groups. The X-axis in Figure 1 shows parent income, with an overall average of $\overline{X}$ = $35,000. However, neither group contains observations at or around $\overline{X}$. The dashed regression lines, extrapolated for the groups and based on the existent observations, are what we use to compare the two groups. The adjusted means of outcomes, $\widehat{Y}_1$ and $\widehat{Y}_2$ represent our best guess

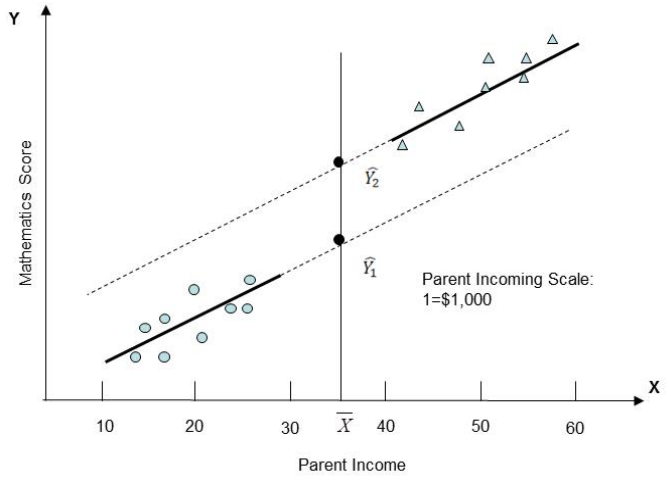

Figure 1. Covariance adjustment for comparing the mathematics performance of students from public and private schools with non-overlapping covariate distributions

on what student average scores would be if the two groups had not differed on parent income. Since these estimates are based on extrapolations, the average group differences adjusted by parents' income ($\widehat{Y}_1$ and $\widehat{Y}_2$) may be incorrect and cannot be trusted for having accurately removed the pre-treatment/pre-test difference. See Zumbo (2008, p. 45) for a similar example in the context of DIF analyses across testing language (English versus French).

It is clear that there is a need to develop more precise methods of DIF that can help to control for confounders. Dorans and Holland (1993) suggested that propensity score matching might be a good solution instead of matching directly on multiple observed variables. Bowen (2011) conducted Mantel-Haenszel DIF analyses after controlling for distributional differences using propensity scores. However, Bowen only used one covariate, total test scores, for estimating propensity scores. Lee and Geisinger (2014) adopted propensity scores to control for the contextual sources when examining gender DIF. Their study shows that the Mantel-Haenszel and logistic regression methods based on propensity scores detected less number of gender DIF items than do the conventional Mantel-Haenszel and logistic regression methods. They suggest that the propensity score approach is a promising strategy for studying the cause of DIF because it can be used for balancing pre-test differences between groups and achieving an effect akin to random assignment if the key covariates are collected. These previous studies, however, did not take into account of the dependence structure of matched pairs or matched sets in their DIF analyses, an issue we will address in the description of conditional logistic regression section.

Review on Conventional Logistic Regression DIF Analysis

A variety of analytical methods have been proposed for detecting DIF. Among them, logistic regression has been highly recommended because of its flexibility and can test both uniform and non-uniform DIF (Swaminathan & Rogers, 1990; Zumbo, 1999, 2007, 2008). Conceptually, the conventional logistic regression DIF analysis is a procedure in which group, ability, and an interaction between group and ability are used to predict the probability of a correct answer to an item of a given sample. Most commonly, the grouping variable is binary, representing a participant's group

membership, while an examinee's total test score is used as a proxy for ability. A typical practice in the DIF literature is to designate a reference group as the group who is suspected to have an advantage over a focal group, though this designation is arbitrary. Group membership is usually defined in terms of a *focal group* (G = 1) and a *reference group* (G = 0).

Two types of DIF are usually distinguished: *uniform DIF* and *non-uniform DIF* (Mellenbergh, 1982). If the regression coefficient of the grouping variable is statistically significant, it suggests that the probability of answering the item correctly is different between these two groups after controlling for the ability; this is the so-called *uniform DIF*. The ability variable should always be statistically significant because examinees with a higher ability should have a higher probability of answering it correctly. If the regression coefficient of the group and ability interaction term is significant, we say that non-*uniform DIF* is present. This scenario suggests that the probability of getting the item correct is different between the two groups and the direction and/or magnitude of the differences may vary depending on participants' abilities. One of the main advantages of using logistic regression for DIF detection is its ability to identify both uniform and non-uniform DIF, a major advantage over other methods, such as Mantel-Haenszel test.

The conventional logistic regression DIF analysis can be conducted in three steps, *null model*, which only has the total test scores, *uniform model*, which includes both the total test scores and the grouping variable, and *non-uniform model*, which adds the interaction of total test scores and grouping variable to the uniform model. Or one can go directly to use the last equation to test both uniform and non-uniform DIF simultaneously. The equations are as follows:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 total$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 total + b_2 group$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 total + b_2 group + b_3 total * group$$

where p_i is the proportion of examinees that answer the item i correctly; *total* indicates the total test scores for each participant; *group* is the dummy coded grouping variable (0 = reference group, 1 = focal group); and *total*group* indicates the interaction between the two. The coefficient b_1 indicates the relation between a person total test score and the score on the item; b_2 shows the mean score difference between the two groups on the item; and b_3 shows the interaction between the person's total test score and the group membership.

Description of Propensity Score Methods

In a randomized experimental study, the random assignment tends to make the groups comparable (balanced over both observed and unobserved covariates); hence, any differences between the groups prior to treatment are only due to chance (Rosenbaum, 2002). However, quasi-experimental or observational studies are widely used to look for cause-effect relationships in psychology, education, social behavioral sciences, biology, and economics, whenever randomized experiments are not ethical or not feasible. When the assignment mechanism is non-random, it is difficult to judge whether differences in the outcomes are due to the treatment or pre-existing differences between groups.

Propensity score matching was first proposed by Rosenbaum and Rubin (1983) and has become a popular method used in medical and economic research, lately extending its popularity into the fields of social, psychological and educational research (Austin, 2008; Thoemmes & Kim, 2011). Propensity score approach 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. That is, the purpose of using propensity score is to balance the characteristics of non-equivalent groups, so that treatment and control groups with the same value of propensity score have the same multivariate distribution of the observed covariates (e.g., Rosenbaum, 1995, 2002, 2010; Rosenbaum & Rubin, 1983; 1985; Rubin, 2001; Schafer & Kang, 2008). To solve the sparseness problem raised by the conventional matching methods, propensity score methods create a single composite score from all observed covariates and match observations from two groups on the basis of one dimensional propensity scores alone.

Formally, propensity score is defined as the conditional probability of assigning an individual to the treatment condition given a set of observed covariates. The expression for the propensity score is:

$$\boldsymbol{e(X_i) = P(Z_i = 1|X_i)}$$

where $e(X_i)$ denotes the propensity score for each individual *i*; Z_i is an indicator for group variable/treatment conditions, and Z_i =1 refers to participants belonging to the treatment group or focal group in DIF context; X_i is a vector of scores on the observed covariates (Rosenbaum & Rubin, 1983). Using conventional notation, the propensity scores are usually estimated by logistic regression:

$$P(Z_i = 1|X_i) = \frac{e^{\beta_0+\beta(x_i)}}{1 + e^{\beta_0+\beta(x_i)}} \quad (1)$$

where β_0 is an intercept and β is a vector of coefficients of covariates (e.g., D' Agostino, 1998; Rosenbaum, 2010, p. 167; Rosenbaum & Rubin, 1983).

The adjustment of group differences using estimated propensity scores is usually accomplished by using one or a combination of the four commonly used statistical methods: (a) matching, (b) stratification, (c) weighting, and (d) covariate/regression adjustment. The first two methods are conducted in two stages, adjusting for the covariates first, and then calculating the group difference/treatment effect, whereas the other two methods are used for the actual adjustment while determining the group difference/treatment effect. Both matching and stratification methods are described herein because they share some similarity, but only propensity score matching is demonstrated in the present study.

Propensity Score Matching

Propensity score matching is a good strategy when a great number of covariates are collected, especially the key ones that can assist in approximating the random assignment mechanism. There are a variety of methods for matching, but the most widely used are greedy (e.g., nearest neighbour) matching and optimal matching (Guo & Fraser, 2014; Pan & Bai, 2015). In greedy matching, a treated unit usually is first selected at random, and a control unit whose propensity score is closest to that of this treated unit is chosen for matching the treated unit. The process is then repeated until all treated units are matched with control units. After matching a pair is not considered for further matching in a manner similar to stepwise regression by forward selection. Optimal matching is similar to greedy matching, but matches can be redone in optimal matching if a more satisfactory match is found for a case. This algorithm helps to minimize the overall global propensity score distance by going back and forth to adjust at the pair level (Rosenbaum, 1991). The optimal matching was adopted for the present demonstration, but one should be aware of other options. Theoretically, optimal matching should perform better than greedy matching in terms of the overall global propensity score distance, but not necessarily the case in terms of achieving the minimum distance for each individual covariate. Researchers are encouraged to compare greedy and optimal matching methods in their practice as some researchers have shown that both of them may work well, but one may perform better than the other in certain conditions (e.g., Austin, 2014; Gu & Rosenbaum, 1993).

The most commonly used optimal matching methods are optimal pair matching and optimal full matching. Both of them also have certain limitations. With pair matching, subjects are matched in pairs and the unmatched subjects are excluded from the analysis after matching, leading to a reduction in sample size. The sample size could be substantially reduced for the final analysis when the focal group is much smaller than the reference group, which may result in under-representation of the original sample and lower power for significance tests.

Since Rosenbaum (1991) introduced optimal full matching, many researchers have tried to examine its performance, applied this method to empirical data, and develop software program to extend its use. For example, Gu and Rosenbaum (1993) conducted a simulation study, Marcus (2000) applied it to the evaluation of the Head Start compensatory education program, and Hansen and Klopfer (2006) extended the full optimal matching using *optmatch* *R* package (Hansen, 2004; Hansen, Fredrickson, Buckner, Errickson, & Solenberger, 2016). In an optimal full matching, matched sets may contain a single treated unit and multiple matched control units (one-to-many) or many treated units with a single matched control unit (many-to-one). The propensity score distance in the one-to-many or many-to-one cases will be adjusted by weights based on the number of matched cases included (Rosenbaum, 1991; Rosenbaum, 2010, p.179-183). With full matching, all subjects are used for matching, but the matching criterion may be looser than that of a pair matching, and hence the balance of the group distributions sometimes may not be as good as that obtained from a pair matching. So in practice, we should compare the balance of the group distributions obtained from both optimal matching

methods and choose the one with a relatively better balance.

We also want to briefly describe stratification approach as it share some similarity with the full matching method. Using propensity score stratification method, subjects are ranked according to their estimated propensity scores and then are categorized into homogenous strata with similar propensity scores. A common approach is to divide subjects into five equally-sized subgroups using the quintiles of the propensity scores. The group difference/treatment effect is estimated within each stratum and then the overall group difference is computed based on a combination of the results from all the strata using weights. Rosenbaum and Rubin (1984) showed that approximately 90% of the selection bias due to confounders can be eliminated by stratifying on the quintiles of the propensity score when estimating a linear treatment. It should be noted that full matching can be considered as a special case of stratification with either one treated unit or one control unit in all the possible matched sets; the matched sets in full matching have been called in different terms, matched sets, matched clusters, subsets, strata or subclasses, in the literature. We do not include a demonstration of propensity score stratification in this paper because of space limitations.

Description of Conditional Logistic Regression

As we mentioned earlier, the regular logistic regression analysis is not appropriate for matched data, which has been largely documented in the literature of case-control studies. The matched sets (matched pairs or matched clusters) are analogous to paired, nested, or multilevel data. Hence, it is important to take account of this nested relationship or dependence structure in one's data analysis. Unfortunately, a lot of previous studies neglected this dependence structure in their matched data and simply conducted regular regression analysis that assumes data independence. For data matched by pairs (e.g., optimal pair matching, greedy matching) or by sets/clusters (e.g., full matching), *conditional logistic regression* is more appropriate because it takes into account of the dependence structure in the data due to matched pairs or matched sets.

The fundamental difference between the conventional/regular and the conditional logistic regression models is that the parameters in the conditional logistic regression are estimated using paired or clustered sample. The discordant pairs or clusters are used for the conditional likelihood estimations, while concordant pairs[2] or clusters are disregarded as they cannot provide any information for the conditional likelihood estimation. Following Hosmer et al.'s notation, the conditional likelihood function for the pair matching is provided as follows (Hosmer, Lemeshow, & Sturdivant, 2013, p.247):

$$l(\beta) = \prod_{k=1}^{K} \left\{ \left[\frac{e^{\beta^T (x_{1k} - x_{0k})}}{1 + e^{\beta^T (x_{1k} - x_{0k})}} \right]^{y_k} \left[\frac{1}{1 + e^{\beta^T (x_{1k} - x_{0k})}} \right]^{1 - y_k} \right\}^{\delta_k}$$

where k indicates the pairs (k=1, 2, ⋯, K); δ_k indicates whether the kth pair is discordant (0/1); y_k is an item score for a treated unit in the kth pair (0 or1) whereas $1 - y_k$ is an item score for a control in the kth pair (0 or 1); β^T is the transpose of β, which is a vector of coefficients of covariates; $x_{1k} - x_{0k}$ is a data vector/matrix of covariate(s), which is equal to the value of the treated minus that of the control. To apply this function in the context of DIF analysis, a treated unit is regarded as a subject from the focal group whereas a control is regarded as a subject from the reference group; the matrix $\beta^T(x_{1k} - x_{0k})$ can be specified in the following expression, $\beta_1(total_{1k} - total_{0k}) + \beta_2(group_{1k} - group_{0k}) + \beta_3(total_{1k} * group_{1k} - total_{0k} * group_{0k})$. Please note that this conditional likelihood function can be generalized for full matching.

The conditional logistic regression allows one to take account of matched pairs or matched clusters while factoring out the nuisance parameters— the varying intercepts of the matched units of cases and controls. Some studies have shown that the use of a matched study design and conditional logistic regression analysis can increase efficiency of parameter estimates, compared to an unmatched design with regular logistic regression analysis (e.g., Breslow et al., 1978; Hosmer, Lemeshow, & Sturdivant, 2013, pp.227-267; Langholz, & Goldstein, 2001). Pike, Hill and Smith (1980) showed that the unconditional likelihood method, i.e., the regular logistic regression, might give biased estimates of odds ratios which were severely inflated compared to the conditional likelihood

[2]A concordant pair is a pair of participants matched on the propensity scores, one from the focal group and the other from the reference group, whose outcome scores on an item are the same.

method, i.e., conditional logistic regression. Breslow and Day (1980) also illustrated how the unconditional likelihood analysis of matched data could produce biased parameter estimates. The application of conditional logistic regression is illustrated in the step-3 of the demonstration.

A Demonstration of Conditional Logistic Regression DIF Analysis Using Propensity Score Approach

Data Sources

The data were retrieved from the TIMSS 2007. Canada was chosen for this demonstration as it is a bilingual country and students were allowed to choose the test language, either in English or French. Three provinces (British Columbia, Ontario, & Quebec) in Canada participated in TIMSS 2007. Quebec is a French speaking province, and British Columbia and Ontario are English speaking provinces even both English and French are official languages in Canada. Booklet one of the TIMSS 2007 Grade-8 mathematics test was used in this demonstration with 25 dichotomous *items* and 4 polytomous items. A detailed demonstration was provided using items #13 and #5. Propensity score DIF methods will be explained in detail in the data analysis section. These two items were chosen because they demonstrated two scenarios: (i) propensity score methods agreed with the conventional DIF methods on the results, (ii) contradictory conclusions on DIF results showed between propensity score and conventional DIF methods.

A total of 822 students were included in the final analysis; 54% are girls with a mean age of 14 (SD=0.49). More students chose to write the English version of the test (541 English vs. 281 French). *Language* (English vs. French) was used as a grouping variable for the DIF analysis, which is called grouping variable or language variable interchangeably in this demonstration. In order for readers to follow the terms used in the output of *MatchIt R* package (Ho, Imai, King, & Stuart, 2011), we also used the terms, control and treatment groups, in the demonstration. Readers should connect these terms with the terms used in DIF analyses: English test takers were considered as a *reference group*, referred to a *control group* in the output of following analyses, while French test takers were considered as a *focal group*, referred to a *treatment group* in the output. All the test and questionnaires were developed in English and then translated to French. A detailed description of these variables can be found in TIMSS 2007 User Guide (Foy & Olson, 2009). The data can be accessed from TIMSS website http://timssandpirls.bc.edu/timss2007/idb_ug.html .

Data analysis

The R packages for propensity score matching require complete data set, with no missing values. In this demonstration, there was only a single student with missing values on the outcome variables (mathematics items), so that this student's data were discarded from the analyses. The missing values in covariates were imputed using multiple imputations, but for the demonstration purpose, only one imputed data set was used for this study. Detailed information about how to deal with missing data can be found in Rubin (2006) and Little and Rubin (2014). Due to limited space, missing data issue was not addressed in this paper.

Software program R 3.1.3 was used for all analyses. The procedures of conditional logistic regression DIF analyses based on the propensity score approach includes four steps: (i) selecting covariates, (ii) estimating propensity scores and then matching data, (iii) running conditional logistic regression DIF analyses using matched data, and (iv) conducting a sensitivity analysis to examine hidden bias. Optimal pair and full matching methods are demonstrated and reported. Appendix A provides the R-code of the demonstration. The 4-step procedure for Propensity Score Optimal Matching is described below. Items #13 and #5 used in the demonstration were released by TIMSS and are described in Appendix B, so that readers can see what these items are and have a better understanding from the content.

Step-1 Selecting covariates. The decision of which covariates to include in an analysis is mainly based on researchers' experiences, expert opinions, and literature review. It is a crucial step because the selection of covariates has a major impact on how well the propensity scores uncover the unknown mechanism of self-selection into groups. Propensity score approach has an underlying assumption, *strong ignorability of treatment assignment*, that is, treatment assignment and people's responses are conditionally independent after controlling for the effects of a collection of covariates that determine the assignment mechanism (Rosenbaum & Rubin, 1983). Effective covariates are those that are more likely to balance out

the pre-test group differences, thus allowing for the possibility of making legitimate causal claim, as does with a randomized experiment. However, in reality one never knows the true causal effect and is unable to include all possible important covariates, so we can only obtain estimates of causal effect.

There are some controversial issues surrounding the selection of covariates, including the belief that all available covariates should always be used for propensity score estimation, and that over-parameterization is not a problem for propensity score estimation. However, some researchers have shown that the selection of covariates is a critical matter. Zhao (2008) found that over-parameterization can bias the parameter estimate of the grouping variable 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. Based on his findings, Cuong suggested to not include covariates that are only related to the grouping variable. Here, we recommend researchers to be aware of these issues when choosing covariates for propensity score estimation.

In this demonstration, nine originally collected variables or derived indices by TIMSS were chosen from students' background questionnaire and were used as the observed covariates for estimating the propensity scores. These covariates include *number of books at home (nbook), use of calculator (calculator), parents' education (parentEdu), availability of computer(computer), time on mathematics homework (timehw), positive affect to mathematics (affect), valuing mathematics (valuing), self-confidence (slfconf), and perception about school safety (safty).* These variables have been shown to be important factors related to student academic achievement in the literature (e.g., Shen, 2002; Robitaille & Garden, 1988; Wu & Erciken, 2006). For example, Leder and Grootenboer (2005) discussed how students' affect (e.g., values, attitudes) is related to mathematics education. Liu, Wu, & Zumbo (2006) reported that most of these variables listed above were correlated to student mathematics achievement across six countries and the correlations varied across countries using TIMSS data. Similarly, Teodorović (2011) found that student individual variables as well as school factors, such as average parent education, time on tasks, classroom climate, school size, and school climate, had statistically significant effects on students' achievement tests on mathematics.

Step-2 Estimating propensity score and matching. For optimal matching, *MatchIt* R package (Ho, Imai, King, & Stuart, 2011) was used to estimate the propensity scores and to match the data. For optimal matching, the covariate balance was examined using two strategies: (1) graphs of propensity score distributions, and (2) percent bias reduction *PBR* $PBR = \frac{Bias_{pre} - Bias_{post}}{Bias_{pre}} 100\%$ where *Bias* = | *Mean*(X_1)-*Mean*(X_0) |, $Bias_{pre}$ refers to Bias computed before matching, $Bias_{post}$ refers to Bias computed after matching, X_0 denotes covariates before matching and X_1 denotes covariates after matching. To use *PBR*, researchers need to calculate the mean difference between two groups before matching as well as that after matching in terms of each covariate and then compare these two mean differences for each covariate. *PBR* suggests the balance between two groups on a particular covariate is improved if the mean difference between two groups after matching becomes smaller.

Step-3 Running conditional logistic regression DIF analyses. For optimal pair and full matching methods, *conditional logistic regression* models were used for the DIF analyses to take into account matched pairs obtained from pair matching or matched clusters obtained from full matching. It is important to note that the algorithm used in conditional logistic regression DIF for matched case-control studies differs from the regular logistic regression DIF as described in the review of conditional logistic regression. The R package E_{pi} was used for the analyses (Carstensen, Plummer, Laara, & Hills, 2016).

Step-4 Sensitivity analysis. The sensitivity analysis is conducted to check the hidden bias due to unobserved covariates that are related to treatment assignment mechanism. This analysis also indirectly tests the underlying assumption of propensity score approach, strong ignorability of treatment assignment. R package *rbounds* was used for the analyses (Keele, 2014). The purpose of sensitivity analysis is to investigate how inferences about the treatment effects/group differences would be altered by hidden bias, and how large the differences would have to be in order to change the conclusion of the study.

Rosenbaum (1995, 2002) developed sensitivity tests for matched data, which can be expressed by the odds ratio of two subjects assigning to treatment groups

$$\frac{1}{\Gamma} \leq \frac{\pi_j/(1-\pi_j)}{\pi_k/(1-\pi_k)} \leq \Gamma.$$

Rosenbaum's tests rely on the sensitivity parameter Γ, which measures the degree of departure from random assignment of treatment. Two subjects *j* and *k* with the same observed covariate X_i as well as the unobserved covariates U_i should have same probabilities of assignment (π_j, π_k), i.e., $\Pr(Z_j = 1 \mid X_j, U_j) = \Pr(Z_k = 1 \mid X_k, U_k)$. Correspondingly, a Γ value of one indicates that the study is free of hidden bias. However, with the same observed X_i but different unobserved U_j, the two subjects would have different probabilities of assignment and Γ would be a value that deviates from one. In a sensitivity analysis, one inquires how large can Γ be when the obtained conclusion begins to change. In other words, the question researchers want to ask is how much hidden bias there would need to alter our conclusion. A study is highly sensitive to hidden bias if the conclusion changes for Γ just rarely larger than one, and it is relatively insensitive to hidden bias if the conclusion changes for quite large values of Γ.

A sensitivity analysis will consider a range of possible values of Γ, starting from one, and show how the conclusion will be changed when Γ is increased and reaches a certain value. The range of Γ values is usually examined with respect to an interval of the *p*-value. For $\Gamma = 1$, one obtains a single *p*-value, namely the *p*-value for a randomized experiment. For each $\Gamma > 1$, one obtains not a single *p*-value, but rather an interval of *p*-values reflecting the uncertainty due to hidden bias. The particular Γ of interest is the value that turns the upper bound *p*-value from significant to non-significant at alpha = 0.05. The larger the upper Γ value is, the more robust the result is to hidden bias. However, there is no criterion to determine how large the Γ value is required for being considered a good cut-off. Keele (2010) suggested values between 1 and 2 for Γ in social sciences as most findings in social sciences are not robust to hidden bias with a larger magnitude. Detailed information about the algorithm for calculating upper bound *p*-value can be found in Rosenbaum (1995, 2002) and Keele (2010).

In application, Rosenbaum's sensitivity analysis has some limitations. The method works well for dichotomous and continuous variables, but has not been generated to ordinal categorical variables. In addition, the current R packages can handle either pair matching or one-to-many matching, but the number of matched subjects needs to be a constant for all matched sets. In the optimal full matching method, the number of matched subjects can vary from case to case. Sensitivity analysis with currently available R packages still cannot handle the full matching case. Hence, we only demonstrated sensitivity analysis for the dichotomous variables with the optimal pair matching method in this demonstration.

Results

For the purpose of demonstration, the 4-step propensity score DIF analysis was illustrated using two items (items #13 and #5) from the grade 8 mathematics test. A student's total score (the proxy variable for ability) was calculated by adding up all mathematics item scores except the one used as the outcome variable for the DIF analysis.

Step-1 selecting covariates. As we mentioned in the description of this step, one should be cautious about the selection of covariates, which may affect the conclusion of DIF analysis. In this demonstration, we included nine covariates (see *step-1* of data analysis section). These covariates were chosen based on the findings from the literature, which have been shown to be influential factors on students' mathematic academic performance. The detailed description is provided in the step-1 of data analysis section. The purpose of DIF investigation in this study is to examine if the translation of the test language gave rise to DIF, had the two groups been comparable. In other words, we investigated whether the test translation caused differences on student mathematics performance given that students from two equally capable groups were comparable on their background variables. It should be noted that the importance of this step is not only about the validity of causal inference we are making, but also about the social consequences of the inference, which can affect education policy (e.g., dealing with achievement gaps if the translation was shown not an issue) or decisions on the test development (e.g., throwing or rewriting DIF items with a high financial cost).

Step-2 Estimating propensity scores and matching data. In step-2, the propensity scores were estimated and then the data were matched (English vs.

French groups) by either optimal pair matching or optimal full matching. For optimal matching, the propensity score estimation is embedded in the R package *MatchIt* (Ho, Imai, King, & Stuart, 2011). An add-on package *optmatch* (Hansen, 2004; Hansen, Fredrickson, Buckner, Errickson, & Solenberger, 2016) will be automatically loaded when performing optimal matching in *MatchIt*. The R code for conducting optimal matching is provided in Appendix A.

Optimal pair matching. Because this step is very important for the propensity score matching methods, the R code of *MatchIt* is also provided in Figure 2 in addition to Appendix A. Optimal pair matching in this demonstration is performed with *MatchIt* by setting *method="optimal"* and *"ratio=1"* in the R code (Figure 2.a). In addition, *distance = "logit"* indicates that logistic regression is used for this analysis because the outcome variable is dichotomous. Figure 3 and the upper body of Table 1 present the balance check for the optimal pair matching method. Figure 3 shows the distributions of estimated propensity scores before as well as after matching using both histogram and jitter graph. In the histogram, the distributions of two groups were not comparable before matching and many English test takers (denoted by "raw control" in the graph) had lower propensity scores. After matching, there was still some noticeable discrepancy between the distributions of two groups though the distribution of English group ("matched control") became similar to that the French group ("matched treated"), which indicates that the covariate balance was less than satisfactory.

In the jitter graph, each circle represents a case's propensity score. The absence of cases in the uppermost "unmatched treatment units" class (i.e., French group) indicates that there were no unmatched treatment units. The two middle classes, "matched treatment units" and "matched control units", showed a close match between French and English groups. The last class shows the unmatched control units (English group); these units were excluded from the further analyses. Among a total of 541 subjects who took the English version test, only 281 subjects were matched with the French group (treatment group).

The upper part of Table 1 presents the percentage of bias reduction (PBR) for optimal pair matching. It shows that nearly half of the covariates had a large magnitude of bias reduction (above 70% reduction) and a few covariates had a medium level of reduction

```
a. Optimal pair matching
m.Out <- matchit ( language ~ nbook + calculator + parentEdu + computer + timehw + affect +
valuing  + slfconf + safty, data = timss, distance = "logit", method = "optimal", ratio = 1)

b. Optimal full matching (one-to-many)
m.Out <- matchit ( language ~ nbook + calculator + parentEdu + computer + timehw + affect +
valuing  + slfconf + safty, data = timss, distance = "logit", method = "full",  min.controls = 1,
max.controls = 5)

c. Optimal full matching (one-to-many & many-to-one)
m.Out <- matchit ( language ~ nbook + calculator + parentEdu + computer + timehw + affect +
valuing  + slfconf + safty, data = timss, distance = "logit", method = "full",  min.controls = 1/5,
max.controls = 5)
```

Figure 2. R code for step-2: optimal pair matching and optimal full matching

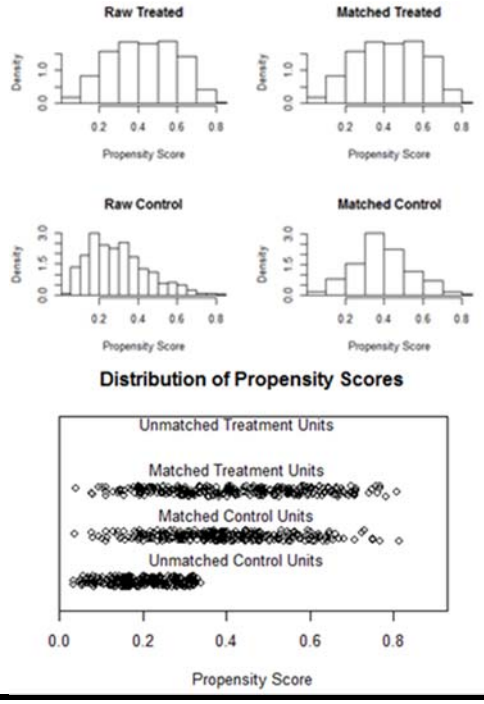

Figure 3. Propensity score distributions before and after the pair optimal matching
Note. "Treated" denotes French test takers; "Control" denotes English test takers

(40%-70% reduction). One covariate, *computer*, had a negative PBR value indicating that the differences between the two groups became even larger after matching. However, the increase was fairly small in magnitude. One covariate, *selfconf*, has a small magnitude of increase in bias, but its PBR cannot be computed because the bias before matching is zero and hence cannot be used as the denominator for computing PBR.

Table 1. Percentage of Bias Reduction (PBR) Using the Optimal Pair and Full matching

Optimal Pair Matching						
		Before Matching		After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	Bias Percentage Reduction (%)
distance	0.430	0.296	0.134	0.390	0.040	70.3
nbook	1.722	2.381	-0.658	1.843	-0.121	81.6
calculator	2.466	2.141	0.326	2.413	0.053	83.6
parentEdu	3.238	3.218	0.020	3.221	0.018	12.4
computer	3.626	3.669	-0.043	3.673	-0.046	-8.1
timehw	0.989	1.198	-0.209	1.100	-0.110	47.1
affect	1.231	1.100	0.132	1.196	0.036	72.9
valuing	1.765	1.784	-0.019	1.776	-0.011	42.6
slfconf	1.392	1.392	0.000	1.377	0.014	-
safty	1.463	1.390	0.073	1.424	0.039	46.1
Optimal Full Matching with One-to-Many						
		Before Matching		After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	Bias Percentage Reduction (%)
distance	0.430	0.296	0.134	0.388	0.043	68.2
nbook	1.722	2.381	-0.658	1.883	-0.161	75.6
calculator	2.466	2.141	0.326	2.395	0.071	78.1
parentEdu	3.238	3.218	0.020	3.226	0.012	40.2
computer	3.626	3.669	-0.043	3.670	-0.043	-0.9
timehw	0.989	1.198	-0.209	1.080	-0.091	56.6
affect	1.231	1.100	0.132	1.178	0.053	59.5
valuing	1.765	1.784	-0.019	1.766	-0.001	97.5
slfconf	1.392	1.392	0.000	1.378	0.014	-
safty	1.463	1.390	0.073	1.418	0.044	39.1
Optimal Full Matching with a Combination of One-to-Many & Many-to-One						
		Before Matching		After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	Bias Percentage Reduction (%)
distance	0.430	0.296	0.134	0.427	0.003	97.6
nbook	1.722	2.381	-0.658	1.729	-0.007	99.0
calculator	2.466	2.141	0.326	2.484	-0.018	94.5
parentEdu	3.238	3.218	0.020	3.288	-0.049	-142.3
computer	3.626	3.669	-0.043	3.617	0.010	77.5
timehw	0.989	1.198	-0.209	1.051	-0.062	70.2
affect	1.231	1.100	0.132	1.199	0.032	75.3
valuing	1.765	1.784	-0.019	1.768	-0.003	86.6
slfconf	1.392	1.392	0.000	1.388	0.003	-
safty	1.463	1.390	0.073	1.478	-0.016	78.4

Note. The full names of nine covariates are as follows: number of books at home (nbook), use of calculator (calculator), parents' education (parentEdu), availability of computer(computer), time on mathematics homework (timehw), positive affect to mathematics (affect), valuing mathematics (valuing), self-confidence (slfconf), and perception about school safety (safty).

Optimal full matching. Optimal full matching can be performed with *MatchIt* by setting *method="full"* in the R code (Figure 2). Researchers can choose one-to-many (one treated unit to multiple controls) or a combination of one-to-many and many-to-one (multiple treated unites to one control unit). An example of the matched sets in a full matching (one-to-many & a combination) can be found in Appendix C. For instance, in Appendix C the matched set #4 included one treated unit and five control unites when using one-to-many full matching, but includes five treated units and one control when using full matching (a combination of one-to-many and many-to-one).

In this demonstration, when setting max.controls=5, we put an upper restriction on the number of controls to include in any matched set. When setting min.controls=1 and max.controls=5, users will get one-to- many matching, which allows matched sets with different ratios, 1:1. 1:2, 1:3, 1:4 or 1:5. The R code is provided in Figure 2.b. When setting *min.controls=1/5* and *max.controls=5*, one can get a combination of one-to-many and many-to-one matched units and put an upper restriction of five on the maximum treated and control units in this example, which allows matching sets with different ratios, 1:1. 1:2, 1:3, 1:4, 1:5, 2:1, 3:1, 4:1 or 5:1. The R code is included in Figure 2.c.

Without defining *max.controls* and *min.controls* in R code, the default is a combination of one-to-many and many-to-one, but there are no upper restrictions. Hansen and Klopfer (2006) recommended to set the upper restrictions because researchers could control the variability of an estimate on the matching and the estimation algorithm would be faster. Researchers should decide on the upper restrictions based on the sample characteristics (e.g., sample sizes) and compare balance results using different ratios of treated units and controls.

Figure 4 and the middle part of Table 1 present the balance check for the one-to-many matching. Figure 5 and the lower part of Table 1 present the balance check for the optimal full matching method with a combination of one-to-many and many-to-one matching. Because all the data points were used for

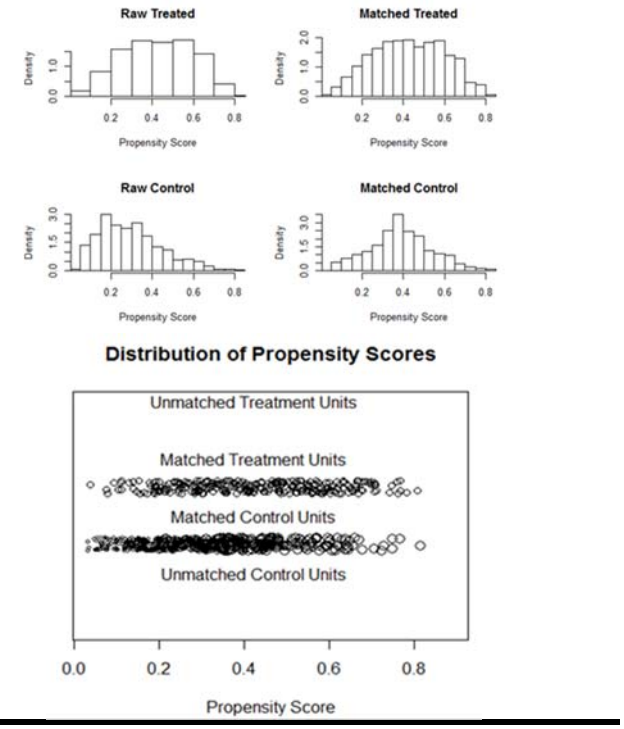

Figure 4. Propensity score distributions before and after the full optimal matching with one-to-many matched cases

Note. "Treated" denotes French test takers; "Control" denotes English test takers

matching in the optimal full matching method, there are no instances of the "unmatched control units" class. The jitter graphs show that all subjects were matched, with the lower propensity scores more piled up among the matched control units (English group).

Although both Figures 4 and 5 show a great deal of improvement in covariate balance between groups after matching, Figure 5 shows a better match when using a combination of one-to-many and many-to-one matching, which is also echoed the percentage of bias reduction (PBR) in the lower part of Table 1. The results of PBR showed that six out of nine covariates had a large magnitude of bias reduction (above 70% reduction), two of them had above 90% reduction (99.0% for *nbook*; 94.5% for calculator), one covariate, *parentEdu*, had a small magnitude of decrease though the PBR value looks large. Similar to pair matching, the PBR could not be computed for *selfconf* as the bias before matching is zero. Hence, optimal full matching with a combination of one-to-many and many-to-one was adopted in our following DIF analyses.

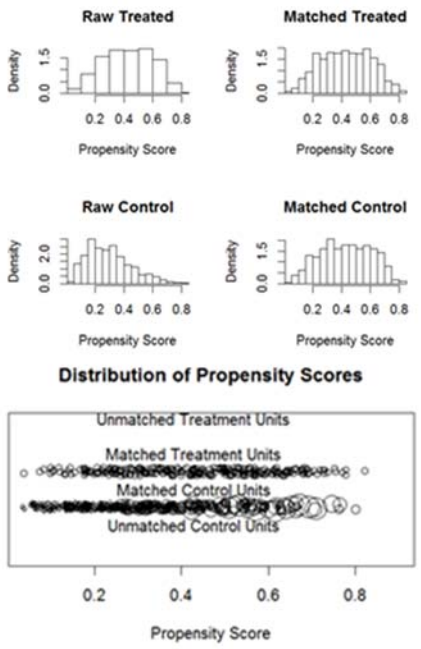

Figure 5. Propensity score distributions before and after the full optimal matching with both one-to-many and many-to-one matched cases

Note. "Treated" denotes French test takers; "Control" denotes English test takers

In summary, the results of balance check suggested that optimal full matching with a combination of one-to-many and many-to-one performed much better than optimal pair matching and reduced more biases on all covariates. In real practice, researchers could choose the optimal full matching in the following data analyses. However, we included both matching methods herein for the purpose of demonstration. Researchers should be aware that the results of pair matching may differ from those of full matching because of its less satisfactory balance. In the next step, DIF analyses were conducted to demonstrate the two scenarios described earlier: consistent and inconsistent results between conventional DIF and propensity score DIF methods.

Step-3 Run conditional logistic regression DIF analyses. For the matched data, the conditional logistic regression method was conducted for the DIF analyses using *Epi* R package (Carstensen, Plummer, Laara, & Hills, 2016). A detailed example of R code and output are provided in Figure 6. In Figure 6.a, the conditional logistic regression is conducted by the code "clogistic()"; "Y5" is the name of the outcome variable; the model is specified by "language * total", which is equivalent to "language + total + language * total" in R code of Figure 6.b; "subclass" is the indicator of matched sets of matched units; "match.data" is the name of the matched data set. The variable "subclass" was generated during the process of matching using *MatchIt* R package and was automatically include in the matched data "match.data". In Figure 6.c, the output of conditional logistic regression is provided: the estimates of regression coefficients are provided in second column, the odds ratios are in the third column, and then followed by the standard error of the estimates, z-scores and *p*-values. In addition, the conventional logistic regression DIF analysis and the logistic regression DIF analysis with covariance adjustment were also conducted in this illustration in order to compare them with the propensity score methods.

a. Simplified R code

```
C5 <- clogistic ( Y5~ language * total, strata = subclass, data = match.data)
```

b. R code with the names of all variables

```
C5 <- clogistic (Y5~ language + total + language * total, strata = subclass, data = match.data)
```

c. Output:

	coef	exp(coef)	se(coef)	z	p
language	0.0796	1.083	0.214	0.372	0.71
total	1.3793	3.972	0.155	8.879	0
language*total	-0.059	0.943	0.267	-0.221	0.82

Figure 6. An example of R code for conditional logistic regression analysis

Note. R code from (a) and (b) are equivalent; Users can choose one of them in practice.

Senario-1: Consistent results among all DIF methods. To provide a visualization tool for DIF analysis, we showed how to plot logistic curves to compare two groups. Figure 7 shows logistic curves generated using the original data in a conventional logistic regression analysis. The left panel of Figure 7 shows that the French group has a higher probability of getting the correct answer on item #13, and that the two logistic curves do not appear to interact within the score range, which may indicate a uniform DIF.

Table 2. Results of DIF Analyses for Item #13 with Raw and Matched Data

	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
Conventional Logistic Regression (Raw Data, n=822)						
language	0.979	2.662	0.175	5.598	< 0.001	***
total	1.056	2.875	0.110	9.620	< 0.001	***
language*total	0.379	1.461	0.225	1.689	0.091	
Conditional Logistic Regression DIF (pair matching, n=306)						
language	0.981	2.67	0.221	4.43	< 0.001	***
total	0.783	2.19	0.206	3.80	< 0.001	***
language*total	0.410	1.51	0.324	1.26	0.210	
Conditional Logistic Regression DIF (full matching, n= 714)						
language	0.687	1.99	0.195	3.53	< 0.001	***
total	0.947	2.58	0.127	7.45	< 0.001	***
language*total	0.321	1.38	0.251	1.28	0.2	

Note. Significance codes: *** = p-value ≤ 0.001; ** = p-value ≤ 0.01; * = p-value < 0.05; n denotes the actual sample size used for the analyses.

Table 2 presents the DIF results for the same item using original data and matched data. The results showed that all the methods agreed with one another. The regression coefficient of "language" was statistically significant across all methods, suggesting the existence of uniform DIF.

Senario-2: Inconsistent results between conventional and propensity score DIF methods.

The right panel of Figure 7 shows that, for item #5, the two logistic curves of a conventional logistic regression are only slightly apart, which suggests that this item is less likely to be DIF. Table 4 presents the DIF results for item #5 using original data and matched data by both pair and full matching methods. Table 5 presents the results using covariance adjustment.

The results obtained from a conventional logistic regression indicate a uniform DIF (*language: beta*=0.359, p=0.039; *language*total*: *beta*=-0.127, p=0.566). The results obtained from the covariance adjustment method indicate no DIF (language: beta=0.339, p=0.078; language*total: beta=-0.112, p=0.618).

Table 3. Results of DIF Analysis for Item #13 Using Logistic Regression with Covariance Adjustment

	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
language	0.984	2.675	0.191	5.160	< 0.001	***
total	0.962	2.617	0.121	7.945	< 0.001	***
language*total	0.415	1.514	0.228	1.821	0.069	
nbook	0.131	1.139	0.074	1.767	0.077	
calculator	0.217	1.243	0.096	2.267	0.023	*
parentEdu	-0.053	0.948	0.093	-0.571	0.568	
computer	0.233	1.262	0.147	1.582	0.114	
timehw	-0.215	0.807	0.130	-1.656	0.098	
affect	0.013	1.013	0.108	0.123	0.902	
valuing	0.223	1.250	0.187	1.191	0.234	
slfconf	0.071	1.073	0.134	0.527	0.598	
safty	-0.010	0.990	0.123	-0.083	0.934	

Note. Significance codes: *** = p-value ≤ 0.001; ** = p-value ≤ 0.01; * = p-value < 0.05 The full names of nine covariates are as follows: number of books at home (nbook), use of calculator (calculator), parents' education (parentEdu), availability of computer(computer), time on mathematics homework (timehw), positive affect to mathematics (affect), valuing mathematics (valuing), self-confidence (slfconf), and perception about school safety (safty).

Table 4. Results of DIF Analyses for Item #5 with Raw Data and Matched Data

	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
Conventional Logistic Regression (Raw Data, n=822)						
language	0.359	1.432	0.174	2.063	0.039	*
total	1.418	4.129	0.129	10.985	< 0.001	***
language*total	-0.127	0.881	0.222	-0.574	0.566	
Conditional Logistic Regression DIF (pair matching, n=268)						
language	0.534	1.707	0.266	2.01	0.045	*
total	1.865	6.455	0.327	5.70	< 0.001	***
language*total	-0.405	0.667	0.372	-1.09	0.28	
Conditional Logistic Regression DIF (full matching, n=691)						
language	0.080	1.083	0.214	0.372	0.71	
total	1.379	3.972	0.155	8.879	< 0.001	***
language*total	-0.059	0.943	0.267	-0.221	0.82	

Note. Significance codes: *** = p-value ≤ 0.001; ** = p-value ≤ 0.01; * = p-value < 0.05; n denotes the actual sample size used for the analyses.

Table 5. Results of DIF Analysis for Item #5 Using Logistic Regression with Covariance Adjustment

	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
language	0.339	1.403	0.192	1.765	0.078	
total	1.351	3.863	0.141	9.609	<0.001	***
language*total	-0.112	0.894	0.225	-0.498	0.618	
nbook	-0.040	0.961	0.077	-0.526	0.599	
calculator	0.111	1.118	0.100	1.116	0.265	
parentEdu	0.062	1.064	0.098	0.638	0.524	
computer	0.090	1.095	0.150	0.604	0.546	
timehw	0.085	1.089	0.134	0.636	0.525	
affect	-0.134	0.875	0.113	-1.187	0.235	
valuing	-0.331	0.718	0.187	-1.773	0.076	
slfconf	0.398	1.489	0.143	2.793	0.005	**
safty	-0.024	0.976	0.128	-0.192	0.848	

Note. Significance codes: *** = p-value ≤ 0.001; ** = p-value ≤ 0.01; * = p-value < 0.05
The full names of nine covariates are as follows: number of books at home (nbook), use of calculator (calculator), parents' education (parentEdu), availability of computer(computer), time on mathematics homework (timehw), positive affect to mathematics (affect), valuing mathematics (valuing), self-confidence (slfconf), and perception about school safety (safty).

The overall results for item #5 obtained from conditional logistic regression suggest no DIF (Table 5) though using pair matching the *language* variable showed statistically significant result (*beta*=0.534, *p*=0.045). We treated it as a no DIF case because the *p*-value was on the borderline and the sensitivity analysis in the next step also showed this marginal significance result would be easily changed, had some unobserved covariates included. Using full matching, the results of conditional logistic regression in Table 4 showed no DIF (*language: beta*=0.08, *p*=0.71; *languag*total: beta*=-0.059, p=0.82).

Step-4 Conducting sensitivity analysis to examine hidden bias. As we indicated earlier, the sensitivity analysis for full matching is still not available in the existing R package. Only the sensitivity analysis (Rosenbaum, 2002) was conducted to check hidden bias using *rbounds* R package (Keele, 2010, 2015) for pair matching. The analysis for binary outcome is based on McNemar test. The R code used for sensitivity analysis is provided as follows, *binarysens*(*X*, *Gamma* = 3, *GammaInc* = 0.2), where X contains outcome (Y) and grouping variables (Tr) for the matched pairs, the upper limit of *Gamma* is three and the increment of Gamma is 0.2. In this demonstration, the outcome *Y*

Figure 7. Fitted Logistic Regression Curves in Conventional Logistic Regression DIF Analyses for Items #13 and #5, respectively, for French vs. English Groups of Test-takers

denotes student responses (correct=1, incorrect=0), and the grouping variable *Tr* denotes language groups (English vs. French). Only group difference (English vs. French) was examined with respect to the outcome variable. Normally, one starts from a significant upper bound *p*-value and continue until the Γ value turns the upper bound *p*-value from significant to non-significant. The larger Γ value indicates the group difference (treatment effect in clinical trials) is more resistant to hidden bias. This is based on the assumption that the treatment effect is statistically significant to begin with.

The upper body of Table 6 presents the results for item#13 with Γ values from 1 to 3 in 0.2 unit increments. Because all the results suggest the presence of uniform DIF (i.e., a significant group difference), the sensitivity analysis starts with a significant *p*-value.

Referring to Table 6, the group difference becomes non-significant between Γ=2.0 and Γ=2.2 (two-tailed α=0.05 level). To attribute DIF to unobserved covariates rather than language group difference (i.e., translation effect), the unobserved

Table 6. Results for Sensitivity Analysis with an Increment of 0.2 in Gamma for Item #13 and with an Increment of 0.1 in Gamma for Item #5

Items #13		
Gamma	Lower bound	Upper bound
1.0	0.000	0.000
1.2	0.000	0.000
1.4	0.000	0.000
1.6	0.000	0.003
1.8	0.000	0.014
2.0	0.000	0.045
2.2	0.000	0.106
2.4	0.000	0.201
2.6	0.000	0.320
2.8	0.000	0.449
3.0	0.000	0.574
Items #5		
Gamma	Lower bound	Upper bound
1	0.039	0.039
1.1	0.013	0.098
1.2	0.004	0.191
1.3	0.001	0.312
1.4	0.000	0.447
1.5	0.000	0.578
1.6	0.000	0.693
1.7	0.000	0.785
1.8	0.000	0.855
1.9	0.000	0.906
2	0.000	0.940

Note: Gamma is odds of differential assignment to treatment due to unobserved factors

covariates would need to produce more than 2-fold increase in the odds of language group membership. In other words, a change of around 1.2 on the odds of treatment assignment will change the DIF results from significant to non-significant. This indicates that the conclusion of DIF for item #13 would be relatively hard to be altered by accounting for some presently unobserved covariates.

The lower body of Table 6 presents the results for item #5 with Γ values from 1 to 2 in an increment of 0.1. The group difference becomes non-significant between Γ=1.0 and Γ=1.1 (two-tailed α=0.05 level). A change of less than 0.1 on the odds of treatment assignment will change the DIF results from significant to non-significant. This indicates that DIF for item #5 could be quite easily altered by accounting for some

unobserved covariates. This echoed the conclusion about no DIF obtained from propensity score approaches.

General Discussion

The identification of DIF items is important in the fields of assessment, testing, and psychometrics when developing a new test, adapting a test to another culture or language, comparing students' academic performance across regions or countries, or comparing paper-and-pencil to computerized tests. However, conventional DIF methods can only tell whether DIF exists or not, but cannot rule out other confounding sources of DIF (e.g., students' motivation, parents' income, or other school factors) from our primary focus (e.g., translation or test administration mode). Hence, it is difficult for researchers or test developers to decide whether or not to retain the DIF item or throw it away.

The present paper extended the previous logistic regression DIF method and demonstrated the application of propensity score methods in DIF analysis. In educational tests, for example, there are many factors related to students' academic performance, which can be potential sources of DIF in addition to translation. Using propensity score matching techniques, we can match students on a variety of confounding variables. While these matches may not be exhaustive, we were at least able to control a great deal of confounding sources of DIF and focus on the DIF effect of our interest.

Propensity score methods were used for making two groups more comparable in terms of a variety of confounding variables before the DIF analysis. Propensity score methods (optimal pair and full matching) were demonstrated step by step and the R code for each method was provided in the Appendix A. The demonstration was conducted to investigate whether the translation of an English test to a French test resulted in DIF. These results were compared to those produced by the conventional logistic regression DIF analysis as well as the logistic regression DIF analysis with covariance adjustment.

Two items are chosen to demonstrate two scenarios: (i) consistent results among all DIF analysis methods, and (ii) inconsistent results between the conventional and propensity score DIF analysis methods. The results obtained from propensity score approaches allowed us to approximate the causal effect of DIF given that two groups were more comparable after matching. However, propensity score approach may not work well in some situations. For example, in this demonstration the pair matching did not achieve a good balance of covariates between two groups and, hence, resulted in a different conclusion on DIF from that of full matching for item #5. Remember that the pair matching showed much less satisfactory balance than did the full matching, so the uniform DIF result obtained from the pair matching might be due to a relatively less balance in covariates. We demonstrate this complexity of DIF results with a purpose to remind researchers of being aware of two issues in practice: (a) researchers may reach different conclusions by using different propensity score matching methods, and (b) unsatisfactory balance of matching may result in questionable results. Hence, including important covariates and achieving a good balance of covariates between two groups are essential to estimate causal effects.

In addition, there is an important issue that has not been fully discussed in the literature for the use of propensity score matching on observational study. That is, what kind of grouping variables should be used for estimating causal DIF? In our demonstration, it makes sense for us to match groups on covariates and make groups comparable before examine DIF because our primary interest is whether the test translation leads in DIF when two groups are comparable on all other factors. However, it may not make much sense for researchers to match on covariates to investigate, for instance, gender DIF or ethnicity group DIF. The purpose of matching on covariates is to eliminate pre-test group differences, to purify the sources of DIF, and make a causal claim about DIF. However, a grouping variable, such as gender or ethnicity, is a characteristic of groups that cannot be manipulated. In addition, gender and ethnicity are proxy variables, which encompass a large number of characteristics of individuals as well as social and/or cultural factors. Researchers probably expect to see some gender differences on a particular test and may want to know what factors result in gender differences on the test instead of matching on these factors that are part of the characteristics of gender group. Therefore, researchers should be cautious about the constitution of the grouping variable of interest when using propensity score matching to make causal claims.

There remain issues regarding propensity score approach to DIF. First of all, the present demonstration did not consider the multilevel structure inherent in the data collection when estimating the propensity scores. Most data collected for international assessments are multilevel, students are nested in schools or are nested in their neighborhood. However, the application of propensity score methods in multilevel models is more complicated and the existent statistical programs still cannot handle multilevel data for propensity score DIF analysis. We could have written our own program for conducting multilevel matching methods, but the data set used for the demonstration had a special issue, which made this moot: the assignment of the language version of tests was done at the school level for most schools; hence, the school indicator (cluster id) used in multilevel modeling would be a perfect predictor for the propensity score estimation. Therefore, we did not consider multilevel models in this context. We note, however, that these models should be considered if the assignment had been done at individual level.

Second, the algorithm of conditional logistic regression for the polytomous outcome variables has not been developed yet, so conditional logistic regression can only be applied to dichotomous outcome variables. The conditional logistic regression analysis can provide results with more precision because it can take account of the dependence structure of pairs or matched sets when using the pair or full optimal matching methods, which is analogous to multilevel modeling. Further research on workable algorithms for the implementation of conditional logistic regression for polytomous variables is encouraged.

Third, sensitivity analysis program is still not available for polytomous outcome variables. Rosenbaum's sensitivity analysis methods work well for dichotomous or continuous variables only. In addition, the existent sensitivity analysis programs still cannot handle the situation when the ratio of the number of subjects between two groups varies across matched sets as in the full optimal matching.

Despite it is still in the stage of development, the propensity score DIF approach can provide researchers and test developer with a more precise tool for examining causal DIF. In particular, it can aid in making a more accurate decision about retaining or removing possible biased items.

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.

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.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*, 1057-1069.

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., Halvorsen, K.T., Prentice, R. L., & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. American *Journal of Epidemiology, 108*, 299-307.

Breslow, N.E., & Day, N.E. (1980). *Statistical Methods in Cancer Research: Volume I—The Design and Analysis of Case-Control Studies. Vol.32 of IARC Scientific Publications.* Lyons: International Agency for Research on Cancer.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test (Research Bulletin 64-61).* Princeton, NJ: Educational Testing Service.

Carstensen, B., Plummer, M., Laara, E., Hills, M. (2016). *Epi: A Package for Statistical Analysis in Epidemiology.* R package version 2.0. R package documentation retrieved from: http://CRAN.R-project.org/package=Epi

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

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A, 128(2)*, 234-266.

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

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*, 2265-2281.

Dorans, N. J., & Holland, P.W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item*

Functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Flanders, W.D. (1986). Conditional logistic analyses of matched case-control studies. *American Journal of Epidemiology, 123(4),* 756-757.

Foy, P., & Olson, J.F. (Eds.), (2009). *TIMSS 2007 International Database and User Guide (Trends in International Mathematics and Science Study).* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Geyer, C.J. (2003). *Generalized linear models in R.* R package documentation retrieved from: http://www.stat.umn.edu/geyer/5931/mle/glm.pdf

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2, 405-420.

Guo, S., & Fraser, W.M. (2014). *Propensity Score Analysis: Statistical Methods and Applications, 2nd Edition*. Thousand Oaks, CA: Sage Publications, Inc.

Hansen, B.B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*, 609-618.

Hansen, B.B., & Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*, 609–627.

Hansen, B.B., Fredrickson, M., Buckner, J., Errickson, J., & Solenberger, P. (2016). *Package: 'optmatch'*. R package version 0.9-6. R package documentation retrieved from: https://cran.r-project.org/web/packages/optmatch/

Hastie, T.J. & Pregibon, D. (1992). *Generalized linear model.* In J. M. Chapman & T. J. Hastie (Eds.), Statistical Models in S. Wadsworth & Brooks/Cole, Pacific Grove, California.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). *MatchIt: Nonparametric preprocessing for parametric causal inference*. R package documentation retrieved from: http://r.iq.harvard.edu/docs/matchit/2.4-20/matchit.pdf

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

Hosmer, D.W., Lemeshow, Jr. S., & Sturdivant, R. X., (2013). *Applied Logistic Regression. 3rd Ed.* New Jersey: John Wiley & Sons, Inc.

Johansone, I. & Malak, B. (2008). Translation and national adaptations of the TIMSS 2007 assessment and questionnaires. In J.F. Olson, M.O. Martin, & I.V.S. Mullis, (Eds.), *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Keele, L.J. (2010). *An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data.* R document is retrieved from http://www.personal.psu.edu/ljk20/rbounds%20vignette.pdf

Keele, L. J. (2014). *rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data. R package version 2.1.* R package documentation retrieved from: https://CRAN.R-project.org/package=rbounds

Langholz, B., & Goldstein, L. (1997). Fitting logistic models using conditional logistic regression when there are large strata. *Computing science and Statistics, 29*, 551-555.

Langholz, B., & Goldstein, L. (2001). Conditional logistic analysis of case-control studies with complex sampling. *Biostatistics, 2*, 63-84.

Le, C. T. & Lindgren, B. L. (1988). Computational implementation of the conditional logistic regression model in the analysis of epidemiologic matched studies. *Computers and Biomedical Research, 21(1)*, 48-52.

Leder, G. & Grootenboer, P. (2005). Affect and mathematics education. *Mathematics Education Research Journal, 17(2)*, 1-8.

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*, 313-338.

Lienhardt, C., Fielding, K., Sillah, J. S., Bah, B., Gustafson, P., et al. (2005). Investigation of the risk factors for tuberculosis: a case–control study in three countries in West Africa. *International Journal of Epidemiology, 34(4)*, 914-923.

Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data (2nd ed.)*. New York, NY: Wiley Press.

Liu, Y., Wu, A. D., & Zumbo, B. D. (2006). The Relation between Outside-of-School Factors and Mathematics Achievement: A Cross-country Study among the U.S. and Five Top-performing Asian Countries. *Journal of Educational Research and Policy Studies, 6*, 1-35.

Marcus, S.M. (2000). Estimating the long-term effects of Head Start. In S. Oden, L.J. Schweinhat, D.P. Weikart, S.M. Marcus, Y. Xie, (Eds.), *Intro Adulthood: A Study of the Effects of Head Start.* Ypsilanti, MI: High/Scope Press.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7(2)*, 105-118.

Pan, W., & Bai, H. (2015). *Propensity Score Analysis: Fundamentals and Developments*. New York, NY: The Guilford Press.

Pike, M.C., Hill, A.P., & Simith, P.G. (1980). Bias and Efficiency in logistic analyses of stratified case-control studies. *International Journal of Epidemiology, 9*, 89-95

Robitaille, P. & Garden, R. A. (1988). *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics*. Oxford- New York- Toronto: Pergamon Press.

Rosenbaum, P.R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*, 1024-1032.

Rosenbaum P.R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, 53*, 597-610.

Rosenbaum, P.R. (1995). *Observational studies (1st ed.)*. New York, NY: Springer.

Rosenbaum, P.R. (2002). *Observational studies (2nd ed.)*. New York, NY: Springer.

Rosenbaum, P.R. (2010). *Design of Observational Studies*. New York, NY: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

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

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.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge, UK: Cambridge University Press.

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

Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: a cross-national analysis based on TIMSS 1999 data. *Assessment in Education, 9(2)*, 161-184.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias* (pp. 9-30). Baltimore: John Hopkins University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Teodorović, J. (2011). Classroom and school factors related to student achievement: what works for students? *School Effectiveness and School Improvement, 22(2)*, 215-236.

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.

Wu, A.D. & Erciken, K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. *International Journal of Testing, 6(3)*, 287-300.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications, *Economics Letters, 98*, 309–319.

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. URL: http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B.D. (2008). *Statistical Methods for Investigating Item Bias in Self-Report Measures, [The University of Florence Lectures on Differential Item Functioning]*. Universita degli Studi di Firenze, Florence, Italy. URL: http://faculty.educ.ubc.ca/zumbo/papers/Zumbo_Univ_Firenze.pdf

Appendix A. R-code for Estimating Propensity Score and DIF Analysis Using Optimal Propensity Score Matching

```
# Upload R packages.
install.packages(c("MatchIt","Epi", "rbounds","ggplot2"))
install.packages("optmatch") #If having trouble to open "MatchIt", download this package

library(MatchIt) # used for optimal matching
library(Epi)   # used for conditional logistic regression analysis
library(rbounds) # used for sensitivity analysis
library(ggplot2) # used for plotting logistic curves
library(optmatch) # used only when having trouble to open "MatchIt"

#~~~~~~~~~~~~~~~~#
##  Read data into R ##
#~~~~~~~~~~~~~~~~#
setwd("C:/Dropbox")  # set up your own working directory
timss<-foreign::read.spss("timss.sav", to.data.frame=TRUE)   # read SPSS data into R; name the data "timss"
table(timss$ITLANG)   # ITLANG = language (English=0; French=1)
                      # We followed the names of variables used in the original TIMSS data.
## We only provided R code for item #5 here.
##  item #13 has the same procedure, so one only needs to change Y5 and Ztot5 to Y13 and Ztot13.

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
##    Conventional Logistic Regression DIF Analysis   ##
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
## In the following DIF analysis, raw total scores were transformed to z-scores before analyses.
## Ztot5 denotes transformed total scores of item #5.
ft5raw<-glm(Y5~ITLANG*Ztot5, data=timss, family=binomial)
          # glm() is a R package for generalized linear modeling
          # ITLANG*Ztot5 is equivalent to ITLANG+Ztot5+ITLANG*Ztot5.
          # "family=binomial" indicates that the outcome variable is binary and logistic regression is used here.
summary(ft5raw)  # The code provides output.

## Covariance adjustment Logistic regression DIF analysis
ft5cv<-glm(Y5~ITLANG*Ztot5+BS4GBOOK+BS4MHCAL+BSDGEDUP+BSDGCAVL
     +BSDMTMH+BSDMPATM+BSDMSVM +BSDMSCM+BSDGPBSS, data=timss, family=binomial)
summary(ft5cv)

#~~~~~~~~~~~~~~~~~~~~~~~~#
##       plot logistic curves       ##
#~~~~~~~~~~~~~~~~~~~~~~~~#
# giving label to "language"
timss$Language <- factor(timss$ITLANG, labels = c("English", "French"))
          # Change the numeric value (0,1) to labels ("English", "French") for grouping variable
          # and rename it from "ITLANG" to "Language".

# saving predicted probability values from the conventional DIF analysis
fit5<-fitted(ft5raw)

# plot item #5
# note that there should be an underscore between goem and line; between scale and linetype; linetype and manual
# note that underscore between scale and y and continuous
ggplot(timss, aes(x=Ztot5, y=fit5, colour=Language, linetype = Language)) +
        geom_line(size = 1.2)  +                              # "size" is to decide on the thickness of the line
        ylab(expression("Pr (" * Y[5] == 1 * ")"))   +    # This will give the label on y-axis.
        xlab("Standardized Total Scores (Item #5)") +          # This will give the label on x-axis.
        scale_linetype_manual(values = c(French = "solid", English = "dashed")) +
```

```
                                                  # making the solid and dashed lines
        theme(legend.justification = c(-1.1,2.2), legend.position = c(0.3, 0.7)) +
                                                  # fixing the legend position in the graph
        scale_y_continuous(limits = c(0, 1.0))    # setting up the scale of 0-1 for y-axis


##############################################
# #      Propensity Score Optimal Matching      ##
##############################################

#  Step-2: Match Data and check balance #
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
## Run MatchIt R package
## optimal pair matching (one-to-one)
m.out<-matchit(ITLANG~ BS4GBOOK+BS4MHCAL+BSDGEDUP+BSDGCAVL+BSDMTMH+BSDMPATM
        +BSDMSVM+BSDMSCM+BSDGPBSS, data=timss, method="optimal", distance="logit", ratio=1)

## optimal full matching (one-to-multiple)
# m.out<-matchit(ITLANG~BS4GBOOK+BS4MHCAL+BSDGEDUP+BSDGCAVL+BSDMTMH
#              +BSDMPATM+BSDMSVM+BSDMSCM+BSDGPBSS, data=timss,
#                  distance = "logit", method="full", min.controls=1, max.controls=5)

## optimal full matching (a combination of one-to-multiple and multiple-to-one)
# m.out<-matchit(ITLANG~BS4GBOOK+BS4MHCAL+BSDGEDUP+BSDGCAVL+BSDMTMH
#              +BSDMPATM+BSDMSVM+BSDMSCM+BSDGPBSS, data=timss,
#                  distance = "logit", method="full", min.controls=1/5, max.controls=5)

summary(m.out)                # the output will provide the percentage of bias reduction (PBR)
match.data<-match.data(m.out)  # save the matched data to a file named "match.data"

# graphic check on the distribution balance for propensity score matching:
plot(m.out,type="jitter")
plot(m.out, type="hist")

## check the matched sets
match.data$subclass <- as.factor (match.data$subclass)
table(match.data$ITLANG, match.data$subclass)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
#   Step-3: Run conditional logistic regression DIF analysis for matched data  #
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
# Using Epi R pacakge
c5<- clogistic(Y5~ ITLANG * Ztot5, strata = subclass, data = match.data); c5
          # "subclass" is an indicator/variable of matched sets.
          # "subclass" is generated by MatchIt R package and automatically included in match.data.


#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
#    Step-4 : Sensitivity Analysis      #
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#

match.data$Tr<-match.data$ITLANG    # change the grouping variable name ITLANG to Tr in order to
                                    # fit to this package
match.data$Y<-match.data$Y5     # Similarly, change the outcome name Y5 to Y

X<-list(mdata = match.data, x=1)
    # x=1 is an arbitrary code to make sure X with enough elements to meet the requirement of rbound package.

binarysens(X, Gamma = 2, GammaInc = 0.1)
    # set up the upper limit for Gamma = 2 and the increment value = 0.1
    # Researchers can change these values.
```

Appendix B. Mathematics Items Used in the Demonstration

(IEA, 2007, TIMSS User Guide for the International Database: Released Items, Mathematics – Eighth Grade)

Item #5.

What is the perimeter of a square whose area is 100 square meters? Answer: _____.

(Original item id in TIMSS: M022055; Content domain: geometry; Cognitive domain: Applying)

Item#13

The figure shows a shaded triangle inside a square.

What is the area of the shaded triangle? Answer: _____.

(Original item id in TIMSS: M022243; Content domain: geometry; Cognitive domain: Apply)

Appendix C. An Example of the Number of Matched Units in Each Matched Set Using Full Matching (one-to-many & a combination of one-to-many and many-to-one)

	One-to-many		Combination	
Matched set	Treatment	Control	Treatment	Control
1	1	1	1	5
2	1	1	1	5
3	1	1	1	3
4	1	5	5	1
5	1	1	1	5
6	1	2	2	1
7	1	5	1	5
8	1	5	1	5
9	1	1	1	4
10	1	1	1	2
⋮	⋮	⋮	⋮	⋮

Acknowledgement:

We would like to thank Bendix Carstensen, senior statistician at Steno Diabetes Center and one of authors of R package *Epi* for conditional logistic regression analysis. We also want to thank Dr. Luke, J. Keele for consulting the application of sensitivity analysis using his *rbounds* R package. Bruno Zumbo's work was supported by the UBC-Paragon Research Initiative.

We would like to thank the anonymous reviewers for their very detailed, informative, and constructive feedback that improved the paper.

Citation:

Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., Wu, A. (2016). Investigating Causal DIF via Propensity Score Methods. *Practical Assessment, Research & Evaluation*, 21(13). Available online: http://pareonline.net/getvn.asp?v=21&n=13.

Corresponding Author:

Bruno D. Zumbo
Paragon UBC Professor of Psychometrics & Measurement
Measurement, Evaluation & Research Methodology Program
The University of British Columbia

email: bruno.zumbo [at] ubc.ca