International Journal of

Research in Education and Science

IJRES

ISSN: 2148-9955

International Journal of Research in Education and Science (IJRES)

www.ijres.net

Propensity Score Matching Helps to Understand Sources of DIF and Mathematics Performance Differences of Indonesian, Turkish, Australian, and Dutch Students in PISA

Serkan Arıkan[1], Fons J. R. van de Vijver [2 3 4], Kutlay Yagmur[2]

[1] Mugla Sitki Kocman University
[2] Tilburg University
[3] North-West University
[4] University of Queensland

To cite this article:

Arikan, S., Van de Vijver, F.J. R. & Yagmur, K. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education and Science (IJRES), 4*(1), 69-81. DOI:10.21890/ijres.382936

IJRES International Journal of Research in Education and Science

Volume 4, Issue 1, Winter 2018 DOI:10.21890/ijres.382936

Propensity Score Matching Helps to Understand Sources of DIF and Mathematics Performance Differences of Indonesian, Turkish, Australian, and Dutch Students in PISA

Serkan Arikan, Fons J. R. van de Vijver , Kutlay Yagmur

Article Info

Article History

Received:
06 May 2017

Accepted:
07 July 2017

Keywords

Propensity score matching
DIF
PISA
Mathematics performance

Abstract

We examined Differential Item Functioning (DIF) and the size of cross-cultural performance differences in the Programme for International Student Assessment (PISA) 2012 mathematics data before and after application of propensity score matching. The mathematics performance of Indonesian, Turkish, Australian, and Dutch students on released items was compared. Matching variables were gender, an index of economic, social and cultural status, and opportunity to learn, in exact, nearest neighbor, and optimal matching. Logistic regression and structural equation modeling were used to identify DIF. If propensity scores were used in the DIF analyses as performance predictors, much less DIF was found than in the original data; similarly, when in tests of country differences in mathematics performance, propensity scores were used as covariates, effect sizes of tests of country differences were reduced substantially. We concluded that propensity scoring provided us with a new tool to better control sources of DIF and country differences in PISA mathematics performance.

Introduction

In this study we aim to evaluate effects of various propensity score matching methods on DIF results and on the size of cross-cultural achievement differences in educational tests. The background of the study is that the nature of DIF has turned out to be elusive in many domains, including educational testing (cf. Holland & Wainer, 1993; Van de Vijver & Leung, 1997). Almost half a century of DIF studies (started with the seminal work of Cleary & Hilton, 1968) has not produced a commonly agreed set of recommendations about how to write items with little or no bias. One of the problems is the poor replicability of findings of item bias studies. We do not seem to have much control over sources of item bias. New approaches to design and analysis are needed to advance the bias field. Propensity score matching has the potential to be such a procedure to shed light on item bias in cross-cultural research. Propensity score matching can be used to produce comparable sample groups by equating groups on relevant background variables. Bias detection procedures and propensity matching procedures share an important characteristic in that they look for matches in different ethnic groups/countries on the basis of some background or psychological characteristic, such as socioeconomic status or total test score. The main difference, however, is that unlike bias detection procedures, propensity matching allows for multiple background variables to be factored in at the same time and that the matching variables do not need to be derived from the target instrument that is scrutinized for bias, such as an educational achievement test, which is typically the case in DIF studies. As a consequence, propensity scoring may provide us with a better tool to control sources of item bias. We examined the impact of propensity matching by comparing DIF and the size of cross-cultural differences before and after matching on student background variables, using PISA 2012 mathematics data.

Differential Item Functioning

DIF occurs and threatens the comparability when students with the same ability level on the underlying construct but coming from different groups (e.g., females and males) show dissimilar mean scores on an item (Van de Vijver & Leung, 1997; Zumbo, 2007). DIF analyses are used for many purposes such as fairness and equity in testing, dealing with a possible threat to internal validity, investigating the comparability of translated and/or adapted measures, trying to understand item response processes, and investigating (lack of) invariance (Zumbo, 2007). Statistical methods are used to detect items showing DIF and these items are removed from the instrument to achieve item bias-free and valid score comparisons. Otherwise, any differences in observed scores could be related to problems based on items rather than true differences in the underlying trait or ability (He & Van de Vijver, 2013). After detecting DIF items statistically, evaluating sources of DIF by expert opinion is often the next step in the procedure. There might be many sources of DIF; examples are poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, influence of culture specific issues such as nuisance factors or connotations associated with the item wording (Van de Vijver & Tanzer, 2004), as well as contextual variables such as classroom size, socioeconomic status, teaching practices or parental socialization styles (Zumbo & Gelin, 2005). By evaluating items, it is necessary to distinguish DIF from item impact and provide explanations for why DIF has occurred in a specific item (Zumbo, 2007). However, by judgmental evaluation alone, it might not be always easy to detect the sources of DIF. For example, Angoff (1993) reported that test developers often had problems to understand why some perfectly reasonable items showed large amounts of DIF. In such cases, it would be helpful to check other relevant factors such as background characteristics of students. Therefore, there is a need to control background variables that might be sources of DIF to make a more informed judgment to reduce the number of DIF items. Matching students of different groups using propensity scores could help to achieve this goal.

Student background variables are generally considered as potential explanations of group differences and sources of DIF. For example, different studies showed that boys are more successful than girls in PISA mathematics test (Areepattamannil, 2014; Liu & Wilson, 2009; Machin & Pekkarinen, 2008). In PISA 2012, boys performed better than girls by 11 points; out of 65 countries, in 38 countries boys performed better than girls whereas in 5 countries girls performed better than boys (OECD, 2014a). Similarly, many studies investigating associations among socio-economic status (SES) and mathematics performance in PISA reported that socio-economically advantaged students perform better (Kilic, Cene, & Demir, 2012; Perry & McConney, 2010). In general terms, SES describes an individual's or a family's ranking on a hierarchy according to access to or control over some combination of valued commodities, such as wealth, power, and social status (Sirin, 2005). SES is also an important variable in understanding the performance differences of students in PISA testing. PISA measure students' socio-economic backgrounds using a continuous scale – the index of economic, social and cultural status (ESCS). PISA data show that there is a significant relationship between students' performance and their socioeconomic background as measured by ESCS although the strength of the relationship differs across countries (Thomson et al., 2013). In the PISA 2012 mathematics test, socio-economically advantaged students scored on average 78 points (effect size of .78) higher than disadvantaged students (OECD, 2013). Opportunity to learn is another variable with a positive relationship to mathematics performance. Opportunity to learn indicators such as exposure to word problems, mathematics topics and applied mathematics problems showed a significant relationship to mathematics performance in PISA 2012 (Schmidt, Zoido, & Cogan, 2014). Among OECD countries, there was a 40 points (effect size of .40) difference on average between students who stated they never encountered applied mathematics problems and students who stated they rarely encountered such problems (OECD, 2014a). Therefore, as these student backgrounds are effective in predicting mathematics performance in PISA, a careful control of these variables by means of propensity score matching could be effective to understand the nature of DIF found in the comparison of students from different countries in the PISA study and to statistically explain country differences in scores.

Propensity Score Matching

When researchers employed randomized experimental designs, the comparison groups are formed to be only randomly different on all background covariates. However, in studies comparing intact groups or nations, randomization is impossible. Matching methods using propensity scores could then be used to compose comparable samples by equating the distribution of covariates in the comparison groups (Stuart, 2010). If the pre-existing achievement differences between countries would disappear after matching, it can be concluded that the country differences in achievement can be attributed to the background differences. If the pre-existing achievement differences would remain intact after matching, the conclusion can be drawn that the differences cannot be reduced to the background differences observed. When used this way, propensity matching can be seen as an advanced kind of covariance analysis (Van de Vijver & Poortinga, 1997).

There are many types of propensity score matching methods. The matching methods that were used in this study are exact, nearest neighbor, and optimal matching. Exact matching is the simplest version of the matching. In this procedure, an individual who has exactly the same values on all covariates is matched with an individual in the comparison group (Ho, Imai, King, & Stuart, 2011). A problem with exact matching is that when matching is done on several background variables, it is probable to end up with a very small sample that is matched on these covariates, but is very dissimilar on other aspects, which can create an even larger bias (Rosenbaum &

Rubin, 1985). As a solution another approach in matching was introduced that constructs matched sets by ensuring similar distributions of the covariates, thereby loosening the need to have exact matches on all the individual variables. Nearest neighbor matching is one of the common matching procedures which selects one individual from a comparison group with the closest matching covariate properties among all available individuals in the group. The unmatched individuals are discarded. As a matched individual is no longer available, the order of matching could change the quality of matches (Stuart, 2010). Optimal matching takes into account global distance instead of individual distance when performing the matching (Rosenbaum, 2002). Therefore, nearest neighbor matching could be used when the aim is to create well-matched pairs, whereas optimal matching could give better results when the aim is to create well-matched groups (Gu & Rosenbaum, 1993). Besides these matching methods, there are other propensity score methods available, such as weighting, full matching, and subclassification. These methods give a weight between 0 and 1 to each individual based on covariates and no individuals are discarded. As the main focus of this study is to form new matched groups and compare results in terms of DIF, weighting, full matching and subclassification methods are beyond the scope of this paper.

After matching is done and new groups are formed, it is necessary to evaluate the quality of the matching by examining the closeness of the covariate distribution of the resulting matched samples, known as balance. A poorly balanced matching means that groups differ considerably in their distributions of matching variables. In order to evaluate balance of the matched groups, Rubin (2001) recommended that propensity score mean differences should be less than half a standard deviation and propensity score variance ratios should be close to one. A matching result that produced imbalanced samples should be rejected and better balanced sample results should be searched (Ho, Imai, King, & Stuart, 2011; Stuart, 2010). There might be cases in which comparison groups are too far apart from each other in the background variables that make it hard to produce adequate estimates using matching (Rubin, 2001).

Studies on Propensity Score Matching in Evaluating DIF

There are few studies that used propensity score matching methods in evaluating DIF. Lee and Geisinger (2014) analyzed an English reading test administered to South Korean college students for gender DIF using Mantel-Haenszel and Logistic Regression DIF detection methods before and after matching students on interest in education. They reported that the propensity score approach using optimal matching reduced the number of biased items. Joldersma and Bowen (2010) compared Language Arts Literacy items translated from Spanish to English using the Mantel-Haenszel DIF detection method before and after matching students on various covariates, such as gender, economic status, and total test score, using nearest neighbor matching. The matching procedure eliminated item bias, which is not surprising given that the authors used the outcome variable, total test score of the students, to match the groups. Wu and Ercikan (2006) used propensity score matching to investigate effects of Extra Lesson Hours After School (ELHAS) on DIF between Taiwanese and U.S. students in TIMSS. In their study, ELHAS was included in a logistic regression model as a main effect and interaction term. They reported that ELHAS was related to a reduction of magnitude and number of DIF items. Similarly, Zumbo and Gelin (2007) used community location and income level of students as a contextual variable to investigate effects on gender DIF using mathematics test of the Foundation Skills Assessment in British Columbia. They named this analysis differential domain functioning and reported that differential domain functioning was present in their study. All of these studies generally implied that contextual variables could be effective in investigating sources of DIF (Lee & Geisinger, 2014; Wu & Ercikan, 2006; Zumbo and Gelin, 2007). What is still missing is a comparison of effects of different propensity score matching methods on DIF using large-scale assessment data.

Present Study

DIF procedures are based on matching on test scores. We argue that matching on additional, potentially bias-relevant background variables would be helpful to identify sources of DIF. What we do here can be seen as a combination of a procedure called thin matching (the use of total score as the matching variable) and thick matching (forming the matching variable by pooling total score levels) (Donoghue & Allen, 1993). In this study, using exact, nearest neighbor, and optimal matching methods, PISA 2012 released mathematics items were analyzed in terms of DIF for Indonesian, Turkish, Australian, and Dutch students. These four countries were selected to represent wide spectrum of countries in terms of mathematics performance such as, Indonesian students were included to represent a low achieving country, Turkish students were included to represent a below average country, Australian students were included to represent an above average country and Dutch students were included to represent a high achieving country according to results of PISA 2012. Additionally, these countries differ in terms of background variables, such as socioeconomic status; therefore, the extent to which these matching methods were effective in comparing culturally different groups could be tested. By using various types of matching methods on data of these differentially achieving countries, we aim to evaluate effects of various matching methods that use propensity score methodology to study DIF results and to understand the nature of bias in the comparison of educational performance of these four countries. So, we examined to what extent propensity score matching methods are effective in understanding nature of bias by reducing or eliminating the sources of bias and to what extent propensity score matching is able to explain cross-national differences in mathematics performance.

Method

Participants

The data of this study were obtained from the PISA 2012 data set. In PISA, the target population is all 15 years-old students of participating countries. This study used all Indonesian, Turkish, Australian, and Dutch students who answered released mathematics items, as the item bias analysis requires that item contents are known (which restricts the number of items that could be included in the analysis). In this study, the data were investigated from 1078 Indonesian students (540 females and 538 males), 951 Turkish students (462 females and 489 males), 2824 Australian students (1398 females and 1426 males) and 839 Dutch students (415 females and 424 males).

Measures

PISA 2012 gathered data on students' mathematics performance and students' characteristics via cognitive items and student questionnaire, respectively. The present study used released sample items of PISA to evaluate DIF. In the PISA 2012 mathematics test, there were 13 released items that were answered by the samples described above.

Student background variables that are considered to be sources of DIF and controlled by propensity scores matching were gender, index of economic, social, and cultural status, and opportunity to learn. The index of economic, social, and cultural status, reported by PISA, is a combination of the highest occupational status of parents, the highest educational level of parents, family wealth, cultural possessions, and home educational resources (OECD, 2014b). Opportunity to learn is defined as student's exposure to subject domain content in school previously and is an important predictor of achievement (Schmidt & Maier, 2009).

Data Analysis

As a first step, DIF analyses were conducted using structural equation modeling (SEM) and logistic regression (LR) DIF detection methods without matching students on contextual variables among Indonesian, Turkish, Australian, and Dutch students. The MPLUS 7.11 and SPSS 19.0 programs were used for SEM and LR DIF analysis, respectively. In the SEM procedure, a Confirmatory Factor Analysis (unifactorial, with all items as indicators of the latent variable) was conducted, assessing configural and scalar invariance. In the logistic regression procedure, total test score, country, and their interaction were used as predictors. Significance of country and their interaction were taken as evidence for uniform bias and non-uniform bias, respectively. Then, for each comparison group, exact, nearest neighbor and optimal matching methods were performed using gender, the index of economic, social and cultural status, and opportunity to learn as contextual variables. Propensity matching does not yet have a single best procedure and there is no guarantee that different procedures yield similar outcomes; therefore, we applied multiple procedures. The MatchIt R package (Ho, Imai, King, & Stuart, 2007) was used to do the matching and to estimate propensity scores. The data derived as a result of each matching method were evaluated in terms of balance. Then, DIF analyses were reconducted using matched group data produced by each matching method. Finally, effects of each matching methods on country mean achievement scores calculated on the basis of the released items were investigated to examine to what extent country score differences could be explained by differences in variables constituting the propensity scores.

Results

Internal Consistency Analysis of the Instrument

Cronbach's alpha reliability coefficients in the PISA 2012 mathematics test were .601 for Indonesian students, .777 for Turkish students, .773 for Australian students, and .738 for Dutch students. These values are satisfactory (Cicchetti, 1994), with the exception of the low value for Indonesia. There are no clear reasons for the lower value in Indonesia.

Descriptive Statistics and Evaluation of Matching

A comparison of the sample sizes of the original and matched data is presented in Table 1 and the means on gender, ESCS, and opportunity to learn are given in Table 2. An exact matching procedure of Turkish and Dutch students yielded only 33 to 32 matches; between Indonesian and Dutch students only 35 to 36 matches could be found. These low numbers of matches suggested that these students (Turkish vs. Dutch; Indonesian vs. Dutch) were very different on these contextual variables. Exact matching resulted in the highest number of matches between Australian and Dutch students, 334 to 286 matches. Nearest neighbor and optimal matching produced new groups based on the minimum number of the students in any group. For instance, when Turkish and Australian students were matched, both groups had 951 students, whereas when Dutch and other countries matched, groups had 839 students.

Table 1. Sample size before and after matching

Comparison Groups	Sample Size			
	Original Data	Exact Matching	Nearest Matching	Optimal Matching
Indonesia	1078	69	951	951
Turkey	951	67	951	951
Indonesia	1078	101	1078	1078
Australia	2824	127	1078	1078
Indonesia	1078	35	839	839
The Netherlands	839	36	839	839
Turkey	951	104	951	951
Australia	2824	131	951	951
Turkey	951	33	839	839
The Netherlands	839	32	839	839
Australia	2824	334	839	839
The Netherlands	839	286	839	839

Indonesian and Turkish students originally had lower mean ESCS values (-1.74 and -1.46, respectively), whereas Australian and Dutch students had a higher mean ESCS (.18 and .28, respectively). Exact matching produced groups with close ESCS values. In both nearest neighbor and optimal matching procedures, ESCS values were very similar between Australian and Dutch students, and between Indonesian and Turkish students. However, after matching, the difference in ESCS diminished only slightly between Indonesian students and both Australian and Dutch students, and between Turkish students and both Australian and Dutch students. For instance, between Dutch and Turkish students, as all Dutch students were kept and number of Turkish student were reduced to the number of Dutch students, the ESCS value of Dutch students remained the same (.28) and the ESCS value of Turkish students only increased from -1.46 to -1.26. Descriptives related to opportunity to learn showed similar findings as found for ESCS in terms of matching. The only difference was that Indonesian students originally had the highest mean score in opportunity to learn. The gender distribution of the original groups was very close (1.50 indicates equal distribution). The gender distributions in the matched groups deviated slightly.

Table 2. Mean of matching variables before and after matching

Comparison Groups	ESCS (Mean)				Opportunity to Learn (Mean)				Gender (Mean)			
	Original Data	Exact M.	Nearest M.	Optimal M.	Original Data	Exact M.	Nearest M.	Optimal M.	Original Data	Exact M.	Nearest M.	Optimal M.
Indonesia	-1.74	-1.71	-1.64	-1.65	3.32	3.21	3.25	3.25	1.50	1.61	1.49	1.49
Turkey	-1.46	-1.69	-1.46	-1.46	2.92	3.20	2.92	2.92	1.51	1.60	1.51	1.51
Indonesia	-1.74	-.33	-1.74	-1.74	3.32	3.31	3.32	3.32	1.50	1.48	1.50	1.50
Australia	.18	-.29	-.59	-.59	3.19	3.32	3.23	3.24	1.50	1.52	1.51	1.51
Indonesia	-1.74	-.21	-1.38	-1.38	3.32	3.37	3.28	3.28	1.50	1.63	1.49	1.49
The Netherlands	.28	-.17	.28	.28	3.15	3.39	3.15	3.15	1.51	1.58	1.51	1.51
Turkey	-1.46	-.35	-1.46	-1.46	2.92	3.06	2.92	2.92	1.51	1.55	1.51	1.50
Australia	.18	-.26	-.63	-.63	3.19	3.10	3.01	3.02	1.50	1.52	1.51	1.51
Turkey	-1.46	-.29	-1.26	-1.26	2.92	3.10	2.98	2.98	1.51	1.52	1.52	1.52
The Netherlands	.28	-.27	.28	.28	3.15	3.09	3.15	3.15	1.51	1.56	1.51	1.51
Australia	.18	.39	.26	.28	3.19	3.22	3.13	3.15	1.50	1.48	1.46	1.49
The Netherlands	.28	.36	.28	.28	3.15	3.22	3.15	3.15	1.51	1.47	1.51	1.51

ESCS = Economic, Social, and Cultural Status. M = Matching.

The quality (balance) of the matching was evaluated by examining propensity score mean differences divided by their standard deviation and propensity score variance ratios (See Table 3). Exact matching produces perfect matches according to balance criteria, which is in line with its definition; however, there is a problem in that there are very few subjects matched on these covariates, which can create very atypical samples. Nearest neighbor and optimal matching produced exactly the same balance evaluation values for each pair of groups. Both matching methods suggested a good match between Australian and Dutch students. Matching results between other groups of students suggested that the balance evaluation values were higher than the expected, which indicated the matching was not entirely adequate as propensity score mean differences were higher than half a standard deviation and propensity score variance ratios were not around one.

Table 3. Evaluation of balance

Comparison Groups	Mean Diff. / Standard Dev.			Variance Ratios		
	Exact M.	Nearest M.	Optimal M.	Exact M.	Nearest M.	Optimal M.
Indonesia - Turkey	.00	.61	.61	1.00	1.46	1.46
Indonesia - Australia	.00	1.23	1.23	1.00	3.15	3.15
Indonesia - the Netherlands	.00	1.45	1.45	1.00	1.48	1.48
Turkey - Australia	.00	.95	.95	1.00	2.97	2.97
Turkey - the Netherlands	.00	1.35	1.35	1.00	1.65	1.65
Australia - the Netherlands	.00	.00	.00	1.00	1.00	1.00

M. = Matching.

Effects of Matching on DIF Results

In this section, results based on SEM and LR DIF detection methods are presented first using original data and then using matched data (See Table 4). Related to SEM original data results, detailed statistics of configural and scalar invariance were given in Appendix A.

Original Data

Between Indonesian and Turkish students, and Indonesian and Dutch students, only item 3 was flagged as having DIF in both SEM and LR. Between Indonesian and Australian students none of the items was flagged as having DIF by both SEM and LR, although there were other items that were flagged as DIF by only one method. Between Turkish and Australian students, no items were flagged by both methods. Between Turkish and Dutch students, items 2 and 4 were flagged as having DIF in both SEM and LR, whereas additional items were flagged by SEM and LR separately. Between Dutch and Australian students, the SEM method identified two items as having DIF, whereas LR identified none. These findings were used as a basis to evaluate results found by matching methods.

Exact Matching

As stated previously, exact matching produced a very small sample size. Therefore, DIF analysis could not be conducted using exact matching data except between Australian and Dutch students. Between Australian and Dutch students, SEM method identified item 6 as having DIF. Originally, SEM detected two different items (7 & 10), so we got fewer, yet different DIF items with exact matching. For LR neither the original data nor the exact matching data revealed any items as having DIF.

Nearest Neighbor and Optimal Matching

Nearest neighbor and optimal matching methods produced almost the same results in terms of descriptives and evaluation of balance. They also produced the same results in the DIF analysis. DIF results between Indonesian students and other three groups of students suggested that among SEM and LR methods there was no convergence. When we compared items flagged in the original data and in nearest neighbor and optimal matching data, there was no clear pattern of diminishing DIF. Between Turkish and Australian students, no items were identified as having DIF in SEM or LR, in line with original data. Between Dutch and Turkish students, the same items in original data were identified as having DIF in SEM and in LR. As the balance evaluation suggested that the matching of the data was not adequate, finding the same results in the matched data as in the original data was not surprising. Therefore, unbalanced matched data between Turkish and Australian students and between Dutch and Turkish students did not produce any change (reduction) in DIF results. The SEM method identified only item 4 as having DIF in the comparison between Dutch and Australian students. The number of items detected as having DIF reduced from two to one. As none of the item was identified as having DIF in LR, no improvement could be found. Therefore, if the balance of matching is not good in nearest neighbor and optimal matching, matching is not helpful to find the sources of DIF.

Using Propensity Scores as a Predictor

As nearest neighbor and optimal matching could not produce a good match, especially for countries that were very different on background variables, the actual propensity scores estimated by matching methods were used as covariate to flag DIF items. As nearest neighbor and optimal matching all produced the same values, only the propensity scores estimated by the former method was used. Propensity scores and interaction between propensity score and group membership were added to the equation in logistic regression to compute DIF; in SEM, the propensity score were added as predictor of the outcome of each item. For LR, results involving Indonesian data showed the same pattern and there was no reduction in the number of biased items flagged. However, when comparing Turkish and Dutch student data, all the items originally flagged as DIF no longer showed DIF. For SEM, using propensity score as predictor eliminated all DIF in the Indonesian-Turkish, Turkish-Dutch, and Australian-Dutch comparisons. Additionally, in the Indonesian-Dutch comparison, the number of items showing DIF decreased from five to two, and in the Indonesian-Australian comparison, decreased from two to one. Overall, using propensity scores as predictor in DIF detection was found to be an effective method to reduce the number of items showing DIF. Our analyses strongly suggest that the estimated propensity scores should be added to the DIF analysis, as only matching data and examining DIF in these matched samples (without using propensity scores) may not be effective to reduce DIF, especially in poorly balanced matches.

Effects of Matching on Mean Score Differences

The mathematics PISA performance differences among these four countries were evaluated in this part of the study. Table 5 shows that, in the original data, the mathematics performance difference between Indonesia and Netherlands had the largest effect size in terms of Cohen's *d* (1988), whereas comparisons of Turkey and Australia, and Australia and Netherland yielded the smallest performance difference. Comparing performance differences after matching with various methods suggested that the differences did not change in effect size dramatically. Only in the Turkish and Australian student comparison, the effect size was decreased moderately. One of the reasons of this finding could be that as the matching did not produce a good balance, matching did not have an effect on performance differences between groups.

The same analyses were repeated by using propensity scores as covariate to understand the remaining performance differences, especially for group comparisons with a poor balance. Using propensity scores as the covariate, there were no huge changes in effect size when Indonesia was compared with other countries. This could be due to high level of opportunity to learn values of Indonesian students which might cancel out effectiveness of propensity scores. The background of the high Indonesian scores on self-reported opportunity to learn is unclear. When comparing Australian and Dutch students, almost the same effect sizes were obtained as these groups had good balance in their match. Using propensity score as covariate did not add new information for well-balanced data. However, using propensity scores as covariate helped to reduce the mathematics performance difference between Turkish-Dutch and Turkish-Australian students' comparisons, suggesting that ESCS and opportunity to learn were important predictors of the country differences in mathematics performance in these countries.

Table 4. Items flagged in DIF analysis

Comparison Groups	SEM					LR				
	Original Data	Exact M.	Nearest M.	Optimal M.	Nearest M. with Covariate	Original Data	Exact M.	Nearest M.	Optimal M.	Nearest M. with Covariate
Indonesia – Turkey	3, 9	NA	1, 2, 7	1, 2, 7	No DIF	3	NA	3	3	3
Indonesia – Australia	1, 6	NA	1, 5, 7, 8	1, 5, 7, 8	7	12	NA	3, 12	3, 12	12
Indonesia – the Netherlands	3, 7, 9, 11, 13	NA	1, 2, 7	1, 2, 7	3, 9	3, 12	NA	3, 12	3, 12	3, 12
Turkey – Australia	No DIF	NA	No DIF	No DIF	No DIF	No DIF	NA	No DIF	No DIF	No DIF
Turkey – the Netherlands	2, 4, 6, 11	NA	2, 4, 6, 11	2, 4, 6, 11	No DIF	2, 4, 7	NA	2, 4, 7	2, 4, 7	No DIF
Australia – the Netherlands	7, 10	6	4	4	No DIF	No DIF	No DIF	No DIF	No DIF	No DIF

NA: Not applicable because sample size was too small for DIF analysis. M. = Matching. SEM = Structural Equation Modeling. LR = Logistic Regression.

Table 5. Effect sizes of total score difference before and after matching

Comparison Groups	Original Data	Exact M.	Nearest M.	Optimal M.	Nearest M. with Covariate	Optimal M. with Covariate
Indonesia – Turkey	-.74***	-.81***	-.72***	-.71***	-.68***	-.68***
Indonesia – Australia	-1.01***	-.94***	-.85***	-.85***	-.69***	-.69***
Indonesia – the Netherlands	-1.36***	-.86***	-1.31***	-1.31***	-1.34***	-1.34***
Turkey – Australia	-.40***	.11	-.07	-.07	.23***	.23***
Turkey – the Netherlands	-.81***	-.36	-.76***	-.76***	-.36***	-.36***
Australia – the Netherlands	-.42***	-.38***	-.41***	-.43***	-.42***	-.45***

***$p < .001$. M. = Matching.

Discussion

In this study we aimed to answer the question to what extent propensity score matching methods are effective in understanding the nature of DIF in PISA 2012 in the comparison of mathematics performance and to explain cross-national differences in mathematics performance. After detecting DIF, evaluating sources of DIF is necessary to distinguish DIF from naturally occurring and valid group differences. However, experience has shown that this evaluation can be cumbersome and that it is often far from easy to detect the sources of DIF. It would be helpful to control other relevant factors, such as background characteristics of students, to address their potential impact on reducing the number of DIF items and eliminate background characteristics to interfere with the results. This study hypothesized that matching students of different groups using propensity scores could help to achieve this goal.

If the balance of the matching is good, we found that matching could reduce the number of items showing DIF as found in the Australian-Dutch comparison. To understand the nature of this (positive) finding, it is important to note that in the Australian-Dutch comparison countries were close to each other in terms of background variables and mathematics performance. However, our results also suggested that if the balance of the matching is not good, using matched groups data did not provide much additional information or aid in understanding the nature of the bias. Therefore, evaluation of balance before conducting further analysis is necessary. Rubin's (2001) recommendation of balance evaluation criteria that propensity score mean differences should be less than half a standard deviation and propensity score variance ratios could be very useful. The feasibility of matching would be enhanced if it can be also used to compare groups that are not very similar in performance or background characteristics. In the study, the solution of poor balance on matching is salient in procedures using actual propensity scores in DIF analysis. DIF detection methods, using propensity scores as covariate in DIF detection is found to be an effective method in both LR and SEM to reduce the number of items showing DIF. This finding is congruent with the finding of Wu and Ercikan (2006) who included Extra Lesson Hours after School in a logistic regression model. They reported that this variable was related to a reduction of the magnitude and number of DIF items. We demonstrated how propensity score matching methods that can deal with multiple matching variables at once can be used not only by LR but also by SEM DIF detection analysis.

One of the perennial problems of DIF studies is to identify sources of DIF. Our study showed that propensity score matching when it is used as a predictor or covariate has the potential to address this problem. Propensity scoring provided us with a new tool to better control sources of item bias. In our case, controlling gender, socioeconomic status and opportunity to learn were effective in identifying sources, as they were important predictors of mathematics performance in PISA (Areepattamannil, 2014; Kilic, Cene, & Demir, 2012; Liu & Wilson, 2009; Machin & Pekkarinen, 2008; Perry & McConney, 2010; Schmidt, Zoido, & Cogan, 2014). At a conceptual level, the findings suggest that diminished DIF items are induced by these background variables and not by other cultural differences. Propensity score matching might be able to focus our search for bias. Interestingly, the bias sources employed in this study (gender, socioeconomic status, and opportunity to learn) may also create bias within countries.

Comparison of cross-national mathematics performance differences before and after matching showed that when the balance was not good, matching did not decrease the size of the country differences. This result could be due to a poor balance in matching very different countries. However, when actual propensity scores were used as covariate, differences between countries tended to be reduced. Large decrements in mathematics achievement difference were achieved between Turkish and Dutch students, and between Turkish and Australian students when we used propensity score as covariate. For Turkish and Australian students, it was found that Turkish students would be more successful than Australian students if they had same ESCS and opportunity to learn. This finding implies that if two groups are different in terms of achievement level and background variables, using propensity score as covariate could explain an important part of achievement difference between these groups. In our case, mathematics performance differences between Turkish and Dutch students as well as Turkish and Australian students could be largely attributed to differences in ESCS and opportunity to learn. The implication of this finding would be that if Turkish students would have higher ESCS and more educational exposure to the educational activities that are related to proficiencies measured in PISA, the difference would be smaller. Remarkably, comparisons involving Indonesian students did not show huge changes in effect size before and after controlling for propensity scores. We can speculate that the high level of opportunity to learn scores of Indonesian students (even higher than those of Dutch students) challenged the effectiveness of propensity scores. Clearly, if matching variables are not fully comparable across countries, any correction procedure will have a questionable value.

When the matching methods are compared, nearest neighbor and optimal matching produced the same results in terms of sample size, mean of matching background variables, evaluation of balance, DIF results, and performance differences measured by effect size. Therefore, selecting one of these methods to produce actual propensity scores most probably will not change the outcome. Exact matching, on the other hand, can decrease sample size drastically, especially when the matched groups differ in matching characteristics. Paradoxically, exact matching may not help to equate groups when it is needed most (namely when groups are most dissimilar). Additionally, we recommend that other matching methods, such as optimal full matching, could be tried to test whether they are effective in reducing bias. In order to test the generalizability of the results, similar procedures might be followed using other educational achievement such as data from PIRLS (Progress in International Reading Literacy Study) or TIMSS (Trends in International Mathematics and Science Study). Overall, this study is novel in that it investigated various propensity score matching methods to understand the nature of DIF by using PISA educational performance data of countries that represents diverse results. This study is important for showing the effectiveness of propensity score in explaining sources of DIF and cross-national differences in mathematics performance. Additionally, presenting various outcome of these three matching methods result using four different countries helped our study to have a good coverage in terms of matching methods and sample. Propensity score matching approach is a novel topic that is expected to gain more popularity in educational and psychological research.

Limitations

Finally, our study has some limitations. First of all, findings are based on selected background variables that are considered to be linked with DIF. These variables are selected based on prospective candidates of sources of DIF. Selecting other sets of variable could produce different results. Another limitation is that this study only focused on comparing matching methods using released items. The other limitation is that choice of PISA data to flag the DIF. As PISA items are prepared with great care and as most probably the best functioning items are released, finding patterns of DIF could be more difficult than any regular set of items. Repeating a similar study with other sets of data could be helpful to see the similarities of the results.

Acknowledgements

This paper is a part of a research project supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with program no. 2219. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the TUBITAK.

References

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W.

Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Areepattamannil, S. (2014). International Note: What factors are associated with reading, mathematics, and science literacy of Indian adolescents? A multilevel examination. *Journal of adolescence*, *37*(4), 367-372.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290.

Cohen, J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale,
NJ: Lawrence Erlbaum Associates.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics*, *18*, 131-154.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405-420.

He, J., & Van de Vijver, F.J.R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G.A.D. Liem & A.B.I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39-56). Charlotte, NC: Information Age Publishing.

Ho, D., 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.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for Parametric Causal Inference. Retrieved from http://r.iq.harvard.edu/docs/matchit/2.4-20/matchit.pdf.

Joldersma, K., & Bowen, D. (2010). *Application of Propensity Models in DIF Studies To Compensate For Unequal Ability Distributions*. Paper presented at the annual meeting of National Council on Measurement in Education, Denver, CO.

Kiliç, S., Çene, E., & Demir, İ. (2012). Comparison of learning strategies for mathematics achievement in Turkey with eight countries. *Educational Sciences: Theory and Practice*, *12*(4), 2594-2598.

Lee, H., & Geisinger, K. F. (2014). The Effect of Propensity Scores on DIF Analysis: Inference on the Potential Cause of DIF. *International Journal of Testing*, *14*(4), 313-338.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164-184.

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science*, *322*, 1331-1332.

OECD (2013), *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)*, PISA, OECD Publishing.

OECD (2014a). *PISA 2012 Results: What students know and can do – Student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014). Paris, France: OECD Publishing.

OECD (2014b). *PISA 2012 Technical Report.* Paris, France: OECD Publishing.

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *The Teachers College Record*, *112*(4), 7-8.

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The *American Statistician, 39*(1), 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*(3-4), 169-188.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A Meta-analytic review of research. *Review of Educational Research*, 75(3), 417-453.

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of Education Policy Research* (pp. 541–559). New York, NY: Routledge for American Educational Research Association.

Schmidt, W., P. Zoido and L. Cogan (2014), *Schooling Matters: Opportunity to Learn in PISA 2012*, OECD Education Working Papers, No. 95. Paris, France: OECD Publishing.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1-21.

Thomson, S., De Bortoli, L., & Buckley, S. (2013). *PISA 2012: How Australia measures up*. Camberwell, Australia: Australian Council for Educational Research.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119-135.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*(3), 287-300.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies, 5*(1), 1-23.

Author Information

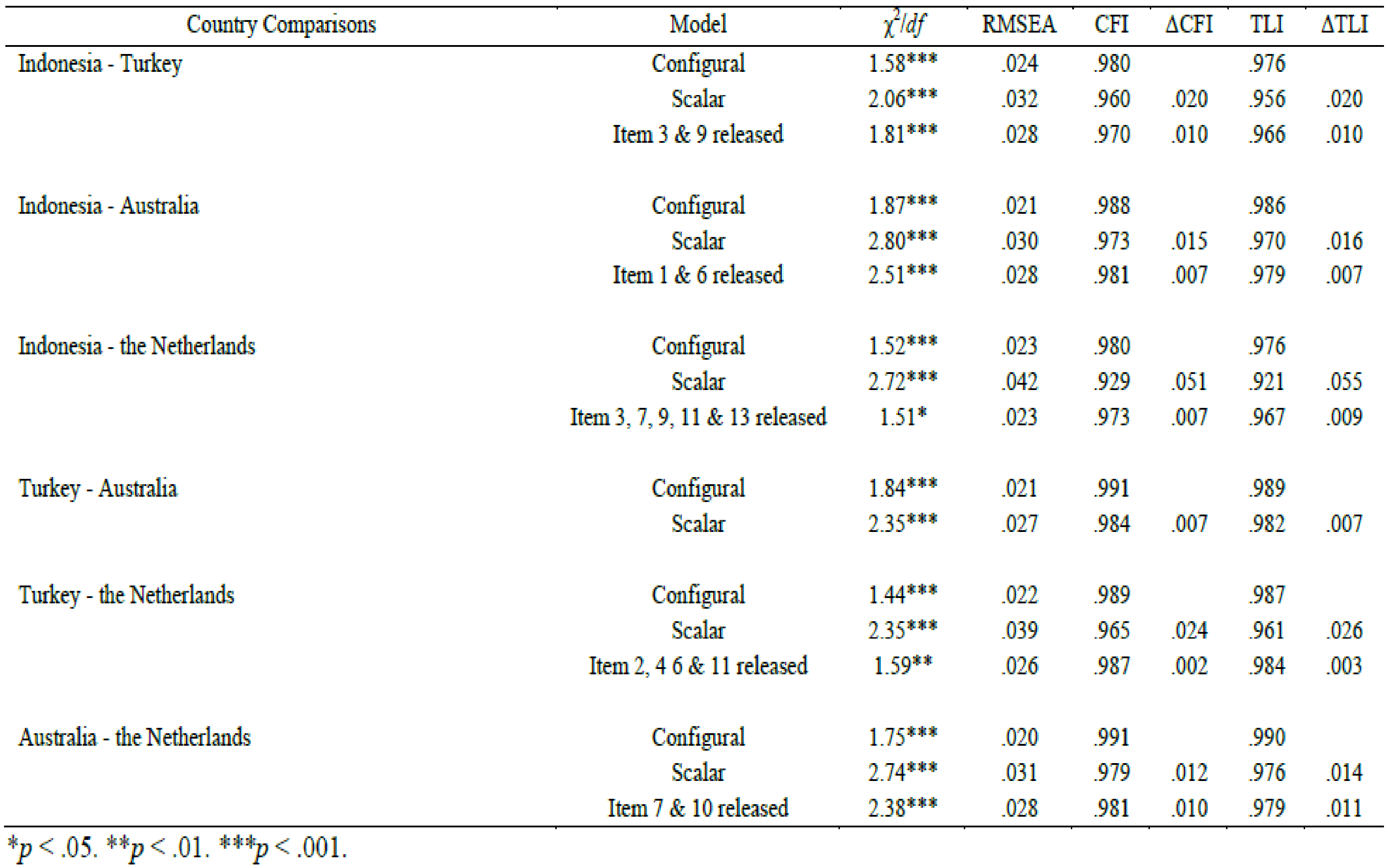
Serkan Arikan
Mugla Sitki Kocman University,
Mugla, Turkey
Contact e-mail: *serkanarikan@mu.edu.tr*

Fons J. R. van de Vijver
Tilburg University, the Netherlands
North-West University, South Africa
University of Queensland, Australia

Kutlay Yagmur
Tilburg University,
Tilburg, the Netherlands

Appendix A.

Item Bias Analysis by Structural Equation Modeling – Original Data



Country Comparisons	Model	χ^2/df	RMSEA	CFI	ΔCFI	TLI	ΔTLI
Indonesia - Turkey	Configural	1.58***	.024	.980		.976	
	Scalar	2.06***	.032	.960	.020	.956	.020
	Item 3 & 9 released	1.81***	.028	.970	.010	.966	.010
Indonesia - Australia	Configural	1.87***	.021	.988		.986	
	Scalar	2.80***	.030	.973	.015	.970	.016
	Item 1 & 6 released	2.51***	.028	.981	.007	.979	.007
Indonesia - the Netherlands	Configural	1.52***	.023	.980		.976	
	Scalar	2.72***	.042	.929	.051	.921	.055
	Item 3, 7, 9, 11 & 13 released	1.51*	.023	.973	.007	.967	.009
Turkey - Australia	Configural	1.84***	.021	.991		.989	
	Scalar	2.35***	.027	.984	.007	.982	.007
Turkey - the Netherlands	Configural	1.44***	.022	.989		.987	
	Scalar	2.35***	.039	.965	.024	.961	.026
	Item 2, 4 6 & 11 released	1.59**	.026	.987	.002	.984	.003
Australia - the Netherlands	Configural	1.75***	.020	.991		.990	
	Scalar	2.74***	.031	.979	.012	.976	.014
	Item 7 & 10 released	2.38***	.028	.981	.010	.979	.011

*$p < .05$. **$p < .01$. ***$p < .001$.