Comparing Apples to Oranges: Using Propensity Score Matching to Reduce Bias in Cross-Country Comparisons in PISA

Research Report

Annemarie Timmers (6238106)

December 2021

Supervisors: Martina Meelissen (University of Twente) and Remco Feskens (CITO, University of Twente)

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

Word count: 2502

1 Introduction

When the results of the *Programme for International Student Assessment* (PISA) come out, many news articles report on the performance of the respective countries on the country league table (see for example Coughlan (2019)). Often, a call for policy reform is made in less-performing countries as they turn to model their education systems after those of the countries at the top of the league (Sellar & Lingard, 2013). However, numerous issues have been raised regarding the validity of PISA's methodology, questioning whether the PISA results should be valued so much in informing education policy reform (Araujo et al., 2017; Fernandez-Cano, 2016; Rutkowski & Rutkowski, 2016). One concern regards the handling of differential item functioning (DIF) in the psychometric model used to analyze the data. DIF occurs when students from different groups with the same ability level have different response probabilities on an item (Ackerman, 1992). Therefore, DIF in PISA items might jeopardize the validity of the country rankings.

This study focuses on one potential cause of DIF: mode effects. Mode effects exist when DIF occurs due to the test administration mode, which arises because of differences in test-taking experience between modes (e.g., ease of reading item texts or ease of reviewing and changing answers) (Kolen & Brennan, 2014). In 2015, PISA transitioned from an exclusively Paper-Based Assessment (PBA) to a mainly Computer-Based Assessment (CBA) while remaining paper-based in some countries, thus evoking the possibility of mode effects (OECD, 2017).

Another change in 2015 was PISA's treatment of DIF. Up until 2012, PISA treated DIF by deleting items, hence not considering DIF in the psychometric model as that was uniformly applied to the remaining items (OECD, 2014; Zwitser et al., 2017). In 2015, DIF due to mode and country effects was considered by allowing item parameters to vary in the countries with severe item misfit (Feskens et al., 2019; OECD, 2017). So, all DIF was taken into account in the psychometric model, which is also undesirable as not all DIF will bias the results (Zumbo & Gelin, 2005; Zwitser et al., 2017).

Propensity score matching (PSM) offers a compromise of methods and might help in disentangling sources of DIF (Liu et al., 2020). Few studies have applied PSM to education assessment data, but those that did, noted a reduction in item bias (e.g., Arikan et al. (2018), Joldersma and Bowen (2010), H. Lee and Geisinger (2014), Puhan et al. (2005), Seo and De Jong (2015), and Wu and Ercikan (2006)). PSM has yet to be used to combat mode effects in PISA and as the cited studies

applied PSM to students from up to four countries, it remains unknown to what extent a reduction in item bias influences country rankings on a larger scale. Therefore, the present thesis will examine the following research question: *to what extent can different propensity score matching methods reduce mode effects in the PISA 2018 science items, and how does this affect the country league table?*

This paper is structured as follows. Section 2 outlines the theoretical background of this study. Section 3 discusses the data and analytical strategy used to analyze the data.

2 Theoretical Background

2.1 The MIMIC Model

The Multiple Indicators, Multiple Causes (MIMIC) model can be used to detect DIF in assessment data. The MIMIC model was first introduced by Jöreskog and Goldberger (1975) with its distinguishing feature being the possibility to regress constructs on covariates. Although the MIMIC model was only used to detect uniform DIF at the beginning of its use, Woods et al. (2009) expanded its properties to also being able to detect nonuniform DIF when they introduced interactions into the model.

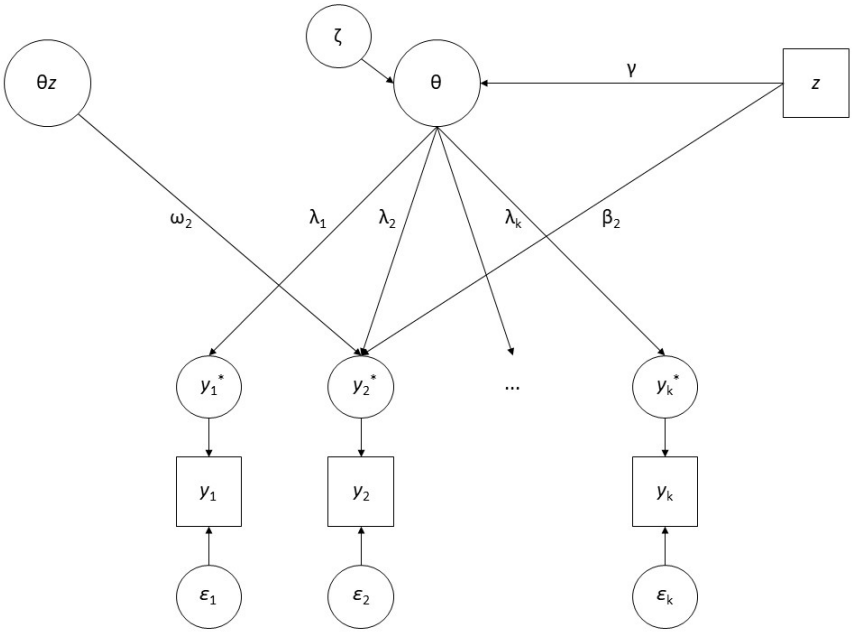

Figure 1: Schematic illustration of the MIMIC model for detecting uniform and nonuniform DIF for item 2.

There are multiple strategies to use the MIMIC model and it is out of the scope of this paper to discuss them all (see J. Lee et al. (2018) for further reading). The constrained-baseline strategy tests one item for DIF at a time, while the other items are assumed to be DIF-free. First, a baseline model is fitted to the data without direct effects. Then, the fit of the baseline model is evaluated against i models (where i is the number of items) that specify a direct effect between one item and the grouping variable, while controlling for group differences by regressing the construct on the grouping variable, and an interaction term between the construct and the grouping variable (J. Lee et al., 2018). Figure 1 displays an example of this item model, where the second item is tested for DIF. The MIMIC model for item i can be written as:

$$y_i^* = \lambda_i\theta + \beta_i z + \omega_i\theta z + \varepsilon_i, \tag{1}$$

where y_i^* is the latent ability underlying the observed item response for item i, λ_i is the factor loading for item i (the item discrimination parameter in IRT models), θ is the latent ability that follows a normal distribution, z is the dummy-coded grouping variable, β_i is the uniform DIF effect for item i, ω_i is the nonuniform DIF effect for item i, and ε_i is the random error for item i (Bulut & Suh, 2017). Uniform and nonuniform DIF for item i are present when $\beta_i \neq 0$ and $\omega_i \neq 0$, respectively.

2.2 Propensity Score Matching

PSM originated in the context of non-randomized experiments as a proxy for random assignment and constructs comparable groups by equating sampling groups on propensity scores based on background characteristics (Rosenbaum & Rubin, 1983). Three steps constitute the matching process: selecting covariates (3.2), estimating propensity scores and matching (2.2.2), and evaluating covariate balance (2.2.3).

2.2.1 Selecting covariates

The first step in PSM is selecting covariates to match the groups on. Although there is some discourse regarding this topic, the consensus is that including covariates that are related to both the construct and the grouping variable generates the best performance (Stuart, 2010).

2.2.2 Estimating propensity scores and matching

After selecting covariates for matching, propensity scores can be estimated. A propensity score is the conditional probability of an individual belonging to the focal group given the selected covariates and can be expressed as:

$$e(\mathbf{X_i}) = P(Z_i = 1|\mathbf{X_i}), \tag{2}$$

where $e(\mathbf{X_i})$ is the propensity score for individual i, Z_i is the indicator of the grouping variable of individual i, $Z_i = 1$ denotes the individuals i belonging to the focal group and $\mathbf{X_i}$ is the vector of scores on the selected covariates for individual i (Rosenbaum & Rubin, 1983). Propensity scores are usually estimated through logistic regression:

$$P(Z_i = 1|\mathbf{X_i}) = \frac{e^{\beta_0+\beta(\mathbf{x_i})}}{1 + e^{\beta_0+\beta(\mathbf{x_i})}}, \tag{3}$$

where β_0 is the intercept and $\beta(\mathbf{x_i})$ is a vector of coefficients for the included covariates (Rosenbaum & Rubin, 1983).

The estimated propensity scores are matched to adjust for group differences. There are multiple ways to do this and it is advised to apply multiple methods as their performance differs based on the data (Liu et al., 2016). Matching methods can broadly be divided into greedy and optimal matching.

Greedy matching Greedy matching methods do not optimize a criterion, which means that units are matched once, without accounting for how other units will be or have been matched. For example, in Nearest Neighbor Matching (NNM), a random unit in the focal group is selected and paired with the closest unit in the reference group, after which the algorithm turns to the next focal unit and repeats this process until all focal units are paired (Liu et al., 2016; Stuart, 2010).

Optimal matching Contrary to greedy matching, matches can be redone in optimal matching until a specified criterion is optimized. The most common optimal matching methods are optimal pair (OPM) and optimal full matching (OFM). OPM matches in pairs and discards the remaining unmatched units, while OFM allows one focal unit to be matched to multiple reference units (one-to-many) or multiple focal units to one reference unit (many-to-one) (Liu et al., 2016; Stuart, 2010).

2.2.3 Evaluating the covariate balance

After matching, the performance of the different methods is evaluated by examining the distance within the matched pairs and the balance of the matched groups (Gu & Rosenbaum, 1993). The distance within pairs implies that the values of the matched units on covariate X should be as close as possible. The difference between the propensity scores of the matched focal and reference units is a measure of this distance (Stuart, 2010).

The matched groups are balanced when the distribution of covariate X in the focal group resembles that of the reference group in the matched subsample. Balance can be assessed visually, by comparing histograms and density plots of the groups before and after matching, or by computing measures that quantify the difference between the distributions. Examples of such statistics are standardized mean differences, which should be close to 0, or variance ratios, which should be close to 1 (Chen et al., 2020).

2.3 Propensity Score Matching in Educational Assessment

Relatively few studies have applied PSM to education assessment data. Seo and De Jong (2015) studied mode comparability in the Michigan Educational Assessment Program. They found the modes to be comparable after matching students on gender, ethnicity, special education, English proficiency, economic disadvantage, and mathematics and readings scores for students and schools. Puhan et al. (2005) reached a similar conclusion when they analyzed mode comparability in a teacher certification test by matching teachers on gender, language, test repeater status, race, GPA, and education level.

Apart from mode effects, other causes of DIF have also been studied using PSM. Wu and Ercikan (2006) looked at the effect of Extra Lesson Hours After School (ELHAS) on DIF between Taiwanese and U.S. students in TIMSS. They found that matching students on ELHAS resulted in a reduction in the number of items displaying DIF. Further, Joldersma and Bowen (2010) examined DIF due to language in a Language Arts Literacy assessment that was translated from English to Spanish. After matching students on gender, economic status, and total test score, they found that item bias was eliminated in both versions of the test. Additionally, H. Lee and Geisinger (2014) studied gender DIF in an English reading test administered to South Korean college students. They found that matching students on interest in education resulted in a reduction in the number of items displaying

DIF. Moreover, Arikan et al. (2018) used the PISA 2012 math items to investigate the effect of different PSM methods on DIF results and the achievement differences among Indonesian, Turkish, Australian, and Dutch students. They matched students on gender, opportunity to learn, and the index of economic, social and cultural status, and found that PSM reduced the number of items flagged for DIF and also diminished country differences.

3 Method

3.1 Data

The publicly available PISA data of 2018 were used in this study. PISA is an international assessment study that takes place every three years, measuring 15-year-old students' skills and knowledge in reading, mathematics and science. Approximately 710,000 students completed the test in 2018, representing over 31 million students in 79 participating countries and economies (OECD, 2019). PISA was computer-administered in the majority of the participating countries. However, for students in nine countries (Argentina, Jordan, Lebanon, Moldova, North Macedonia, Romania, Saudi Arabia, Ukraine and Vietnam), the test was paper-administered. Figure 2 presents a map of the world that shows which countries participated in PISA and the mode of test administration.

As PISA aims to measure a broad scope of knowledge in a limited amount of testing time, an integrated test design is used in which students only make a subset of the total number of items through multiple test versions, called booklets. Each booklet covers two domains and contains four clusters of items, with each cluster covering one domain and taking thirty minutes to complete. In addition to the test, students complete a 35-minute questionnaire in which information on the students themselves, their homes, and their school and learning experiences is gathered. As different booklets are used for PBA and CBA, booklets 1 and 13 were selected for this study. These were regular booklets with an identical fixed unit order for both modes. Moreover, they were administered to a large number of students, specifically 2303 and 20,288 students from 9 and 70 countries, for PBA and CBA respectively.

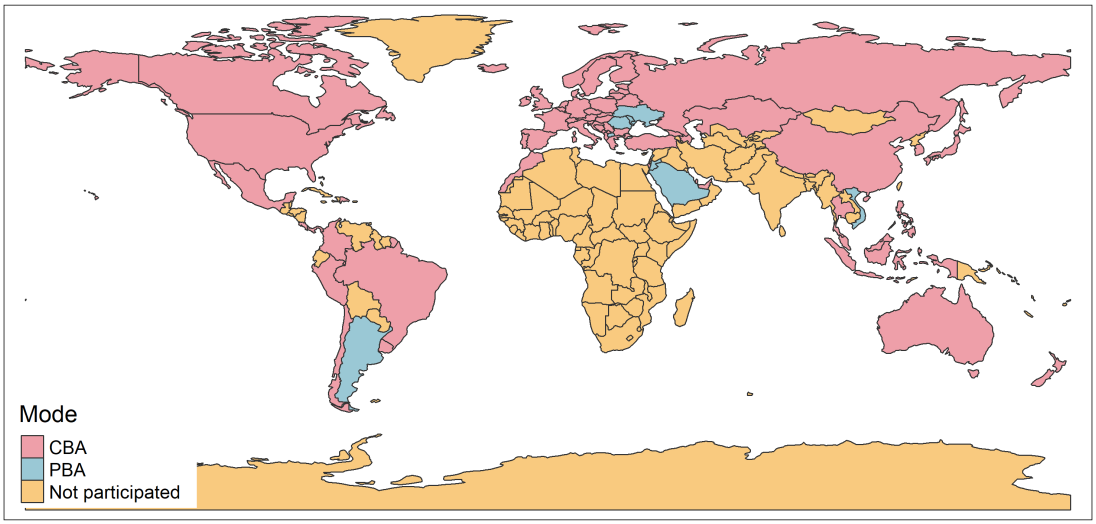

Figure 2: Map of the world showing the participating countries in PISA 2018 and whether the students made the test on a computer (CBA) or on paper (PBA).

3.2 Instruments

Science items Booklets 1 and 13 consist of the first two out of a total of six science clusters that cover 38 of the 115 available science items. Out of the 38 items, three were scored polytomously. All other items were dichotomously scored. Moreover, 13 items were open responses and the remaining 25 items were either simple or complex multiple-choice items. Additional information on the items can be found in Annex A of the PISA 2018 technical report.

Covariates Apart from test data, PISA collects an abundance of background information, which is stored in a questionnaire dataset. The index of economic, social and cultural status (ESCS) is probably the most important variable. This variable is constructed by PISA as a composite measure, including highest parental occupation, an index representing parental education, and an index of household possessions which functioned as a proxy for family income. Apart from ESCS, the other covariates included the effort the students have put in the PISA test, the effort they would have put in a graded test, the number of class periods in science, and gender.

3.3 Data Analysis

3.3.1 Pre-processing

First, the items of the first two science clusters in the selected booklets were extracted from the test data, as well as the country and student ID within the country. The items in the PBA and CBA booklets were combined and a dummy variable was constructed to indicate the mode of administration. Similarly, the questionnaire data was filtered, such that only the data of the selected booklets remained. Next, cases of unit nonresponse were deleted and item nonresponse cases were replaced with a 0 when the student left the question open.

The item responses were recoded into numeric variables where 0 equaled no credit and 1 equaled full credit for binary items and partial credit for polytomous items, for which a 2 equaled full credit. One item (*DS438Q03C*) was polytomously scored, while the second and third categories corresponded to full credit. This item was recoded such that the categories were collapsed into a single 1 full credit score.

3.3.2 Detecting DIF

After pre-processing, the first step was to detect DIF using the MIMIC model under the constrained-baseline strategy as described in Section 2.1. The students' science ability was the construct of interest and the dummy-coded mode-indicator served as the grouping variable. We evaluated the presence of DIF by comparing a model for each item against the baseline model with a χ^2-difference test. The item was subject to DIF if the item model fitted significantly better than the baseline model. Then, the significance of the β_i and ω_i parameters determined the DIF-type, respectively corresponding to uniform and nonuniform DIF. Herein, Bonferroni's correction criterion was applied, such that $p < 0.001$ was considered statistically significant.

3.3.3 Applying Propensity Score Matching

Once DIF was detected, the second step was applying PSM. Section 3.2 discussed the selected covariates. Propensity scores were estimated using logistic regression and matched using three methods: NNM, OPM and OFM (discussed in Section 2.2.2), which resulted in three matched subsamples. Subsequently, the performance of the methods was evaluated by visually comparing histograms and density plots before and after matching, as well as examining standardized mean

differences and variance ratios for each covariate. Then, the MIMIC model as described above was applied again with the matched subsamples as input data to investigate how matching affected the number of items displaying DIF.

3.3.4 Constructing Country League Tables

The last step was to construct country league tables, wherein we followed PISA's methodology (OECD, n.d.). The unidimensional IRT model for dichotomous responses and generalized partial credit model for polytomous responses were fitted to each of the three matched subsamples to estimate a set of item parameters for each matching method (three sets in total). Returning to the full sample, these item parameters were used to draw ten plausible values for each student per Rubin's rules. This resulted in three sets of ten plausible values for each student in the full sample. The plausible values were weighted to account for the survey design and pooled into an estimate of the science ability for each country. Ordering them for high to low resulted in a country league table for each matching method, which were compared to the table originally published by PISA for differences to inspect how the matching methods affected the ranking order.

3.3.5 Software

Due to the file size, the initial extraction of the test and questionnaire data corresponding to the selected booklets was done in SPSS (IBM Corp., 2020). All other analyses were performed in RStudio (RStudio Team, 2020), using the packages "lavaan" (Rosseel, 2012), "MatchIt" (Ho et al., 2011) and "cobalt" (Greifer, 2020). Accompanying syntax and code are available through Github.

4 Results

4.1 Example Table

	Before matching			Nearest Neighbor Matching		
Covariate	Computer	Paper	Difference	Computer	Paper	Difference
ESCS	0	0	0	0	0	0
Effort PISA	0	0	0	0	0	0
Effort graded	0	0	0	0	0	0
Science classes	0	0	0	0	0	0
Gender	0	0	0	0	0	0
	Optimal Pair Matching			Optimal Full Matching		
Covariate	Computer	Paper	Difference	Computer	Paper	Difference
ESCS	0	0	0	0	0	0
Effort PISA	0	0	0	0	0	0
Effort graded	0	0	0	0	0	0
Science classes	0	0	0	0	0	0
Gender	0	0	0	0	0	0

Table 1: Standardized means of matching variables for computer and paper students before and after matching, and their difference.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Araujo, L., Saltelli, A., & Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development*, *19*(1), 20–34. https://doi.org/10.1108/IJCED-12-2016-0023

Arikan, S., van de Vijver, F. J., & 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*, *4*(1), 69–81. https://doi.org/10.21890/ijres.382936

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*, 51. https://doi.org/10.3389/feduc.2017.00051

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and psychological measurement*, *80*(3), 476–498. https://doi.org/10.1177/0013164419878861

Coughlan, S. (2019). Pisa Tests: UK Rises in International School Rankings. *BBC News*, *3*. https://www.bbc.com/news/education-50563833

Fernandez-Cano, A. (2016). A methodological critique of the PISA evaluations. *Relieve*, *22*(1), 1–16. https://doi.org/10.7203/relieve.22.1.8806

Feskens, R., Fox, J.-P., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. *Theoretical and practical advances in computer-based educational measurement* (pp. 231–247). Springer, Cham. https://doi.org/10.1007/978-3-030-18480-3_12

Greifer, N. (2020). Cobalt: Covariate balance tables and plots. r package. version 4.0. 0. https://github.com/ngreifer/cobalt

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. https://doi.org/10.2307/1390693

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://www.jstatsoft.org/v42/i08/

IBM Corp. (2020). *IBM SPSS Statistics for Windows* (Version 27.0). Armonk, NY: IBM Corp. https://hadoop.apache.org

Joldersma, K., & Bowen, D. (2010). Application of propensity models in DIF studies to compensate for unequal ability distributions. *Annual Meeting of National Council on Measurement in Education, Denver, CO*.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *journal of the American Statistical Association*, *70*(351a), 631–639. https://doi.org/10.2307/2285946

Kolen, M. J., & Brennan, R. L. (2014). Practical issues in equating. *Test equating, scaling, and linking* (pp. 283–369). Springer, New York.

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. https://doi.org/10.1080/15305058.2014.922567

Lee, J., Little, T. D., & Preacher, K. J. (2018). Methodological issues in using structural equation models for testing differential item functioning. *Cross-cultural analysis* (pp. 65–94). Routledge, London.

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., Kroc, E., & Zumbo, B. D. (2020). Investigating the performance of propensity score approaches for differential item functioning analysis. *Journal of Modern Applied Statistical Methods*, *18*(1), 18. https://doi.org/10.22237/jmasm/1556669280

Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research, and Evaluation*, *21*(1), 13. https://doi.org/10.7275/ewqz-n963

OECD. (2014). *PISA 2012 technical report*. Paris: OECD Publishing. https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

OECD. (2017). *PISA 2015 technical report*. Paris: OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2019). *Pisa 2018 assessment and analytical framework*. https://doi.org/10.1787/b25efab8-e

OECD. (n.d.). *PISA 2018 technical report*. Unpublished. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Puhan, G., Boughton, K. A., & Kim, S. (2005). Evaluating the comparability of paper-and-pencil and computerized versions of a large-scale certification test. *ETS Research Report Series*, *2005*(2), i–15. https://doi.org/10.1002/j.2333-8504.2005.tb01998.x

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

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Sellar, S., & Lingard, B. (2013). Looking East: Shanghai, PISA 2009 and the reconstitution of reference societies in the global education policy field. *Comparative Education*, *49*(4), 464–485. https://doi.org/10.1080/03050068.2013.770943

Seo, D. G., & De Jong, G. (2015). Comparability of online-and paper-based tests in a statewide assessment program: using propensity score matching. *Journal of Educational Computing Research*, *52*(1), 88–113. https://doi.org/10.1177/0735633114568856

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1. https://doi.org/10.1214/09-STS313

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *Journal of psychopathology and behavioral assessment*, *31*(4), 320. https://doi.org/10.1007/s10862-008-9118-9

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. https://doi.org/10.1207/s15327574ijt0603_5

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.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8