Investigating Measurement Invariance in PISA 2012: Measures of Self-Efficacy and Performance in Mathematics

CSSM502

FALL 2022

Nilüfer Göktaş

49837

Introduction

Measurement invariance refers to the functional equivalence of a measure between different groups (Putnick & Bornstein, 2016). In other words, when a measure is invariant or equivalent, the construct intended to measure has the same meaning for different groups over time (Putnick & Bornstein, 2016).

Before conducting any group comparisons, testing measurement invariance is essential because using a measure that lacks invariance for conducting a group comparison would generate invalid results (Putnick & Bornstein, 2016). More specifically, when the measure is not functionally equivalent, the groups with the precisely same true ability would have different test scores.

In the present study, my aim is to investigate the measurement invariance in PISA 2012 by focusing on mathematics self-efficacy and mathematics performance of two countries: Turkey and the Netherlands. I chose to use the data from PISA because it is a large-scale testing of students from several countries regarding their performance and attitude in mathematics, science, and reading (OECD, 2012a). Since it consists of several measures and allows several group-level analyses across countries, testing measurement invariance is suitable with this data set. As a result of the present study, it is expected that both mathematical self-efficacy and performance have measurement invariance across two countries: Turkey and the Netherlands (e.g., Ding et al., 2022; Güngör & Atalay-Kabasakal, 2020).

Method

Participants

Data from the PISA 2012 Test was accessed from the official webpage of the Organization of Economic Cooperation and Development (OECD, 2012a). I focused on two countries for the testing of measurement invariance: Turkey and the Netherlands. The countries were chosen based on the criteria of being comparable in terms of sample size. For instance, Turkey sample consisted of 4848 participants (1.01% of the overall sample), whereas the Netherlands sample consisted of 4460 participants (0.93% of the overall sample).

After data cleaning (i.e., excluding participants with missing values), a total of 5902 participants were included in the multiple-group confirmatory factor analysis. More specifically, 3130 were from Turkey (49.5% female), and 2779 were from the Netherlands (48.4% female).

Measures

Mathematics Self-Efficacy

The 8-item Mathematics Self-Efficacy Scale aims to assess the confidence of individuals in solving mathematical problems (OECD, 2012a). The items are rated on a 4-point Likert scale (1 = *very confident*, 4 = *not at all confident*). For example, one of the items in the scale asks how confident the individual feels "calculating how much cheaper a TV would be after a 30% discount" (OECD, 2012).

Mathematics Performance

PISA Test aims to directly measure the performance of individuals in mathematics using various complex questions (For the sample items, please see OECD (2012b). Since it is large-scale testing from multiple countries, it is essential to consider the effect of the design on the individuals' scores (e.g., stratification) (OECD, 2012c). Therefore, PISA assigns five "plausible values" for mathematical performance, calculated based on Item Response Theory, for each individual (OECD, 2012c). In this study, instead of individual items, five plausible values for mathematical performance were used.

Analytical Procedure

Data preparation and cleaning were executed in Python (Van Rossum & Drake, 2009). PISA 2012 Data was downloaded from the official webpage of the Organization of

Economic Cooperation and Development (OECD, 2012a). The data file had 480174 observations/participants and 634 variables in total. However, since the file was in ".txt" format, all variables were nested within one row as a long string. Therefore, while I prepared the data frame containing self-efficacy and performance scores, I spliced the variables from particular positions in the long strings. For the positions of the variables, I consulted the "Codebook for PISA 2012 Student Questionnaire" (OECD, 2012a).

For each country (Turkey and the Netherlands), the variables of "identifier", "country", "gender", "age", and "stratum" were spliced. Each participant was assigned to a particular "education" group based on the stratum. Finally, the mathematics self-efficacy scale items and five plausible mathematics performance values were spliced. As the final step, the participants with missing values in the mathematics self-efficacy scale were excluded. In other words, only the participants who fully completed the mathematics self-efficacy scale were included the multiple-group confirmatory factor analysis (MGCFA) to test measurement invariance.

To test measurement invariance, I switched to R (2021) and used *lavaan* (Rosseel, 2012) and *semTools* (Jorgensen et al., 2022) packages. Firstly, confirmatory factor analyses were run for each measure to test a unidimensional model. As a result, the model for self-efficacy was modified for both countries, and the model for performance remained the same. Finally, MGCFA was run for each measure step by step: (1) Configural invariance, (2) Metric invariance, and (3) Scalar Invariance.

Results

Confirmatory Factor Analysis (CFA)

Mathematics Self-Efficacy

For each country (Turkey and the Netherlands), separate confirmatory factor analyses were run in order to test the unidimensional model of self-efficacy items.

The results of CFA for Turkey indicated CFI as .823, TLI as .752, and SRMR as .066. Since CFI and TLI should meet the criterion of >.90 for adequate model fit (Hu & Bentler, 1999), the unidimensional model of self-efficacy items was modified. After adding residual correlation between Item 5 (i.e., Solving an equation like 3x+5=17) and Item 7 (i.e., Solving an equation like 2(x+3)=(x+3)(x-3)), a confirmatory factor analysis was rerun to test the modified unidimensional model. As a result, CFI was .958; TLI was .938; and SRMR was .037. Since CFI and TLI were >.90, and SRMR was <.08 (Hu & Bentler, 1999), the modified model showed adequate fit.

For the Netherlands, CFA results showed CFI as .862, TLI as .807, and SRMR as .063. Since CFI and TLI should meet the criterion of >.90 for adequate model fit (Hu & Bentler, 1999), the unidimensional model of self-efficacy items was modified in the same way as Turkey. After adding residual correlation between Item 5 (i.e., Solving an equation like 3x+5=17) and Item 7 (i.e., Solving an equation like 2(x+3)=(x+3)(x-3)), a confirmatory factor analysis was rerun to test the modified unidimensional model. As a result, CFI was .970; TLI was .956; and SRMR was .028. Since CFI and TLI were >.90, and SRMR was <.08 (Hu & Bentler, 1999), the modified model showed adequate fit. For the testing of measurement invariance, the modified model of self-efficacy was used.

Mathematics Performance

Similar to mathematics self-efficacy, confirmatory factor analysis was run in order to test the unidimensional model of the mathematical performance separately for Turkey and the Netherlands.

For Turkey and the Netherlands, CFA results indicated CFI as 1.00, TLI as 1.00, and SRMR as .001, which may indicate a problem with the model.

Multiple-Group Confirmatory Factor Analysis (MGCFA)

Multiple-group confirmatory factor analysis (MGCFA) was run step by step (i.e., configural, metric, and scalar invariance, respectively) for each measure. The criteria for the goodness of fit statistics and delta values were as follows (Hu & Bentler, 1999; Putnick & Bornstein, 2016): CFI > .90, TLI > .90, SRMR < .08, RMSEA < .08; for metric invariance, ΔCFI < .02, ΔRMSEA < .03; for scalar invariance, ΔCFI < .01, ΔRMSEA < .01.

Mathematics Self-Efficacy

MGCFA results for the mathematics self-efficacy were summarized in Table 1. The mathematical self-efficacy measure showed configural and metric invariance since the goodness of fit statistics and delta values met the criteria above. However, at scalar invariance, since RMSEA was higher than .08, and ΔCFI and ΔRMSEA were higher than .01, the model did not show adequate fit. Overall, mathematics self-efficacy had weak measurement invariance.

Table 1. Multiple-Group Confirmatory Factor Analysis Results for Mathematics Self-Efficacy

	X^2	*df*	X^2 / *df*	SRMR	TLI	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	598.165	38	15.74	0.03	0.947	0.964	0.071		
Metric	656.277	45	14.58	0.037	0.951	0.961	0.068	0.003	0.003
Scalar	1119.781	52	21.53	0.051	0.927	0.932	0.083	0.029	-0.015

Mathematics Performance

The results from MGCFA for the mathematics performance were summarized in Table 2. Similar to the results of confirmatory factor analysis to test the unidimensional model, MGCFA results indicated CFI as 1.00, and TLI as 1.00. These values may indicate a problem with the model or the inappropriateness of plausible values (i.e., an overall performance score of the individual) for testing measurement invariance (i.e., an item-level analysis). Therefore, I think interpreting these results would decrease the statistical conclusion validity of the study.

Table 2. Multiple-Group Confirmatory Factor Analysis Results for Mathematics Performance

	X^2	*df*	X^2 / *df*	SRMR	TLI	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	10.125	10	1.01	0.001	1.00	1.00	0.002		
Metric	12.556	14	0.90	0.003	1.00	1.00	<0.0001	0	0.002
Scalar	16.519	18	0.92	0.004	1.00	1.00	<0.0001	0	0

Conclusion

In the present study, I investigated the measurement invariance in PISA 2012 Data by focusing on the mathematics performance and self-efficacy measures for Turkey and the Netherlands. To test measurement invariance, I conducted multiple-group confirmatory factor analysis (MGCFA). As a result, the mathematics self-efficacy measured showed weak equivalence between Turkey and the Netherlands. However, the results for mathematics performance may not seem statistically valid. One reason may be analyzing plausible values for performance, instead of focusing on the individual test items. Since the test of measurement invariance is an item-level analysis, interpreting an analysis conducted with plausible values may decrease statistical conclusion validity.

References

Ding, Y., Yang Hansen, K., & Klapp, A. (2022). Testing measurement invariance of mathematics self-concept and self-efficacy in Pisa using MGCFA and the alignment method. European Journal of Psychology of Education. https://doi.org/10.1007/s10212-022-00623-y

Güngör, M., & Atalay-Kabasakal, K. (2020). Investigation of measurement invariance of science motivation and self-efficacy model: PISA 2015 Turkey sample. International Journal of Assessment Tools in Education, 207–222. https://doi.org/10.21449/ijate.730481

Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Jorgensen, T. D., Pornprasertmanit, S. Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling. R package version 0.5-6. Retrieved from https://CRAN.R-project.org/package=semTools

OECD. (2012a). *Database - PISA 2012*. Programme for International Student Assessment. Retrieved January 21, 2023, from https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

OECD. (2012b). *PISA Test*. Programme for International Student Assessment. Retrieved January 21, 2023, from https://www.oecd.org/pisa/test/

OECD. (2012c). *PISA 2012 technical report*. Programme for International Student Assessment. Retrieved January 21, 2023, from https://search.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. URL https://www.R-project.org/

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The State of the art and Future Directions for Psychological Research. Developmental Review, 41, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2). https://doi.org/10.18637/jss.v048.i02

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.