ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

Invariance Levels across Language Versions of the PISA 2018 Reading Comprehension Tests in China, France, and the United States

Da Qi, Rushi Yu, Xiaofang Chen, Yi Cheng

School of International Studies, Zhejiang University, Hangzhou, 310058, China

Abstract

Since its emergence in 2000, the Program for International Student Assessment (PISA) has extended its influence to 79 countries. The original English and French language versions have been adapted to many other languages to facilitate PISA's application in different countries and regions. The comparisons between countries are valid only if the different language versions are equivalent to each other. Therefore, it is necessary to investigate whether PISA has construct comparability across its different language versions. The current study thus aimed to analyze the equivalence among three language versions (Chinese, French, and English) of the PISA 2018 Reading Comprehension Test to explore whether there exists item-level non-equivalence between the original versions and the adapted one. After defining the testlet as the unit of analysis, equivalence among the language versions was analyzed using two invariance testing procedures: multiple group mean and covariance structure analyses and ordinal logistic regression. The procedures yielded concordant results supporting metric equivalence across all three language versions. The equivalence thus supports the estimated reading literacy score comparability among the language versions used in China, France, and the United States.

Keywords

Program for International Student Assessment; Reading Comprehension Tests; Item-level Equivalence; Language Testing.

1. Introduction

With the rapid development of globalization, large-scale standardized international assessments have prevailed for the evaluation and comparison of the quality of education across different countries. As one of the most widely known ones, the Program for International Student Assessment (PISA) organized by the Organization for Economic Co-operation and Development (OECD) measures the reading, mathematics, and science abilities of 15-year-old students around the world [1]. Educational policymakers believe that based on these assessments, policy solutions to the shortcomings of their education system can be found. It should be admitted that information provided by these big data assessments has great potential to inspire policy improvement, while the limitations of the comparative results should also be considered. An innate feature of these international assessments is that they are multilingual and cross-cultural, since PISA, which was originally developed in English and French, has been adapted to different language versions in order to test students in different countries. Therefore, to ensure the functional equivalence of these assessments is an integral part of PISA study for the valid comparison among students and for the further utilization of PISA results in governments' educational policies and strategies.

Previous research has revealed that language can play an essential role in international language and literacy assessment measures for adolescent youth [2-6]. It is noteworthy that

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

although many studies have been conducted to examine the equivalence of PISA tests across cultures or languages, most of them are in the mathematics and science domains [5]. Among the several studies carried out in the reading domain, they mainly used the data before 2018, usually from PISA 2009 [5, 7].

Thus, to complement the previous investigations, the current study attempts to analyze the item equivalence levels among the three language versions (Chinese, French, and English) of the reading comprehension test in PISA 2018. In this sense, we can compare whether there exists differential item functioning (DIF) between the original versions (the English and French ones) and the adapted one (the Chinese one). Hence, two research questions are of our primary concerns:

1. Are there any invariances across different languages versions of the PISA 2018 reading comprehension tests in China, France, and the United States?

2. If the answer to Question 1 is 'yes', what are the invariance levels of each item in different languages?

2. Theoretical Basis

Equivalence refers to the use of two or more forms of a test that are alternate with each other. To be interchangeable, the forms must measure the same construct in the same way. Only when construct comparability is achieved through equivalent forms can the differences in test results across countries be regarded as the performance or ability differences between groups [8]. Thus, construct and score comparability is essential to the interpretation of test results for inferencing test validity.

In international assessments, there are multiple factors, such as linguistic and cultural ones, that may influence construct and score comparability. Cultural and linguistic contexts of the assessment may become the sources of incomparability with measures designed for one population and implemented to another. As they represent the culture or language use of those who designed it, they may lack the accurate applicability to the new group [9, 10]. In this case, construct comparability is threatened because the original group the test is designed for may have different degrees of familiarity with the test language. Furthermore, the effects of culture and language on testing were proved to be greater for tests with higher linguistic requirements, e.g., reading tasks.

Since the current study is oriented to the international reading literacy studies of PISA 2018, sources of incomparability are principally rooted in the different translation versions of the same test. While the texts developed in the original languages turn to translated forms, equivalence must be ensured, otherwise, the measurement of reading proficiency on students from other language backgrounds is biased. To investigate the construct comparability in such cases, examining the invariance or variance among the items of the tests given to students in Chinese, English, and French may contribute to the illustration of the adaptation effect and the problems related to test translation.

There are different approaches that can examine the degree of construct comparability at the item level, among which DIF distinguishes itself. DIF can indicate that the relations between test items may vary across languages or cultures, due to poor item translations, or to specific cultural or linguistic elements [11]. Many previous studies have also applied the methods used in DIF to examine the item-level equivalence or non-equivalence between different versions of the same test. Huang et al. (2016) conducted a study on the cross-language, cross-cultural validity of the PISA 2006 Science assessment through three DIF analyses between the USA and Canada, Chinese Hong Kong and mainland China, and between the USA and mainland China [12]. They found that DIF was serious between US and Mainland Chinese students, but was minimal between English-speaking Canadian students and their US peers. Le (2009) used the PISA cycle

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

3 field trial data to investigate the relationships between gender DIF across countries and test languages for science items and their formats [13]. He found that for each of the test language groups, 5.6% and 2.8% of the items were flagged as substantially favoring males and females respectively. Thus, the testing languages have a significant effect on gender DIF.

3. Methods and Materials

3.1. Participants

As the major domain of assessment in PISA 2018, reading comprehension tests were administered to all the students participating in PISA 2018. The participants chosen in this study were all the students taking the PISA 2018 reading comprehension tests from China (Beijing-Shanghai-Jiangsu-Zhejiang) [henceforth China (B-S-J-Z)], France, and the United States, which respectively included 12,058, 6,308, and 4,838 fifteen-year-old students that were near the end of their compulsory education. The students in the three countries all took the test in the versions of their native languages, i.e., Chinese, French, and English.

3.2. Instrument

PISA has two types of assessment, namely a computer-based assessment (CBA) in most participating countries and a paper-based assessment (PBA) in only a few countries, with different items in them. In China, France, and the United States, PISA 2018 was administered as a CBA with the same items in different language versions. In PISA 2018, multistage adaptive testing (MSAT) was introduced in reading tests to measure the higher and lower ends of students' ability more accurately [1]. The students in the three countries took three stages of tests in succession, namely Core stage, Stage 1, and Stage 2, which consist of 50 units with 249 items. The items given to each student were dynamically determined on the basis of students' performance on the tests in prior stages. Specifically, students first completed 7 to 10 items in the short Core stage, wherein at least 7 items were automatically scored. Scores at this stage affected whether students took comparatively easy or difficult items at Stage 1, and responses to the automatically scored items from both the Core stage and Stage 1 were used to determine the item difficulty students took at Stage 2.

3.3. Procedure and Data Analysis

Analysis of DIF was conducted in this study to assess the validity of cross-lingual comparisons. The specific procedures are as follows.

First, descriptive statistics on the overall situation of each language group (Chinese, French, English) were used to determine the mean and standard deviation of the total reading scores of each group. Next, a Kruskal-Wallis one-way analysis of variances (ANOVA) was applied to find whether there were significant differences among the scores of the three language groups since the data did not normally distribute.

Second, local independence among items was evaluated in order to check the local independence assumption and to define the unit of analysis. First, the local independence assumption was checked because the presence of groups of items related to a single content area could yield misleading results in the application of psychometric models [14]. Specifically, a chi-square independence test was performed through a 2×2 contingency table for the items in all reading units. Second, if most of the items within different reading units could pass the independence test, testlets, a set of items that are analyzed as a unit, would be defined based on the units given by the designers of PISA 2018 reading comprehension tests [15-17]. It should be noted that following the methods given by Oliden and Lizaso (2013), the 22 open-response items (8.8% of the total 249 items) that were coded on scores ranging from 0 to 2 were dichotomized before forming the testlets to give the same weight to all the items, by assigning a 1 to the 2-point scores, and a 0 to the 0- and 1-point score [7]. In this step, all the missing

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

values in the dataset were imputed with the method of k-nearest neighbor model (KNN), which was realized by the yaImpute package in R [18].

Third, two DIF detection procedures, namely, Mean and Covariance Structure Analysis (MACS) and Ordinal Logistic Regression (OLR) were carried out to assess the item level equivalence among the three language versions. The two methods were both adopted because they were complementary to each other, with MACS having strong assumptions difficult to meet, and the OLR less restrictive but the parameter values more difficult to interpret than MACS [19]. In this step, MACS evaluates factorial invariance based on the linear factor model to determine whether the same measurement model fits across samples, while OLR assesses the effect of the grouping variable (language) and the interaction of language and reading literacy through the application of different regression models to each of the testlets. In OLR, the Chinese version of the reading tests was used as the reference sample compared with the other two language versions in pairs. The dependent ordinal variable was the scores obtained in the testlet, and the predictor variable was the eight levels of reading literacy as measured by PISA [20]. The details of the eight levels are shown in Table 1. It should be noted that there are 9 cases in the three language samples that got scores lower than the 1c level. To include them into the OLR models, we added them as the lowest level in the reading fluency scales and recoded all the 9 levels as Levels 1 to 9 to facilitate the OLR analysis.

Table 1. Summary description of the eight levels of reading proficiency in PISA 2018.

Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average)
6	698	1.30%
5	626	8.70%
4	553	27.60%
3	480	53.60%
2	407	77.40%
1a	335	92.30%
1b	262	98.60%
1c	189	99.90%

For each testlet, a baseline model with only one independent predictor was assessed, and then two more parameters were added and assessed, i.e., language and the interaction between language and reading competency. DIF is concluded if the chi-square value is significant and the R2 difference is great enough. According to Jodoin and Gierl (2001), a cutoff value of .07 indicates a severe lack of invariance, and .03 moderate differential functioning [21].

Last, a progressive measurement of invariance test for all data was conducted using confirmatory factor analysis (CFA) with Mplus-8 [22]. Basically, multiple sets of CFA were conducted to establish baseline unidimensional models and to estimate the reliability of the scores. Then various levels of invariance, namely configural invariance, metric invariance, and scalar invariance, were examined successively and jointly across the three language groups [23, 24]. Configural invariance (equality of factor pattern matrices) was the simplest model, based on which metric invariance (equality of the loadings) and scale invariance (equality of the intercepts) were assessed by adding constraints to the configural invariance. The statistical significance of the likelihood ratio test ($p < .01$) and the changes in CFI values were used simultaneously as the criteria to compare the nested models [25].

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

4. Results

4.1. Descriptive Statistics

The highest average reading scores were attained by the students from China (B-S-J-Z) who completed the test in Chinese ($M_{reading}$ = 561.03, SD = 90.34). The lowest average reading scores were found among the students from France who took the test in French ($M_{reading}$= 484.27, SD = 105.40). We then applied the Kolmogorov-Smirnov test to explore whether the reading scores of the students followed the normal distribution. The results showed that the scores of the students in the three countries all failed to show a normal distribution ($p = .000 < .05$). Therefore, to accurately assess the statistical significance of differences within the reading scores in the three language versions, we employed the Kruskal-Wallis one-way ANOVA. The results demonstrated that the hypothesis of equality of the competency means related to testing language cannot be accepted [$F_{reading}$ (2, 23203) = 1490.95, $p = .000 < .001$; partial $\eta^2 = 0.114$]. According to the results of pairwise comparisons, the reading scores in all of the three countries are significantly different from each other. The box plots in Figure 1 further displayed that the reading scores obtained by Chinese students are much higher than those got by the American and French students. In this sense, the reading scores of the two original versions (the English and French ones) of PISA 2018 are in stark contrast with those of the translated version (the Chinese one). It is thus necessary to examine whether such a difference was caused by DIF so as to provide the educational policymakers with more valid test results for the improvement of the education system. In the later sections of Results, we would investigate in depth the item level equivalence between the three language versions.

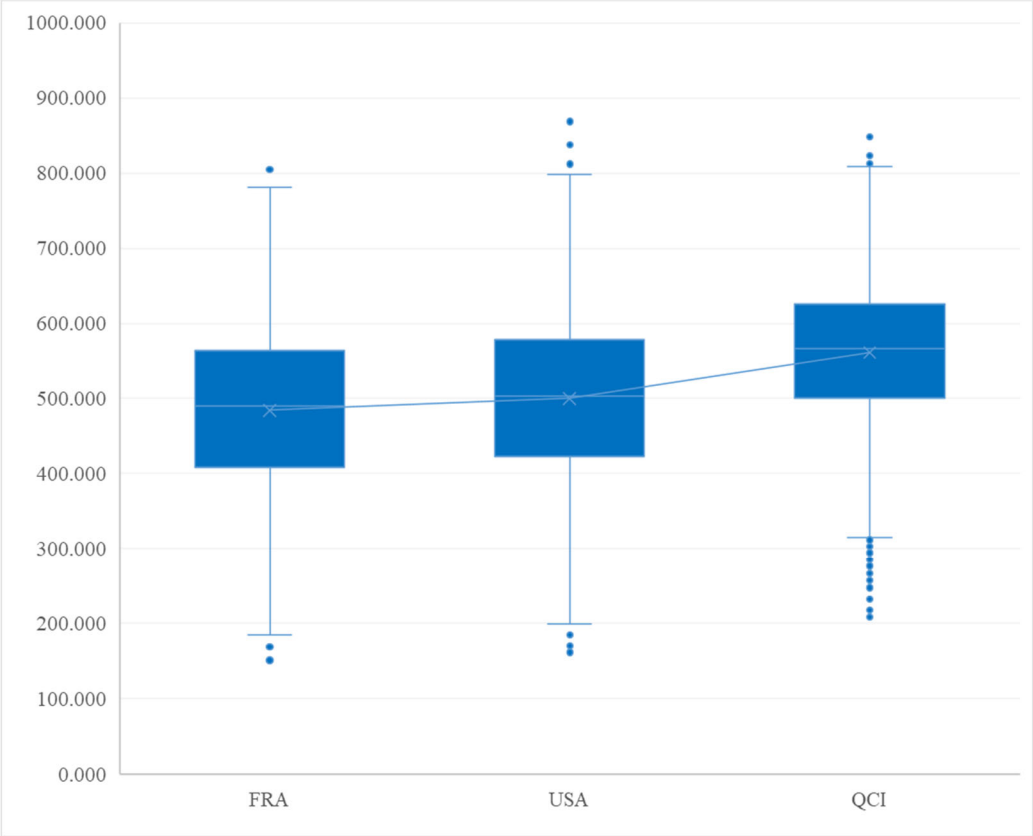

Figure 1. The reading score distributions in China (B-S-J-Z), France, and the United States. FRA is referred to as France, USA is the United States, and QCI is China (B-S-J-Z).

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

Table 2. Pairwise comparisons of the reading scores in the three countries.

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
FRA-USA	-943.653	128.015	-7.371	0.000*	0.000
FRA-QCI	-4768.868	104.089	-45.815	0.000*	0.000
USA-QCI	3825.214	114.000	33.555	0.000*	0.000

*p < .01

4.2. Local Independence

Before diving into the analysis of item-level equivalence, we first examined whether the items all obeyed the local item independence to make sure that the design of the reading units (henceforth testlets) are reasonable and the scores obtained can authentically show the reading ability of the students. Local item independence was examined using 556 two-way contingency tables. The hypothesis of local independence was accepted in 98.20% (546 out of 556) of the cases ($p < .01$). The specific cases that did not pass the independence test are shown in Table 3. Since the items that did not obey local independence were small in number (only 1.8% of all the cases), we carried out further analysis based on the current division of testlets. The number of items in each testlet ranged from 2 to 8. Two of the testlets contained two items, ten testlets had three items, nine contained four items, ten had five items, seven contained six items, ten had seven items, and one testlet contained eight items.

Table 3. The testlets that have dependent item pairs.

Testlet	Number of two-way contingency tables	Number of dependent item pairs	Dependent item pairs
Alfred Nobel	21	1	CR543Q04S & DR543Q15C
Chocolate and Health	6	1	DR455Q02C & DR455Q03C
Cliff Palace	15	1	CR550Q06S & DR550Q07C
FestiRock	21	1	CR552Q09S & DR552Q03C
Microlending	21	1	CR556Q04S & CR556Q10S
Narcissus	3	1	CR437Q01S & DR437Q07C
Opening Night	10	1	DR562Q06C & CR562Q07S
Rapa Nui	21	2	DR551Q11C & CR551Q08S DR551Q05C & CR551Q08S
Work Right	3	1	CR466Q03S & CR466Q06S

4.3. Unidimensionality and Reliability

The Internal consistency was tested using Cronbach's alpha coefficient. The goodness-of-fit indexes (CFI) for the Chinese (CFI = .941), English (CFI = .969) and French (CFI = .955) samples were all greater than .9. The RMSEA values were optimal across all groups; none of them exceeded the cutoff point of .06 [26]. As shown in Table 4, the internal consistency coefficients were greater than .9 in the three samples assessed.

From the results in Sections 4.2 and 4.3, we can see that except for the reading ability of the students, there seems to exist no internal factors that make the results of PISA 2018 reading comprehension tests non-reliable and invalid. We then analyzed whether the items demonstrated significant differences by conducting OLR and MACS analyses.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

Table 4. Descriptive statistics, unidimensionality and internal consistence.

Group	N	M	SD	χ^2	df	CFI	RMSEA	Cronbach's α
Chinese	12,058	561.03	90.34	2053*	269	.941	.031	.954
English	4,838	500.15	108.4	985*	185	.969	.021	.970
French	6,308	484.27	105.40	1167*	191	.955	.026	.968

* significant values $p < .01$

4.4. Ordinal Logistic Regression

Logistic regression models were estimated for each of the 50 testlets, with the Chinese reference sample (the translated one) compared with the English and French groups (the original ones). As shown in Table 5, although the chi-square values obtained for some of the comparisons were significant, i.e., 27 out of 100, the effect size associated with the language did not reach the preset limit ($R^2_{Mod2\text{-}Mod1} = 0.07$) in any of the comparisons. Thus, from the results of OLR, the PISA reading comprehension tests do not bias towards a certain linguistic group of people.

Table 5. The results of ordinal logistic regression.

Testlet	Chinese/English		Chinese/French		Testlet	Chinese/English		Chinese/French	
	$G^2_{Mod2\text{-}Mod1}$	$R^2_{Mod2\text{-}Mod1}$	$G^2_{Mod2\text{-}Mod1}$	$R^2_{Mod2\text{-}Mod1}$		$G^2_{Mod2\text{-}Mod1}$	$R^2_{Mod2\text{-}Mod1}$	$G^2_{Mod2\text{-}Mod1}$	$R^2_{Mod2\text{-}Mod1}$
Machu Picchu	11.214*	0.016	8.792*	0.017	The Cleanup	0.115	0.034	0.312	0.017
South Pole	18.448*	0.008	2.792	0.007	Aesop	0.267	0.021	8.490*	0.021
Great Pacific Garbage Patch	0.372	0.009	6.806*	0.011	Teen Health Forum	0.119	0.023	1.057	0.022
Fair Trade	19.551*	0.006	0.488	0.008	Biscuits	0.074	0.029	0.001	0.026
Nalini Nadkarni	7.578	0.005	17.717*	0.006	Job Vacancy	0.365	0.028	0.019	0.024
Sleep	11.690*	0.023	0.850	0.029	Summer Job	0.084	0.022	0.006	0.022
Sitting Disease	0.545	0.018	0.004	0.021	FestiRock	0.193	0.003	0.049	0.003
Book Survey	6.129	0.023	2.194	0.027	Microlending	0.073	0.003	1.346	0.005
Microwave Ovens	4.202	0.019	4.834	0.022	Space Debris	3.871	0.004	0.004	0.005
Narcissus	7.686*	0.027	2.712	0.027	Plastic	0.001	0.005	0.265	0.005
Chocolate and Health	0.000	0.025	10.061*	0.030	Alfred Nobel	31.464*	0.040	12.091*	0.044
Building a Legend	0.170	0.022	3.662	0.021	Nikola Tesla	0.429	0.045	8.126*	0.046
Bulletin Board	4.472	0.024	5.425	0.024	Question of the Week	8.785*	0.061	0.010	0.054
Olympic Flag	0.890	0.025	4.807	0.027	The Favour	0.029	0.056	0.684	0.051
Sebastiao Salgado	1.993	0.022	5.969	0.020	Rapa Nui	4.155	0.043	93.254*	0.050
Message in a Bottle	1.472	0.020	0.633	0.020	The Skellig Rocks	14.689*	0.040	48.882	0.043
Drugged Spiders	1.150	0.019	0.312	0.017	Smoke Jumpers	0.426	0.061	0.201	0.051
Exchange	2.736	0.029	22.992*	0.040	Employment	2.131	0.051	9.255*	0.037
Work Right	0.265	0.023	0.078	0.029	Gulf of Mexico	0.481	0.048	20.568*	0.039
World Languages	15.475*	0.019	14.442*	0.023	The Portrait	0.545	0.005	13.378*	0.007
Telephone	3.995	0.031	7.472	0.037	Optician	5.689	0.005	1.175	0.008
Making News Travel	0.140	0.020	1.498	0.026	Kokeshi Dolls	9.984*	0.007	18.011*	0.012
Cliff Palace	5.715	0.018	5.550	0.022	Literary Magazine	0.870	0.003	1.902	0.003
Opening Night	2.496	0.019	13.39	0.020	Shirts	8.431*	0.006	23.853*	0.008
Children's Futures	0.047	0.022	2.739	0.020	About a book	7.833*	0.009	12.244*	0.010

* $p < 0.01$

4.5. Multiple Group Mean and Covariance Structure

In the analysis of MACS, we adopted the progressive assessment of invariance, which began with the configural invariance model and further developed with more restrictions added on that model. After processing the data, we found that the goodness-of-fit values (CFI = .925; RMSEA = .033) supported the baseline invariance model. With restrictions added on the regression coefficients, the data was tested against the metric invariance hypothesis. Although the difference in chi-square values between the configural and metric models was statistically significant, $\chi^2(98) = 16900$, $p < .001$, the CFI did not change substantially ($\Delta CFI = .005 < .01$). Next, the scale invariance was assessed by placing restrictions on the response thresholds. The difference in chi-square values between this model and the previous one was significant, $\chi^2(98) = 3854$, $p < .001$; However, the CFI value showed that the differences across the three versions were scale-invariant.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

Table 6. The results of progressive assessment of factorial invariance.

Model	Goodness-of-fit indexes				Difference test	
	χ^2	df	CFI	RMSEA	χ^2	df
Configural invariance	875852*	3525	.925	.031		
Metric invariance	892752*	3623	.930	.028	16900*	98
Scale invariance	896606*	3721	.937	.028	3854*	98

* $p < .01$

In a brief summary, with the two methods employed in the current study, no item-level DIF was found between the Chinese, French, and English versions of the PISA 2018 reading comprehension tests. Thus, the two research questions raised in the Introduction can be answered by the OLR and MACS results, which suggest that PISA 2018 ensured considerable reliability and validity by accountably adapting the reading items and texts to different languages.

5. Discussions

With a large number of countries participating in PISA, the diversity of the samples has posed challenges to the refinement of the content and instruments to better compare students from different regions or countries with various cultural and linguistic backgrounds. To find out whether PISA can accurately and fairly compare different groups of people, this study attempted to investigate one of the basic hypotheses underpinning the comparability of PISA results: item-level equivalence, which concerns whether there exists DIF between different linguistic groups. Given that there have been few studies examining the differences between PISA's Chinese version and its two original versions – the English and the French ones, the purpose of this work was to evaluate the equivalence among the three language versions used in the 2018 edition of PISA to assess reading literacy.

The reference sample in this study was the group that completed the test in Chinese, which was compared with the groups consisting of students who took the test in the English and French language versions. The reading comprehension tests in PISA 2018 adopted different testlets, each of which consists of a set of items. Hence, it is necessary to first assess local item independence and internal consistency to exclude the possibility that the differences between the scores in different groups were due to internal factors. After the hypothesis of local independence and internal consistency was accepted, the items designed for each of the 50 testlets in the reading comprehension tests were then converted to 50 polytomous variables.

Two methods to assess invariance were applied, ordinal logistic regression and multiple-group mean and covariance structure models. By using more than one procedure, cross information can be gathered to support the results obtained. Ordinal logistic regression was applied to pairs, using the Chinese language as the reference sample. Equivalence across the three versions could be assessed simultaneously with multiple-group mean and covariance structure models, offering information for all possible comparisons. This characteristic extends the generalization of results to inter-linguistic comparisons. The results obtained using both procedures were congruent and positive, supporting the hypothesis of estimated reading literacy score comparability among the Chinese, English, and French language versions, and between the translated version (the Chinese one) and the original versions (the English and French ones).

The complexity and linguistic wealth attached to different social environments make the testing language a variable to be controlled in every educational assessment process. The adaptation of tests and the verification of equivalence means that a check must be performed to ensure that no bias can invalidate comparisons between scores obtained in different language versions of the same test. If the internal structure of the tests was not equivalent in the different language

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

groups, students with the same level of competence would obtain different scores. This would lead to erroneous conclusions in studies based on the hypothesis of equivalence between scores.

The importance of the cross-linguistic study is clear. Among other aspects, the countries differ in terms of gross domestic product, spending on education, language, culture, and even philosophy in education. In this differential context, comparisons associated with the PISA results or any other educational assessment project are only valid if no bias is present in the instrument used. Currently, few studies have been conducted to explore whether the Chinese version is comparable with the versions of the languages in other language families, e.g., the Indo-European family. Huang et al.'s (2016) study on the cross-language, cross-cultural validity of the PISA 2006 Science assessment through three DIF analyses showed that DIF was serious between U.S. and Mainland Chinese students, but was minimal between English-speaking Canadian student and their U.S. peers [12].

The study carried out by Huang (2010) also showed that the number of DIF items is the smallest between Canadian and U.S. students and the largest between U.S. and Chinese students [27]. However, their content analysis revealed that language difference only accounted for a small proportion of DIF between U.S. and Chinese students, whereas differential curriculum coverage was found to be the most serious cause of DIF in both the Hong Kong-Mainland and the U.S.-Chinese comparisons. In addition, they found that differential content familiarity is also a potential cause of DIF. Substantive studies such as those cited here are both important and basic for improving the education system; nevertheless, they depend on the measurement equivalence, which is a condition that must be evaluated.

With so few studies such as those carried out in the present study, any project comparing school effectiveness among countries with different languages might be negatively affected; the effect would be even more extreme in the study of the countries where more than one language is spoken and different language versions of the same test are used. PISA 2018 was also administered in different languages in certain countries such as Spain and Sweden. In the presence of bias, if we want to further explore the construct comparability between these bilingual or multilingual countries, the comparisons carried out in each of them could be called into question if statistical procedures are not used to adjust for differences between scores.

Given these circumstances, it is important to have studies, such as the one presented in this study, which provide an in-depth analysis of the psychometric structure of the tests used. This kind of work delves into the origin of the differences and experts to develop measurement instruments that meet the conditions required by the goals of any assessment project.

6. Conclusions and Implications

Using the methods for analyzing DIF, the present study examined the degree of construct comparability at the item level. We found that there was no significant difference across different languages versions of the PISA 2018 reading comprehension tests in China, France, and the United States. It is based on the responsible translation that the validity of test score comparisons across countries can be ensured, which makes possible the further utilization of PISA reading results in governments' educational policies and strategies. Since there is no doubt that such policies cannot be grounded in poor-quality and non-equivalent translations, the current study contributed to the administrations' scientific decision-making process by confirming the validity of the PISA 2018 reading comprehension tests.

However, it is worth noting that the Chinese students participating in PISA 2018 were all from Beijing, Shanghai, Jiangsu, and Zhejiang, all of which enjoy much higher levels of economic development compared with other regions in China. In this regard, the conclusion drawn in the current study might only be applicable in the comparison between the well-developed regions in China with France and the United States. We hope that the educational policymakers would

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

take into consideration such a factor when adjusting the current educational strategies. Moreover, involving students from other regions of China could be of great help to accurately demonstrate the regional differences across China and better explore whether the item level equivalence found in the present study still holds when those students from under-developed areas participate in PISA.

More than anything, however, the current study indicated the necessity of conducting more research on the translation of international assessments to examine the score comparability across the tests administered in different languages. Further investigations may take into account more language versions and enlarge the sample size in statistical analysis to make the results more solid and comprehensive. Moreover, apart from the quantitative methods adopted in this study, qualitative approaches like expert coding of the translated texts and interviews of translators can also be carried out to make a more accurate evaluation of construct comparability [28]. We hope that the questions briefly touched on in this study will be examined in more depth in the future.

References

[1] OECD: PISA 2018 Results. Information on https://www.oecd.org/pisa/publications/pisa-2018-results.htm

[2] R. K. Hambleton: Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, C. D. Spielberger (Eds.): Adapting Educational and Psychological Tests for Cross-cultural Assessment (Lawrence Erlbaum, the United States 2005), p. 3-38.

[3] K. Ercikan, J. Lyons-Thomas: Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, M. C. Rodriguez (Eds.): APA Handbooks in Psychology®. APA Handbook of Testing and Assessment in Psychology, Vol. 3. Testing and Assessment in School Psychology and Education (American Psychological Association, the United States 2013), p. 545–569.

[4] T. Sukin, S. Sireci, S. L. Ong: Using Bilingual Examinees to Evaluate the Comparability of Test Structure across Different Language Versions of a Mathematics Exam. Actualidades en Psicología. Vol. 29 (2015) No. 119, p. 131-139.

[5] M. Asil, G. T. Brown: Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of Non-invariance. International Journal of Testing. Vol. 16 (2016) No. 1, p. 71-93.

[6] P. Smith, P. Frazier, J. Lee, R. Chang: Incongruence between Native and Test Administration Languages: Towards Equal Opportunity in International Literacy Assessment. International Journal of Testing. Vol. 18 (2018) No. 3, p. 276-296.

[7] P. E. Oliden, G. M. Lizaso: Invariance Levels across Language Versions of the PISA 2009 Reading Comprehension Tests in Spain. Psicothema. Vol. 25 (2013) No. 3, p. 390-395.

[8] R. K. Hambleton, P. F. Merenda, C. D. Spielberger: Adapting Educational and Psychological Tests for Cross-cultural Assessment (Lawrence Erlbaum, the United States 2005).

[9] H. Landrine, E. A. Klonoff: The African American Acculturation Scale II: Cross-validation and Short Form. Journal of Black Psychology. Vol. 21(1995) No. 2, p. 124-152.

[10] M. E. Oliveri, K. Ercikan: Do Different Approaches to Examining Construct Comparability in Multilanguage Assessments Lead to Similar Conclusions? Applied Measurement in Education. Vol. 24 (2011) No. 4, p. 349-366.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202112_4(12).0073

[11] R. K. Hambleton, A. L. Zenisky: Translating and Adapting Tests for Cross-cultural Assessments. In D. Matsumoto, F.J.R. van de Vijver (Eds.): Cross-cultural Research Methods in Psychology. (Cambridge University Press, the United States 2010), p. 46-70.

[12] X. Huang, M. Wilson, L. Wang: Exploring Plausible Causes of Differential Item Functioning in the PISA Science Assessment: Language, Curriculum or Culture. Educational Psychology (Dorchester-on-Thames). Vol. 36 (2016) No. 2, p. 378-390.

[13] L. T. Le: Investigating Gender Differential Item Functioning across Countries and Test Languages for PISA Science Items. International Journal of Testing. Vol. 9 (2009) No. 2, p. 122-133.

[14] H. Wainer, R. Lukhele: How Reliable are TOEFL Scores? Educational & Psychological Measurement. Vol. 57 (1997), p. 741-758.

[15] H. Wainer, G. L. Kiely: Item Clusters and Computerized Adaptive Testing: A Case for Testlets. Journal of Educational measurement. Vol. 24 (1987) No. 3, p. 185-201.

[16] H. Wainer, C. Lewis: Toward a Psychometric for Testlest. Journal of Educational Measurement. Vol. 27 (1990), p. 1-14.

[17] H. Wainer, S. G. Sireci, D. Thissen: Differential Testlet Functioning: Definitions and Detection. Journal of Educational Measurement. Vol. 28 (1991), p. 197-219.

[18] N. L. Crookston, A. O. Finley: yaImpute: An R Package for KNN Imputation. Journal of Statistical Software. Vol. 23 (2008) No. 10, p. 1-16.

[19] P. Elosua, C. S. Wells: Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. Psicológica. Vol. 34 (2013), p. 327-342.

[20] I. Peña-López: PISA 2018 Results. What Students Know and Can Do. Information on https://www.oecd-ilibrary.org/docserver/5f07c754-en.pdf?expires=1625039194&id=id&accname=oid008303&checksum=1A0CFEAC3973113D04132576E4AF4AD3

[21] M.G. Jodoin, M.J. Gierl: Evaluating Type I Error and Power Rates Using an Effect Size Measure with Logistic Regression Procedure for DIF Detection. Applied Measurement in Education. Vol. 14 (2001), p. 329-349.

[22] L.K. Muthén, B.O. Muthén: Mplus User's Guide. Eighth Edition (Muthén & Muthén, the United States 2017).

[23] B. M. Byrne: Testing for Multigroup Equivalence of a Measuring Instrument: A Walk through the Process. Psicothema. Vol. 20 (2008), p. 872-882.

[24] P. Elosua, J. Muñiz: Exploring the Factorial Structure of the Self-concept: A Sequential Approach Using CFA, MIMIC and MACS Models, across Gender and Two Languages. European Psychologist. Vol. 15 (2010), p. 58-67.

[25] G. W. Cheung, R. B. Rensvold: Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method. Journal of Management. Vol. 25 (1999), p. 1-27.

[26] L. T. Hu, P. M. Bentler: Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. Structural Equation Modeling: A Multidisciplinary Journal. Vol. 6 (1999) No.1, p. 1-55.

[27] X. Huang: Differential Item Functioning: The Consequence of Language, Curriculum, or Culture? (Ph.D., University of California, Berkeley, the United States 2010).

[28] I. Arffman: Equivalence of Translations in International Reading Literacy Studies. Scandinavian Journal of Educational Research. Vol. 54 (2010) No.1, p. 37-59.