Psychological Test and Assessment Modeling, Volume 64, 2022 (2), 158-185

An Apple in These Countries is an Orange in Those: Differential Item Functioning of Items in the PISA 2018 Home Possessions Scale

Selene S. Lee[1], Lale Khorramdel[2], & Kentaro Yamamoto[3]

Abstract

In the Programme for International Student Assessment (PISA), students' socio-economic status (SES) is measured with parents' education, parents' occupation, and home possessions. An important assumption in comparative research using these SES scores is that they are comparable across countries, an assumption which needs to be tested. This study focuses on the home possessions (HOMEPOS) scale, a component of PISA's SES scale, and finds that some items in the scale function differently across countries when measuring family wealth – in other words, they exhibit differential item functioning (DIF). The study also finds that there are associations between DIF and a country's level of economic development, geographic location, and socio-cultural characteristics of a group (such as language and religion). This paper provides several recommendations that can potentially improve the comparability of the HOMEPOS scale across countries in the future, and by extension, the cross-country comparability of the SES scale of PISA. This is an important contribution, considering the increasing diversity of countries that are participating in PISA as well as the increasing focus on educational equity in many countries.

Keywords: differential item functioning (DIF), home possessions (HOMEPOS) scale, socioeconomic status (SES), Programme for International Student Assessment (PISA), international large-scale assessment (ILSA)

[1] Educational Testing Service (Princeton, NJ, USA), Correspondence concerning this article should be addressed to: Selene S. Lee, PhD, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA; email: slee003@ets.org

[2] Boston College (Chestnut Hill, MA, USA), The main work on this paper was conducted while the author was working at ETS.

[3] Kyamamo10 (Lawrenceville, NJ, USA), The main work on this paper was conducted while the author was working at ETS.

Introduction

In the Programme for International Student Assessment (PISA), an international large-scale assessment (ILSA) administered every three years, students' socio-economic status (SES)[1] is measured through parents' education, parents' occupation, and home possessions. The rationale for selecting these three variables is that SES has been traditionally measured with education, occupational status, and income (Organization for Economic Co-operation and Development [OECD], 2020), using the framework proposed by Duncan, Featherman, and Duncan in 1972 (Sirin, 2005). This framework for measuring SES in PISA has been used ever since the OECD first launched PISA in 2000, when two-thirds of the participating countries[2] were members of the OECD (OECD, n.d.-a). In every subsequent cycle of PISA, the participating countries have become more diverse, with non-OECD members accounting for more than half of the countries that participated in PISA 2018. This study investigates the extent of the cross-country comparability of the home possessions (HOMEPOS) scale, a component of PISA's SES scale, and also provides recommendations that can potentially improve the comparability of this scale across countries in future rounds of PISA. These recommendations are timely, as the OECD has acknowledged the need to revise the SES scale in response to the increasing diversity of the participating countries (OECD, 2019).

SES and Educational Equity

The goal of achieving social equity in education, which means that all students have equal learning opportunities and that differences in their educational outcomes are not related to their socio-economic background (OECD, 2018), has been advocated in some countries for over a century (Massachusetts Board of Education, 1849). The importance of achieving educational equity is also being recognized worldwide – it is clearly highlighted in the Sustainable Development Goals (SDGs), adopted by the United Nations in 2015, the fourth goal of which is to "ensure inclusive and equitable quality education (United Nations, 2015, p. 17)." The Education 2030 Framework for Action, the framework which outlines the strategies for achieving SDG Goal 4, also states that "inclusion and equity in and through education is the cornerstone of a transformative education agenda (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2016, p. 7)." In spite of these efforts, SES is still an important determinant of students' educational achievement in many countries, as shown in numerous studies (Broer et al., 2019; Coleman, 1968; Hattie, 2008; Sirin, 2005).

[1] In PISA, the SES scale is called the economic, social and cultural status (ESCS) scale.

[2] In this paper, "countries" refer to "countries and economies."

With such a strong interest among educational practitioners and researchers to establish policies that can promote educational equity, many have turned to ILSAs for insights from countries that have made progress in achieving this goal. For example, a report by the OECD used data from PISA 2006 to analyze the profile of resilient students (defined as students who achieved high scores on PISA in spite of their disadvantaged socio-economic background) and suggested that increasing class time for students with low SES and boosting their self-confidence could help them improve their academic performance (OECD, 2011). Also, Broer et al. (2019) analyzed data from the Trends in International Mathematics and Science Study (TIMSS) from 2003 to 2015 and noted that a reduction in government spending on education and having a decentralized education system were associated with an increase in the achievement gap between low- and high-SES students, although they did not claim that this relationship was causal. These examples show that data from ILSAs can be used to support the advancement of social equity in education by helping researchers investigate performance differences across and within countries and also by helping policy makers establish performance-targeted interventions to improve educational outcomes.

Measurement Invariance in International Large-Scale Assessments

An important assumption in comparative research using SES scores from ILSAs, such as PISA, is that they are comparable (i.e., invariant) across countries, an assumption which needs to be tested (van de Vijver, 2019). Measurement invariance, in this context, means that the items used to establish the SES scale are measuring the same construct, independent of the group (or country) a respondent belongs to, thereby ensuring that the measurement of a construct is fair across the different groups (Wu et al., 2007). It should be noted that there are several levels of measurement invariance. In a multiple-group confirmatory factor analysis (MG-CFA) framework, configural invariance (i.e., the lowest level of invariance) requires the factor structure to be identical across groups, while metric invariance requires the factor loadings to also be identical across groups, and scalar invariance requires the intercept of the regression equations to also be identical across groups (Jöreskog, 1971).

Recently, many studies have shown that scalar invariance cannot be established for either the cognitive assessment or the background questionnaires of ILSAs. For example, Sandoval-Hernandez et al. (2019) examined the different SES scales used in PISA, TIMSS, and the Third Regional Comparative and Explanatory Study on Education Quality (TERCE) and found that none of the scales were scalar invariant across the participating countries when using an MG-CFA framework. Also, according to the technical report of PISA 2015 (OECD, 2017a, chapter 16, p. 340), the three components used to measure SES in PISA – parents' education, parents' occupation, and home possessions – had different factor loadings across countries, implying that scalar invariance could not be established for the SES scale across the participating

countries. In PISA 2018, the relationship between each of the three indicators and SES was constrained to be the same across all countries by making SES the mean of the three indicators (OECD, 2020, chapter 16, p. 9). However, the cross-country comparability of each of the three indicators is still a concern, as explained in the next section.

Cross-country Comparability of the PISA SES Scale

Parents' education, one components of SES in PISA, is measured by the years of schooling received by the most highly educated parent of the student (OECD, 2020, chapter 16). This information is provided by the students, not the parents themselves, and is later mapped onto the International Standard Classification of Education (ISCED) framework established by UNESCO in 1997. Subsequently, the ISCED level of the most highly educated parent is converted into years of schooling using the median value – across all countries that participated in PISA 2015 – of the years of schooling for each ISCED level. This method may have compromised the cross-country comparability of parents' education, not only because the same years of schooling were uniformly assigned to each level of education for all countries regardless of the actual education system, but also because equivalent levels of education in different countries had been classified into different ISCED levels to begin with (Schneider & Kogan, 2008).

Information on parents' occupational status, another component of SES in PISA, is also provided by students (OECD, 2020, chapter 16). These responses are subsequently coded using the International Standard Classification of Occupations (ISCO) framework established by the International Labour Organization (ILO) in 2008, then mapped onto the International Socio-Economic Index of Occupational Status (ISEI) framework. A higher ISEI score is assumed to indicate a higher occupational status (Ganzeboom & Treiman, 2003; OECD, 2020, chapter 16), and the higher ISEI score of either the mother or father is used to calculate students' SES scores. An obvious threat to the cross-country comparability of this indicator is that the relative prestige and income of occupations can vary significantly across countries (Avvisati, 2020).

The third component of SES in PISA is home possessions, a proxy for family wealth in the absence of questions in the student background questionnaire that directly ask about family wealth. As with the other two indicators of SES, this information is provided by students (OECD, 2020, chapter 16). The HOMEPOS scale, which included 25 items in PISA 2018, is scaled using item response theory (IRT). In recent years, many studies have raised concerns about the cross-country comparability of this scale. For example, Rutkowski and Rutkowski (2013) analyzed data from the HOMEPOS scale in PISA 2009 and concluded that scalar invariance could not be established across countries for the three HOMEPOS subscales – the family wealth possessions scale (WEALTH), the cultural possessions scale (CULTPOS), and the home

education resources scale (HEDRES) – when using an MG-CFA framework. In addition, Pokropek et al. (2017) examined the 15 items included in the HOMEPOS scale in PISA 2012 and found that none of the items met all three criteria they had established to assess cross-country comparability – the modification indices (MI), the expected parameter change (EPC), and the root mean square deviation (RMSD).[3] Similarly, Lee and von Davier (2020) analyzed data from the HOMEPOS scale from 2000 to 2015 and found that none of the items were fully comparable across countries when using the RMSD and the mean deviation (MD).[4] While the studies cited above used different methodologies, the findings were generally consistent with regards to concerns about the cross-country comparability of the HOMEPOS scale.

Operationally, starting with PISA 2015, differential item functioning (DIF) in the HOMEPOS scale was detected and addressed using multiple-group concurrent calibration with partial invariance constraints (OECD, 2017a, chapter 16). For each HOMEPOS item, this procedure allowed country-by-language groups exhibiting DIF to receive unique item parameters,[5] while constraining the other country-by-language groups to have the same item parameters. Essentially, this created a partial invariance model which aimed to maximize the comparability of the scale across the country-by-language groups while ensuring the accuracy of the item parameters for each country-by-language group. In a nutshell, this modeling approach allowed the treatment of measurement invariance so more comparable scales could be established.

This paper aims to contribute to the important research on the cross-country comparability of the HOMEPOS scale by analyzing data from PISA 2018 – the most recent administration of PISA – to examine whether DIF in this scale is related to a country-by-language group's characteristics such as the economic development level, geographic location, and language. While other studies on this topic examined the measurement invariance of the HOMEPOS scale as a whole (e.g., Rutkowski & Rutkowski, 2013; Sandoval-Hernandez et al., 2019), the current study focuses on DIF of items in the HOMEPOS scale and the possible reasons for DIF. It also expands on the research by Pokropek et al. (2017) and Lee and von Davier (2020) by using data from the PISA 2018 HOMEPOS scale which has not yet been analyzed extensively. The results of this paper can provide information to PISA test developers on how to improve the comparability of the HOMEPOS scale across countries, thereby reducing the occurrence of DIF before it has to be treated later in the scaling process. The specific research questions addressed by this paper are:

1) Study 1: Which items in the PISA 2018 HOMEPOS scale are most impacted by DIF and, thus, least comparable across the country-by-language groups?

[3] The criteria Pokropek et al. (2017) used to flag an item-country pair for differential item functioning (DIF) was MI > 50, EPC > 0.35, and RMSD > 0.10.

[4] The criteria Lee and von Davier (2020) used to flag an item-group pair for DIF was RMSD > 0.15 or absolute MD > 0.15.

[5] The criterion used to detect DIF for an item-group pair was RMSD > 0.30 (OECD, 2017a).

2) Study 2: For which items in the PISA 2018 HOMEPOS scale is the level of DIF associated with a country's economic development level?

3) Study 3: Which country-by-language groups have the highest number of items exhibiting DIF in the PISA 2018 HOMEPOS scale?

4) Study 4: Are there clusters of country-by-language groups that have a similar pattern of DIF across the items in the PISA 2018 HOMEPOS scale?

Methods

Data

Data from the PISA 2018 HOMEPOS scale, a scale included in the student background questionnaire, were used in this research. PISA is an ILSA administered by the OECD every three years since 2000 to assess 15-year-old students' ability to use their knowledge and skills in reading, mathematics, and science to solve real-life challenges (OECD, n.d.-b). In 2018, 77 countries participated in PISA, including all 36 OECD countries (at the time PISA 2018 was administered) and 41 non-OECD countries. This research excluded samples that were not nationally representative, such as samples from specific regions or cities within a country.[6] It also excluded the Ultra-Orthodox group in Israel, because many of the HOMEPOS items had not been administered to this group. Following the operational procedure used for the cognitive assessment in PISA 2018, within each country, minority languages that were used as the language of examination by at least 5% of the test takers (using sampling weights) were considered to be independent groups in the analyses (OECD, 2020, chapter 9). This grouping method was used because it was hypothesized that socio-cultural factors, which are partially captured by the language of examination, can affect how an item functions when it is used to measure family wealth. A total of 97 groups (called country-by-language groups) including 588,879 students (in unweighted sample size) were used in the analyses.

In PISA, the student background questionnaire was administered directly to students for 30 to 35 minutes after the cognitive assessments (OECD, 2020, chapter 2). The HOMEPOS scale, one of the 66 scales included in the student background questionnaire, consisted of 25 items, including 16 dichotomous items (i.e., it asked whether the student's family owned an item or not) and nine polytomous items (i.e., it asked how many of each item the student's family owned; OECD, 2020, chapter 16). The dichotomous items were: a desk to study at; a room of your own; a quiet place to

[6] The four samples that were excluded were Beijing-Shanghai-Jiangsu-Zhejiang (B-S-J-Z), Hong Kong, and Macao in China; and Baku in Azerbaijan.

study; a computer you can use for school work; educational software; a link to the internet; classic literature; books of poetry; works of art; books to help with your school work; technical reference books; a dictionary; books on art, music, or design; and three items that each country could specify. The three country-specific items were excluded from this research because they could not be compared across countries. The polytomous items were: televisions, cars, rooms with a bath or shower, cell phones with internet access, computers, tablet computers, e-book readers, musical instruments, and books.[7]

Using the unweighted sample sizes, the percent of missing data across all items in the HOMEPOS scale ranged from 0.6% in Korea to 14.1% for the Arabic-speaking group in Israel. For 84 out of the 97 country-by-language groups, the percent of missing data across all items in the scale was less than 5%.

Scaling

The item parameters for the items in the HOMEPOS scale were estimated using IRT. Specifically, the two-parameter logistic (2PL) model (Birnbaum, 1968) was used to scale the dichotomous items, while the generalized partial credit model (GPCM; Muraki, 1992) was used to scale the polytomous items.[8] The following is the formula for the GPCM, which reduces to the 2PL model for dichotomous items, where $P(X_{vi} = k)$ is the probability of person v scoring k on item i out of the m_i possible scores on the item, θ_v is the person's latent trait, b_i is the general location of the item on the latent continuum, a_i is the slope of the item, d_i is the additional step parameters of the item, and D is the scaling constant 1.7:

$$P(X_{vi} = k \mid \theta_v, b_i, a_i, d_i) = \frac{\exp\{(\sum_{r=0}^{k} D a_i(\theta_v - b_i + d_{ir})\}}{\sum_{u=0}^{m_i} \exp\{\sum_{r=0}^{u} D a_i(\theta_v - b_i + d_{ir}\}} \quad (1)$$

For the scaling, each student was weighted so the sample would be nationally representative, and the weights were adjusted so the sum of the weights would be 5,000 for each country, ensuring that each country contributed equally to the estimation of the item parameters. The scaling was conducted using the software mdltm (von Davier, 2005) and missing data were treated as ignorable missing data.[9]

[7] The response categories for the polytomous items were zero, one, two, and three or more, except for the number of books. The response categories for the number of books were zero to 10, 11 to 25, 26 to 100, 101 to 200, 201 to 500, and more than 500.

[8] To solve the indeterminacy of the IRT scale, the average of the item discrimination parameters across the items was constrained to one, while the average of all the intercepts across the items was constrained to zero.

[9] The software mdltm provides marginal maximum likelihood (MML) estimates obtained using customary expectation-maximization (EM) methods with optional acceleration.

Study 1: Differential Item Functioning – Analysis at the Item Level

In recent years, several methods have been proposed to detect DIF in ILSAs, due to the difficulties of establishing scalar invariance in an MG-CFA framework when using data from many heterogenous groups with large sample sizes. These methods include multiple group concurrent calibration (Glas & Jehangir, 2014; Oliveri & von Davier, 2011), a robust method for detecting item misfit in large-scale assessments (von Davier & Bezirhan, 2021), Bayesian structural equation modeling (Muthén & Asparouhov, 2012), and the alignment method (Muthén & Asparouhov, 2014). This study used multiple group concurrent calibration to detect DIF, in line with the operational procedures used in PISA 2018 (OECD, 2020, chapter 9) and PISA 2015 (OECD, 2017a, chapter 9), although the criterion for detecting DIF was different from the operational procedure.[10]

With multiple group concurrent calibration, for each item, the observed item characteristic curve (ICC) for each country-by-language group (estimated with data only from that group) and the model-based ICC (estimated with data pooled across all the country-by-language groups) are estimated concurrently. Subsequently, for each item, the discrepancy between the observed ICC for each country-by-language group and the model-based ICC is quantified using the MD as a measure of DIF. Equation (2) provides the formula for MD, where $P_o(\theta)$ refers to the observed proportion of students that owned an item at a given level of θ, obtained from the pseudo counts from the expectation step in the EM algorithm; $P_e(\theta)$ refers to the proportion of students that is expected to own an item at a given level of θ, obtained from the model-based ICC; and $f(\theta)$ refers to the number of students at a given level of θ:

$$MD = \int (P_o(\theta) - P_e(\theta)) f(\theta) d\theta \qquad (2)$$

The MD values range from –1 to 1, with values further from zero indicating a larger difference between the observed ICC for the country-by-language group and the model-based ICC. In general, the MD is positive when a higher proportion of students own an item than what is predicted by the model, while the MD is negative when a lower proportion of students own an item than what is predicted by the model. For example, the left graph in Figure 1 shows the observed ICC for classic literature for the United States (the dashed line) and the model-based ICC for classic literature (the solid line) estimated with data pooled across all country-by-language groups that participated in PISA 2018. For each value of the HOMEPOS score on the x-axis, a lower proportion of students in the United States owned books of classic literature than what was predicted by the model. For this item, the MD for the United States was –0.19, a negative value. For comparison, the right graph in Figure 1 shows the observed ICC for classic literature for Russia (the dashed line) and the model-based ICC for classic

[10] Operationally, in PISA 2018 and 2015, an RMSD cut-off of 0.3 was used to detect DIF for each item-group pair, while this study used a criterion of absolute MD $\geq$ 0.12 to detect DIF (i.e., MD $\leq$ –0.12 or MD $\geq$ 0.12).

literature (the solid line – note that this line is the same as the solid line in the left graph). For each value of the HOMEPOS score on the x-axis, a higher proportion of students in Russia owned books of classic literature than what was predicted by the model. For this item, the MD for Russia was 0.34, a positive value.

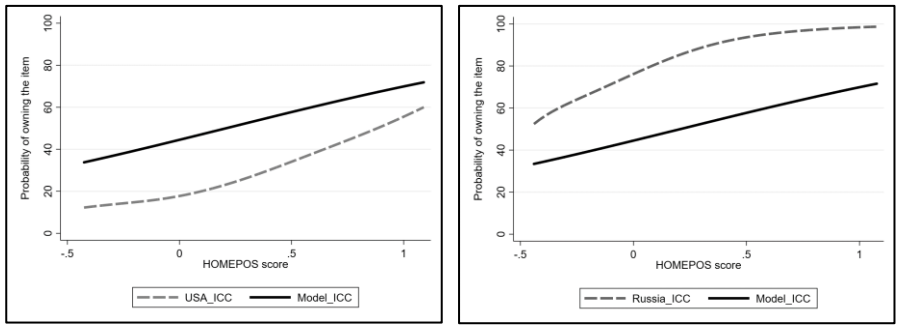

Figure 1

Model-Based and Observed ICC for Classic Literature for the United States (Left) and Russia (Right)

When the absolute MD value for a country-by-language group was greater than or equal to 0.12 for an item (MD ≤ –0.12 or MD ≥ 0.12), it was assumed that DIF had been detected for the country-by-language group for the item.[11] In this context, DIF indicates that the relationship between the probability of owning an item and the HOMEPOS score for a country-by-language group is different from the relationship between the probability of owning the item and the HOMEPOS score for the pooled group. In other words, the item functions differently in that country-by-language group, compared to the other country-by-language groups, when the item is used to measure family wealth. For each item, the more country-by-language groups for which DIF is detected, the less comparable the item is in measuring family wealth across the country-by-language groups, because it means the relationship between ownership of the item and family wealth varies across the country-by-language groups. Through this analysis, it was possible to identify the items in the HOMEPOS

[11] Operationally, in PISA 2018 and PISA 2015, an RMSD cut-off of 0.3 was used to detect DIF for each item-group pair, and country-by-language groups for which DIF had been detected were allowed to receive unique parameters for the item (OECD, 2020, chapter 9; OECD, 2017a, chapter 9). In this research, a stricter criterion of 0.12 was used to detect DIF, as the goal was to detect as many item-group pairs exhibiting DIF as possible. Also, this research used the MD instead of the RMSD to detect DIF, as the MD gives information about the direction and magnitude of DIF, whereas the RMSD only gives information about the magnitude of DIF. The direction of DIF was especially important for the correlational analyses (in Study 2) and the cluster analysis (in Study 4).

scale that were the least comparable across the country-by-language groups in measuring family wealth. Note that the country-by-language groups for which DIF had been detected were not allowed to receive unique parameters for the item, because the focus of the research was to examine the comparability of the scale across the country-by-language groups, not to improve the scale's accuracy for each country-by-language group.

Study 2: Correlation Between DIF and a Country's Economic Development Level – Analysis at the Item Level

The purpose of this analysis was to examine the associations between the MD of each item and a country's level of economic development. For each item, the MD for each country-by-language group was correlated with the country's gross domestic product (GDP) per capita in purchasing power parity (PPP) in 2018 (World Bank, n.d.). Note that the country-level statistics were used for GDP per capita in the absence of information on the GDP per capita for each country-by-language group.

Study 3: Differential Item Functioning – Analysis at the Group Level

In this study, the number of items in the HOMEPOS scale for which DIF had been detected was analyzed for each country-by-language group. Again, the criterion to detect DIF for a country-by-language group for an item was an absolute MD value greater than or equal to 0.12. For each country-by-language group, more items exhibiting DIF meant that more items in the scale functioned differently for the country-by-language group compared to the other country-by-language groups, implying that a higher overall level of misfit was found in the scale for that country-by-language group. Through this analysis, it was possible to identify the country-by-language groups for which the HOMEPOS scale had the highest overall level of misfit.

Study 4: Cluster Analysis – Analysis at the Group Level

In this study, country-by-language groups were classified into clusters using cluster analysis – an unsupervised machine learning technique which uses attributes (in this case, the MD for each item in the HOMEPOS scale) to classify observations (in this case, country-by-language groups) into mutually exclusive clusters so that observations in the same cluster are as similar as possible and observations in different clusters are as dissimilar as possible (Boehmke & Greenwell, 2020). Through this process, country-by-language groups that are similar in the level and direction of DIF across all the items in the HOMEPOS scale were classified into the same cluster. After defining the clusters, it was possible to examine the characteristics of each cluster, in

other words, characteristics that were associated with the level and direction of DIF across all the items in the HOMEPOS scale. Note that Israel was excluded from this analysis because Israel had not administered some items (i.e., a room of your own, rooms with a bath or shower), making it impossible to classify Israel into a cluster. Thus, only 95 country-by-language groups were included in the cluster analysis.

The cluster analysis was conducted in SAS (version 9.4) using Ward's minimum-variance clustering method (Ward, 1963), an agglomerative method in which each country-by-language group is placed in its own cluster at the beginning of the clustering process. At each subsequent step, the two clusters that minimize within-cluster variance and maximize between-cluster variance are combined to form a new cluster, and this is repeated until there is only one cluster containing all the country-by-language groups. Subsequently, the cluster solution with the smallest number of clusters that satisfies the following criteria is determined to be the ideal cluster solution: (a) equal within-cluster variance as the preceding step, (b) lower within-cluster variance than the subsequent step, and (c) a higher pseudo-F statistic (which indicates the separation among all clusters; Caliński & Harabasz, 1974) than the pseudo-t^2 statistic (which indicates the separation of the two clusters that were most recently combined; Duda et al., 1973), as demonstrated by Cooper and Milligan (1988).

Results and Discussion

Scaling

The item parameters for all of the items in the HOMEPOS scale are presented in Table 1. The a-parameters ranged from 0.46 (for books of poetry) to 2.37 (for computer you can use for school work). The b-parameters ranged from –1.80 (for dictionary) to 0.39 (for books of poetry) for dichotomous items, and from –0.82 (cell phones with internet access) to 1.58 (e-book readers) for polytomous items. These item parameters were subsequently used to estimate the IRT-based HOMEPOS score, as weighted likelihood estimates (WLE; Warm, 1989), for each student as well as the average HOMEPOS score for each country-by-language group.

Table 1

Item Parameters for Items in the HOMEPOS Scale

Item	Parameters						
	a	b	d				
A desk to study at	1.23	–0.90					
A room of your own	0.85	–0.89					
A quiet place to study	0.89	–1.20					
A computer you can use for school work	2.37	–0.33					
Educational software	1.13	0.18					
A link to the Internet	2.32	–0.68					
Classic literature	0.63	0.21					
Books of poetry	0.46	0.39					
Works of art	0.87	–0.05					
Books to help with your school work	0.60	–1.26					
Technical reference books	0.91	0.06					
A dictionary	0.59	–1.80					
Books on art, music, or design	0.72	0.13					
Televisions	0.65	–0.75	1.75	–0.67	–1.08		
Cars	1.01	0.32	0.69	–0.06	–0.63		
Rooms with a bath or shower	0.97	0.21	1.30	–0.41	–0.89		
Cell phones with internet access	0.88	–0.82	0.36	–0.71	0.35		
Computers	2.10	0.11	0.54	–0.12	–0.42		
Tablet computers	0.98	0.54	0.43	–0.25	–0.18		
E-book readers	0.63	1.58	–0.31	–0.26	0.57		
Musical instruments	0.65	0.71	0.09	–0.28	0.19		
Books	0.54	0.68	0.61	0.76	–0.47	–0.26	–0.63

Note: a refers to the slope parameter, b refers to the general location parameter, and d refers to the step parameters for polytomous items in the GPCM.

It is interesting to note that the average HOMEPOS score for the country-by-language groups does not have a linear relationship with the countries' per-capita wealth, even though the HOMEPOS scores were meant to be a proxy for family wealth. Figure 2 presents a scatterplot of the average HOMEPOS score for each country-by-language group using PISA 2018 data and countries' GDP per capita in PPP from 2018 (World Bank, n.d.). The correlation is 0.67, and the relationship between the two variables is

better explained by a logarithmic function than a linear function, as indicated by the trendline in the figure. The logarithmic relationship between the two variables suggests that the marginal contribution of per-capita income in improving the HOMEPOS score diminishes at higher levels of income.

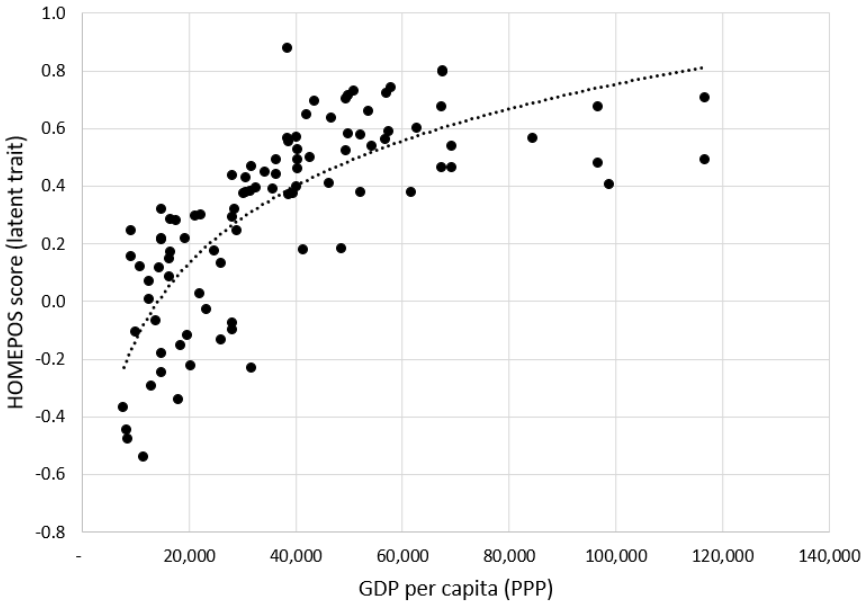

Figure 2

Scatterplot Between Each Country-by-Language Group's Average HOMEPOS Score (WLE) and the GDP per Capita in PPP

For comparison, Figure 3 below presents a scatterplot of the average HOMEPOS score for each country-by-language group using PISA 2018 data and countries' Human Development Index (HDI) in 2018 (United Nations Development Programme [UNDP], n.d.). The HDI, developed by the UNDP, is a composite index which is calculated by taking the geometric mean[12] of the following: a) the logarithm of the gross national income (GNI) per capita, representing a decent standard of living; b) the expected years of schooling for school-age children or the mean years of schooling for adults aged 25 years or more, representing education; and c) the life expectancy at birth, representing a long and healthy life. The correlation between the average

[12] The geometric mean is calculated by taking the cube root of the product of the three components.

HOMEPOS score for each country-by-language group in 2018 and countries' HDI in 2018 is 0.83, which is higher than the correlation between the average HOMEPOS scores and the GDP per capita. Also, the relationship between the two variables is relatively linear, perhaps because the HDI already takes into account the diminishing marginal contribution of income in improving the HDI at higher levels of income (by taking the logarithm of the GNI per capita when calculating the HDI).

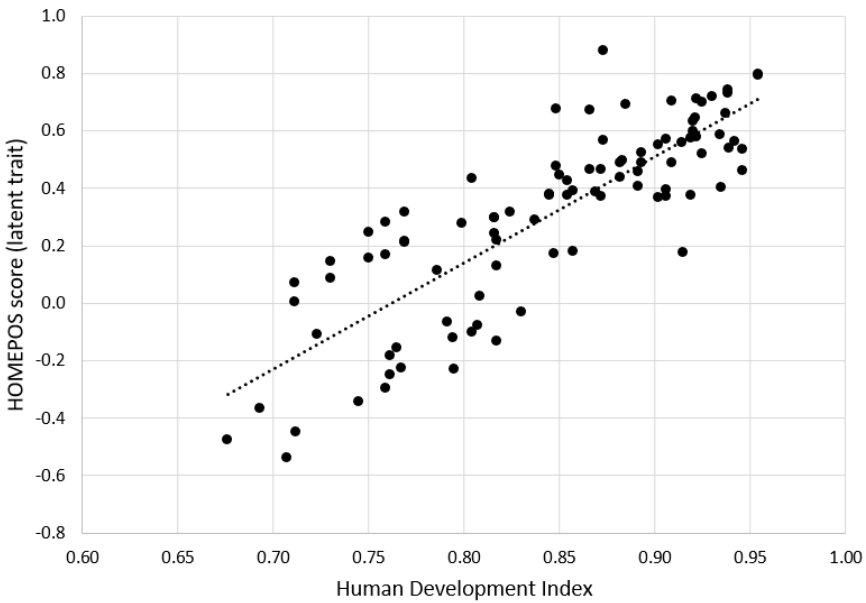

Figure 3

Scatterplot Between Each Country-by-Language Group's Average HOMEPOS Score (WLE) and the Human Development Index

Study 1: Differential Item Functioning – Analysis at the Item Level

Figure 4 presents, for each item, the percent of country-by-language groups for which DIF had been detected using a cut-off of 0.12 for the absolute MD value. DIF had been detected for over 50% of the country-by-language groups for classic literature and books of poetry, implying that these items functioned very differently across the country-by-language groups when they were used to measure family wealth. In other words, these items were not very comparable across the country-by-language groups

in measuring family wealth. Issues with the cross-country comparability of both of these items had already been raised in studies by Lee and von Davier (2020) and Pokropek et al. (2017). Operationally, to address the high proportion of country-by-language groups in which DIF had been detected for these two items, all country-by-language groups received unique item parameters for these items in PISA 2018 and PISA 2015 (OECD, 2020, chapter 9; OECD, 2017a, chapter 9). While this operational procedure may have increased the accuracy of the item parameters within each country-by-language group, it does not address the non-comparability of the items across the country-by-language groups. Therefore, it is recommended that these items be excluded from the HOMEPOS scale (and potentially be replaced by new items with a better fit) in the future to improve the comparability of the scale across country-by-language groups.

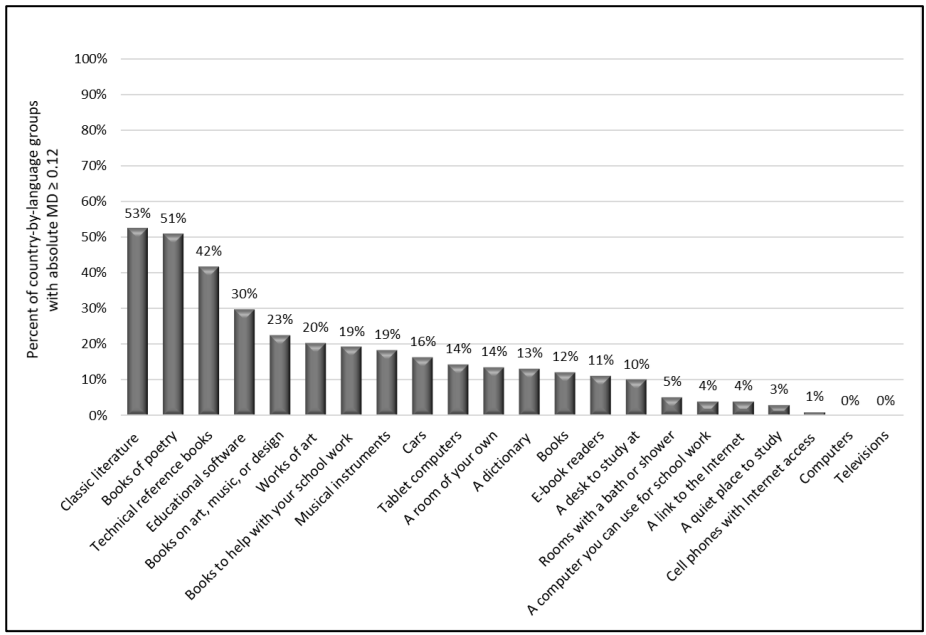

Figure 4

Percent of Country-by-Language Groups With Absolute MD Value Greater Than or Equal to 0.12

Study 2: Correlation Between DIF and a Country's Economic Development Level – Analysis at the Item Level

For each item, the correlation between the MD for each country-by-language group and the countries' GDP per capita in PPP in 2018 is presented in Table 2. The correlations ranged from –0.56 (for books of poetry) to 0.52 (for tablet computers).

Table 2

Correlation Between MD of Each Country-by-Language Group and Country's Economic Development Level

	Correlation between MD & GDP per capita (PPP)
Books of poetry	–0.56
Classic literature	–0.55
Books on art, music, or design	–0.47
Televisions	–0.37
A computer you can use for school work	–0.35
Technical reference books	–0.34
Books to help with your school work	–0.33
Works of art	–0.32
A quiet place to study	–0.31
A dictionary	–0.26
A desk to study at	–0.24
Educational software	–0.22
A room of your own	–0.20
Books	–0.15
Musical instruments	–0.10
E-book readers	–0.06
Cell phones with internet access	–0.04
A link to the internet	–0.03
Computers	0.01
Rooms with a bath or shower	0.12
Cars	0.31
Tablet computers	0.52

The lowest correlation between the MD and GDP per capita in PPP was for books of poetry which had a correlation of –0.56. The scatterplot for this item is presented in Figure 5. A negative correlation generally means that in countries with a low GDP per capita, the MD is positive (i.e., a higher proportion of students own the item than what is predicted by the model), while in countries with a high GDP per capita, the MD is negative (i.e., a lower proportion of students own the item than what is predicted by the model). It could be that these items are symbols of status or cultural capital in countries with a low GDP per capita, resulting in families buying more books of poetry in these countries compared to families with a similar level of wealth in other countries.

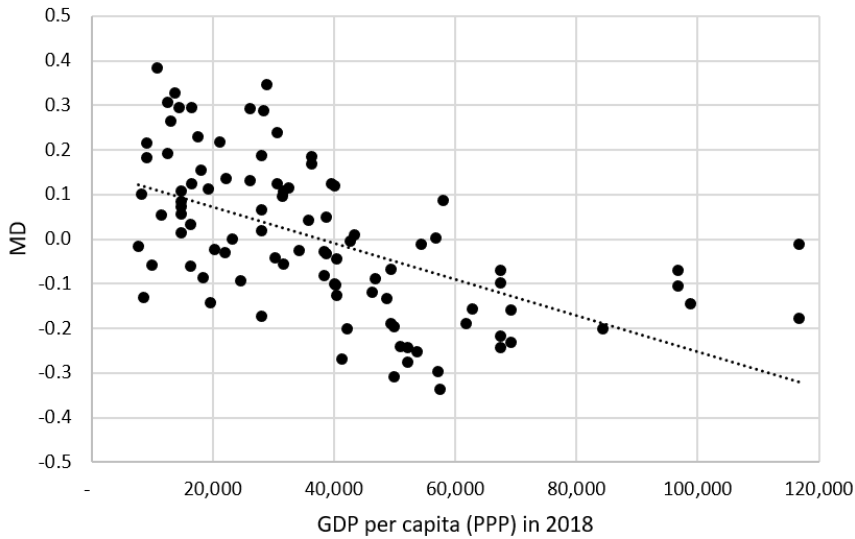

Figure 5

Scatterplot Between Each Country-by-Language Group's MD and GDP per Capita in PPP for Books of Poetry

In contrast, the highest correlation between the MD and GDP per capita in PPP was for tablet computers, which had a correlation of 0.52. The scatterplot for this item is presented in Figure 6. A positive correlation generally means that in countries with a low GDP per capita, the MD is negative (i.e., a lower proportion of students own the item than what is predicted by the model), while in countries with a high GDP per capita, the MD is positive (i.e., a higher proportion of students own the item than what is predicted by the model). It may be that tablet computers are a status symbol in countries with a high GDP per capita, resulting in families in these countries buying more tablet computers compared to families with a similar level of wealth in other

countries. Alternatively, it could mean that tablet computers are simply more accessible in wealthier countries because the marketing of this product is targeted at those countries or the international supply chain makes it easier for tablet computers to be sold in those countries. Whatever the reason, items for which the level of DIF is associated with a country's economic development level should be excluded from the HOMEPOS scale, as it can create bias in the HOMEPOS scores across countries. As more developing countries are planning to participate in the next cycle of PISA in 2022 (OECD, 2017b), it is more important than ever that the HOMEPOS scores are not biased by the economic development level of the participating countries.

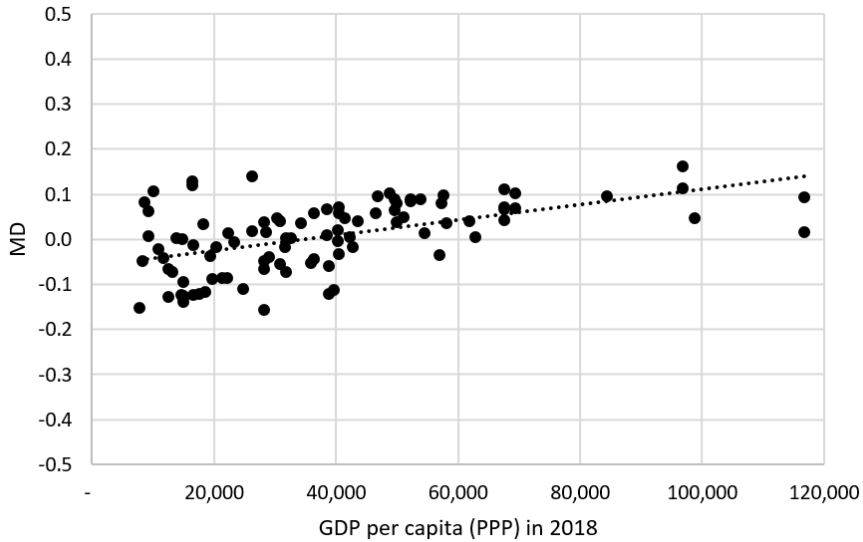

Figure 6

Scatterplot Between Each Country-by-Language Group's MD and GDP per Capita in PPP for Tablet Computers

Study 3: Differential Item Functioning – Analysis at the Group Level

Figure 7 presents, for each country, the percent of items in the scale for which DIF had been detected. Note that only the result for the main language group of each country is shown on the map. In Qatar (Arabic), Saudi Arabia (Arabic & English), the Philippines (English), the United Arab Emirates (Arabic), Brunei (English), Indonesia (Indonesian), Kazakhstan (Kazakh), Japan (Japanese), and Malaysia (Malay), DIF had been detected in over a third of the items. The high percent of items for which DIF

had been detected for these country-by-language groups implied that for these groups, many of the items in the HOMEPOS scale functioned differently than in the other groups when the items were used to measure family wealth. In other words, a high overall level of misfit was found in the scale for these country-by-language groups.

Figure 7

Percent of Items With Absolute MD Greater Than or Equal to 0.12

Note. Only the result for the main language group of each country is shown on the map. Also, countries that were not included in the analysis are colored in white.

It is interesting to note that among the top nine country-by-language groups with the highest number of items for which DIF had been detected, seven are from countries with a large Muslim population (i.e., Qatar, Saudi Arabia, the United Arab Emirates, Kazakhstan, Brunei, Indonesia, and Malaysia) – these countries are in the Middle East, Central Asia, and Southeast Asia. The high proportion of items exhibiting DIF in Muslim-majority countries indicate that many of the items in the HOMEPOS scale function differently in countries with a large Muslim population compared to other countries. This suggests that how an item functions in measuring family wealth may be related to the socio-cultural characteristics of a group, such as religion.

As noted by Brese and Mirazchiyski (2013) and Yang and Gustafsson (2004), a major challenge in measuring family wealth with household items is that some items may be valued differently in different societies and cultures, or some items may be less accessible in certain countries due to external circumstances, making the relationship

between the ownership of an item and family wealth vary across countries. For example, Russian students may be more likely to have books of classic literature at home compared to students with a similar level of wealth in other countries, simply because Russians have a culture of valuing classic literature (or at least classic Russian literature). Also, students in countries where public transportation is widely available may have less cars at home compared to students with a similar level of wealth in other countries. When many of the items in the HOMEPOS scale function differently for a country compared to other countries, a high overall level of misfit will be found in the scale – indicated by a darker shade in the map above.

Study 4: Cluster Analysis for Country-by-Language Groups – Analysis at the Group Level

The results from the cluster analysis indicated that classifying the 95 country-by-language groups into five clusters produced the best solution, with cluster sizes ranging from 51 country-by-language groups (for Cluster 1) to just three country-by-language groups (for Cluster 5). These results are presented in Figure 8 (with only the result for the main language group of each country presented on the map).

Figure 8

Results of the Cluster Analysis

Note. Only the result for the main language group of each country is presented on the map. Also, countries that were not included in the analysis are colored in white.

The map above revealed some clear geographic patterns – Cluster 1 mostly included countries in North/Central/South America, Western Europe, Scandinavia, and Australia. Interestingly, 26 of the 35 OECD member countries were in Cluster 1.[13] In contrast, Cluster 3 mostly consisted of countries that had been member states or satellites of the Soviet Union (e.g., Russia, Belarus, Georgia, Kazakhstan, Moldova, Romania, and Ukraine). Cluster 2 also included several countries that had been member states or satellites of the Soviet Union (e.g., Estonia, Latvia, Lithuania, and Slovakia), but the countries in Cluster 2 are geographically located between the countries in Cluster 1 and countries in Cluster 3. It is also interesting to note that Cluster 2 included the nine remaining OECD member countries, suggesting that the countries in Cluster 2 are geographically and economically distinct from the countries in Cluster 3, in spite of their shared recent history. Lastly, Cluster 4 included countries that are mainly in Southeast Asia (e.g., Indonesia, Malaysia, Philippines, Thailand, and Vietnam), while Cluster 5 consisted of three countries in the Middle East (e.g., Qatar, Saudi Arabia, and the United Arab Emirates). These results suggest that there are regional similarities in how an item functions in measuring family wealth, perhaps because countries in the same region are similar in terms of their cultural values, level of economic development, geopolitical situation, etc.

Some differences were noted even within countries. Among the 20 countries with multiple language groups, in 17 countries, the main and minority language groups were classified into the same cluster. However, in three countries, the minority language group was classified into a different cluster from the main language group – in the United Arab Emirates and Qatar, the Arabic-speaking group was classified into Cluster 5, while the English-speaking group was classified into Cluster 1. In Malaysia, the Malay-speaking group was classified into Cluster 4, while the English-speaking group was classified into Cluster 1. This suggests that how an item functions in measuring family wealth may also be related to the socio-cultural characteristics of a group which is captured by the language of examination of the group.

For each item, the average MD for each cluster is presented as a radar chart in Figure 9. Classic literature had the highest variability in the average MD across the clusters, with the average MD ranging from –0.21 (for Cluster 5) to 0.31 (for Cluster 3). The item with the second highest variability in the average MD across the clusters was books of poetry. Similar to the findings in Study 1, these results show that classic literature and books of poetry are not comparable across the clusters in measuring family wealth, and thus are recommended for exclusion from the HOMEPOS scale.

[13] The 35 countries exclude Israel which is a member of the OECD but was not included in the cluster analysis.

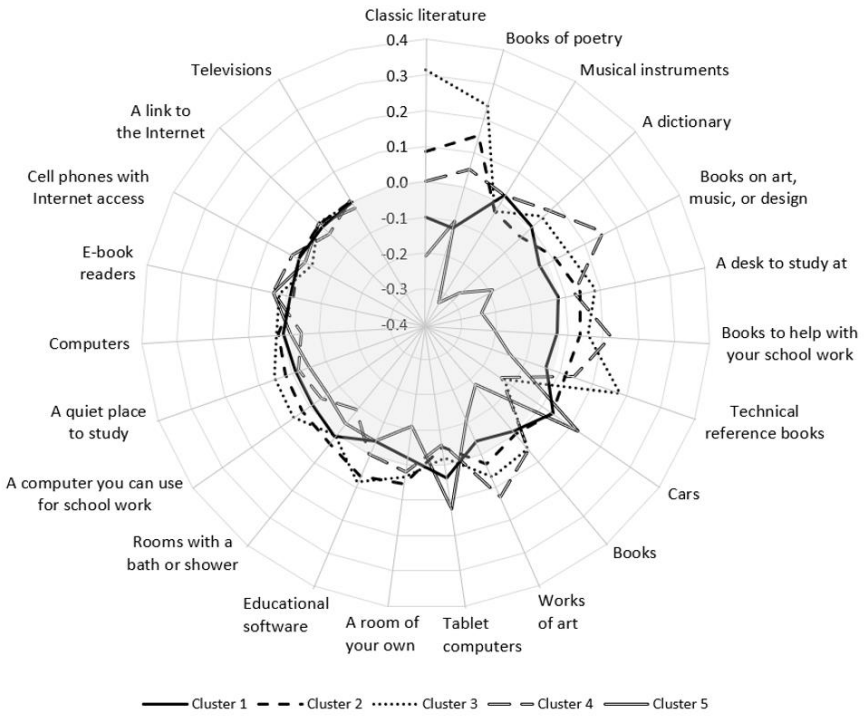

Figure 9

Average MD for Each Item and Cluster

Note. The items are presented in descending order of variability in the cluster-level average MD.

Conclusion

This paper contributes to studies on the cross-country comparability of the measurement of SES in ILSAs by analyzing data from the HOMEPOS scale of PISA 2018. While other studies on this topic have examined the measurement invariance of the HOMEPOS scale as a whole (e.g., Rutkowski & Rutkowski, 2013; Sandoval-Hernandez et al., 2019), this study focused on DIF of items in the HOMEPOS scale and the possible reasons for DIF. The insights provided by this study can help PISA test developers improve the comparability of the HOMEPOS scale across countries, thereby reducing the occurrence of DIF before it has to be treated later in the scaling process.

In Study 1, it was found that classic literature and books of poetry were the least comparable across the country-by-language groups in measuring family wealth, with DIF being detected for over 50% of the country-by-language groups for these items. This is in line with the findings by Lee and von Davier (2020) and Pokropek et al. (2017). In Study 2, the MD for these two items were also found to have a strong negative correlation with countries' GDP per capita in PPP in 2018, providing further support for excluding these items from the HOMEPOS scale to improve the comparability of the scale across countries in the future. Study 2 also found that the MD for tablet computers had a strong positive correlation with countries' GDP per capita in PPP in 2018 – this item is also a candidate for exclusion from the HOMEPOS scale, especially since more economically diverse countries are planning to participate in PISA 2022.

Analyses from Study 3 showed that DIF was detected for a high percentage of items in certain countries, especially countries with a large Muslim population, suggesting that how an item functions in measuring family wealth may be related to the socio-cultural characteristics of a group. In Study 4, it was found that there are regional similarities in how the items in the HOMEPOS scale function in measuring family wealth, perhaps because countries in the same region have similar characteristics such as cultural values, level of economic development, and geopolitical situation. Both studies provide evidence that the HOMEPOS scale functions differently across countries in measuring family wealth, a serious issue considering the increasing diversity of countries planning to participate in future rounds of PISA.

A practical solution for increasing the cross-country comparability of the HOMEPOS scale while maintaining the relevance of the scale for the participating countries could be to use different sets of items for different socio-cultural groups which can be defined by geographic region, language, religion, etc. For this recommendation to work, a core set of items that function very similarly across all groups in measuring family wealth should be administered to all groups so that the scale can be linked across the groups. Items such as a quiet place to study, a link to the internet, a computer you can use for school work, a desk to study at, e-book readers, a room of your own, and musical instruments are good candidates for the core set of items, as DIF was detected in less than 20% of the country-by-language groups for these items using both the absolute MD and RMSD (with a cut-off of 0.12). Also, keeping a sub-set of items that are common across the PISA cycles will allow the scale to be linked across cycles, making trend analyses possible. More research is needed to identify additional items that function similarly across time and across countries in measuring family wealth regardless of the socio-cultural characteristics, religion, economic development level, and the geopolitical situation of the countries participating in PISA.

It should be noted that there are some limitations of this study that should be taken into account when interpreting its results. One limitation is the reliability of self-reported data in low-stakes assessments such as PISA. Thus, it may be unrealistic to assume that all students replied accurately to the student background questionnaire. Also, considering that students had to take the PISA cognitive assessment for two

hours before responding to the background questionnaire, respondent fatigue may have affected the accuracy of students' responses to the HOMEPOS scale. Another limitation is that an MD cut-off of 0.12 may have been too strict to detect DIF, inflating the number of country-by-language groups for which DIF had been detected for each item. However, in the absence of definitive research on the ideal cut-off for detecting DIF in the background questionnaires, this study used a relatively strict criterion, as the goal was to detect as many existing item-group pairs exhibiting DIF as possible and to not overlook potential invariance issues. In the future, more research should be conducted on the ideal MD cut-off to detect DIF in the background questionnaires.

Future research could also examine the usefulness of mixture IRT models (cf., Rost et al., 1999; von Davier et al., 2007) instead of cluster analysis for a more model-based approach. Mixture IRT models – which assume different item parameters for different latent classes – and the hybrid model (Yamamoto, 1987; Yamamoto, 1989; Yamamoto & Everson, 1997) – which assumes different models for different latent classes – could be used as alternative methods for detecting cross-country invariance of item parameters and DIF. It would be interesting to compare the results of these alternative approaches to the findings in this paper. Even if similar results are found, there might be differences in the robustness and accuracy of the estimations (which could be revealed through simulation studies).

In spite of the limitations mentioned above, the insights and recommendations from this research can help improve the comparability of the HOMEPOS scale across countries in the future, and by extension, the cross-country comparability of the SES scale of PISA. This would be an important contribution, since the SES variable is one of the most widely used variables in PISA-related research (Avvisati, 2020) and the non-comparability of the SES scores would threaten the validity of international comparative research using these scores. This research is also very timely, considering the increasing diversity of countries that are participating in PISA, the acknowledgement by the OECD of the need to make the SES scale more comparable across the participating countries, and the increasing focus on educational equity in many countries around the world.

References

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-scale Assessments in Education, 8:8*, 1–37. https://doi.org/10.1186/s40536-020-00086-x

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Boehmke, B., & Greenwell, B. (2020). *Hands-on machine learning with R*. CRC.

Brese, F., & Mirazchiyski, P. (2013). *Measuring students' family background in large-scale international education studies* (IERI Monograph Series – Special Issue 2). IEA-ETS Research Institute (IERI). https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/Special_Issue_2/10_IERI_Special_Issue_2_complete.pdf

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS*. Springer Nature. https://doi.org/10.1007/978-3-030-11991-1_1

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Coleman, J. (1968). The concept of equality of educational opportunity. *Harvard Educational Review, 38*(1), 7-22. https://doi.org/10.17763/haer.38.1.m3770776577415m2

Cooper, M. C., & Milligan, G. W. (1988). The effect of error on determining the number of clusters. In W. Gaul & M. Schader (Eds.), *Data, expert knowledge and decisions* (pp. 319–328). Springer. https://doi.org/10.1007/978-3-642-73489-2_27

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis* (Vol. 3, pp. 731–739). Wiley.

Duncan, O. D., Featherman, D. L. & Duncan, B. (1972). *Socioeconomic background and achievement*. New York, NY: Seminar Press.

Ganzeboom, H. B., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison* (pp. 159–193). Springer.

Glas, C. A., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97–115). CRC Press.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. https://doi.org/10.1007/BF02291366

Lee, S. S., & von Davier, M. (2020). Improving measurement properties of the PISA home possessions scale through partial invariance modeling. *Psychological Test and Assessment Modeling, 62*(1), 55–83. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/04_Lee.pdf

Massachusetts Board of Education. (1849). *Twelfth annual report of the Board of Education, together with the twelfth annual report of the Secretary of the Board (1848)*. https://archives.lib.state.ma.us/handle/2452/204731

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. https://doi.org/10.1177/014662169201600206

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 166–172. https://doi.org/10.3389/fpsyg.2014.00978

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333. https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Organization for Economic Co-operation and Development (OECD). (n.d.-a). *PISA participants.* http://www.oecd.org/pisa/aboutpisa/pisa-participants.htm

Organization for Economic Co-operation and Development (OECD). (n.d.-b). *What is PISA?* https://www.oecd.org/pisa/

Organization for Economic Co-operation and Development (OECD). (2011). *How do some students overcome their socio-economic background?* (PISA in Focus #5). https://www.oecd-ilibrary.org/education/how-do-some-students-overcome-their-socio-economic-background_5k9h362p77tf-en

Organization for Economic Co-operation and Development (OECD). (2017a). *PISA 2015 technical report*. https://www.oecd.org/pisa/data/2015-technical-report

Organization for Economic Co-operation and Development (OECD). (2017b). *PISA for Development* (PISA for Development Brief 19). https://www.oecd.org/pisa/19-Mainstreaming-PISA-D-into-PISA.pdf

Organization for Economic Co-operation and Development (OECD). (2018). *Equity in education: Breaking down barriers to social mobility*. https://doi.org/10.1787/9789264073234-en

Organization for Economic Co-operation and Development (OECD). (2019). *PISA 2021 context questionnaire framework (Field trial version)*. https://www.oecd.org/pisa/sitedocument/PISA-2021-questionnaire-framework.pdf

Organization for Economic Co-operation and Development (OECD). (2020). *PISA 2018 technical report*. https://www.oecd.org/pisa/data/pisa2018technicalreport

Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the cross-country comparability of indicators of socioeconomic resources in PISA. *Applied Measurement in Education, 30*(4), 243–258. https://doi.org/10.1080/08957347.2017.1353985

Rost, J., Carstensen, C., & von Davier, M. (1999). Sind die Big five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data]. *Diagnostica, 45*(3), 119-127. https://doi.org/10.1026/0012-1924.45.3.119

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278. https://doi.org/10.2304/rcie.2013.8.3.259

Sandoval-Hernandez, A., Rutkowski, D., Matta, T., & Miranda, D. (2019). Back to the drawing board: Can we compare socioeconomic background scales? *Revista de Educación, 383*, 37–61. https://doi.org/10.4438/1988-592X-RE-2019-383-400

Schneider, S. L., & Kogan, I. (2008). The International Standard Classification of Education 1997: Challenges in the application to national data and the implementation in cross-national surveys. In. S. L. Schneider (Ed.), *The International Standard Classification of Education 1997* (pp. 13–46). University of Mannheim. https://www.mzes.uni-mannheim.de/publications/misc/isced_97/schn08b_the_international_standard_classification_of_educa.pdf

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453. https://doi.org/10.3102/00346543075003417

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development* [General Assembly Resolution]. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

United Nations Development Programme (UNDP). (n.d.). *Human Development Index (HDI)*. http://hdr.undp.org/en/indicators/137506

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2016). *Education 2030: Incheon Declaration and Framework for Action for the implementation of Sustainable Development Goal 4*. http://uis.unesco.org/sites/default/files/documents/education-2030-incheon-framework-for-action-implementation-of-sdg4-2016-en_2.pdf

van de Vijver, F. (2019). Introduction. In F. van de Vijver (Ed.), *Invariance analyses in large-scale studies* (OECD Education Working Paper No. 201, pp. 9–12). OECD. https://doi.org/10.1787/19939019

von Davier, M. (2005). *Multidimensional discrete latent trait models (mdltm)* [Computer software].

von Davier, M., & Bezirhan, U. (2021, December 23). A robust method for detecting item misfit in large scale assessments. https://doi.org/10.31234/osf.io/mnsdg

von Davier, M., Rost, R., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 1-12). Springer.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236–244. https://doi.org/10.2307/2282967

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

World Bank. (n.d.). *GDP per capita, PPP*. https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?end=2015&name_desc=false&start=1960

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*(1). https://doi.org/10.7275/mhqa-cd89

Yamamoto, K. (1987). *A model that combines IRT and latent class models* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Yamamoto, K. (1989). *Hybrid model of IRT and latest class models.* (Research Report 89-41). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds)., *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). Waxmann Publishing.

Yang, Y., & Gustafsson, J. E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation, 10*(3), 259–288. https://doi.org/10.1076/edre.10.3.259.30268