Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 55-83

Improving measurement properties of the PISA home possessions scale through partial invariance modeling[1]

Selene S. Lee[2] & Matthias von Davier[3]

Abstract

This paper analyzes the longitudinal and cross-country measurement invariance of the home possessions scale, one of the three components used to measure socioeconomic status (SES) in the Programme for International Student Assessment (PISA). It finds that most of the items in the scale are invariant over time but not invariant across countries. Another finding is that using multiple-group concurrent calibration with partial invariance will make the home possessions scale more comparable over time and across countries, compared to the original home possessions scales that had been generated in each cycle. Moreover, using the new method based on the two-parameter logistic (2PL) model and the generalized partial credit model (GPCM) maintained, and in some cases, even improved the within-country accuracy of the scores, as indicated by an increased regression coefficient when the home possessions scores were used to predict the cognitive proficiency scores for reading, math, and science.

Keywords: Measurement invariance; Home possessions (HOMEPOS) scale; Socioeconomic status (SES); Programme for International Student Assessment (PISA); International large-scale assessment (ILSA)

[1] Please note that this paper is a shortened version of the first author's Ph.D. dissertation. The full dissertation can be found here: https://repository.upenn.edu/edissertations/3310/

[2] *Correspondence concerning this article should be addressed to:* Selene S. Lee, PhD, Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19103, USA; email: selene_lee@alumni.upenn.edu

[3] National Board of Medical Examiners (NBME), Philadelphia, USA

Introduction

In international comparative research in education, it is important to collect reliable and valid information on students' socioeconomic status (SES), because researchers use SES to contextualize the results of an assessment or use it as a control variable when analyzing the relationship between academic achievement and other variables (Sirin, 2005). In recent decades, international large-scale assessments, such as the Programme for International Student Assessment (PISA), have made it possible to conduct research on SES in a wide range of countries and also over time. However, as noted by Rutkowski and Rutkowski (2013), any cross-country and longitudinal comparisons using SES data from international large-scale assessments should be preceded by a careful examination of the psychometric properties of the scale used to measure SES, a topic which is rarely addressed by researchers. The current paper is designed to fill the gaps in this area of research by analyzing the measurement properties of the home possessions scale, one of the three components used to measure SES in PISA (along with parents' education and parents' occupation; OECD, 2017, p. 339). More specifically, this paper provides a careful examination of the measurement and invariance properties of the PISA home possession scale across countries and over cycles.

Information on home possessions (i.e., the items that a person owns at home) has often been used as a proxy for family wealth (Filmer & Scott, 2008; Montgomery, Gragnolati, Burke, & Paredes, 2000), due to the difficulties of measuring family wealth directly through surveys (Brese & Mirazchiyski, 2013; Tourangeau & Yan, 2007). However, there are challenges to using home possessions as a proxy for family wealth. For example, the ownership of an item does not convey information about the quality of the item that is owned (Falkingham & Namazie, 2002), how accessible the item is in a country due to economic and logistical reasons, or how valued it is due to sociocultural reasons (Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004). This touches upon the issue of measurement invariance which will be explored in detail in this paper.

Before comparing the results of analyses that use home possessions scores from different cycles of PISA or from different countries that participated in PISA, researchers need to make sure that the meaning of family wealth (i.e., the latent trait) is consistent across time and across countries. In item response theory (IRT), the method used to scale the home possessions scale in PISA, the meaning of the latent trait depends on the relationship between the latent trait and the items that are used to measure it. Thus, in order to make meaningful comparisons of the home possessions scores over time and across countries, the relationship between the items in the home possessions scale and family wealth should be consistent over time (i.e., longitudinal measurement invariance of the scale should be established) and across countries (i.e., cross-country measurement invariance of the scale should be established).

A few examples will be used to illustrate how the relationship between an item in the home possession scale and family wealth can vary over time and across countries: In 2000, cell phones were relatively expensive, so only people from wealthy families could afford it. However, by 2015, cell phones became much cheaper, making it more accessible to people from less wealthy families. Since the relationship between the ownership of a cell phone and family wealth changed over time, longitudinal measurement invariance cannot be established for this item. In the case of cross-country measurement invariance, in Vietnam, only people from relatively wealthy families tend to have a car, perhaps because most people do not need to travel far for their daily activities. In contrast, in the United States, even people from less wealthy families tend to own a car, due to the long distances they have to travel every day. Since the relationship between the ownership of a car and family wealth is different across countries, cross-country measurement invariance cannot be established for this item.

Methods to evaluate measurement invariance

Traditionally, measurement invariance has been tested using multiple-group confirmatory factor analysis (Jöreskog, 1971; Meredith, 1993). However, when many groups and items are included in the analysis, for example, when the analysis is conducted with data from international large-scale assessments, scalar or strong invariance (i.e., the same item characteristics – for example, the slope and intercept parameters in IRT analysis – can be assumed across all groups) is rarely achievable in practice (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Davidov, Muthén, & Schmidt, 2018; Marsh et al., 2018). In fact, in a study by Sandoval-Hernandez, Rutkowski, Matta, and Miranda (2019), only configural invariance (i.e., item characteristics differ across groups, while still the same conceptual definition of the latent construct is assumed) could be established for the Economic, Social and Cultural Status scale of PISA 2015 which was administered in 72 countries, as well as for the Home Educational Resources scale of the Trends in International Mathematics and Science Study (TIMSS) in 2015 which was administered in 44 countries.

Due to the difficulties of establishing scalar invariance in practice, in recent years, several alternative methods have been proposed to assess the measurement invariance of scales when many groups and items are included in the analysis. These methods assume that meaningful comparisons can still be made across groups when there are some violations of scalar invariance that threaten the equality of the measurement model across groups. One such method is multiple-group concurrent calibration with partial invariance constraints which establishes partial invariance across populations in the context of IRT. The basic form of this procedure allows some item parameters to vary if large deviations of item functions are detected (Glas & Jehangir, 2014; Glas & Verhelst, 1995; Oliveri & von Davier, 2011; von Davier & von Davier, 2007; Xu & von Davier, 2008; Yamamoto, 1998; Yamamoto, Khorramdel, & von Davier, 2013; Yamamoto & Mazzeo, 1992). Specifically, for each item and group, item fit statistics are computed to quantify the discrepancy between the observed item characteristic curve (ICC) for the group and the model-based ICC estimated with data from all groups. When the item fit statistic exceeds a certain threshold, and no error can be detected in the item content, scoring, and test administration, then differential item functioning (DIF) is assumed for the group – in other words, it is considered that measurement invariance cannot be established across groups for the item. Pokropek, Borgonovi, and McCormick (2017) used a similar method (along with two other methods) to examine the longitudinal measurement invariance of the

PISA home possessions scale between 2000 and 2012 as well as the cross-country measurement invariance of the scale in 2012.

It should be noted that even if measurement invariance cannot be established for the home possessions scale over time and across countries, it is still possible to impose constraints on the items in the scale so they have the same item parameters across time and across countries. This will ensure that the generated scale is consistent over time and across countries. However, since imposing equal item parameters will not accurately describe the relationship between the item and family wealth at each time point and within each country, the accuracy as well as the validity of the home possessions scores will be compromised for some time points and in some countries. In other words, there is a trade-off between the comparability of the scale across time and countries and the accuracy of the scores within each time point and country.

Operationally, different methods have been used over the PISA cycles to scale the home possessions scale. For example, in PISA 2009 and 2012, the IRT scaling was done separately for each country, and a common scale across countries was created through linear transformations of the generated scores. In earlier cycles, different scaling approaches were taken. A summary of the scaling methods for the home possessions scale for each cycle of PISA is presented in Appendix A, and more details can be found in the PISA technical reports (OECD, 2002; OECD, 2005; OECD, 2009; OECD, 2012; OECD, 2014, OECD, 2017). It should be noted that the home possessions scores from different PISA cycles are not comparable, since a different scaling method was used in each cycle.

Research questions

The current paper uses the methods described in von Davier et al. (2019) to analyze the longitudinal and cross-country measurement invariance of the PISA home possession scale. It expands the scope of the research conducted by Pokropek et al. (2017) by analyzing all six cycles of PISA for which data are publicly available (from 2000 to 2015) and by including the majority of countries that participated in these PISA cycles. The research questions for each of the four studies are presented below:

1) Study 1: Which items in the PISA home possessions scale demonstrate measurement invariance over multiple PISA cycles?
2) Study 2: Which items in the PISA home possessions scale demonstrate measurement invariance across the country-by-language groups?
3) Study 3: Within each cycle, are the new home possessions scores (generated from Study 2) more comparable across countries than the original home possessions scores (obtained from the public dataset)?
4) Study 4: Within each cycle, are the new home possessions scores (generated from Study 2) a more accurate measure of family wealth within countries compared to the original home possessions scores (obtained from the public dataset)?

Methods

Data

Data for this research were drawn from publicly available datasets of PISA, an international large-scale assessment coordinated by the Organization for Economic Cooperation and Development (OECD) to assess the knowledge and skills of 15-year-old students in reading, math, and science. PISA has been administered every three years since 2000 ("About PISA," n.d.). Participation in PISA is voluntary, and the number of participating countries has steadily increased over the years. In the first cycle of PISA administered in 2000, 29 OECD and 14 non-OECD countries participated ("PISA 2000," n.d.), but these numbers increased to 34 OECD and 38 non-OECD countries in the sixth cycle of PISA administered in 2015, the last cycle for which data are currently available ("PISA 2015," n.d.).

While the analyses tried to include as many countries as possible, there were some criteria for exclusion. For example, countries that had been excluded from the public dataset due to data adjudication issues, political issues, or other issues were naturally excluded from the analyses. In addition, data from samples that were not nationally representative, such as data from specific regions or cities within a country, were also excluded from the analyses. Lastly, since the datasets from 2000 and 2003 did not have information on the language of examination, while country-by-language groups (rather than countries) were used as the unit of analysis in Study 2, countries with multiple language groups (defined as a country in which a minority language was used as the language of examination by at least 5 % of the test takers, based on data from PISA 2006 and beyond) were excluded from the PISA 2000 and 2003 datasets in the current analyses. As a result, 10 countries (out of 43 countries) were excluded from the PISA 2000 dataset, while nine countries (out of 41 countries) were excluded from the PISA 2003 dataset used in the analyses. Appendix B lists the 75 countries included in the analyses as well as the unweighted sample size for each country.

In the datasets, the student weights were adjusted so that the sum of the student weights for each cycle would be equal, regardless of the number of countries that participated in each cycle. Furthermore, within each cycle, the sum of the student weights was adjusted to be equal for all countries, regardless of the size of the target population in each country. This weighting method ensured that each cycle contributed equally to the analyses, while in each cycle, each country contributed equally.

Measures

To measure the family wealth of students who participated in PISA, students were asked whether they owned or had access to certain items at home. This information was collected through the home possessions scale which was included in the student background questionnaire. The student background questionnaire was administered directly to students and was designed to take no longer than 35 minutes, with 30 minutes allocated to the international questionnaire and an additional five minutes for any country-specific questions (OECD, 2017, p. 36).

Some items in the home possessions scale were dichotomous (i.e., it asked whether the student's household owned the item or not), while other items had polytomous ordinal response options (i.e., it asked how many of the item the student's household owned). In almost every cycle of PISA, several items were added to or dropped from the home possessions scale, taking into account the social, economic, and technical changes in the participating countries (OECD, 2017, p. 341). Each country was also allowed to include up to three country-specific items in the scale, but these items were excluded from the analyses, because they were not administered in every country. Table 1 presents the items that were included in the home possessions scale in each cycle. For all items and cycles, data were missing for 5 % or less of the sample across the countries.

Analyses

Study 1: Measurement invariance across cycles. The purpose of Study 1 was to determine which items demonstrated measurement invariance across the six cycles of PISA for which data were publicly available. The operational method that was used to examine measurement invariance and item fit in PISA 2015 (OECD, 2017, p. 149) and the software mdltm (von Davier, 2005) for multidimensional discrete latent traits models was used for this study. The software provides marginal maximum likelihood (MML) estimates obtained using customary expectation-maximization (EM) methods with optional acceleration.

To scale the items, the two-parameter logistic (2PL) model (Birnbaum, 1968) was used for the dichotomous items, while the generalized partial credit model (GPCM; Muraki, 1992) was used for the polytomous items. All items were calibrated concurrently, placing them on a common scale measuring the latent trait, family wealth. Missing data, whether it was because an item was not administered in a cycle, a country did not participate in a cycle, or a student did not respond to an item, were treated as ignorable missing data (i.e., missing data that do not provide information with regard to the latent trait and, therefore, are not included in the likelihood estimation).

In line with the operational IRT scaling used in PISA 2015, DIF across cycles was detected using multiple-group concurrent calibration with partial invariance constraints (OECD, 2017, p. 150). More specifically, for the multiple-group IRT model, each cycle was defined as a group, so data from all countries that participated in a cycle were pooled together to form one group. For each item, common item parameters across all groups (cycles) were estimated, and a model-based ICC for each item was produced. Then, for each item-by-cycle combination, the root mean square deviation (RMSD) and the mean deviation (MD) were calculated. Both fit statistics quantify the difference between the model-based ICC for each item and the observed ICC. RMSD values are always positive or zero, and they indicate the root of the average squared differences between the model-based ICC and the observed ICC, weighted by the proficiency distribution. A higher

Table 1:
Items Included in the PISA Home Possessions Scale

	PISA Cycle						# of cycles
	2000	2003	2006	2009	2012	2015	
Desk to study at	✓	✓	✓	✓	✓	✓	6
Room of your own	✓	✓	✓	✓	✓	✓	6
Quiet place to study	✓	✓	✓	✓	✓	✓	6
Computer you can use for school work		✓	✓	✓	✓	✓	5
Educational software	✓	✓	✓	✓	✓	✓	6
Link to the internet	✓	✓	✓	✓	✓	✓	6
Classic literature	✓	✓	✓	✓	✓	✓	6
Books of poetry	✓	✓	✓	✓	✓	✓	6
Works of art	✓	✓	✓	✓	✓	✓	6
Books to help with your school work	✓	✓	✓	✓	✓	✓	6
Technical reference books				✓	✓	✓	3
Dictionary	✓	✓	✓	✓	✓	✓	6
Books on art, music or design						✓	1
Your own calculator	✓	✓	✓				3
Dishwasher	✓	✓	✓	✓	✓		5
DVD player				✓	✓		2
Television *	✓		✓	✓	✓	✓	5
Car *	✓		✓	✓	✓	✓	5
Room with a bath or shower *	✓			✓	✓	✓	4
Cellular phone *	✓		✓	✓	✓	✓	5
Computer *	✓		✓	✓	✓	✓	5
Tablet computer *						✓	1
E-book reader *						✓	1
Musical instrument *	✓					✓	2
Books * [a]		✓	✓	✓	✓	✓	5

Note. Polytomous items are indicated with an asterisk. Except for the number of books (the last item), the response categories for the polytomous items were zero, one, two, and three or more.

[a] The response categories for this item were zero to 10, 11 to 25, 26 to 100, 101 to 200, 201 to 500, and more than 500.

value of RMSD indicates a higher level of misfit between the model-based ICC and the observed ICC. The MD is similar to RMSD in that it quantifies the difference between the model-based ICC and the observed ICC. However, MD also takes into account the direction of the deviation, so values can be either positive or negative, with values further from zero indicating a larger misfit between the model-based ICC and the observed ICC.

In the first round of the item calibration, item-by-cycle combinations with an RMSD value greater than 0.40 were considered to exhibit substantial misfit, which means that the data for these item-by-cycle combinations were not sufficiently described by the common item parameters. In the subsequent rounds of item calibration, unique item parameters were estimated for these item-by-cycle combinations showing misfit. The process of assigning unique item parameters to item-by-cycle combinations that exhibited DIF was repeated using progressively lower RMSD cut-off values until all item-by-cycle combinations had RMSD values at or below 0.15. While the operational procedure for the PISA 2015 background questionnaire used an RMSD cut-off of 0.30 to detect DIF (OECD, 2017, p. 296), this study used a lower RMSD cut-off in order to identify more item-by-cycle combinations that exhibited DIF. Specifically, the cut-off used in this study was the criterion used for comparing different scaling methods in preparation for PISA 2015 (OECD, 2017, p. 174), and it is close to the RMSD cut-off of 0.12 used for detecting DIF in the PISA 2015 cognitive items (OECD, 2017, p. 151). A fully automated algorithm for this method was implemented in the software mdltm (von Davier, 2005) which was applied for the analysis of the Programme for the International Assessment of Adult Competencies (PIAAC) data as well as for PISA 2015 and 2018. Details on the application of this method to cognitive data from PISA cycles 2000 to 2012 is explained in von Davier et al. (2019).

At the end of the process, each item's model-based ICC for each cycle adequately fit the observed ICC for the cycle. This essentially created a partial invariance model (Byrne, Shavelson, & Muthén, 1989) in which for most items, all cycles were constrained to have the same item parameters, while for some items, certain cycles were assigned unique (cycle-specific) item parameters.

Study 2: Measurement invariance across country-by-language groups. The purpose of Study 2 was to determine which items demonstrated measurement invariance across the country-by-language groups. To take into account DIF across cycles, the final model of Study 1 was used as the initial model of Study 2 by making a new column in the dataset for the item-by-cycle combination that had exhibited DIF in Study 1. This is in line with the explanation by Glas and Jehangir (2014) that defining group-specific item parameters (or cycle-specific item parameters in this study) is equivalent to defining an incomplete design in which the item exhibiting DIF across groups is split into a number of virtual items, and each virtual item is considered to have been administered to a specific group. Again, the software mdltm (von Davier, 2005) was used for the analysis.

In Study 2, students within countries were grouped by the language in which they took the PISA cognitive assessment, because it was hypothesized that the relationship between an item and family wealth depended on sociocultural factors which were partially captured by the language of examination. Languages that were used as the language of examination by at least 5 % of the test takers in the country were considered to be independent country-by-language groups, while languages that were used as the language of examination by less than 5 % of the test takers were combined with the majority language group of the country. This created a total of 96 country-by-language groups.

The procedure for detecting DIF across groups and assigning unique item parameters to groups exhibiting DIF was the same as in Study 1, but in Study 2, the country-by-language groups (rather than the cycles) were defined as the groups in the multiple-group IRT model.

Study 3: Cross-country comparability of the new home possessions scale. The purpose of Study 3 was to assess whether, within each cycle, the new home possessions scores (generated from Study 2) were more comparable across countries than the original home possessions scores (obtained from the public dataset). To assess this, the cross-country variation in the relationship between the new home possessions scale and an external criterion was compared to the cross-country variation in the relationship between the original home possessions scale and the same external criterion. A lower variation in the former would imply that the new home possessions scale is more similar across countries than the original home possessions scale.

The external criterion that was chosen for this study was SES, since the relationship between the home possessions scale and SES could be measured by the component loadings of the home possessions scale on the first principal component, representing SES, in a principal component analysis (PCA) conducted with the home possessions scale, parents' education, and parents' occupation (the three components used to measure SES in PISA; OECD, 2017, p. 339). For each cycle and country, PCAs were conducted twice – first using the original home possessions scores from the public dataset, then using the new home possessions scores from Study 2.

To allow for comparisons across cycles, only the 50 countries that participated in all cycles of PISA from 2006 to 2015 were included in the study. Data from 2000 and 2003 were not included, because it would have reduced the number of countries included in the analysis to 26. Stata (version 15) was used for this study.

Study 4: Within-country accuracy of the new home possessions scale. The purpose of Study 4 was to assess whether, within each cycle and country, the new home possessions scores were a more accurate measure of family wealth than the original home possessions scores. As mentioned above, there is a trade-off between the comparability of the scale across countries and the accuracy of the scores within countries. Therefore, constraining the items to have the same item parameters across most of the country-by-language groups through concurrent calibration may have decreased the accuracy of the home possessions scores within countries, compared to when all countries were allowed to estimate their own item parameters for each item.

To assess the accuracy of the home possessions scores within countries, for each cycle and country, bivariate linear regressions were run twice to predict students' PISA scores in reading, math, and science – first using the original home possessions scores from the public dataset, then using the new home possessions scores from Study 2. Since many studies have found that family wealth was a statistically significant predictor of students' academic achievement (Sirin, 2005), if the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores, this was taken as evidence that the new home possessions scale was more accurate in measuring students' family wealth than the original home possessions scale.

For reasons explained above, only the 50 countries that participated in all cycles of PISA from 2006 to 2015 were included in the study. The International Database Analyzer (IDB Analyzer; version 4.0.23) developed by the International Association for the Evaluation of Educational Achievement (IEA) was used for the analysis.

Results

Study 1: Measurement invariance across cycles

Longitudinal measurement invariance could not be established for four items – educational software, internet, cell phone, and computer (as a polytomous item). For most of these items, the item discrimination parameter (α) decreased, indicating that the relationship between the item and family wealth became weaker over time, or the item endorsement parameter (β) decreased, implying that the item became more accessible over time. This is not surprising, considering that technological advances made these items more affordable with time. To assess whether the results had been affected by the differences in the countries that participated in each cycle of PISA, the analysis was conducted again with only the 26 countries that participated in all six cycles of PISA, but only the four items mentioned above exhibited DIF across cycles.

Appendix C presents the item discrimination parameter (α), the item endorsement parameter (β), and the step endorsement parameters (δ_j) for all the items in the home possessions scale at the end of Study 1.

Study 2: Measurement invariance across country-by-language groups

Results by item. Figure 1 presents, for each item, the percent of country-by-language groups that required unique item parameters.

Items for which over 50 % of the country-by-language groups required unique item parameters included bathroom, classic literature, poetry books, TV, and dishwasher. The high percentage of country-by-language groups that required unique item parameters for these items indicated that the relationship between these items and family wealth varied across the country-by-language groups.

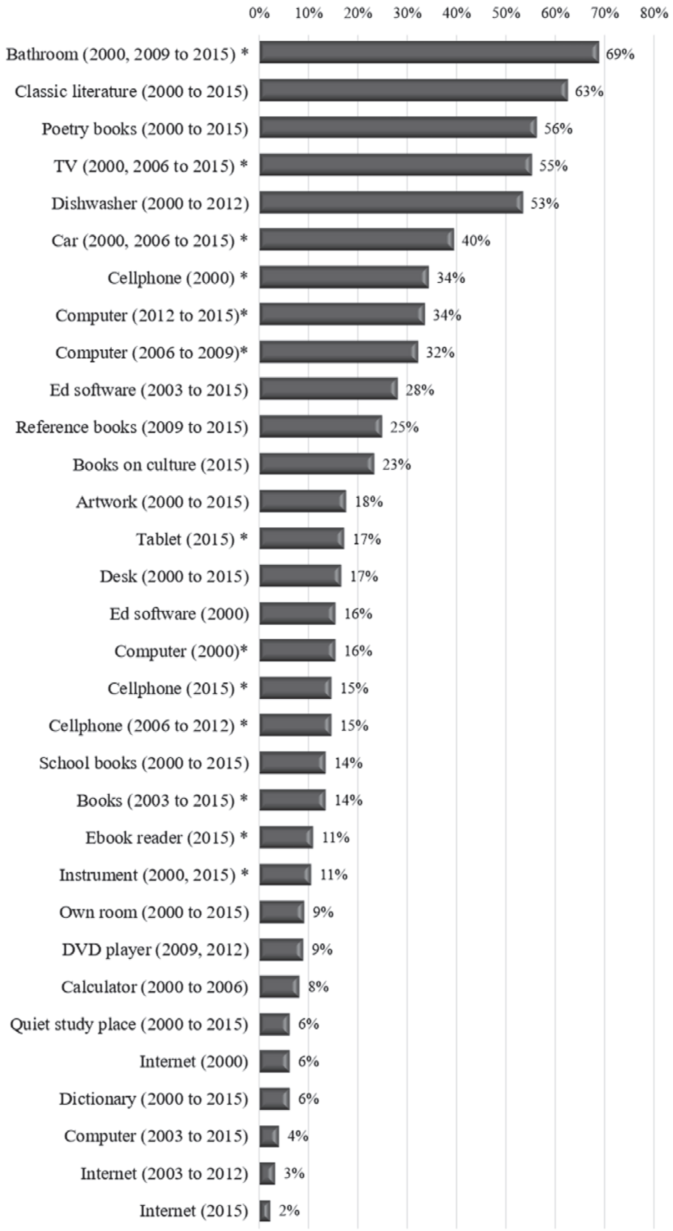

Figure 1:
Percent of country-by-language groups that required unique item parameters, by item. For each item, each cycle that required unique item parameters in Study 1 was counted as a separate item. Note that not all items were administered in every cycle of PISA. Polytomous items are indicated with an asterisk.

Items for which under 10 % of the country-by-language groups required unique item parameters included internet, computer (as a dichotomous item), dictionary, quiet study place, calculator, DVD player, and own room. The low percentage of country-by-language groups that required unique item parameters for these items indicated that the relationship between these items and family wealth were similar across the country-by-language groups.

Results by country-by-language group. Appendix D presents, for each country-by-language group, the percent of items in the scale that required unique item parameters in Study 2. Country-by-language groups that required unique item parameters for over 50 % of the items included Kyrgyzstan (Uzbek), Qatar (Arabic), Kyrgyzstan (Kyrgyz), Kyrgyzstan (Russian), and the United Arab Emirates (Arabic). The high percentage of items that required unique item parameters in these country-by-language groups indicated that for many of the items in the scale, the relationship between the item and family wealth in these country-by-language groups was different from the relationship between the item and family wealth in the other country-by-language groups. In other words, in these country-by-language groups, a high level of overall misfit was found in the scale.

Country-by-language groups that required unique item parameters for less than 10 % of the items included Iceland (Icelandic), Croatia (Croatian), Greece (Greek), Germany (German), Spain (Spanish, Galician, Valencian, Basque), Spain (Catalan), Slovak Republic (Slovak), Luxembourg (German, English), Portugal (Portuguese), Mexico (Spanish), Hungary (Hungarian), Brazil (Portuguese), and Austria (German, English). The low percentage of items that required unique item parameters in these country-by-language groups indicated that for many of the items in the scale, the relationship between the item and family wealth in these country-by-language groups was similar to the relationship between the item and family wealth in the other country-by-language groups. In other words, in these country-by-language groups, a low level of overall misfit was found in the scale.

Study 3: Cross-country comparability of the new home possessions scale

Figure 2 presents, for each cycle, the standard deviation of countries' component loadings for home possessions, parents' education, and parents' occupation on SES, the first principal component of the PCA.

For all of the cycles included in the study, the standard deviation of countries' component loadings for the home possessions scale on SES was lower when the new home possessions scores were used in the PCA instead of the original home possessions scores. This indicated that the relationship between the home possessions scale and SES was more consistent across countries when the new home possessions scores were used in the PCA, implying that the new home possessions scale was more similar across countries compared to the original home possessions scales generated in each cycle (the methods of which are presented in detail in Appendix A).

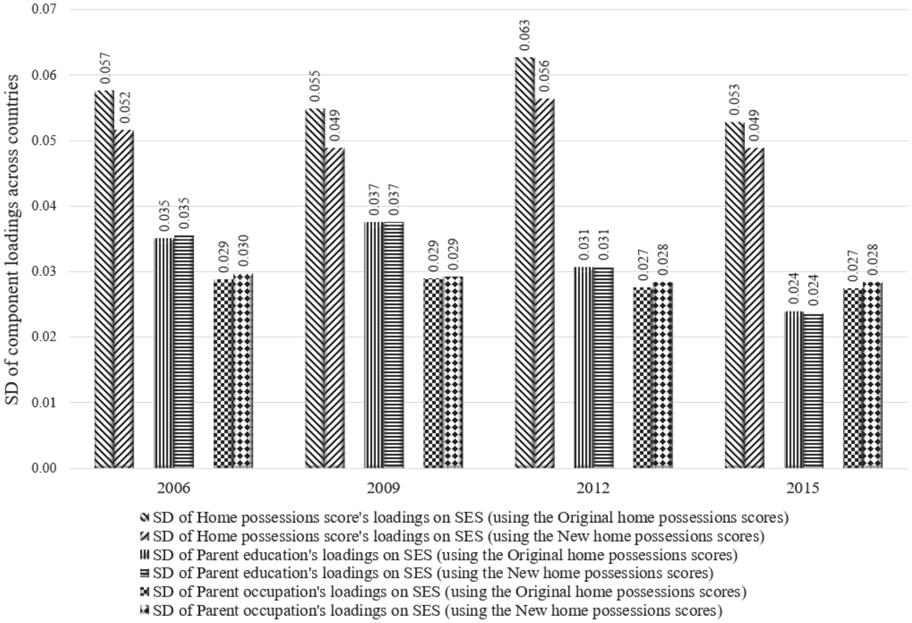

Figure 2:
Standard deviation of countries' component loadings for home possessions, parents' education, and parents' occupation on SES, using the original and new home possessions scores.

For the cycles from 2006 to 2012, the increase in the comparability of the new home possessions scale across countries may have been due to the fact that in the new scale, the default was to constrain the item parameters to be equal across the country-by-language groups, and only the item-by-country-by-language-group combinations for which the observed ICC exhibited substantial misfit with the international ICC were allowed to estimate unique item parameters. This is in contrast to the original method which estimated item parameters separately for each country.

For the 2015 cycle, the new home possessions scale was also more comparable across countries than the original home possessions scale. This was unexpected, because in 2015, the original home possessions scale used an RMSD cut-off of 0.3 to assign unique item parameters to item-by-country-by-language-group combinations, which is higher than the cut-off of 0.15 used in this study, so more item-by-country-by-language-group combinations were constrained to have the same item parameters in the original home possessions scale than the new home possessions scale.

Study 4: Within-country accuracy of the new home possessions scale

Figure 3 presents, for each cycle, the average r^2 across countries of the bivariate regressions predicting students' scores on the PISA math, science, and reading assessments with the original and new home possessions scores.

For the cycles from 2006 to 2012, for all three domains, the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores. This implied that within countries, the new home possessions scores were a more accurate measure of family wealth than the original home possessions scores. Although the accuracy of the new home possessions scores within countries may have decreased for these cycles because the item parameters were constrained to be equal across the country-by-language groups by default (instead of having the item parameters estimated separately for each country), using the 2PL model and the GPCM to calibrate the items (instead of the Rasch model [Rasch, 1960] and the partial credit model [PCM; Masters, 1982], respectively) may have counter-balanced this by increasing the accuracy of the new home possessions scores within countries.

For the 2015 cycle, for all three domains, the new home possessions scores explained as much of the variation in the PISA cognitive scores as the original home possessions scores. This implied that within countries, the new home possessions scores were as

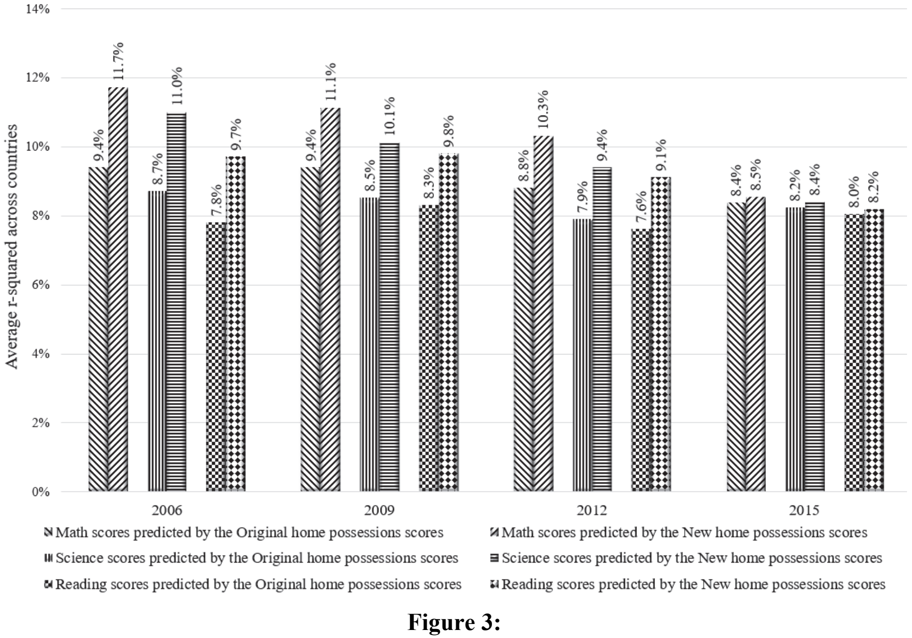

Figure 3:
Average r^2 across countries of the bivariate regressions predicting students' scores on the PISA cognitive assessments with the original and new home possessions scores.

accurate a measure of family wealth as the original home possessions scores. While the accuracy of the new home possessions scores for 2015 may have increased because more item-by-country-by-language-group combinations were allowed to estimate unique item parameters than in the original home possessions scale, this may have been balanced out by calibrating the items parameters with data from all cycles instead of data only from the 2015 cycle.

The results of this study provide evidence that the new home possessions scores are at least as accurate as the original home possessions scores in measuring family wealth within countries, even though the new home possessions scale is a more comparable measure of family wealth across countries than the original home possessions scale. In other words, the accuracy of the scores within countries was not compromised by the increased comparability of the scale across countries.

Summary and conclusion

The home possessions scale was first developed in the PISA 2000 cycle when only 43 countries participated, of which two-thirds were members of the OECD. Very little changes have been made to the scale since then, even though in the most recent cycle of PISA in 2018, 79 countries participated, of which less than half were members of the OECD. With more countries planning to participate in future cycles of PISA, the participating countries will become increasingly heterogeneous, presenting further challenges to the comparability of the home possessions scale over time and across countries. The current paper aimed to assess and improve the longitudinal and cross-country comparability of the home possessions scale, a topic which has been rarely addressed by researchers until now, but is important when comparing the differences in the relationship between family wealth and educational achievement across time and across countries.

Study 1 found that among the 25 items included in the home possessions scale, longitudinal measurement invariance could not be established for only four items, all related to technology – educational software, internet, cell phone, and computer (as a polytomous item). These results are important, because it revealed that most of the items in the scale were in fact invariant (i.e., comparable) over time. This means that it is possible to create a home possessions scale that is longitudinally comparable, which will allow researchers to compare the results of analyses that use home possessions scores from different cycles of PISA.

In Study 2, it was found that for some items in the scale (i.e., bathroom, classic literature, poetry books, TV, and dishwasher), the relationship between the item and family wealth was heterogenous across the country-by-language groups, while for other items (i.e., internet, computer [as a dichotomous item], dictionary, and quiet place to study), the relationship between the item and family wealth was relatively similar across the country-by-language groups. Excluding the former items from the scale would improve the comparability of the scale across countries, which would allow researchers to make more appropriate comparisons of the home possessions scores from different countries. Study 2 also found that in some country-by-language groups, the relationship between many of the items in the scale and family wealth was different from the relationship between these items and family wealth in the other country-by-language groups. More research should be conducted on why such a high level of overall misfit was found in the scale in these country-by-language groups and what other items can be included in the scale to make the scale more comparable across different country-by-language groups.

Study 3 found that for all cycles included in the study, the new home possessions scale was more comparable across countries than the original home possessions scale. The increase in the cross-country comparability of the new home possessions scale may have been a result of constraining the item parameters to be equal across the country-by-language groups by default and only allowing the item-by-country-by-language-group combinations for which the observed ICC exhibited substantial misfit with the international ICC to estimate unique item parameters.

In Study 4, it was found that for most of the cycles included in the study, the new home possessions scores were a more accurate measure of family wealth within countries than the original home possessions scores, even though some of the methods used to increase the comparability of the home possessions scale across countries decreased the accuracy of the scores within countries. The increase in the accuracy of the new home possessions scores within countries may have been a result of using the 2PL model and the GPCM (instead of the Rasch model and the PCM, respectively) to calibrate the items in conjunction with multiple-group concurrent calibration with partial invariance constraints.

In sum, the paper found that most of the items in the scale were invariant over time but not invariant across countries, and it also identified items that should be excluded from the scale to increase the cross-country comparability of the scale. In addition, the paper found that using multiple-group concurrent calibration with partial invariance constraints can make the scale more comparable over time and across countries, compared to the original home possessions scales generated in each cycle, by utilizing common parameters for most item-by-group combinations. Of course, when this method is used, the within-country accuracy of the scores is reduced compared to a scaling model which assigns only unique item parameters and no common item parameters to countries. (Note that this latter model would not produce scores that are comparable across countries.) However, by assigning unique item parameters only to country-by-language groups that exhibit substantial misfit with the common parameters, multiple-group concurrent calibration with partial invariance constraints obtains the highest possible within-country accuracy while also maintaining the highest possible cross-country comparability. In other words, the approach finds the best possible balance between comparability and accuracy of the scores.

Limitations

As with any research, the findings of this paper should be interpreted in light of its limitations. The first limitation is that the language of examination was used to divide students within a country into subgroups, based on the assumption that the relationship between an item and family wealth may be different for different sociocultural groups

(Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004). However, if the language in which a student is assessed is not reflective of the sociocultural group to which the student belongs, using the language of examination will not be an effective way to divide the population into different sociocultural groups.

Second, all countries with a sizeable minority language population were excluded from the datasets for 2000 and 2003 used in the analyses. Nevertheless, for most of the items, the item parameters estimated with data from all countries that participated in 2000 and 2003 were very similar to the item parameters estimated with data excluding countries with a sizeable minority language population from these cycles, implying that excluding these countries from the dataset in 2000 and 2003 did not have a large effect on the observed and model-based ICCs for 2000 and 2003.

Third, assigning unique item parameters to certain item-by-country-by-language-group combinations may have resulted in parameters that conformed to the random variability of the sample, especially for small samples. However, considering that all of the country-by-language groups had a sample size of at least 1,000 students, with the exception of the Azerbaijan-Russian speaking group (which had a sample size of 473 students), it was assumed that the country-by-language groups were large enough to make the item parameter estimates robust to the idiosyncrasies of the samples (Lim & Drasgow, 1990).

In spite of these limitations, this paper provided many insights into the longitudinal and cross-country measurement invariance of the PISA home possessions scale. The results of this paper can help improve the comparability of the PISA home possessions scale over time and across countries so that researchers can more appropriately compare the results of analyses that use the home possessions scores from different cycles of PISA and from different countries that participated in PISA.

Acknowledgement

We would like to thank Lale Khorramdel, Robert Boruch, Paul McDermott, Michael Rovine, and two anonymous reviewers for providing constructive feedback on earlier drafts of the manuscript.

References

About PISA. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/

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). Reading, MA: Addison-Wesley.

Brese, F., & Mirazchiyski, P. (2013). *Measuring students' family background in large-scale international education studies* (IERI monograph series - Special issue 2). Princeton, NJ: IEA-ETS Research Institute (IERI).

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. doi: 10.1037/0033-2909.105.3.456

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*. 55-75. doi: 10.1146/annurev-soc-071913-043137

Davidov, E., Muthén, B., & Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods & Research, 47*(4), 631-636. doi:10.1177/0049124118789708

Falkingham, J., & Namazie, C. (2002). *Measuring health and poverty: A review of approaches to identifying the poor*. London, UK: DFID Health Systems Resource Centre.

Filmer, D., & Scott, K. (2008). *Assessing asset indices* (Policy Research Working Paper No. 4605). Washington, DC: World Bank.

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). Boca Raton, FL: CRC Press.

Glas, C. A., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-95). New York, NY: Springer.

IDB Analyzer [Computer software]. Hamburg, Germany: IEA. Downloaded from the IEA website: https://www.iea.nl/data

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

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. doi: 10.1037/0021-9010.75.2.164

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524-545. doi:10.1037/met0000113

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. doi: 10.1007/BF02296272

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi: 10.1007/BF02294825

Montgomery, M. R., Gragnolati, M., Burke, K. A., & Paredes, E. (2000). Measuring living standards with proxy variables. *Demography, 37*(2), 155-174. doi: 10.2307/2648118

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

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.

Organization for Economic Co-operation and Development (OECD). (2002). *PISA 2000 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2012). *PISA 2009 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2014). *PISA 2012 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2017). *PISA 2015 technical report*. Paris, France: OECD.

PISA 2000 list of participating countries/economies. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/pisa2000listofparticipatingcountrieseconomies.htm

PISA 2015 list of participating countries/economies. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/pisa2000listofparticipatingcountrieseconomies.htm

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. doi: 10.1080/08957347.2017.1353985

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

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. doi: 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. doi:10.4438/1988-592X-RE-2019-383-400

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

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. doi: 10.1037/0033-2909.133.5.859

von Davier, M. (2005). Multidimensional discrete latent trait models (mdltm) [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*(3), 115-124. doi: 10.1027/1614-2241.3.3.115

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000-2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466-488. doi: 10.1080/0969594X.2019.1586642

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series, 2008*(1), 1-18. doi: 10.1002/j.2333-8504.2008.tb02113.x

Yamamoto, K. (1998). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey* (pp. 161-178), Washington, DC: National Center for Education Statistics.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Chapter 17: Scaling PIAAC cognitive data. In OECD (Ed.) *Technical report of the survey of adult skills (PIAAC)* (pp. 408-440). Paris, France: OECD.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155-173. doi: 10.2307/1165167

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

Appendix A:
Original Scaling Methods for the PISA Home Possessions Scale

PISA Cycle	Scaling Method
2000	The Rasch model was used to scale the dichotomous items, while the PCM was used to scale the polytomous items. Instead of generating a single score for home possessions, separate scores were generated for household wealth, cultural possessions, and home educational resources. The item endorsement parameter (β) for each item was estimated on the combined OECD sample (OECD, 2002).
2003	The Rasch model was used to scale the dichotomous items, while the PCM was used to scale the polytomous items. A single score was generated for home possessions, again estimating the item endorsement parameters (β) on the combined OECD sample (OECD, 2005).
2006	The Rasch model was used to scale the dichotomous items, while the PCM was used to scale the polytomous items. The item endorsement parameters (β) were estimated separately for each country (OECD, 2009).
2009	The Rasch model was used to scale the dichotomous items, while the PCM was used to scale the polytomous items. The item endorsement parameters (β) were estimated within each country using data from all cycles the country had participated in, with each cycle weighted equally. Subsequently, a linear transformation was applied to each country's parameters to place them on a common scale (OECD, 2012).
2012	The Rasch model was used to scale the dichotomous items, while the PCM was used to scale the polytomous items. The item endorsement parameters (β) were estimated separately for each country using data from all previous cycles. The relative position of each country was estimated on a joint scale (OECD, 2014).
2015	The 2PL model was used to scale the dichotomous items, while the GPCM was used to scale the polytomous items, resulting in a discrimination parameter (α) and an endorsement parameter (β) for each item. These parameters were estimated using data only from the 2015 cycle. To address DIF across countries, multiple group concurrent calibration with partial invariance constraints was used, allowing an item-by-country-by-language group combination to estimate its own item parameters if the RMSD was over 0.3 (OECD, 2017).

Appendix B:
Countries Included in the Analyses

Country	Sample Size (Unweighted)							# of Cycles
	2000	2003	2006	2009	2012	2015	Total	
Albania	2,783			4,596	4,743		12,122	3
Algeria						5,519	5,519	1
Argentina	2,230		4,339	4,774	5,908		17,251	4
Australia	2,859	12,551	14,170	14,251	14,481	14,530	72,842	6
Austria	2,640	4,597	4,927	6,590	4,755	7,007	30,516	6
Azerbaijan			5,184	4,691			9,875	2
Belgium			8,857	8,501	8,597	9,651	35,606	4
Brazil	2,717	4,452	9,295	20,127	19,204	23,141	78,936	6
Bulgaria	2,615		4,498	4,507	5,282	5,928	22,830	5
Canada			22,646	23,207	21,544	20,058	87,455	4
Chile	2,721		5,233	5,669	6,856	7,053	27,532	5
Colombia			4,478	7,921	9,073	11,795	33,267	4
Costa Rica				4,578	4,602	6,866	16,046	3
Croatia			5,213	4,994	5,008	5,809	21,024	4
Czech Republic	3,066	6,320	5,932	6,064	5,327	6,894	33,603	6
Denmark	2,382	4,218	4,532	5,924	7,481	7,161	31,698	6
Dominican Republic						4,740	4,740	1
Estonia			4,865	4,727	4,779	5,587	19,958	4
Finland			4,714	5,810	8,829	5,882	25,235	4
France	2,597	4,300	4,716	4,298	4,613	6,108	26,632	6
Georgia				4,646		5,316	9,962	2
Germany	2,830	4,660	4,891	4,979	5,001	6,504	28,865	6
Greece	2,605	4,627	4,873	4,969	5,125	5,532	27,731	6
Hungary	2,799	4,765	4,490	4,605	4,810	5,658	27,127	6
Iceland	1,882	3,350	3,789	3,646	3,508	3,371	19,546	6
Indonesia	4,089	10,761	10,647	5,136	5,622	6,513	42,768	6
Ireland	2,128	3,880	4,585	3,937	5,016	5,741	25,287	6
Israel			4,584	5,761	5,055	6,598	21,998	4
Italy	2,765	11,639	21,773	30,905	31,073	11,583	109,738	6
Japan	2,924	4,707	5,952	6,088	6,351	6,647	32,669	6
Jordan			6,509	6,486	7,038	7,267	27,300	4
Kazakhstan				5,412	5,808		11,220	2

Country	Sample Size (Unweighted)							# of Cycles
	2000	2003	2006	2009	2012	2015	Total	
Korea (South)	2,769	5,444	5,176	4,989	5,033	5,581	28,992	6
Kosovo						4,826	4,826	1
Kyrgyzstan			5,904	4,986			10,890	2
Latvia			4,719	4,502	4,306	4,869	18,396	4
Lebanon						4,546	4,546	1
Liechtenstein	175	332	339	329	293		1,468	5
Lithuania			4,744	4,528	4,618	6,525	20,415	4
Luxembourg			4,567	4,622	5,258	5,299	19,746	4
Macedonia						5,324	5,324	1
Malaysia				4,999	5,197		10,196	2
Malta				3,453		3,634	7,087	2
Mauritius				4,654			4,654	1
Mexico	2,567	29,983	30,971	38,250	33,806	7,568	143,145	6
Moldova				5,194		5,325	10,519	2
Montenegro			4,455	4,825	4,744	5,665	19,689	4
Netherlands	1,382	3,992	4,871	4,760	4,460	5,385	24,850	6
New Zealand	2,048	4,511	4,823	4,643	4,291	4,520	24,836	6
Norway	2,307	4,064	4,692	4,660	4,686	5,456	25,865	6
Panama				3,969			3,969	1
Peru	2,460			5,985	6,035	6,971	21,451	4
Poland	1,976	4,383	5,547	4,917	4,607	4,478	25,908	6
Portugal	2,545	4,608	5,109	6,298	5,722	7,325	31,607	6
Puerto Rico						1,398	1,398	1
Qatar			6,265	9,078	10,966	12,083	38,392	4
Romania			5,118	4,776	5,074	4,876	19,844	4
Russia	3,719	5,974	5,799	5,308	5,231	6,036	32,067	6
Serbia			4,798	5,523	4,684		15,005	3
Singapore				5,283	5,546	6,115	16,944	3
Slovak Republic			4,731	4,555	4,678	6,350	20,314	4
Slovenia			6,595	6,155	5,911	6,406	25,067	4
Spain			19,604	25,887	25,313	6,736	77,540	4
Sweden	2,464	4,624	4,443	4,567	4,736	5,458	26,292	6
Switzerland			12,192	11,812	11,229	5,860	41,093	4
Taiwan			8,815	5,831	6,046	7,708	28,400	4
Thailand	2,959	5,236	6,192	6,225	6,606	8,249	35,467	6

Country	Sample Size (Unweighted)							# of Cycles
	2000	2003	2006	2009	2012	2015	Total	
Trinidad and Tobago				4,778		4,692	9,470	2
Tunisia		4,721	4,640	4,955	4,407	5,375	24,098	5
Turkey		4,855	4,942	4,996	4,848	5,895	25,536	5
United Arab Emirates				10,867	11,500	14,167	36,534	3
United Kingdom	5,195	9,535	13,152	12,179	12,659	14,157	66,877	6
USA	2,135	5,456	5,611	5,233	4,978	5,712	29,125	6
Uruguay		5,835	4,839	5,957	5,315	6,062	28,008	5
Vietnam					4,959	5,826	10,785	2
Total	**83,333**	**188,380**	**389,345**	**492,327**	**463,231**	**456,917**	**2,073,533**	
Total # of countries	**32**	**30**	**55**	**68**	**61**	**65**	**75**	

Appendix C:

Final Item Parameters for Study 1

Item (Cycles)	Item discrimination (α)	Item endorsement (β)	Step endorsement (δ_j)				
Desk (2000 to 2015)	0.82	-1.57					
Own room (2000 to 2015)	0.64	-1.26					
Quiet study place (2000 to 2015)	0.63	-1.75					
Computer (2003 to 2015)	2.96	-0.48					
Ed software							
Ed software (2000)	1.53	0.30					
Ed software (2003 to 2015)	0.89	0.29					
Internet							
Internet (2000)	2.06	0.56					
Internet (2003 to 2012)	2.42	-0.24					
Internet (2015)	2.20	-0.74					
Classic literature (2000 to 2015)	0.35	0.00					
Poetry books (2000 to 2015)	0.28	-0.13					
Artwork (2000 to 2015)	0.60	-0.07					
School books (2000 to 2015)	0.44	-2.22					
Reference books (2009 to 2015)	0.65	-0.05					
Dictionary (2000 to 2015)	0.74	-2.27					
Books on culture (2015)	0.61	0.08					
Calculator (2000 to 2006)	0.93	-1.82					
Dishwasher (2000 to 2012)	0.85	0.15					
DVD player (2009, 2012)	0.92	-1.35					
TV (2000, 2006 to 2015) *	0.59		-3.09	-0.40	0.35		
Car (2000, 2006 to 2015) *	0.74		-0.52	0.58	1.44		
Bathroom (2000, 2009 to 2015) *	0.72		-1.57	0.89	1.68		
Cellphone *							
Cellphone (2000) *	0.68		-0.01	0.41	0.56		
Cellphone (2006 to 2012) *	0.67		-1.65	-0.77	-1.49		
Cellphone (2015) *	0.73		-1.19	-0.22	-0.88		
Computer *							
Computer (2000) *	2.00		0.07	1.02	1.38		
Computer (2006 to 2009) *	1.95		-0.54	0.61	1.02		
Computer (2012 to 2015) *	1.55		-0.72	0.23	0.63		
Tablet (2015) *	0.63		0.15	1.22	1.07		
Ebook reader (2015) *	0.48		2.51	2.52	1.51		
Instrument (2000, 2015) *	0.43		0.82	1.34	0.61		
Books (2003 to 2015) *	0.29		-0.48	-0.73	1.62	1.22	1.89

Note. Not all items were administered in every cycle of PISA. Polytomous items are indicated with an asterisk.

Appendix D:
Percent of Items that Required Unique Item Parameters, by Country-by-Language Group

Country (Language)	# of items administered to the group	# of items that required unique item parameters	% of items that required unique item parameters
Albania (Albanian)	27	7	26
Algeria (Arabic)	22	4	18
Argentina (Spanish)	27	4	15
Australia (English)	32	8	25
Austria (German, English)	32	3	9
Azerbaijan (Azerbaijani) *	21	7	33
Azerbaijan (Russian) *	21	10	48
Belgium (Dutch, German) *	28	5	18
Belgium (French) *	28	4	14
Brazil (Portuguese)	32	3	9
Bulgaria (Bulgarian)	32	6	19
Canada (English) *	28	9	32
Canada (French) *	28	8	29
Chile (Spanish)	32	4	13
Colombia (Spanish)	28	8	29
Costa Rica (Spanish)	27	5	19
Croatia (Croatian)	28	1	4
Czech Republic (Czech)	32	5	16
Denmark (Danish)	32	10	31
Dominican Republic (Spanish)	22	8	36
Estonia (Estonian) *	28	3	11
Estonia (Russian) *	28	7	25
Finland (Finnish) *	28	7	25
Finland (Swedish) *	28	6	21
France (French)	32	6	19
Georgia (Georgian, Azerbaijani, Russian)	27	10	37
Germany (German)	32	2	6
Greece (Greek)	32	2	6

Country (Language)	# of items administered to the group	# of items that required unique item parameters	% of items that required unique item parameters
Hungary (Hungarian)	32	3	9
Iceland (Icelandic)	32	1	3
Indonesia (Indonesian)	32	10	31
Ireland (English, Irish)	32	7	22
Israel (Hebrew, English, French, Spanish) *	28	5	18
Israel (Arabic) *	28	3	11
Italy (Italian, German, Slovenian)	32	7	22
Japan (Japanese)	32	13	41
Jordan (Arabi)	28	8	29
Kazakhstan (Kazakh) *	21	10	48
Kazakhstan (Russian) *	21	9	43
Korea, South (Korean)	32	12	38
Kosovo (Albanian)	22	9	41
Kyrgyzstan (Kyrgyz) *	21	11	52
Kyrgyzstan (Russian) *	21	11	52
Kyrgyzstan (Uzbek) *	21	13	62
Latvia (Latvian) *	28	6	21
Latvia (Russian) *	28	7	25
Lebanon (French) *	22	5	23
Lebanon (English) *	22	3	14
Liechtenstein (German)	27	5	19
Lithuania (Lithuanian, Russian, Polish)	28	5	18
Luxembourg (German, English) *	28	2	7
Luxembourg (French) *	28	4	14
Macedonia (Macedonian, Turkish) *	22	7	32
Macedonia (Albanian) *	22	4	18
Malaysia (Malay) *	21	6	29
Malaysia (English) *	21	6	29
Malta (English)	27	4	15
Mauritius (English)	20	4	20

Country (Language)	# of items administered to the group	# of items that required unique item parameters	% of items that required unique item parameters
Mexico (Spanish)	32	3	9
Moldova (Romanian) *	27	8	30
Moldova (Russian) *	27	9	33
Montenegro (Montenegrin, Albanian)	28	8	29
Netherlands (Dutch)	32	9	28
New Zealand (English)	32	4	13
Norway (Norwegian)	32	8	25
Panama (Spanish)	20	6	30
Peru (Spanish)	32	9	28
Poland (Polish)	32	5	16
Portugal (Portuguese)	32	3	9
Puerto Rico (Spanish)	22	9	41
Qatar (Arabic) *	28	15	54
Qatar (English) *	28	7	25
Romania (Romanian) *	28	13	44
Romania (Hungarian) *	28	4	14
Russia (Russian)	32	8	25
Serbia (Serbian, Hungarian, Slovak, Romanian)	22	6	27
Singapore (English)	27	10	37
Slovak Republic (Slovak) *	28	2	7
Slovak Republic (Hungarian) *	28	4	14
Slovenia (Slovenian, Italian)	28	3	11
Spain (Spanish, Galician, Valencian, Basque)*	28	2	7
Spain (Catalan) *	28	2	7
Sweden (Swedish, English)	32	7	22
Switzerland (German, Italian) *	28	7	25
Switzerland (French) *	28	5	18
Taiwan (Chinese)	28	7	25
Thailand (Thai)	32	7	22
Trinidad and Tobago (English)	27	8	30

Country (Language)	# of items administered to the group	# of items that required unique item parameters	% of items that required unique item parameters
Tunisia (Arabic)	28	3	11
Turkey (Turkish)	28	8	29
United Arab Emirates (Arabic) *	27	14	52
United Arab Emirates (English) *	27	9	33
United Kingdom (English, Welsh)	32	7	22
United States of America (English)	28	3	11
Uruguay (Spanish)	32	13	41
Vietnam (Vietnamese)	26	12	46

Note. Each cycle that required unique item parameters in Study 1 was counted as a separate item. Countries with more than one country-by-language group are indicated with an asterisk.