Assessing and Improving Measurement Properties of ESCS in PISA 2018

Research report

Student: Kirsten van Kessel
Program: Methods and Statistics for the Behavioral, Biomedical and Social Sciences
Supervisors: Remco Feskens and Marieke van Onna
Candidate journal: Large-scale Assessments in Education
FETC case number: 23-1790
Word count: 2451

Utrecht University
22-12-2023

1 Introduction

Socio-economic status (SES) is an important construct in educational research. It provides insight into the background of students and whether students from different backgrounds have equal opportunities in school (Yetişir and Kaan, 2021; Pritchett and Viarengo, 2021; Harju-Luukkainen et al., 2017; Tang et al., 2021). Data for such studies generally comes from International Large-Scale Assessments (ILSAs) that aim to compare the educational systems of multiple countries (Desa et al., 2018). Results from ILSAs are used for policy making and to inform a general public (Hastedt and Rocher, 2020). As ILSAs aim to make cross-national comparisons, it is of importance that the scales of the latent constructs are measurement invariant (Davidov et al., 2018). This means that the construct should have the same meaning in different countries. To test this, the factor model that measures the concept is compared on factor loadings, intercepts and residual variances (Van de Schoot et al., 2012). If none of these are significantly different between countries, it is shown that the questions and the underlying latent construct can be interpreted the same way across countries (Van de Schoot et al., 2012). Research on measurement invariance in ILSAs is usually focused on variables that measure cognitive abilities of students, such as reading skills. However, background variables like SES should also be measurement invariant. SES is an important variable for two reasons.

Firstly, SES is used to assess whether all students have equal learning opportunities. This is done by investigating the relation between SES and cognitive outcome variables, such as student's reading, mathematics, or science performance (Yetişir and Kaan, 2021; Pritchett and Viarengo, 2021; Harju-Luukkainen et al., 2017) or problem solving (Tang et al., 2021).

Secondly, SES is used in the calculation of the cognitive outcome variables (OECD, 2017). Values on the cognitive outcome variables are generated as multiple plausible values (Little and Rubin, 2019; Wu, 2005; Von Davier et al., 2009). These plausible values are taken out of a posterior distribution for each student and each cognitive outcome variable. SES serves as an important prior in the creation of this posterior distribution, even as the correlations between the cognitive outcome variables. By having multiple plausible values the measurement error can be accounted for better when the relation between cognitive outcome variables and background variables is researched. Since SES is used in the calculation of cognitive variables, it is important that the measure itself is also measurement invariant. Even more so, because most students do not complete every cognitive test. These missing cognitive outcome variables are purely based on the background variables and the correlation between the cognitive outcome variables (OECD, 2017).

One example of an ILSA where this is the case is the Programme for International Student Assessment (PISA). PISA investigates the performance of 15-year-olds in reading, mathematics and science every three years (OECD, 2023). PISA's measure of SES is called ESCS: the index of economic, social and cultural status. In the 2018 edition of PISA, ESCS is constructed using three indicators (OECD, 2017). These indicators are highest parental occupation,

parental education, and household possessions.

There are two main limitations in the composition of ESCS in PISA. One limitation is that the indicators do not reflect ESCS equally in each country (Downey, 2023). Currently, ESCS is computed by giving all indicators a weight of one in the principle component analysis (PCA) (OECD, 2017). Avvisati (2020) recommended these weights to simplify the procedure and make the measure of ESCS more stable over the years. However, Downey (2023) found that this approach is performing worse than approaches that allow the weights to differ between countries. This is a sign of measurement variance, which means that the same score on ESCS has a different meaning in different countries. Also, while the modelling of ESCS in PISA has been researched multiple times, it is not investigated whether a different modelling can lead to a different relation between ESCS and performance of students.

Another limitation is that the data imputation method is outdated (Avvisati, 2020). Currently, when one of the three indicators is missing, the value is imputed by stochastic regression imputation. The values on the other indicators are used to predict the missing value and a random error term is added. If two or three of the indicators are missing, the value is not imputed. This means that valuable information is not used for computing ESCS. A method that does use all available information is multiple imputation (Rubin, 1987, 1996; van Buuren and Groothuis-Oudshoorn, 2011). This is already used for cognitive variables in PISA (OECD, 2017), but not in background variables such as ESCS. It has not been studied before how multiple imputation can improve the measurement model of ESCS.

This leads to the following research question: Can the measurement of ESCS be advanced? To figure this out, there are three sub-questions.

1. To what extent is ESCS subject to measurement variance?
2. Can ESCS be modelled differently to reduce measurement variance and to better handle missing data?
3. Does this improved model of ESCS lead to a different estimate of the relation between ESCS and cognitive outcome variables?

We expect to find measurement variance in ESCS. We also expect that modelling ESCS differently can help to reduce this. Moreover, we expect that a difference in modelling can lead to a different estimate of the relation of ESCS and cognitive outcome variables.

2 Methods

2.1 Data

For the analyses, the data from PISA 2018 (OECD, 2018) will be used. This dataset includes responses of 612,004 students from 79 countries. All analyses will be performed in R (R Core Team, 2023). The variables HISEI and

PAREDINT are used as given in the PISA 2018 data. The variables HOMEPOS and ESCS are recreated according to the PISA 2018 technical report (OECD, 2017). See Section 2.2 for further explanation of the variables.

2.2 Variables

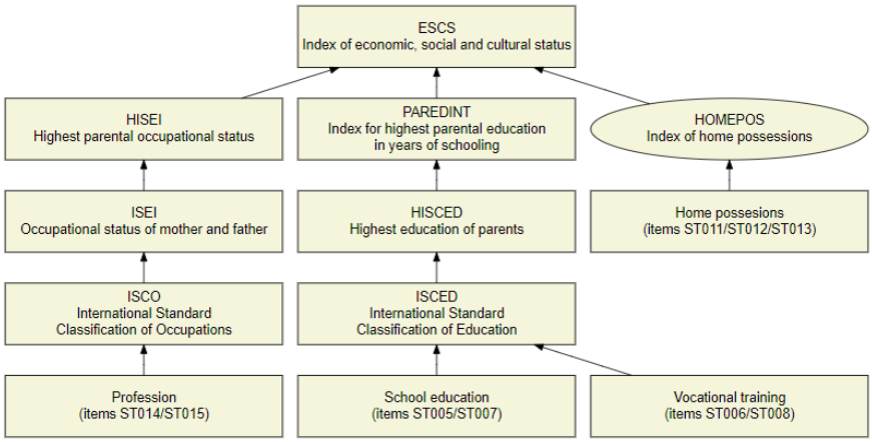

Figure 1: Computation of ESCS in PISA 2018.

The ESCS measurement model as can be seen in Figure 1 will be replicated. ESCS exists of three indicators (OECD, 2017).

The first indicator is highest parental occupation (HISEI). HISEI was measured using two open-ended questions about the occupation of the student's mother (ST014) and the student's father (ST015). The open answers were then coded into ISCO codes. To make these codes meaningful on an international level, the ISCO codes were transformed into scores on the international socio-economic index of occupational status (ISEI). The highest ISEI score of either the mother or the father became the final HISEI score. For this score, a higher value indicates a higher level of occupational status.

The second indicator is parental education (PAREDINT). PAREDINT was measured using four questions. Two questions were about the school education of the mother (ST005) and the father (ST007) and two question were about the vocational training of the mother (ST006) and the father (ST008). The answers were then constructed into one of six education categories for both the mother (MISCED_D) and the father (FISCED_D). This lead to the final HISCED_D, which is the highest of the two scores. The HISCED_D was then recoded into the final PAREDINT, which is the number of year the parent studied, using an internationally standardized transformation of HISCED_D. For this standardization, the median number of year the parent studied according to the 2015 version of PISA are used.

The third indicator is household possessions (HOMEPOS). This was obtained using several questions. The first question (ST011) was a lists of possible home possessions, including three items that differed per country. Students marked a 1 when they did have a certain item in their home and a 2 when they did not. ST011 was then reverse-coded, so a higher level indicates the possession of that item. The other two questions are about the amount of possessions (ST012) and the amount of books in the home (ST013). All questions were then scaled into an index using a 2 parameter logistic model (2PL) (Birnbaum, 1968; Muraki, 1992) with parameters for all sub questions of question ST011, ST012, and ST013 (25 sub questions). The parameters of some items were estimated for all countries separately. The parameters of some items were estimated separately for only some countries. The parameters for two items for Hong Kong were estimated for the English and Chinese assessment separately. See Table 1 in the abstract. for an overview of the parameter estimation method of every item.

After estimating the item parameters, the index for household possessions was estimated for each student using weighted likelihood estimation (WLE) (Warm, 1989).

After obtaining the student's values on HOMEPOS, stochastic regression imputation was performed per country on each indicator. This means that every indicator is regressed on the values of the other two indicators. For example, the missing values on the HISEI indicator are imputed as $HISEI_i = \beta_0 + \beta_1 PAREDINT_i + \beta_2 HOMEPOS_i + \epsilon$, where ϵ is a random value taken from a distribution with a mean of 0 and a standard deviation of σ, the residual standard error of the regression model.

After imputation, each indicator was standardized so it has a mean of 0 and a standard deviation of 1 in the pooled data of all countries. Then, a PCA model is created in which each indicator has a loading of 1 on the latent construct ESCS. That is, ESCS is the mean of the three indicators.

2.3 Research question 1: Testing for measurement invariance

The HOMEPOS indicator will be recreated according to the steps described in Section 2.2. For the 2PL and WLE, the R package dexterMML (Koops et al., 2020) is used. The computed HOMEPOS values are compared to PISA's HOMEPOS values using the descriptive statistics, correlation and a plot (see Figure 2 and Table 2 in the abstract).

ESCS is also recreated according to Section 2.2. The computed ESCS values are compared to PISA's ESCS values using the descriptive statistics, correlation and a plot (see Figure 3 and Table 2 in the abstract).

To check for measurement variance in the (recreated) data, a multigroup confirmatory factor analysis will be performed using the R package lavaan (Rosseel, 2012). Four different models are run and compared (Van de Schoot et al., 2012). 1) A configural invariance model, in which the factor loadings and intercepts are allowed to differ between countries. 2) A metric invariance model, in which the

factor loadings are constrained to be equal across countries, but the intercepts are allowed to differ. 3) A scalar invariance model, in which the factor loadings and intercepts are constrained to be equal across countries. 4) A full uniqueness model, in which the factor loadings, the intercepts and the residual variances are constrained to be equal across countries. Note that this may differ from the PISA-modelling. The models are compared using a χ^2-test to see if they are significantly different. As the χ^2-test is sensitive to sample size, the models are also compared using fit indices, such as the TLI, CFI, RMSEA, AIC and BIC (Van de Schoot et al., 2012). We take into account that scalar invariance or full uniqueness are rarely achieved (Davidov et al., 2018). If a stricter model does not perform better than its former model, the next stricter model will be omitted.

2.4 Research question 2: Improving measurement of ESCS

To improve the measurement of ESCS, a method called alignment optimization is performed using the R package sirt (Robitzsch, 2023). Alignment optimization tries to search for an optimal set of measurement parameters from the configural invariance model (Kim et al., 2017). This model is a factor model with ESCS as the latent (unobserved) factor and the three indicators as the observed variables. First, the configural invariance model is estimated, in which the factor loadings and intercepts are allowed to differ between countries, but the factor means are fixed to 0 and the variances are fixed to 1. In the next step, the factor means and variances are freed, meaning they are allowed to be any value. This value is chosen based on the total loss function F:

$$F = \sum_{p} \sum_{g_m < g_n} w_{g_m . g_n} f(\lambda_{p g_m} - \lambda_{p g_n}) + \sum_{p} \sum_{g_m < g_n} w_{g_m . g_n} f(\tau_{p g_m} - \tau_{p g_n}) \quad (1)$$

in which p is the number of indicators, g_m and g_n indicate country m and n for every pair of countries in the data, $\lambda_{p g_m}$ and $\lambda_{p g_n}$ indicate the factor loadings of country m and n, and $\tau_{p g_m}$ and $\tau_{p g_n}$ indicate the intercepts of country m and n. Within the total loss function F the component loss function f scales the difference between the parameters for every pair of countries and for every measurement parameter:

$$f(x) = \sqrt{\sqrt{x^2 + 0.01}} \quad (2)$$

$w_{g_m . g_n}$ is a weight that represents the size of the countries. It is defined as:

$$w_{g_m . g_n} = \sqrt{N_{g_n} N_{g_m}} \quad (3)$$

in which N_{g_n} and N_{g_m} are the number of participants from country n and m.

The value chosen based on F should minimize the noninvariance for every pair of countries and every intercept and loading. Based on these values, countries are divided into invariant and noninvariant countries. Invariant countries

of which a measurement parameter is not statistically different from the mean value of the invariant countries are selected. For the countries that do not belong to this category, the parameter is statistically different from the mean value of the invariant countries. Now the most invariant and noninvariant items or indicators of the measurement model are identified. The R^2 index is investigated to discover how invariant the parameters are (Kim et al., 2017). Based on the R^2 index of the parameters, the model is optimized by choosing to fix parameters that are deemed invariant to be the same across countries and free parameters that are deemed noninvariant to vary across countries.

To further improve the measurement of ESCS, the imputation of missing data is investigated. A subset of the cases that have a value for HISEI, PAREDINT, and HOMEPOS is made. In this subset, artificial missing values are made. On this dataset, we try three different methods of handling the missing data: 1) Not imputing, 2) Imputing using the current method (stochastic regression imputation) and 3) Imputing using multiple imputation using the R package mice (van Buuren and Groothuis-Oudshoorn, 2011). In every imputed dataset, ESCS is calculated as the mean of the three indicators. This ESCS is compared to the ESCS in the original subset of complete cases using the mean squared error (James et al., 2013). The imputation method with the lowest mean squared error is chosen as the best imputation method.

2.5 Research question 3: Relation between ESCS and cognitive variables

Based on the findings of research questions 1 and 2, a new measurement model of ESCS is created. As ESCS is used in the computation of the plausible values of the cognitive outcome variables, we replicate the plausible value generation model using the new values of ESCS (OECD, 2017). An IRT model is fit on each of the cognitive outcome variables. Item parameters are estimated to make sure the scales are comparable across countries. All variables from the background questionnaire, including ESCS, are contrast coded (to be able to include missing values) and reduced to a smaller number of variables using principle component analysis. Then, using the item parameters as dependent variables and reduced background variables as independent variables, latent multivariate regression models are used to estimate regression coefficients and the residual variance matrix. A posterior distribution of the cognitive outcome variables is created using the latent multivariate regression models and the student data. Then for each student ten plausible values are drawn out of the posterior distribution (OECD, 2017).

The relation between ESCS and one of the cognitive outcome variables (reading, science or mathematics) is investigated using regression analysis. The first regression used ESCS and the cognitive outcome variable as in the original model made by PISA. The second regression used ESCS and the cognitive outcome variable as in the best adapted model. Note that in these analyses, plausible values are used for the cognitive outcome variables to address uncertainty (Wu, 2005; Von Davier et al., 2009; Little and Rubin, 2019). This means that ten

models are estimated per regression analysis. The final models are the combination of these ten models. The regression coefficients and the p-values of the two final models are compared to discover if the models lead to different conclusions on the relation between ESCS and the cognitive outcome variables.

3 Appendix

Table 1: Items divided into three parameter estimation methods.

Estimated separately for all countries	Estimated separately for some countries	Estimated separately for assessment language
ST011Q07TA	ST011Q01TA	ST012Q01TA
ST011Q08TA	ST011Q05TA	ST012Q02TA
ST011Q17TA	ST011Q09TA	
ST011Q18TA	ST011Q11TA	
ST011Q19TA	ST011Q12TA	
ST012Q03TA	ST011Q16NA	
	ST012Q01TA	
	ST012Q02TA	
	ST012Q05NA	
	ST012Q06NA	
	ST012Q07NA	

Table 2: Descriptive statistics of the original PISA and recreated variables HOMEPOS and ESCS.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
HOMEPOS PISA	-10.20	-1.12	-0.37	-0.44	0.30	5.92
HOMEPOS computed	-9.55	-0.67	0.03	-0.01	0.69	6.99
ESCS PISA	-8.17	-1.01	-0.19	-0.28	0.59	4.21
ESCS computed	-4.87	-0.53	0.06	0.00	0.63	3.46

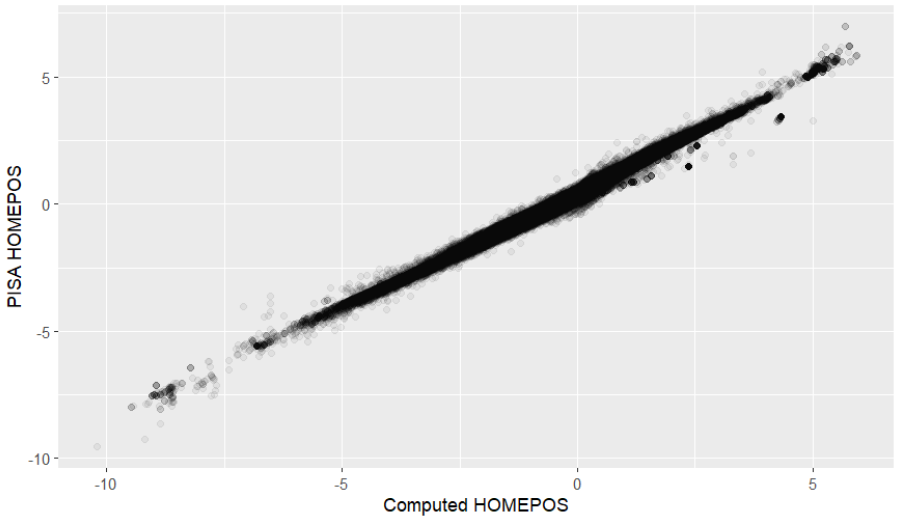

Figure 2: Correlation between the PISA values of HOMEPOS and the computed values of HOMEPOS: r = 0.997.

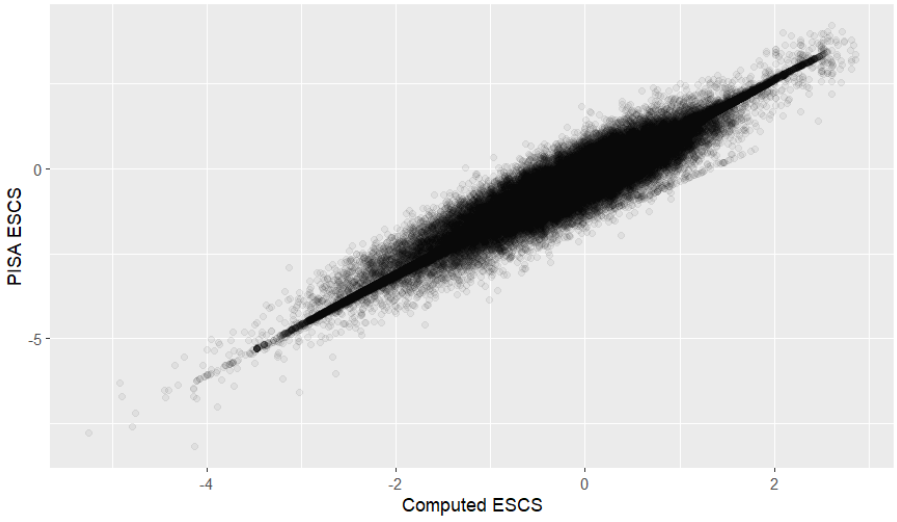

Figure 3: Correlation between the PISA values of ESCS and the computed values of ESCS: r = 0.991.

References

Avvisati, F. (2020). The measure of socio-economic status in pisa: A review and some suggested improvements. *Large-Scale Assessments in Education*, 8(1):1–37.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Davidov, E., Muthen, B., and Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones.

Desa, D., van de Vijver, F., Carstens, R., and Schulz, W. (2018). Measurement invariance in international large-scale assessments: Integrating theory and method. *Advances in comparative survey methodology*, pages 881–910.

Downey, J. A. (2023). *A Validation of the Measurement of Socioeconomic Status in PISA*. PhD thesis, University of California, Santa Barbara.

Harju-Luukkainen, H., Tarnanen, M., Nissinen, K., and Vettenranta, J. (2017). Economic, social and cultural status (escs) and mathematics performance of immigrant students in the finnish metropolitan area in pisa 2012. *Nordic dialogues on children and families*, pages 120–137.

Hastedt, D. and Rocher, T. (2020). International large-scale assessments in education: A brief guide. iea compass: Briefs in education. number 10. *International Association for the Evaluation of Educational Achievement*.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kim, E. S., Cao, C., Wang, Y., and Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4):524–544.

Koops, J., Bechger, T., Partchev, I., and Maris, G. (2020). *dexterMML: MML estimation supplement to dexter*. R package version 0.4.1.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, 16(2):159–176.

OECD (2017). Pisa 2018 technical report.

OECD (2018). Pisa 2018 student questionaire.

OECD (2023). About pisa.

Pritchett, L. and Viarengo, M. (2021). Learning outcomes in developing countries: Four hard lessons from pisa-d.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robitzsch, A. (2023). *sirt: Supplementary Item Response Theory Models*. R package version 3.13-228.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rubin, D. (1996). Multiple imputation after 18+ years", with discussion journal of the american statistical assocation.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse.

Tang, P., Liu, H., and Wen, H. (2021). Factors predicting collaborative problem solving: based on the data from pisa 2015. In *Frontiers in Education*, volume 6, page 619450. Frontiers Media SA.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *European journal of developmental psychology*, 9(4):486–492.

Von Davier, M., Gonzalez, E., and Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*, 2(1):9–36.

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

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in educational Evaluation*, 31(2-3):114–128.

Yetişir, M. İ. and Kaan, B. (2021). The effect of school and student-related factors on pisa 2015 science performances in turkey. *International Journal of Psychology and Educational Studies*, 8(2):170–186.