Title
A Validation of the Measurement of Socioeconomic Status in PISA

Author
Downey, Jonathan

Publication Date
2023

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

A Validation of the Measurement of Socioeconomic Status in PISA

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Education

by

Jonathan A. Downey

Committee in charge:

Professor Andrew Maul, Chair
Professor Stefan Gries
Dr. Diego Carrasco

June 2023

The dissertation of Jonathan A. Downey is approved.

Professor Stefan Gries

Dr. Diego Carrasco

Professor Andrew Maul, Committee Chair

June 2023

A Validation of the Measurement of

Socioeconomic Status in PISA

by

Jonathan A. Downey

Dedicated to my father for always supporting curiosity.

Acknowledgements

Thank you to everyone who helped me in this undertaking. In particular, I would like to recognize:

- Dr. Andy Maul for introducing me to the field of psychometrics and his feedback on several drafts of this paper.
- Dr. Stefan Gries for many patient hours helping me improve my understanding of research methods and statistical computing skills.
- Dr. Diego Carrasco for lending his expertise in international large-scale assessments.
- Dr. Chris Ozuna for introducing me to PISA and his collaboration on research that would motivate this project.
- Drs. Niklas Griessbaum, Jeff Inglis, and Daniel Katz for their insights and stimulating discussions.
- My family and (Dr.) Marisa for the love and support over these past years. The completion of this project would not have been possible without you.

Curriculum Vitæ

Jonathan A. Downey

Education

2023	Ph.D. in Education, University of California, Santa Barbara, CA. Interdisciplinary Emphasis: Quantitative Methods in Social Science.
2013	M.A. in Language Acquisition in Multilingual Settings, University of the Basque Country, Vitoria-Gasteiz, Spain.
2009	B.A. in Economics, Tufts University, MA.
2009	B.A. in Spanish Language and Literature, Tufts University, MA.

Awards

2021	Morrison Fellowship. Gevirtz Graduate School of Education (UCSB).
2020	Linguistic Data Consortium Scholarship. University of Pennsylvania.

Research Experience

2017	Project manager and principal investigator of *Visual Feedback Lab*. Gevirtz Graduate School of Education, UC Santa Barbara: - Led a team of five student researchers in the development of a signal processing algorithm integrated in a language learning web app.
2017	Lead curriculum and materials designer for *Basque 154: A Sociocultural Approach to Learning Basque Online*. UC Santa Barbara: - Coordinated, conducted, and recorded 20 video interviews in the Basque Country for incorporation in a sociocultural pedagogical framework.
2016	Member of Finnish Educational Technology Collaboration. UC Santa Barbara, University of Tampere (Finland), University of Oulu (Finland): - Participated in exploratory working group sessions with Finnish universities.

Professional Experience

2017 - 2023	Instructional Design Assistant. UC Santa Barbara: Letters and Science IT Department, CA.

2016 - 2017	Teaching Assistant: *Introduction to Literary Studies* and *Public Speaking*. UC Santa Barbara: English Department, CA.
2015 - 2016	Spanish Teacher. Arrupe Jesuit High School, Denver, CO.
2014 - 2015	Technical Writing Instructor. King Faisal University, Engineering Department, Al-Ahsa, Saudi Arabia.
2012 - 2014	Director of Programs: Spain and Morocco. Walking Tree Travel. Vitoria-Gasteiz, Spain; Tangier, Morocco.

Publications

2018	Ibrahim, A., Huynh, B., Downey, J., Höllerer, T., Chun, D., O'donovan, J. (2018). Arbis pictus: A study of vocabulary learning with augmented reality. IEEE transactions on visualization and computer graphics, 24(11), 2867-2874.
2018	Supporting content for *Language Learning and Technology Journal*, *22*(1).

Presentations

March 2021	*A Rasch-based Approach to Validating Essay Ratings*. AERA Research In-Progress Gala. Virtual.
February 2021	*Measuring Productive Vocabulary Using Item Response Theory*. International Objective Measurement Workshop (IOMW). UC Berkeley.
February 2020	*Hidden Dangers in L2 Assessment: Issues of Reliability and Validity*. CATESOL Los Padres conference. UC Santa Barbara.
November 2018	*Encouraging Learner Engagement with Peer Feedback*. Foreign Language Education Symposium (FLEDS) conference. Middlebury Institute of International Studies. Monterey, CA.
April 2018	*Sociocultural Strategies and Digital Tools for Minority Language Instruction*. Verbal Kaleidoscope Conference. UC Santa Barbara.
April 2017	*Visual Feedback: An Easier Way to Improve Pronunciation*. University of California GradSlam. UC Santa Barbara.

Affiliations

2019 - current	Member of American Educational Research Association.
2016 - current	Board member of UCSB GGSE IT Committee.

Abstract

A Validation of the Measurement of
Socioeconomic Status in PISA

by

Jonathan A. Downey

Socioeconomic status (SES) is a key covariate in analyses by the Programme for International Student Assessment (PISA). In its technical documentation, PISA offers evidence for the Economic, Social, and Cultural Status instrument (ESCS) as a valid measure of SES. This dissertation, however, offers both conceptual and empirical arguments that ESCS does not meet the realist and pragmatic requirements set forth by modern measurement and validity theories. I further demonstrate how the adoption of non-measurement models in ESCS has undermined the trustworthiness of PISA's most recent headline findings. Finally, I offer guidance for improving the validity of SES measurement in future PISA cycles.

Use was made of the computational facilities administered by the Center for Scientific Computing at the CNSI and MRL (an NSF MRSEC; DMR-1720256) and purchased through NSF CNS-1725797.

Contents

Chapter 1

PISA and SES

Standardized testing is essential for informed educational policymaking, allowing for the systematic observation and comparison of educational systems and policies by placing scholastic outcomes on a common scale. This offers policymakers a scientific basis for decisions regarding curricula development and the allocation of budgetary resources, minimizing speculation and political influence. Due to their scale and scope, standardized tests grant researchers the power to identify contextual factors that may be predictive of academic success that might go undetected in smaller and more homogeneous contexts. These include affective variables such as student attitudes and motivations, as well as demographic information like socioeconomic status (SES).

While domestic assessments like the National Assessment of Educational Progress (NAEP) are useful for reviewing the effectiveness of educational programs that are already in place, international assessments are especially valuable because, not only can they can help national education systems compare the success of their programs with that of similar programs in other countries, they can also inform policymakers about the potential of new programs that are already in place internationally. Therefore, these assessments offer the advantage of showing national policymakers, not only what is working now, but what is possible in the future. They also allow researchers to make inferences regarding attributes that can be difficult to examine on a national or sub-national basis, due to a small population size or under-sampling.

The most prominent international large-scale educational assessments are the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA). PISA, in particular, has increasingly come to the forefront of public attention since its inception in 2000. Every three years, PISA releases a series of reports describing the state of national educational systems in OECD-member and OECD-partner countries. It assesses three main subject matters: mathematics, science, and reading. Each cycle, one of these emphases rotates into focus as the "major domain." Each cycle also incorporates a one-off "inactive domain" — "global competence" in the most recent, 2018 cycle. PISA distinguishes itself from other assessments like TIMSS and PIRLS, which assess grades four and eight, by its assessment of 15-year-old students who are finishing their compulsory education. In doing so, it seeks to gauge the holistic efficacy of national educational systems rather than just elementary and middle grade education. Moreover, PISA aims to be "different from traditional assessments" (OECD, 2019d, p. 3) by going "beyond assessing whether students can reproduce what they have learned in school." PISA looks to assess the extent to which students can "extrapolate from what they know, think across the boundaries of subject-matter disciplines, apply their knowledge creatively in novel situations and demonstrate effective learning strategies," and thus, determine the extent to which they "have acquired the knowledge and skills that are essential for full participation in modern societies" (OECD, 2013, p. 14).

PISA's usefulness depends upon the public perception that its findings are objective and trustworthy. PISA currently enjoys a considerable degree of trust, as evidenced by the wide publicity of its reports. Findings are often reported in major news outlets.

For example, in the United States, the reports receive coverage in the New York Times, Washington Post, LA Times, Time Magazine, The Economist, and The Atlantic, among others. Politicians refer to PISA often. The previous three U.S. secretaries of Education have acknowledged the significance of national PISA scores and the disappointing performance of the United States (this excludes Miguel Cardona, whose tenure as of this writing has yet to coincide with a PISA report release). This visibility gives PISA significant influence over public perceptions and opinions of what public education should look like and what its goals should be. The influence of PISA on national public policies is often quite direct. First, PISA results can be the impetus for national governments to set strategic educational policy goals. PISA gives the examples of "performance targets" set by Mexico in 2006, and by the UK and Japan in 2010 (Breakspear, 2012; OECD, 2012a). Suboptimal results in Germany from the 2000 cycle led to the implementation of national curricula standards. Similarly, concern over "stagnation" in PISA standings was a driver for the United States to implement a range of new policies including No Child Left Behind, Race to the Top, the Common Core State Standards Initiative, and the Every Student Succeeds Act (Goldstein, 2019). Also, "PISA shock" — the political pressure generated from unexpectedly-poor PISA results — occasionally pushes participating nations into significantly increasing spending on public education. In reforms touted by PISA on its homepage, PISA results in the early 2000s were partially responsible for an almost-doubling of federal education spending in Germany (OECD, n.d.-c) and the continuation of minimum per-pupil spending and teacher salaries in Brazil (The Economist, 2010).

While public trust in PISA and its visibility tend to be mutually reinforcing, the widespread perception of validity in the assessment appears to be slipping. In the last decade, criticisms have called into question PISA's sampling methodologies (e.g., high exclusion rates and unrepresentative sampling in China, Singapore, and Argentina; see

Sahlberg and Hargreaves, 2015; L. Rutkowski and Rutkowski, 2016), the validity of its claim of going beyond assessing whether students can reproduce what they have learned in school (e.g., Hanushek and Woessmann, 2012; Zhao, 2020), and the social benefit of its testing model. Criticisms of the latter have especially focused on PISA's contribution to the increased global importance of preparing for standardized testing in scholastic curricula, as well as its casting of math, science, and traditional literacy skills as the core elements of a modern education due to their role in boosting national economic productivity (e.g., Andrews et al., 2014; Sahlberg and Hargreaves, 2015; Engel and Rutkowski, 2020). These issues have led to increasing skepticism of the project in the eyes of educators, parents, politicians, and the news media around the world (e.g., Wuttke, 2007; Rhodan, 2013 [Time Magazine]; Andrews et al., 2014; Heim, 2016 [Washington Post]; HP, 2018, March 17 [Luxembourg Times]; Auld and Morris, 2016; The Economist, 2019).

Long-term support for PISA also depends upon the validity of its measurement instruments. If confidence is lost in PISA's measures, not only will public support for PISA-based public policy be undermined, but policymakers may also be incentivized to reference whichever alternate assessment paints their national educational system in a light that reinforces a preferred narrative or fits a certain political agenda.

Measuring SES is especially important in standardized testing for a variety of reasons. First, decades of widely cited research have reported correlations between SES and academic outcomes (e.g., Duncan et al., 1972; see Sirin, 2005). Second, the constituent constructs of SES — typically education, occupation, and income — can be explicitly targeted by policy interventions such as continuing education tax credits (e.g., in the United States, the American Opportunity Tax Credit or Lifetime Learning Credit), employment subsidies (e.g., the Trade Adjustment Assistance program), tax breaks, and stimulus payments. Third, it is difficult to meaningfully interpret national scores of academic performance as a reflection of the efficacy of national educational systems without controlling for SES. Directly comparing raw test scores is not very useful because a difference in curricula is rarely the only differentiating factor between groups of students. For example, without controlling for socioeconomic status, Singapore's top PISA reading score is difficult to meaningfully interpret in terms of policy direction. The secretary of education of the Philippines (a country with a comparatively low expenditure per capita on education) might be hesitant to adopt Singaporean educational policies given the possibility that the latter's high scores may be largely a product of the general wealth of educational resources available to students in that country, both inside and outside of the classroom, rather than of the initiatives themselves. Wealthy students have more access to tutors and quiet study spaces, as well as more time to dedicate to their studies than poor students. PISA's own analysis finds that 17 of 23 countries with above-average expenditure per student obtained higher than average reading scores in 2018 and that expenditure per student can also explain up to 49% of variance in national performance (OECD, 2019d, p. 23). Therefore, the Philippines might be interested in specifically examining outcomes to low-SES Singaporean students.

Similarly, in within-country analyses, it is important to determine whether educational policies benefit students in a country to a similar degree, regardless of socioeconomic status. A curriculum that leads to a higher average score, is not necessarily superior if it also leads to greater test score inequity across the student population. From a macroeconomic perspective, education is more relevant to poor students. Wealthy students are likely to succeed economically regardless of their education. Education offers students from disadvantaged backgrounds the opportunity for economic success by teaching the skills necessary for high-skilled occupations and opening the doors to institutions of higher learning where valuable professional contacts can be forged. It is also important to society overall that the lower and middle class is well-educated, as it can be a driver of national productivity growth. Education is also a moderating influence on the intergenerational concentration of wealth in the upper class.

Unfortunately, the importance of SES is not always fully appreciated. For example, while PISA's "Excellence and Equity in Education" report (OECD, 2016) describes a reduction in the power of SES as a predictor of science achievement scores between 2006 and 2015 in the United States (i.e., a gain in equity), this finding was overshadowed in the press by drops in overall science scores (Porter, 2015). The former finding is arguably more important, as it suggests that available resources are being allocated more efficiently in the United States. Also, the latter could simply be indicative of a decrease in the overall quantity of scholastic resources during that period. Indeed, from 2006 to 2009, the United States experienced the steepest reduction in gross national income per capita since 1970. Therefore, findings that do not control for SES may tell us more about the state of the U.S. economy than about the efficacy of its educational policies.

PISA attempts to measure SES with the Economic, Social, and Cultural Status (ESCS) instrument. Example findings using ESCS measures include the conclusion that U.S. scores of financial literacy in 2018 were "strongly associated with the socio-economic status of students" (United States Country Profile, OECD, n.d.-b) and that the relationship between science achievement and SES in Chile, Denmark, Mexico, Slovenia, and the United States weakened significantly between 2006 and 2015 (OECD, 2016).

This dissertation, however, raises concerns regarding the validity of ESCS as a measure of SES — a challenge that calls into question the relevance of all PISA findings related to educational equity. In Chapter 2, I review the composition of the ESCS instrument and its components, PARED, HISEI, and HOMEPOS, as well as the evidence offered by PISA for its validity. In Chapter 3, I explain that PISA's interpretation of ESCS as a measure of SES is incompatible with a modern conception of validity and why the validity evidence they offer is unsatisfactory. In Chapter 4, I examine the validity of the individual components of ESCS: PARED, HISEI and HOMEPOS. I argue that these are not measures of education, occupational status, and income, respectively, but rather are operationalizations that lack appropriate theoretical and empirical justification. In Chapter 5, I demonstrate the practical impact of constructing ESCS with non-measurement models, finding that it significantly affects the trustworthiness of top-level findings published in PISA's 2018 executive summary report. In Chapter 6, I provide some actionable suggestions to PISA for improving the measurement of student socioeconomic advantage, specifically regarding the overall choice of measurement attribute, the recognition of country-specific attributes, the specification of models to reflect a measurement framework, and reconceptualizing the overall process of validation.

Chapter 2

An Overview of ESCS

2.1 The Components of ESCS

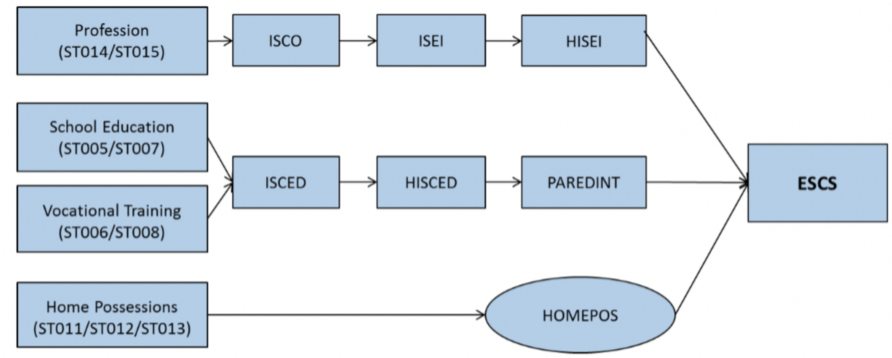
PISA attempts to control for SES with the ESCS instrument, a composite summary of a student's parental education, parental occupational status, and family income. Education, occupational status, and income are, in turn, operationalized by PARED, HISEI, and HOMEPOS (Figure 2.1), each of which is derived from data provided by students on the contextual questionnaire that accompanies the cognitive exam:

> The ESCS is a composite score based on three indicators: highest parental occupation (HISEI ["Highest International Socio-Economic Index"]), parental education (PARED), and home possessions (HOMEPOS) including books in the home [. . .] The rationale for using these three components was that socio-economic status has usually been seen as based on education, occupational status and income. (OECD, 2019c)

PARED (PAREDINT in the 2018 cycle) scores describe the greatest amount of education attained by a parent of the surveyed student. Students are prompted for the highest level of schooling and formal vocational training that each parent has received (Figure 2.2). PARED refers to the highest student-reported parental ISCED (Interna-

Figure 2.1

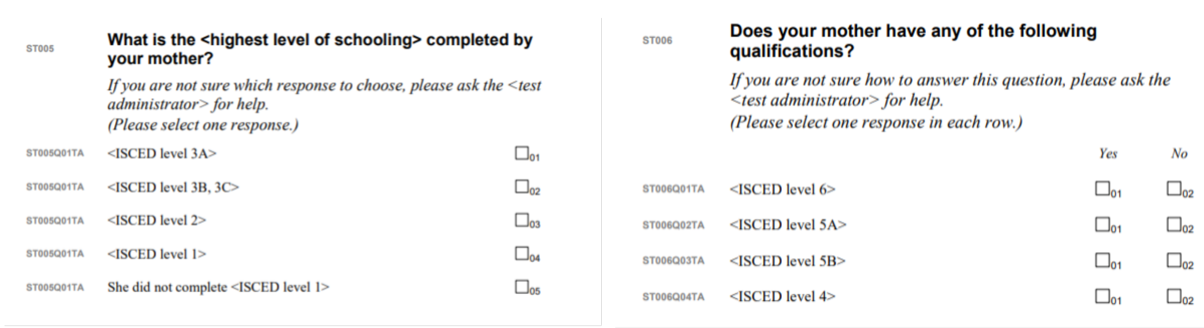
PISA's representation of ESCS in the 2018 Technical Report.



tional Standard Classification of Education) level transformed into years of schooling in a manner that varies by country (detailed in Annex D of the 2018 Technical Report, see Table 2.1). PAREDINT, transforms the highest ISCED value to an internationally-standardized estimated number of years of education by taking the median 2015 values for years of schooling for each category across all countries. PAREDINT value are used in the calculation of ESCS in the 2018 cycle.[1]

Figure 2.2

Examples of questions from student questionnaire


ST005 **What is the <highest level of schooling> completed by your mother?**

If you are not sure which response to choose, please ask the <test administrator> for help.
(Please select one response.)

ST005Q01TA	<ISCED level 3A>	☐01
ST005Q01TA	<ISCED level 3B, 3C>	☐02
ST005Q01TA	<ISCED level 2>	☐03
ST005Q01TA	<ISCED level 1>	☐04
ST005Q01TA	She did not complete <ISCED level 1>	☐05

ST006 **Does your mother have any of the following qualifications?**

If you are not sure how to answer this question, please ask the <test administrator> for help.
(Please select one response in each row.)

		Yes	No
ST006Q01TA	<ISCED level 6>	☐01	☐02
ST006Q02TA	<ISCED level 5A>	☐01	☐02
ST006Q03TA	<ISCED level 5B>	☐01	☐02
ST006Q04TA	<ISCED level 4>	☐01	☐02

[1] Both instruments will be referred to as "PARED" from this point forward because this term is used in all cycles before 2018 and is more widely referenced in outside literature.

Table 2.1

PARED coding scheme

PISA PARED level	ISCED 1997	Description
0	-	None
1	1	Primary
2	2	Lower Secondary
3	3B or 3C	Vocational /pre-vocational upper secondary
4	3A and/or 4	General upper secondary or non-tertiary post-secondary
5	5B	Vocational tertiary
6	ISCED 5A and/or ISCED 6	Theoretically oriented tertiary and post-graduate

In the contextual questionnaire, students are prompted to provide the job titles of their mother and father (Figure 2.3). HISEI scores correspond to the occupation of the parent with the highest occupational score on the International Socio-Economic Index of occupational status (ISEI; Ganzeboom et al., 1992). ISEI assigns numeric values to occupational categories of the ISCO-08 framework (International Labour Office, 2008), or to the ISCO-88 framework (International Labour Office, 1988) prior to the 2008 cycle. ISEI scores are derived from regression coefficients in a path model that describes the contributions of age, education, and occupation to income in 31 datasets (Figure 2.4).

The path model is comprised of three regression equations:

$$Income = \beta_{41} * Age + \beta_{42} * Education + \beta_{43} * Occupation + \epsilon \tag{2.1}$$

$$Occupation = \beta_{31} * Age + \beta_{32} * Education + \epsilon \tag{2.2}$$

$$Education = \beta_{21} * Age + \epsilon \tag{2.3}$$

Each occupation's ISEI score is the estimated regression coefficient for each interven-

Figure 2.3

Vocational questions from the student questionnaire

ST014 **The following two questions concern your mother's job:**

(If she is not working now, please tell us her last main job.)

ST014Q01TA What is your mother's main job?
(e.g. school teacher, kitchen-hand, sales manager)

*Please type in the job title.*___________________________ 01

ST014Q02TA What does your mother do in her main job?
(e.g. teaches high school students, helps the cook prepare meals in a restaurant, manages a sales team)

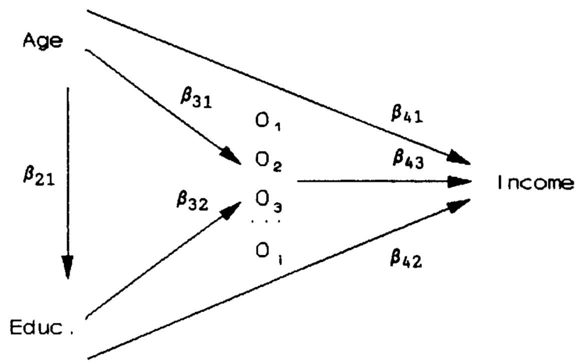
Please use a sentence to describe the kind of work she does or did in that job.

____ 01

ing occupational category when the model is constrained to "minimize the direct effect of education on income and maximize the indirect effect of education on income through occupation" (Ganzeboom et al., 1992, p. 11). As the full model is saturated, model parameters cannot be estimated directly. Rather, an alternating least squares procedure (de Leeuw et al., 1976; de Leeuw, 1988) is applied, where path coefficients of a non-saturated model (dropping the coefficient for the direct effect of education on income, i.e., β_{42}) with arbitrary occupational weights are estimated such that the total residual sum of squares is minimized. Subsequently, the weights of the occupational categories are estimated by keeping the path coefficients fixed while again minimizing the total residual sum of squares. These two steps repeat iteratively until the residual sum of squares stabilizes, at which point the direct effect of education on income in the original saturated model will also stabilize at a roughly minimal point. The ISEI score for each occupational category is then linearly transformed so that each score lies between 10 and 90. The lowest scoring occupations (10 points) are "Cooks Helper" (ISCO code 5312) and

Figure 2.4

Ganzeboom et al.'s (1992) ISEI path model



"Occupations enter this system in the form of a large set of dummy variables, represented as $\emptyset, ..., \emptyset_i$, which represent detailed occupational categories." (Ganzeboom et al., 1992, p. 11)

"Agricultural Worker" (ISCO code 6210). The highest scoring occupation (90 points) is "Judge" (ISCO code 1220).

HOMEPOS is proposed as a measure of income (OECD, 2017, p. 339). HOMEPOS values are derived from the presence and/or quantity of certain items in the family home (Table 2.2). The index of items has been composed of between 18 (2000 cycle) and 25 (2015 and 2018 cycles) distinct objects. The index includes both dichotomous items (which ask whether a given item is present in the home) as well as polytomous items (which ask how many of a given item is present). For example, students are typically asked whether they have "a link to the internet" and "a room of your own." They are also asked how many televisions and cars are at their family home (Figure 2.5). Each country is also given the freedom to choose up to three custom items that will be included in the next cycle's questionnaire for that country's students. These country-specific items are intended to account for the unique representations of wealth due to sociocultural context.

The household possessions in HOMEPOS are grouped into several non-exclusive categories: family wealth possessions (WEALTH), cultural possessions (CULTPOSS), home

Figure 2.5

Examples of dichotomous and polytomous HOMEPOS items from the 2018 contextual questionnaire

ST011 **Which of the following are in your home?**

(Please select one response in each row.)

		Yes	*No*
ST011Q01TA	A desk to study at	☐01	☐02
ST011Q02TA	A room of your own	☐01	☐02
ST011Q03TA	A quiet place to study	☐01	☐02
ST011Q04TA	A computer you can use for school work	☐01	☐02

ST012 **How many of these are there at your home?**

(Please select one response in each row.)

		None	*One*	*Two*	*Three or more*
ST012Q01TA	Televisions	☐01	☐02	☐03	☐04
ST012Q02TA	Cars	☐01	☐02	☐03	☐04
ST012Q03TA	Rooms with a bath or shower	☐01	☐02	☐03	☐04
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	☐01	☐02	☐03	☐04

educational resources (HEDRES), and information communication technology resources (ICTRES). PISA does not explicitly define selection criteria or a selection procedure for the home possession items. Likewise, certain items are dropped from one cycle to the next, but no case-by-case justifications are given in the technical documentation.

PISA describes HOMEPOS as a "scale index," meaning that students' HOMEPOS scores are estimated by applying an Item Response Theory (IRT) model to their response patterns to the itemset. All scores are on a logit scale, with zero representing the expected "ability" score of an individual randomly sampled from a normally distributed population. Each household possession in the itemset also receives a "difficulty" score on the same scale (see de Ayala [2009] for a review of IRT methods). Specifically, all HOMEPOS item and person parameter estimates are concurrently estimated in a Generalized Partial Credit Model (GPCM; Muraki, 1992) using the mdltm software package (von Davier, 2005):

$$P(X_{ji} = k|\theta_j, \beta_i, \alpha_i, \delta_i) = \frac{exp(\sum_{r=0}^{k} D\alpha_i(\theta_j - (\beta_i + \delta_{ir})))}{\sum_{u=0}^{m_i} exp(\sum_{r=0}^{k} D\alpha_i(\theta_j - (\beta_i + \delta_{ir})))} \tag{2.4}$$

... where $P(X_{ji} = k)$ is the probability that person j receives a score of k on item i, out of m_i possible scores. θ_j and β_i represent the respective person ability and item difficulty

Table 2.2

2018 home possession items

Household possession	Response type	Sub-index
A desk to study at	Dichotomous	HEDRES
A room of your own	Dichotomous	WEALTH
A quiet place to study	Dichotomous	HEDRES
A computer you can use for schoolwork	Dichotomous	HEDRES
Educational software	Dichotomous	HEDRES, ICTRES
A link to the Internet	Dichotomous	ICTRES, WEALTH
Classic literature (e.g., <Shakespeare>)	Dichotomous	CULTPOSS
Books of poetry	Dichotomous	CULTPOSS
Works of art (e.g., paintings)	Dichotomous	CULTPOSS
Books to help with your schoolwork	Dichotomous	HEDRES
<Technical reference books>	Dichotomous	HEDRES
A dictionary	Dichotomous	HEDRES
Books on art, music, or design	Dichotomous	CULTPOSS
<Country-specific wealth item 1>	Dichotomous	WEALTH
<Country-specific wealth item 2>	Dichotomous	WEALTH
<Country-specific wealth item 3>	Dichotomous	WEALTH
Televisions	Polytomous	WEALTH
Cars	Polytomous	WEALTH
Rooms with a bath or shower	Polytomous	WEALTH
<Cell phones with Internet access > (e.g., smartphones)	Polytomous	ICTRES, WEALTH
Computers (desktop computer, portable laptop, or notebook)	Polytomous	ICTRES, WEALTH
<Tablet computers> (e.g., <iPad®>, <BlackBerry® PlayBook™>)	Polytomous	ICTRES, WEALTH
E-book readers (e.g., <Kindle™>, <Kobo>, <Bookeen>)	Polytomous	ICTRES, WEALTH
Musical instruments (e.g., guitar, piano)	Polytomous	CULTPOSS
Books	Polytomous	-

Items in brackets represent country-specific item instantiations.

parameters, α_i represents a discrimination parameter (allowing for varying slopes on each item's logistic curve), and δ_{ir} represents the category step parameter of category r in item i (allowing for the differentiation of various difficulty parameters in polytomous items).[2] Also, a sampling weight is applied to each student response set in the estimation process to compensate for over- or under-sampling of his/her region. The GPCM model is derived from the Partial Credit Model (PCM, Wright and Masters, 1982):

$$P(X_{ji} = k|\theta_j, \beta_i, \delta_i) = \frac{exp(\sum_{r=0}^{k}(\theta_j - (\beta_i + \delta_{ir})))}{\sum_{u=0}^{m_i} exp(\sum_{r=0}^{k}(\theta_j - (\beta_i + \delta_{ir})))} \tag{2.5}$$

. . . with the inclusion of the discrimination parameter for each item. The PCM, in turn, is an extension of the Rasch latent variable measurement model (Rasch, 1960):

$$P(X_{ji} = 1|\theta_j, \beta_i) = \frac{exp(\theta_j - \beta_i)}{1 + exp(\theta_j - \beta_i)} \tag{2.6}$$

...with the inclusion of a distinct difficulty parameter for each polytomous item category, δ_{ir}.

From 2000 until 2012, the PCM was used in the estimation of HOMEPOS; the GPCM was only introduced in 2015. In justifying the change, PISA cites "concerns over the insufficiencies of the Rasch model to adequately address the complexity of the PISA data [that] have been raised in the past (Kreiner and Christensen, 2014; Oliveri and Von Davier, 2011, among others)" (OECD, 2017). Oliveri and Von Davier (2011), in particular, compare the difference in HOMEPOS item fit from the 2006 PISA cycle between the Rasch model and a 2PL IRT mixture model (von Davier, 2007). They find that fit is significantly improved using the 2PL IRT model and, therefore, recommend for its adoption in future PISA cycles. Similarly, PISA states:

> . . . research literature (especially Glas and Jehangir, 2014) suggests that a generalisation of [the partial credit] model, the generalised partial credit model (GPCM) (Muraki, 1992), is more appropriate in the context of PISA since it allows for the item discrimination to vary between items within any given scale. (OECD, 2019c)

[2] D is a scaling factor of 1.7, minimizing the discrepancy with estimates on a probit scale derived from a normal ogive link function, rather than a logistic link function.

PISA also cites the use of generalized IRT models such as the 3-PL IRT and GPCM in other large-scale assessments, such as NAEP, TIMSS, and PIRLS:

> Other national and international studies utilise more general IRT models (Mazzeo and von Davier, 2014; Von Davier and Sinharay, 2013). The National Assessment of Educational Progress (NAEP), for example, uses the three-parameter IRT model and the generalised partial credit model (GPCM; Allen et al., 2001) as does the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin et al., 2000). (OECD, 2017, p. 142)

As opposed to the common set of items, in which items are constrained to equality across all countries, a unique difficulty parameter is estimated for each national instance of country-specific items (even if the same object is chosen by multiple countries for an item). PISA also assigns unique item parameters to item-by-country-by-cycle instances that demonstrate poor model fit. For example, in 2015, all countries received unique parameters for two items, *Classic Literature* and *Books of Poetry*, and Albania, Japan, and Puerto Rico were assigned unique parameters for the items, *Books on art, music, or design*; *Educational software*; and *Books to help with your schoolwork*, respectively (OECD, 2017, pp. 302-304). Also, some countries do not present every item to students. For example, in 2015, Lebanon and Malaysia did not ask students whether they had a *Tablet computer* or *E-book reader*, respectively.

Two different techniques have been used to aggregate PARED, HISEI, and HOMEPOS scores into a final ESCS score. The most recent PISA cycle, administered in 2018, calculated ESCS as the arithmetic mean of the three scores. In each cycle between 2000 and 2015, ESCS was calculated as the first principal component.

2.2 PISA's Validation of ESCS

PISA's validation of ESCS consists of two analyses of cross-country comparability: first, between ESCS values, and second, between constituent HOMEPOS values. PISA offers two types of evidence supporting the comparability of ESCS values across countries. First, they publish country-group factor loadings of ESCS onto PARED, HISEI, and HOMEPOS, to "provide insight into the extent to which relationships of the index [for each country] were similar between the three variables" (OECD, 2017, p. 340; see Table 2.3). Second, they provide Cronbach's alpha statistics of internal consistency (e.g., OECD, 2017, pp. 295-296; OECD, 2019c; see Table 2.3) with the justification that "similar and high values across countries are a good indication of having measured reliably across countries" (OECD, 2019c). Therefore, both the overall magnitude of internal consistency is analyzed, as well as its comparability across countries. These two validity criteria are referred to by PISA as "cross-country comparability" and "internal consistency," respectively.[3] The perceived need for these analyses is understandable: ESCS scores in isolation are meaningless because they do not have an interpretable unit. Therefore, maintaining a high degree of cross-context comparability is essential to ensuring any relative interpretability of ESCS values.

PISA's validation of HOMEPOS follows a similar procedure, reporting the degree to which observed item response patterns fit the GPCM model is comparable across country groups. PISA tests this invariance across groups with the root mean square deviance (RMSD) statistic (OECD, 2017, p. 296):

[3]There may be some terminological confusion here. PISA's technical documentation refers to the validation of "internal consistency" and "cross-country comparability" as two facets of validity. Internal consistency, however, is also referred to as one of the two aspects of cross-country comparability (OECD, 2017, pp. 295-296; OECD, 2019c). Furthermore, PISA at times even appears to consider cross-country comparability to be synonymous with validity — for example, they cite Avvisati et al. (2019), an article specifically about different ways of assessing country invariance, as a general overview of "different methodological approaches for validating questionnaire constructs."

$$RMSD_g = \sqrt{\frac{1}{K+1}\sum_{k=0}^{K}(P_{obs,gk}(\theta) - P_{exp,gk}(\theta))^2 f(\theta)d\theta} \tag{2.7}$$

...where $RMSD_g$ refers to the root mean squared distance between the observed (obs) and expected (exp) item characteristic curves of group g on item category k. PISA's cutoff for an acceptable RMSD statistic is 0.3. Similarly, as evidence of "invariance of item parameters," PISA provides the item difficulty and slope parameters of the HOMEPOS GPCM (e.g., OECD, 2017, p. 301-302). PISA also publishes Cronbach's alpha statistics for country-group estimates of HOMEPOS and each of the household possession subscales: CULTPOSS, HEDRES, WEALTH, and ICTRES (Table 2.4; OECD, 2017, p. 301).

Table 2.3

Example of factor loadings and Cronbach's alpha values as validity evidence for ESCS from the 2015 cycle

Country	HISEI	PARED	HOMEPOS	Reliability
Australia	0.8	0.79	0.67	0.6
Austria	0.81	0.79	0.72	0.66
Belgium	0.84	0.79	0.71	0.68
Canada	0.8	0.79	0.64	0.58
Chile	0.85	0.84	0.77	0.76
Czech Republic	0.82	0.76	0.72	0.65
Denmark	0.83	0.79	0.68	0.65
Estonia	0.83	0.78	0.68	0.63
Finland	0.8	0.76	0.68	0.59
France	0.83	0.78	0.72	0.66
Germany	0.83	0.81	0.74	0.7
Greece	0.83	0.82	0.71	0.7
Hungary	0.85	0.83	0.75	0.74
Iceland	0.75	0.76	0.65	0.53
Ireland	0.81	0.8	0.7	0.65
Israel	0.8	0.79	0.68	0.6
Italy	0.83	0.79	0.72	0.68
Japan	0.74	0.76	0.68	0.54
Korea	0.78	0.79	0.73	0.62
Latvia	0.83	0.82	0.72	0.69
Luxembourg	0.86	0.79	0.75	0.72
Mexico	0.85	0.85	0.8	0.77
Netherlands	0.81	0.78	0.75	0.67
New Zealand	0.81	0.75	0.68	0.58
Norway	0.8	0.78	0.68	0.6
Poland	0.81	0.8	0.71	0.65
Portugal	0.86	0.84	0.76	0.75
Slovak Republic	0.84	0.82	0.74	0.72
Slovenia	0.84	0.82	0.69	0.68
Spain	0.85	0.83	0.74	0.73
Sweden	0.82	0.77	0.66	0.61
Switzerland	0.82	0.81	0.69	0.68
Turkey	0.82	0.79	0.77	0.68
United Kingdom	0.8	0.76	0.73	0.63
United States	0.84	0.81	0.74	0.71

Table 2.4

Example of Cronbach's alpha values as validity evidence for HOMEPOS from the 2015 cycle

Country	HOMEPOS	CULTPOSS	HEDRES	WEALTH	ICTRES
Australia	0.734	0.575	0.647	0.64	0.481
Austria	0.728	0.586	0.507	0.664	0.478
Belgium	0.731	0.624	0.524	0.667	0.523
Canada	0.73	0.584	0.629	0.649	0.52
Chile	0.809	0.571	0.541	0.75	0.626
Czech Republic	0.715	0.626	0.55	0.628	0.48
Denmark	0.684	0.597	0.504	0.559	0.371
Estonia	0.741	0.576	0.493	0.682	0.477
Finland	0.706	0.643	0.544	0.558	0.427
France	0.712	0.657	0.496	0.634	0.487
Germany	0.714	0.601	0.522	0.624	0.501
Greece	0.752	0.581	0.498	0.699	0.562
Hungary	0.78	0.65	0.555	0.711	0.516
Iceland	0.693	0.53	0.581	0.63	0.4
Ireland	0.73	0.582	0.55	0.608	0.465
Israel	0.737	0.634	0.587	0.696	0.545
Italy	0.732	0.557	0.491	0.651	0.523
Japan	0.698	0.588	0.472	0.565	0.524
Korea	0.779	0.631	0.552	0.627	0.482
Latvia	0.723	0.584	0.42	0.646	0.503
Luxembourg	0.761	0.61	0.556	0.698	0.526
Mexico	0.867	0.601	0.574	0.847	0.739
Netherlands	0.678	0.574	0.498	0.57	0.424
New Zealand	0.748	0.561	0.653	0.673	0.549
Norway	0.726	0.621	0.608	0.636	0.445
Poland	0.748	0.598	0.456	0.69	0.496
Portugal	0.771	0.598	0.478	0.672	0.55
Slovak Republic	0.78	0.618	0.675	0.695	0.548
Slovenia	0.72	0.62	0.472	0.634	0.477
Spain	0.755	0.598	0.51	0.656	0.555
Sweden	0.748	0.611	0.608	0.653	0.473
Switzerland	0.702	0.587	0.529	0.616	0.492
Turkey	0.855	0.641	0.65	0.773	0.673
United Kingdom	0.748	0.631	0.629	0.638	0.501
United States	0.802	0.593	0.66	0.692	0.578

Chapter 3

Interpreting ESCS as a Measure

3.1 A Modern Definition of Validity

I argue that PISA's evidence of cross-country comparability and internal consistency is not sufficient for establishing the validity of ESCS because it does not recognize Messick's (1989) "unified" validity theory — the modern standard among educational and social science researchers today. According to this theory, validity refers to an "overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment," (Messick, 1995, p. 6) consolidating evidence of content, substantive, structural, generalizability, and external validities (Messick, 1995, pp. 5-6), while integrating *consequential* validity – the social consequences of test score use. The official definition of validity in the Standards for Educational and Psychological Testing, jointly authored by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), closely follows Messick's:

> The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed. (2014, p. 11)

The choice of Messick's holistic view of validity as the modern standard is not arbitrary. It reflects a broader post-positivist paradigm shift in the social sciences that began in the 1960s with ideas such as Quine's ontological relativity and Kuhn's revolutionary science. The first half of the 20th century was dominated by strong positivist conceptions of validity, most notably formalized in Cronbach and Meehl's foundational 1955 paper, *Construct validity in psychological tests*. Cronbach and Meehl held that validity could only be established empirically, as the extent to which observations that are indicative of constructs are demonstrated to correlate in ways that are hypothesized a priori by a falsifiable theory:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. (Cronbach and Meehl, 1955, p. 290)

A collection of such theoretically linked observations is referred to as a *nomological network*. One of the advantages of the nomological network framework recognized at that time, is to allow the validity of an instrument to be gauged independently of its correlation with an external construct, which was the basis of earlier notions of criterion validity (see Maul, 2018). One problem with relying exclusively on correlation with external criteria is that validity can never be established satisfactorily, as the process of validation entails a never-ending chain of correlation comparisons:

> When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the 'infinite frustration' of relating every criterion to some more ultimate standard. (Gaylord, n.d., as cited in Cronbach and Meehl, 1955)

However, Borsboom et al. (2004) illustrate that the nomological network approach has weaknesses. They point out that, in most applied settings, observation necessarily proceeds theory generation, whereas validation via a nomological network presupposes a theory upon which the relationships between observed data is tested, "[getting] the scientific method backward" (p. 1064). Also, for these theoretical relationships to be meaningful in a measurement context, the theoretical terms should be sufficiently defined such that the relationships between their referents are quantifiable and empirically testable:

> ...few, if any, nomological networks in psychology that are sufficiently detailed to do the job of fixing the meaning of theoretical terms. To fix this meaning requires a very restrictive nomological network. (Borsboom et al., 2004, p. 1064)

Messick's view of validity, however, moves beyond testing static and universal "systems of laws" (Cronbach and Meehl, 1955, p. 290) by incorporating pragmatic validity criteria. Under this perspective, a measure only exhibits sufficient consequential validity if it useful for accomplishing a certain goal (Adams, 1966; Torres Irribarra, 2021). This is especially attractive for the social realities of applied sciences, for example, addressing potential misuse of testing. In fact, just 15 years after the publication of *Construct validity in psychological tests*, Cronbach himself would acknowledge the importance of social relativism, referring to validity in terms of "the soundness of all the interpretations of a test" (Cronbach, 1971, p. 1443).

A pragmatic conception of measurement validity is especially relevant to PISA's context, where the influence of PISA reports on school curricula and educational economic policies is hotly debated. Certainly, the general public is invested in the notion that standardized tests impact students' lives and society at large. Likewise, national educational policymakers who read PISA reports see the assessment as a means for testing hypotheses regarding the current capabilities of actual students and educational systems with the goal of drafting appropriate educational policies. In ESCS, they seek to understand how inequity is distributed in their country, so that policy actions are tailored accordingly. Indeed, in its discussion of construct validation, PISA acknowledges that measured constructs should be interpreted in the context of larger educational systems and policies:

> The development of comparable measures of student background, attitudes and perceptions is a major goal of PISA. Cross-country validity of these constructs is of particular importance as measures derived from questionnaires are often used to explain differences in student performance within and across countries and are, thus, potential sources of policy-relevant information about ways of improving educational systems. (OECD, 2012b, p. 286)

A pragmatic view of validity implies a validation process in which establishing the plausibility of interpretations is fundamental. Kane's (1992) argumentative approach to validation describes identifying candidate interpretations through an "interpretive argument," after which a corresponding "validity argument" can be made using available empirical and logical evidence to support the interpretation in question. Therefore, under a pragmatic view of measurement, the adequacy of evidence in the PISA assessment should be determined by the degree to which instruments can be argued to achieve social and educational goals, not by the congruence of test items to abstract constructs in a nomological network. In the case of invariance testing, for example, it is necessary to gauge how different types and degrees of invariance might ultimately impact the findings in the top-level reports that are presented to policymakers, and how policies might subsequently be plausibly affected. For example, in PISA's case, validity should be gauged in terms of changes to national rankings and national comparisons to international/OECD mean or quartile benchmarks of ESCS and other metrics derived from ESCS (e.g., R-squared metrics of ESCS influence on academic performance), as opposed to whether fit statistics of internal consistency and data-to-model fit meet cut-off points chosen by convention (e.g., chi-square, root mean squared error of approximation [RMSEA; Steiger, 1990], standardized root mean square residual [SRMR; Hu and Bentler, 1998], comparative fit index [CFI; Bentler, 1990]). For example, in PISA's technical report, the acceptability of ESCS is partly determined by whether Cronbach's alpha coefficient reaches arbitrary cut-offs:

> The [Cronbach's alpha] coefficient ranges between 0 and 1, with higher values indicating higher internal consistency. Commonly accepted cut-off values are 0.9 to signify excellent, 0.8 for good, and 0.7 for acceptable internal consistency. (OECD, 2017, p. 295)

These cut-offs lack inherent meaning for policymaking. Moreover, even though many countries do not meet the minimum cutoff of 0.7 (for example, in 2015 the internal consistency of the United Arab Emirates [ARE] is only 0.38), it is difficult to say why the highest value of 0.77 in Mexico is adequate for policymaking purposes. Likewise, it is not easy to justify the minimum level of comparability of factor loadings across countries (i.e., why certain sets of factor loadings might be "equal enough"). Certainly, we would be surprised if income, education, and occupation were not at least moderately correlated, as education and occupation can directly cause, or can be caused by, income.

3.2 A Latent Variable Measurement Interpretation of ESCS

The unified theory of validity carries implications for the ESCS measurement model, or in other words, for how ESCS values relate to what PISA is attempting to measure. ESCS values are interpreted by educational stakeholders as being causally determined by quantities of a personal SES attribute that exists in a realist sense, impacting social outcomes. This ontological commitment, in turn, implies a latent variable model of measurement.

This interpretation of measurement dictates that differences in measured values should be proportional to and caused by corresponding differences in the underlying attribute, even though it is unobservable. This idea is the fundamental concept behind several families of statistical models that seek to quantify latent constructs, including factor analysis, latent class analysis, latent profile analysis, and item response theory. This notion of measurement is also preferred by philosophers of science (e.g., Michell, 1997; Borsboom, 2005). Michell (p. 358) holds that "it is invariably along such lines that measurement is, and always has been, defined in the physical sciences."

PISA often encourages a realist interpretation of ESCS. When PISA reports countries' ESCS values side-by-side in lists and tables (e.g., OECD, 2019b, p. 51), readers are led to interpret these values as quantities of a real attribute, intervals of which can be meaningfully compared. Conversely, they are not easily interpretable strictly as a summary of the variance in PARED, HISEI, and HOMEPOS values. Such a purely descriptive statistical interpretation would be too abstract and arbitrary for the purposes of PISA. Likewise, readers interpret regressions of test scores on ESCS as a quantification of something real – the causal relationship between a socioeconomic quality from which students benefit and their educational achievement. For example, reporting correlations of ESCS scores with measures of educational achievement in calculations using a "socioeconomic gradient" reinforces the implication of ESCS as a causal agent, even though language implying descriptive and predictive modeling intentions is carefully employed to avoid direct claims of causality. For example, in the executive summary of Volume II of 2018 cycle report, "Where All Students Can Succeed," (OECD, 2019b), PISA states:

> In PISA, the socio-economic gradient is traditionally used to examine the relationship between students' socio-economic status and their performance (OECD, 2016). More specifically, the slope of the gradient summarises the differences in performance observed across socio-economic groups, while the strength of the gradient refers to how well socio-economic status predicts performance. (p. 55)

> On average across OECD countries in 2018, a one-unit increase in the PISA index of economic, social and cultural status was associated with an increase of 37 score points in the reading assessment. (p. 55)

Similarly, when PISA uses the slope of the socio-economic gradient to suggest the presence of relative educational inequity in Belarus, Belgium, the Czech Republic, France, Hungary, Israel, the Slovak Republic, and Ukraine in 2018 (OECD, 2019b, p. 55), reporting national R-squared values of the relationship of ESCS to reading performance, a causal relationship is implied.[1] These implications are also present when PISA states:

> On average across OECD countries, 12% of reading performance was accounted for by the PISA index of economic, social and cultural status. (OECD, 2019b, p. 50)
>
> In 11 countries and economies, including the OECD countries Australia, Canada, Denmark, Estonia, Finland, Japan, Korea, Norway and the United Kingdom, average performance was higher than the OECD average while the relationship between socio-economic status and reading performance was weaker than the OECD average. (OECD, 2019b, p. 15)

After all, without the implication of causality, what is the use of these analyses to readers of the PISA reports?

Also, when PISA cites similar factor loadings of PARED, HISEI, and HOMEPOS as evidence of validity, it suggests a latent variable interpretation of SES measurement, in that each of these components is measuring "the same thing" in each country context. Furthermore, the discrepancies between country factor loadings across countries are only

[1] There are also technical issues with drawing inferences from gradient instruments based on R-squared values. Despite acknowledging that a low R-squared value should not necessarily be interpreted as a weak relationship between academic achievement and socioeconomic advantage due to non-linearity and multidimensionality in underlying data (OECD, 2019b, p. 56), PISA does not analyze the residuals of these gradient plots to check for non-linearities, nor do PISA's principal findings in its executive summaries mention these nuances. For example, the executive summary of Volume II of the PISA 2018 results report (OECD, 2019b, pp. 15-17) directly interprets R-squared values in a bivariate analysis of ESCS scores and reading performance scores as the relationship between "socio-economic status and reading." Furthermore, by reporting only the slopes and R-squared statistics of these plots, PISA is neglecting to test for the significance of the potential relationships between academic achievement and socioeconomic advantage. Neither p-values nor confidence intervals are reported.

reconcilable if they are attributed to measurement error of an SES attribute that exists in some quantity for each person, of which education, occupational status, and income are indicators. Also, although the ESCS instrument takes different forms across participating countries and testing cycles, ESCS values or national rankings of ESCS values are consistently compared to an international or OECD average. If ESCS is not a reflection of some underlying objective attribute, then such comparisons to an average are not meaningful.

PISA also generally encourages the interpretation of measured constructs as comparable quantities of a personal attribute in its definitions and descriptions. Even though PISA does not formally define the terms "measurement" and "attribute" in its documentation, PISA consistently refers to "comparable measures" in its official literature, and uses analogies from physical measurement where interval comparability is well-established:

> The development of comparable measures of student background, practices, attitudes and perceptions is a major goal of PISA. (OECD, 2017, p. 295)

> PISA has become the world's premier yardstick for evaluating the quality, equity and efficiency of school systems. (OECD, 2018a, p. 2)

The notion of SES as strictly a "composite" of education, occupational status, and income, provided in the definition adopted from the National Center for Education Statistics (NCES, Cowan et al., 2012) is further undermined when it is alternatively described as "the relative position of a family or individual on a hierarchical social structure" (Cowan et al., p. 16), an analogy that invokes comparable distances between readings. Also, PISA's use of terms in official documentation such as "measure," (Willms and Tramonte, 2015; OECD, 2018b, p. 56) "estimator," (OECD, 2019c, p. 50) or "proxy" (OECD, 2019a, p. 234) suggests the objective nature of SES. Similarly, when PISA discusses ascertaining "construct validity" in its technical documentation, use of the term, "construct" seems to imply that ESCS is a reflection of an independent entity, or at least more so than a summary statistic.

PISA's treatment of SES as an ontologically substantive attribute is a notion present in the broader SES literature, as well. For example, Sirin (2005) alludes to "conceptual meanings" of SES and that parental income, parental education, and parental occupation are the three main "indicators" of SES, rather than elementary components:

> While there is disagreement about the conceptual meaning of SES, there seems to be an agreement on Duncan, Featherman, and Duncan's (1972) definition of the tripartite nature of SES that incorporates parental income, parental education, and parental occupation as the three main indicators of SES (Gottfried, 1985; Hauser, 1994; Mueller and Parcel, 1981). (Sirin, 2005, p. 418)

Perhaps widespread tendency to attribute meaning to SES is simply a product of the "correlation, therefore causality" fallacy. In PISA's case, just because HOMEPOS, PARED, and HISEI are observed to correlate, does not necessarily suggest the existence of a common, causal latent attribute. Correlational relationships are not meaningful in the absence of a testable nomological network:

> Note that neither the idea of implicit definition of constructs nor the idea of construct validity itself can be formulated in the absence of a theory that relates the construct to other constructs. [...] Just like construct validity itself, such attempts do not get off the ground without some kind of nomological network. (Borsboom et al., 2004, p. 1064)

Other causal structures could produce these observations. For example, it could be the case that income, education, and occupational status mutually cause each other, and do so in similar ways across national contexts. Similarly, even if there is a common causal factor of HOMEPOS, PARED, and HISEI, PISA's correlational evidence does not suggest of which attribute they are indicators. For example, HOMEPOS, PARED, and HISEI could all simply be indicators of income. Alternatively, they could indicate the extent of urban development where a student lives. Cities tend to have better professional opportunities and easier access to education than rural areas. Also, the cost of living is usually higher in cities, necessitating a higher income to live there. PISA would be unlikely, however, to endorse the notion that living in a city is synonymous with SES. Similarly, Cronbach and Meehl caution against automatically interpreting reliability as evidence of construct validity:

> It is unwise to list uninterpreted data of this sort under the heading 'validity' in test manuals, as some authors have done. High internal consistency may lower validity. Only if the underlying theory of the trait being measured calls for high item intercorrelations do the correlations support construct validity [. . .] Whether a high degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct. (Cronbach and Meehl, 1955, p. 288)

One can raise similar doubts regarding evidence of cross-country comparability — by what logic would one hypothesize a priori that education, occupation, and income would interact in similar ways in national contexts that are very different economically, historically, demographically, geographically, etc.? Moreover, the suggestion that HOMEPOS, PARED, and HISEI exhibit similar variation in the various OECD member and partner countries is not an obvious interpretation of the data reported in the 2015 cycle, where reported country-group Cronbach's alphas in OECD member countries exhibit considerable range (from 0.36 in the UAE to 0.77 in Mexico).

Likewise, data-to-model fit is not sufficient evidence of construct validity. Good fit could simply be the result of over-fitting, which limits the usefulness of an instrument for inference-making because model parameterizations that fit the data well in one cycle will likely fit the data of the next cycle worse than a more conservative parameterization would.

3.3 SES Is Not an Attribute

A pragmatic interpretation of ESCS implies that SES is an attribute, the quantities of which are represented by ESCS values. There are problems, however, with conceptualizing SES as an attribute. An attribute is a singular quality manifested in a subject (Wright & Masters, 1982), and SES is not commonly considered to be a unitary construct. Indeed, PISA refers to SES in relation to a range of disparate constructs, for example defining it in terms of "financial, social, cultural, and human capital;" "wealth, prestige, and power;" "a wide range of outcomes pertaining to [a person's] physical, economic, and social well-being;" and "objective material living conditions:"

> Socio-economic status is a broad concept that aims to reflect the financial, social, cultural and human-capital resources available to students (Cowan et al., 2012). Socio-economic status may also be referred to as 'the relative position for the family or individual on a hierarchical social structure, based on their access to, or control over, wealth, prestige and power' (see Willms and Tramonte, 2015 quoting Mueller and Parcel, 1981). Socio-economic status is thus a measure of students' access to family resources (financial capital, social capital, cultural capital and human capital) and the social position of the student's family/household. (OECD, 2019b, p. 52)

> ... a person's position on an SES hierarchy is related to a wide range of outcomes pertaining to their physical, economic, and social well-being (Willms and Tramonte, 2015, p. 16).

> PISA measures students' objective material living conditions through a composite index of economic, cultural and social status (ESCS). (OECD, 2019a, p. 267)

Synthesizing these various constructs into a single attribute is problematic from a theoretical standpoint. Bourdieu's (1983) and Coleman's (1988) original theories where financial, social, cultural, and human capital are introduced do not exclusively describe the constituent constructs of ESCS (education, occupational status, and income), nor can these different capital constructs necessarily be aggregated (furthermore, a justification of their aggregation is not specified by PISA). They are broad sociological theories that do not propose criteria for systematic empirical observation. Social, cultural, and human capitals are discussed as qualitative, rather than quantitative, properties in the original writings. Testing the construct validity of a measurement instrument requires a theory of the construct that is detailed enough to preference one instrument over another. Without some intermediate theory of its application in the PISA context, Bourdieu's definition of social capital as "the aggregate of actual or potential resources" (Bourdieu, 1983, p. 21) is too broad to act as a guide for how to best account for the national and generational discrepancies in social advantage for which PISA needs to control. This is a point to which researchers have alluded in the past, for example, pointing out that PISA does not provide a theoretical foundation for its composition of ESCS (e.g., Avvisati, 2020), nor an explicit justification for its use (e.g., D. Rutkowski and Rutkowski, 2013). Willms and Tramonte appear to suggest that financial, social, cultural, and human-capital resources are collectively meaningful as, "a metaphor for the cultural and social assets that families possess which lead to higher levels of physical, economic, and social well-being" (p. 16). However, the notion that these capital types are a metaphor for assets misconstrues key aspects of the respective theories of capital proposed by Bourdieu and Coleman. First, neither theory accounts for all four types of capital. Second, both authors would likely argue that social capital is ontologically real, not a metaphor. Third, the comparison of these capitals with "cultural and social assets" is tautological. What is the difference between capital and assets? Fourth, unlike financial capital, social capital cannot be possessed by an individual, but rather is an aspect of a social structure:

> Social capital is [...] not a single entity but a variety of entities, with two elements in common: they all consist of some aspect of social structures, and they facilitate certain actions of actors — whether persons or corporate actors — within the

structure. (Coleman, 1988, p. 99)

There is no reason to assume the inherent objectivity of SES. It is a social construct, and, therefore, the public conception of it can differ across sociogeographical contexts. Also, the notion of what constitutes wealth and poverty in a given place shifts over time, even if official poverty line metrics may not. This relativism is problematic for quantifying SES. Policymakers with divergent political agendas can operationalize SES differently to provide whatever evidence they find convenient. This could result in the implementation of a suboptimal policy, or even no policy at all. For example, if PISA suggests that increasing SES boosts reading performance, and subsequently, one policymaker interprets SES in a way that places more weight on income, while another interprets in a way that preferences education, it may be the case that no policy gets passed because there is no consensus on what which of the inferences is valid. If "an increase in the number of years of parental education" were found to correlate with higher reading scores of children nine years later, an inference could be drawn that a continuing education credit offered to parents by the government in 2009 caused an increase in reading performance by the children in 2018. This inference, in turn, could promote a renewal of that program or an allocation of additional funding. Alternatively, if "an increase in SES" were found to correlate with higher reading scores of children nine years later, ambiguity in the definition of SES does not easily lend to making real-world public policy because SES can be alternately interpreted as income, education, occupation, etc.

The perception that financial, social, cultural, and human capitals can be meaningfully quantified and aggregated may largely be a linguistic artifact. Perhaps confusion stems from the choice of the term "capital," which might invite unsound analogies from economics and finance, where it refers to distinct classes of assets that are fungible due to a common unit of value. Cultural capital and social capital do not have these properties, however — one cannot aggregate them into a single portfolio whose value can be compared with that of other portfolios. Bourdieu's theory of social capital recognizes financial, social, and cultural capital as constructs that are mutually-reinforcing, but essentially distinct. Prior to his tenure as the former head of OECD's Centre for Educational Research and Innovation (CERI), Tom Schuller acknowledged that these types of capital are not exclusive and that they may overlap (2001, p. 90). Finally, while Willms and Tremont acknowledge differences in the theories, they do not recognize that there simply is no consensus definition of social capital. Even the theories of Bourdieu and Coleman are quite different, especially in terms of the mechanisms governing social capital and the role that social capital plays in society.

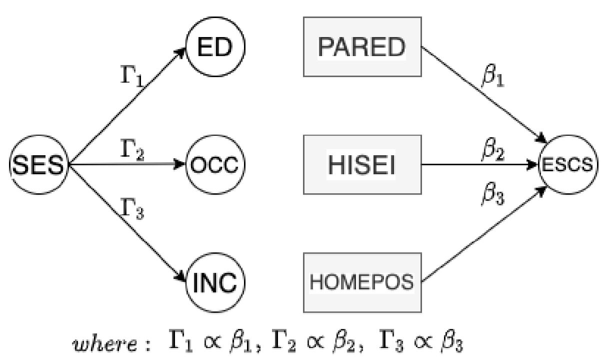
Another linguistic pitfall could be that some definitions of SES refer to a singular predicate. For example, Cowan et al.'s (2012) definition of SES as, "one's access to financial, social, cultural, and human capital resources" refers to SES as a singular entity, "access". However, it is important not to conflate grammatical singularity with ontological singularity. "Access" is not defined or explained as a concept. It is not evident what difference there is between one's "resources" and one's "access to resources." Neither is there a justification for considering access to be a meaningful concept when considering financial capital, social capital, and cultural capital. If access is considered to be a measure of capital, then SES becomes an irrelevant construct: If ESCS is a measure of SES, and SES is, in turn, a measure of access to family resources, then ESCS should be considered a measure of family resources. After all, a "measure of a measure" is either nonsensical or redundant. In either case, we can conclude that, if SES is simply a label for "access to family resources," then the latter should be the focus of definition and validation.

3.4 ESCS Does Not Conserve Attribute Quantities

Another problem with interpreting ESCS as a latent variable measurement model is that, even if SES were an attribute, "measured" ESCS values do not correspond to quantities of any single underlying attribute. Under a measurement model, correlations between ESCS and the values of its indicators, PARED, HISEI, and HOMEPOS, must be proportional to the respective correlations between the SES attribute and its indicators, education, occupational status, and income (Figure 3.1). Without this correspondence, a unit increase in SES would not necessarily cause a proportional increase in ESCS. Specifically, aggregating ESCS as the first principal component (2000-2015 cycles) or the mean (2018 cycle) of PARED, HISEI, and HOMEPOS does not conserve the isomorphism in these two sets of relationships.

Figure 3.1

The component loadings of ESCS must be proportional to the indicator loadings of SES



While factor analytic approaches typically seek to identify common causal factors of observable variables, principal components are descriptive statistics — the most effective descriptors of overall system variance. Unlike latent variables, they are difficult to interpret in a theoretical framework because they summarize the entire variance-covariance matrix of observed data, pooling observed covariance between variables (the off-diagonal elements of the variance-covariance matrix) and observed variance (the diagonal elements of the variance-covariance matrix), as opposed to the variance corresponding to any one theoretical term. The use of PCA to reduce dimensionality is useful for data compression applications but is of limited use for identifying magnitudes of latent attributes. Quantification is not necessarily measurement; there are many alternative, but still arbitrary, aggregations of PARED, HISEI, and HOMEPOS into a single value.

PISA's decision in the 2018 cycle to model ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI, rather than the first principal component appears to be at least partially based on a recommendation proposed by Avvisati:

> In PISA 2018, the ESCS was constructed as the arithmetic mean of the three indicators after their imputation and standardization (Avvisati, 2020). In previous cycles, the ESCS was constructed based on a principal component analysis (PCA) as the component score for the first principal component. However, analysis has shown that factor loadings are quite similar across countries and components. Consequently, the decision was made to set equal arbitrary factor loadings. Each component is assigned the factor loading 1. The theoretical eigenvalue in such a case equals 3 as the eigenvalue is the sum of all squared factor loadings. Using factor loadings of 1 and an eigenvalue of 3 in the usual formula for the computation of ESCS equals the computation of ESCS as mean of all three components. (OECD, 2019c)

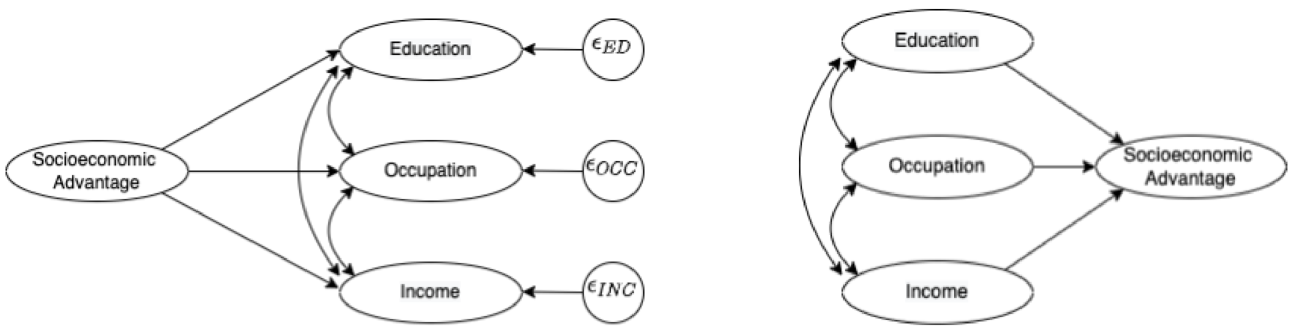
Aggregating ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI, rather than the first PCA component, however, does not remedy the problem of attribute representation. As PISA notes, as simply taking the mean of the three components is the equivalent of conducting a PCA with all factor loadings arbitrarily fixed at one, and therefore, HOMEPOS, PARED, and HISEI can still not be considered to be indicators of SES. In fact, this drifts even further from a latent variable approach because significant changes in the relationships between PARED, HISEI, and HOMEPOS are now undetectable because factor loadings are not calculated at all. PISA claims that mean aggregation is acceptable because factor loadings have been observed to be relatively equal in the past. It is not clear, however, when factor loadings are "similar enough" to justify a mean-aggregation approach, nor does the observation of similar factor loadings in past cycles automatically mean that they will continue to be similar in the current cycle or future cycles. Finally, if PISA's intention with ESCS is to capture the greatest amount of common variance in PARED, HISEI, and HOMEPOS (again, the focus should rather be on measurement, not variance minimization), mean-aggregation comes with an informational cost in comparison to PCA, inflating the loadings of variables that otherwise might be comparatively weak.

Another issue with PISA's choice of aggregation is that PCA and mean aggregation employ a formative structure, which implies that SES is caused by education, occupational status, and income rather the inverse (Edwards and Bagozzi, 2000; Figure 3.2). When PISA explicitly defines ESCS as the mean or first principal component of PARED, HISEI, and HOMEPOS, it follows that education, occupation, and income are not indicators that are influenced by SES, but rather they are the fundamental constituents of SES. Their acceptability as indicators is not testable because the formative framework is not empirical, but rather definitional. In such a relationship, variance in factor loadings between country and cycle instances of ESCS cannot be accounted for as measurement error because there is no latent attribute which ESCS attempts to quantify. Therefore, the fact that the contributions of PARED, HISEI, and HOMEPOS to ESCS vary across country and cycle (which they have in every PISA cycle) means that each instance of ESCS must be considered as a distinct construct, an implication that invalidates ESCS as a meaningful control in all regression analyses of student ability, except those that are within individual country-by-cycle instances.[2]

[2] "When measuring change, do not change the measure." (Beaton, 1990, p. 165)

Figure 3.2

Reflective vs. formative model structures



The left model represents SES as a unidimensional construct to which education, income, and occupational status are related reflectively. The right model depicts the traditional socioeconomic advantage as a multidimensional construct to which education, income, and occupational status are related formatively.

For example, it is natural to compare PISA's 2018 estimate of ESCS in the United States of 0.11 (OECD, 2019b, p. 252) with the 2015 value of 0.10 (OECD, 2016, p. 401). However, the composition of ESCS changed between these two cycles — the factor loadings of PARED, HISEI, and HOMEPOS in the United States changed from 0.76, 0.80, and 0.73 in 2015, respectively, to 1.0, 1.0, and 1.0 in 2018 when PISA aggregated ESCS as the mean rather than the first principal component. Therefore, the decrease of 0.01 cannot be interpreted in terms of a consistent unit because it is not clear whether variation in "measured" values across contexts was due to variation in the quantity of the SES attribute or to inconsistency in instrument design. In order to determine the true difference in amounts of attributes across contexts, PISA would need to recalculate ESCS values with a standardized instrument with fixed factor loadings — an analysis which is not performed.

More generally, the notion of construct validity is not applicable in formative contexts because there is no theory that stipulates the existence of a real, but unobservable attribute:

> ...although the present validity concept can be applied directly to reflective latent variable models used in psychological measurement, it seems that formative models (Bollen and Lennox, 1991; Edwards and Bagozzi, 2000) do not allow for such application.[3] (Borsboom et al., 2004, p. 1069)

Recall that PISA justifies defining ESCS as a composite of PARED, HISEI, and HOMEPOS by pointing to what has "usually" been done in the past:

> The ESCS is a composite score based on three indicators: highest parental occupation (HISEI), parental education (PAREDINT), and home possessions (HOMEPOS)... The rationale for using these three components was that socio-economic status has usually been seen as based on education, occupational status and income. (OECD, 2019c)

While this justification may be acceptable if SES is strictly a composite score where the relationship is definitional, it does not hold as a rationale for PARED, HISEI, and HOMEPOS as indicators of a latent attribute, nor does it justify the notion of PARED, HISEI, or HOMEPOS as measures of education, occupational status, and income, respectively (which will be discussed in the next chapter). If SES does exist as a personal attribute, then PISA should test the appropriateness of using PARED, HISEI, and HOMEPOS as indicators by analyzing, not only the fit of items to a measurement model, but also their theoretical coherence. For example, the quotation above suggests that constructs (e.g., parental education) and measures (e.g., PARED) are the same thing, when in reality, they are not — rather, the latter is proposed as a measure of the former. One should also keep in mind that education, occupational status, and income are not the only proposed indicators of socioeconomic advantage. White cautions that, "although 'everybody knows' what is meant by SES, a wide variety of variables are used as indicators of SES" (1982, p. 462). Even in PISA's own documentation, the definition

[3]Borsboom et al. (2004, p. 1061) view a causal, reflective modeling structure as synonymous with validity itself, claiming that, "a test is valid for measuring an attribute if and only if a) the attribute exists, and b) variations in the attribute causally produce variations in the outcomes of the measurement procedure."

of SES given by Cowan et al. (2012, p. 4) mentions that "an expanded SES measure could include measures of additional household, neighborhood, and school resources." Likewise, Willms and Tramonte recognize the operational relationship between SES and its traditional components of education, occupation, and income:

> In most studies of the effects of families, schools, and communities on children's academic and social-emotional development, SES is operationally defined with measures describing the occupational prestige, educational levels, and the income of the children's parents. (Willms and Tramonte, 2015, p. 16)

Research has also highlighted empirical concerns with the traditional conception of SES for explaining test score variation. For example, O'Connell (2019) cautions that the explanatory power of household income and parental education for predicting student achievement is not constant as one moves along these variables. He finds that with increasing levels of household income, the explanatory power of household income on MARA (Mathematics and Reading Ability) scores decreases. Conversely, with increasing levels of parental education, the explanatory power of parental education on MARA scores increases.

3.5 Alternative Interpretations of ESCS

It is also evident that PISA does not fully embrace a latent variable measurement interpretation of ESCS because, at times, it is treated as different types of models. For example, ESCS is designed as a descriptive statistic, an aggregation of PARED, HISEI, and HOMEPOS as their first principal component or arithmetic mean. As discussed earlier, PCA is a method that attempts to capture the maximum amount of data variance through a fixed number of parameters and does not recognize latent variables from a reflective perspective. Furthermore, one of the components of ESCS, HOMEPOS, has been redesigned between cycles to capture as much variance in observed data as possible. PISA justifies these changes by citing authors (e.g., Kreiner and Christensen, 2014, p. 225; Oliveri and Von Davier, 2011, p. 329) who advocate for replacing the more parsimonious Rasch model (whose fundamental relationship to measurement will be discussed further in the next chapter) with more generalized IRT models, such as the 2PL and GPCM models, suggesting the that validity of HOMEPOS is a function of data-to-model fit, rather than the degree to which it represents quantities of an attribute.

The validity evidence that PISA provides in its technical reports, however, does not reflect a latent variable interpretation of measurement, nor that of a summary statistic. For descriptive modeling, a summary statistic might be considered valid if it explains maximal variance in the observed data. PISA does not validate ESCS as one would typically validate a summary statistic though – by comparing it to alternative statistics. While PCA captures maximum variance by design, PISA does not argue that HOMEPOS is a more informative description of one's household possessions than other model designs.

Of course, it is not surprising that PISA does not engage in such an effort. It is doubtful that neither PISA stakeholders nor PISA itself interprets ESCS strictly as a summary of data variance. To the contrary, ESCS values published in the PISA literature are intended to be interpreted as measures, albeit not according to latent variable theory. Specifically, PISA's validation effort, which consists of an analysis of internal consistency and reliability of ESCS factor loadings across countries using Cronbach's alpha, implies a classical test theory (CTT) perspective of measurement.

CTT is a common framework which conceives of attributes abstractly, as "true scores." True scores are epistemologically inaccessible, and therefore, "observed scores" serve as estimates of true scores obscured by a degree of error. Under CTT, validity and reliability are indistinguishable concepts, both referring to the proximity of an observed score (X) to the true score (T), and is impossible to gauge because both the true score and the amount of error (E) are unobservable:

$$X = T + E \tag{3.1}$$

Therefore, the quality of an instrument is assessed solely by its reliability, in practice determined by the internal consistency of the observed scores which it records. Accordingly, in its validation description, PISA states that, "similar and high values across countries are a good indication about having measured reliably across countries" (OECD, 2017, p. 295). The strict reliability of an instrument is difficult to ascertain in most practical settings because it relies on the application of parallel testing – obtaining repeated measurements under identical testing conditions (Lord and Novick, 1968):

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2 \tag{3.2}$$

...where X' is a theoretically parallel score to X. For this reason, Cronbach's alpha (Cronbach, 1951) is typically employed as an alternative:

$$\alpha = (\frac{k}{k-1})(1 - \frac{\sum_{j=1}^{k} \sigma_{U_j}^2}{\sigma_X^2}) \tag{3.3}$$

...where $\sum_{j=1}^{k} \sigma_{U_j}^2$ is the sum of the variances of k individual items, u_j, $j = 1, ..., k$, in test form U, and σ_X^2 is variance in overall test scores. Values of the statistic range from 0 to 1, where values closer to 1 reflect a higher degree of variance in within-question responses, as opposed to variance in the between-question response patterns.

There are several problems with CTT as a measurement theory, however. First, CTT operationally equates the true score with the attribute being measured, simply defining SES as the expected score produced by the ESCS instrument. In doing so, it denies a realist interpretation of the attribute (Borsboom, 2005). A classic example of operationalism is the "measurement" of one's Intelligence Quotient (IQ), where a score on the IQ test is often subsequently treated as an inherent quality of a person. The trustworthiness of this procedure, however, is based only on social perception, rather than objective properties of IQ as an attribute. The concept of operationalism was first introduced by physicist Percy Bridgman (1922), who did not foresee the consequences that his new concept of measurement would hold for the social sciences, most notably psychology. At the time, there was broad uncertainty regarding the nature of psychological measurement, and indeed, whether it was at all possible. This uncertainty led to the establishment of the Ferguson Committee by the British Association of the Advancement of Science in 1932 with the goal of deciding whether psychological assessment could be considered measurement, and if so, under what circumstances (Markus and Borsboom, 2013, p. 27). Ultimately in its 1940 Final Report, the committee decided that given the absence of evidence for additivity for psychological constructs, "any law purporting to express a quantitative relation between sensation intensity [i.e., observation] and stimulus intensity [i.e., attribute] is not merely false, but it is in fact meaningless" (Ferguson et al., 1940, p. 245).

Stevens (1946), however, became a vocal advocate for operationalism as a means of preserving existing paradigms of psychological measurement, famously stating that measurement is "the assignment of numbers according to a rule," a definition that even Bridgman would ultimately disavow. The rule-based assignment process has been criticized by many metrologists and philosophers of science. As Michell (2005, p. 286) puts it, the "central principle [of operationalism] was that the concepts investigated in science are constituted by the operations used to measure them, thereby confusing what is measured with how it is measured and denying the logical independence of what is known from the process of knowing it." McGrane (2015) describes the acceptance of operationalism in the field of psychology as a substitute for systematic approaches to measurement, such as those that have lent such credibility to the physical sciences.

Classical test theory suffers from other conceptual failings, as well. For example, there are issues with the statistical properties of the "platonic" true score (i.e., the singular objective true score) because true scores at the extreme limits of the testing range (e.g., 0 and 100 for a 100-point test scale) can only have positive and negative errors applied to them, respectively, to produce the corresponding observations. This means that the correlation between true score and error is not zero, as larger true scores will tend to have more negatively skewed error and smaller true scores will have more positively skewed error. This, in turn, violates the stipulation of the CTT model that true scores and errors be uncorrelated (Lumsden, 1976). Furthermore, if the true score is defined as the expected value of observed test scores (Lord and Novick, 1968), the variance of the observed scores will be higher in the middle of the observation range than at the extremes. This violates the assumption of homoscedasticity, which is a requisite for all linear regression models, including the CTT model (Lumsden, 1976).

Also, the notion of completely independent parallel testing upon which CTT relies – where the true score is obtained when independent observations are averaged over an infinite number of repetitions – is a fallacy. Proponents of CTT illustrate this idea with the so-called "Mr. Brown" thought experiment (Lazarsfeld, 1959; Lord and Novick, 1968), in which repeated, but independent, observations would be produced if the testing subjects were "brainwashed" between each testing application, completely removing any memory or learning effect that they would have acquired through the initial testing (and without incurring any other neurological side-effects!). This thought experiment, however, cannot justify real-world applications of CTT. Not only is the "brainwashing" mechanism not testable because it does not exist, multiple observations can never be made in practice because testing is inherently dynamic — each observation occurs at a different moment in time under slightly different circumstances. Also, if the Mr. Brown brainwashing treatment is supposed to recreate the observational conditions under which the initial measurement was carried out, then the exact sources of error confounding the original measurement would be reproduced as well (Borsboom, 2005), meaning that the expected value of error is not zero.

Furthermore, the nature of error under CTT is unclear. In many real-world testing situations, the degree of error that we actually observe is not plausible under CTT. Take for example, the item format, "Do you agree with the following statement? — Yes/No": if hypothetical observations of "yes" and "no" were collected under the thought experiment, how would these be averaged? Often, the notion of error does not even make sense in such situations. Can a respondent believe that he agrees with the statement when he "actually" does not? Also, the separation between different sources of error is not considered — should mechanical errors (e.g., accidentally filling in the wrong bubble on a test form) warrant equal consideration in the estimation of the true score as errors due to misunderstanding?

It is noteworthy that a realist interpretation of measurement was actually compatible with the original applications of CTT in the astronomy context under which the law of averages was originally applied (e.g., the work of the 19th century Belgian astronomer and sociologist Adolphe Quetelet), as compared to the psychological contexts to which the logic of CTT has been applied in the past century. In averaging repeated measurements of planetary locations, CTT was not at odds with a realist perspective of an attribute. Presumably, it was understood that planets' physical locations existed in a realist sense. While the "true" location of a planet was unknown from an epistemic standpoint, its ontological existence was never in question. Also, in the vacuum of space, a repeated measurements scenario is less problematic – there are far fewer threats to the independence of planetary measurements than with psychological measurements, as the action of observation does not affect the subject of measurement (planets do not know or care that they are being measured), and observational conditions are more uniform. The CTT framework only abandons its realist ontological commitments when applied to psychology where, unlike the existence of planets, the reality of many popular constructs is questionable. For example, intelligence can alternatively refer to an objective, but hypothetical, physical property of the brain, an "emergent" representation of various component cognitive processes, or to a loose bundle of separate abilities and behaviors that is socially convenient to subsume under one label. Each of these possibilities entails a distinct degree of "realness," but this is rarely considered to disqualify intelligence as a measurable attribute in applied psychological contexts.

There are concerns that are more directly relevant to PISA, as well. In refusing a realist interpretation of the attribute, CTT methods both prevent the testability of the relationship between ESCS and SES (as SES is defined as the expected value of ESCS) and the interpretation of SES as a causal agent on student achievement (the mean is not "caused" by the values averaged, after all). Once again, PISA stakeholders perceive SES as being able to causally impact student achievement, which is why it is useful as a control variable in the first place. Also, Cronbach's alpha describes the ratio of covariance between item scores across tested individuals and variance within item scores across individuals, and therefore, does not recognize error at an individual level, but rather at the population level, compromising the interpretation of SES as a property of an individual.

Regardless of the appropriateness of CTT validation methods, the differences in factor loadings and reliabilities of PARED, HISEI, and HOMEPOS between countries and cycles are often substantial, casting doubt onto whether sufficient internal consistency can be claimed at all. In the 2009, 2012, and 2015 cycles, the reliabilities of HISEI, PARED, and HOMEPOS across OECD countries have a standard deviation 0.065 (with a minimum reliability of 0.53 in Iceland in 2009 and a maximum reliability of 0.80 in Mexico in

2012) — three and a half times higher than the standard deviation of the reliabilities within each country of 0.020 during the same period. Similarly, we can also see that the differences in HISEI, PARED, and HOMEPOS factor loadings between countries, when observed across cycles, are greater than the differences within individual countries across cycles (Table 3.1). This suggests that variation in factor loadings between countries is related to country-specific characteristics, calling into question whether low variation can universally be considered to be indicative of instrument quality.

Table 3.1

Variance in HISEI, PARED, and HOMEPOS factor loadings between 2012, 2015, and 2018

Statistic	**HISEI**	**PARED**	**HOMEPOS**
Between-country variance	0.0023	0.0027	0.0054
Within-country variance	0.0002	0.0003	0.0005
F-statistic	0.1050	0.1213	0.0946

Also, the consistency is not stable within certain countries. For example, the factor loadings for HISEI and PARED in Poland generally rise from 2009 to 2015 (Table 3.2).

Table 3.2

Factor loading drift in Poland between 2009 and 2015

Cycle	HISEI	PARED	HOMEPOS	Reliability
2009	0.81	0.8	0.71	0.65
2012	0.87	0.87	0.74	0.75
2015	0.89	0.88	0.71	0.75

It should be noted that, aside from latent variable theory and CTT, the other well-known measurement theory is the representational theory of measurement, (Krantz et al., 1971) which views measurement as the mapping of a set of observations to a numerical system. According to this theory, the quantity of a construct corresponds to a ratio between two observations (Borsboom, 2005, p. 4). Representational theory is seldom considered in applied social sciences due to the difficulty in discovering a mathematical function that perfectly describes the numerical relationships between empirical observations. Furthermore, it makes no metaphysical claims regarding constructs that are of interest to most consumers of PISA reports: the nature of student attributes, observed test scores, and the causal mechanisms between them. Nor does representational theory recognize the concept of measurement error; rather, any inconsistency in the isomorphism between observations and predicted values of a model is the result of an insufficiently specified mapping function. For these reasons, ESCS values are not naturally interpretable under representational theory, and it will not be discussed further.

Chapter 4

PARED, HISEI, and HOMEPOS

A thorough validation effort of ESCS necessarily entails separate validity analyses of each of its three component instruments: PARED, HISEI, and HOMEPOS. In this chapter, I argue that these constituent instruments are not valid measures of education, occupational status, and income, respectively, but rather operationalizations where attribute and measure are equated by definition.

4.1 The Interpretation of PARED, HISEI, and HOMEPOS as Measures

PISA treats PARED, HISEI, and HOMEPOS as measures of education, occupational status, and income when it assigns numerical values to them. As with ESCS values, interval differences between these values are intended to be meaningful, as opposed to being interpreted only as categorical descriptions or ordinal rankings. Specifically, the difference between two scores implies that one individual has "that amount more" education, occupational status, or income than another individual.

The interpretation of HISEI and HOMEPOS as measures is also reflected in their design. As the contribution of a specific ISCO occupational category to income, HISEI is a linear transformation of income. As income satisfies the conditions for measurability (being a quantitative attribute with an accepted unit), HISEI is interpretable in terms of measurement, as well. HOMEPOS, on the other hand, is interpretable as a measure because items in the contextual questionnaire are generally considered to be "indicators" of a latent income attribute:

> Collecting information about household possessions as indicators of family wealth has received much attention in international studies in the field of education (Spiezia, 2011; Traynor and Raykov, 2013). Household assets are believed to capture wealth better than income because they reflect a more stable source of wealth.[1] (OECD, 2014, p. 316)

HOMEPOS items are supposedly interchangeable without implying the measurement of a distinct construct, contrasting with items which constitute an index, a formative structure where household items are exclusive and exhaustive components. Under a formative model, rather than indicating a hypothesized income attribute, a given item set is defined to be income (Figure 4.1), and substituting index items changes the fundamental nature of the construct, rendering different indexes incomparable:

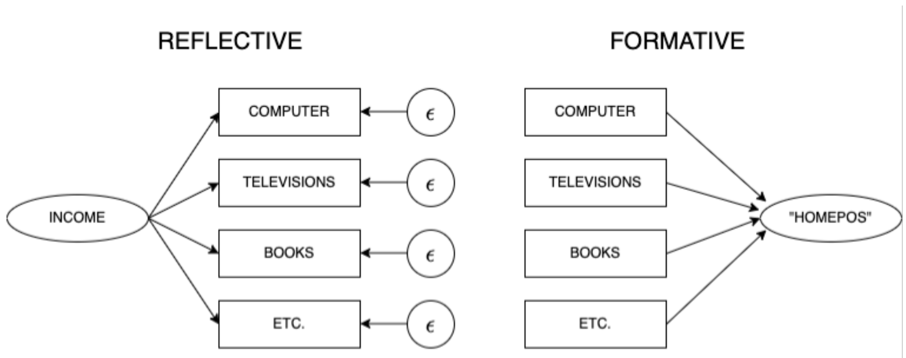
> "Altering the indicators of an index changes the definition of the variable being indexed, whereas changing the indicators for a measure will not alter the latent variable (although precision of measurement and or unit size may be affected). So, if midline girth is added to height and weight as indicators of stature or all electronic commodities are eliminated from the Consumer-Product-Index (CPI) market basket, the definition of what is being indexed changes." (Stenner et al., 2008)

PISA clearly intends HOMEPOS items as indicators, rather than as fundamental

[1]Although PISA claims that HOMEPOS is a measure of income, it sometimes claims to capture wealth. Note that wealth is not synonymous with income in this case, as PISA claims that HOMEPOS "captures wealth better than income" (OECD, 2014, p. 316). Wealth cannot automatically be synonymous with income because to do so would be to claim that HOMEPOS is a better indicator of income than reported income — a strong claim that needs to be supported (and is not argued for by Spiezia, 2010; nor Traynor and Raykov, 2013).

Figure 4.1

Contrasting formative and reflective structures of HOMEPOS



components of an index because ESCS estimates are almost exclusively used for cross-country and cross-cycle comparison, contexts across which item membership is variable. When the base item set and country-specific items are updated between cycles (which has occurred between every PISA cycle to date), PISA intends to incorporate indicators that are more theoretically appropriate or are better targeted to differentiate between higher or lower quantities of the underlying income attribute. The reflective approach is also necessary because PISA intends to take the entirety of family wealth into account, not just the portion described by the household items in the questionnaire. Presumably, PISA would not claim that the specific collection of household items that appears in the questionnaire (books, desks, E-readers, etc.) in each cycle actually defines wealth to the exclusion of items that do not appear on the questionnaire. There are other alternative household possessions that also contribute to wealth (e.g., property, financial assets in investment accounts, private health insurance plans, etc.). The reason PISA does not ask students about other possessions is not that they are not representative of wealth, but rather that they are less identifiable and/or quantifiable to the students who complete the questionnaire. PISA even explicitly acknowledges, albeit passingly, that the nature of ESCS fundamentally changes when items are substituted under an index approach, but considers this a technicality:

> "For some scales, some countries opted to delete one or two items. Strictly speaking, this constituted a different scale and, therefore, a footnote was added in the tables to note which item had been deleted." (OECD, 2019c)

Despite PISA's interpretation of HOMEPOS as a set of indicators, it is at times treated as an index, thereby conflating income as both an index and a latent variable. This kind of confusion of formative and reflective structures is common in the broader applied measurement literature, as well (Edwards and Bagozzi, 2000; Stenner et al., 2009). Not only is HOMEPOS consistently referred to as an "index" in PISA literature (e.g., OECD, 2019c), PISA's division of the home possession item set into multiple sub-indices (family wealth possessions [WEALTH], cultural possessions [CULTPOSS], home educational resources [HEDRES], and information communication technology resources [ICTRES]; see Table 2.2) contradicts the notion that HOMEPOS items are indicators of a single latent attribute. Also, the latter three sub-indices are distinct from conventional conceptions of income. Moreover, CULTPOSS items like musical instruments and books of poetry to income are presumably related to income through the other SES indicators, education, and occupation — a connection which violates the independence of income, education, and occupation as indicators of SES.

4.2 PARED, HISEI, and HOMEPOS are Operationalizations

Despite the interpretation of PARED, HISEI and HOMEPOS as measures of education, occupation, and income, in reality, they are only operationalizations — they equate the object and method of observation, disregarding the existence of an attribute in the realist sense (see Chapter 3 for a discussion of operationalism). PISA's designation of PARED and HISEI as "simple indices," defined as "variables that are constructed through the arithmetic transformation or recoding of one or more items in exactly the same way across assessments" (OECD, 2019b, p. 212) closely follows Stevens's (1946) definition of operational measurement as, "the assignment of numbers according to a rule." Operationalizations cannot be validated because the relationship between the instrument and a hypothesized latent attribute is not testable, as a latent attribute is not recognized. Accordingly, no theoretical explanation is given as to why they would be valid measures of any latent attribute nor why they might be preferable to alternative quantifications. However, adopting an operationalist approach to measurement does not do away with the need for validation, as the structure of the operationalizing instrument is still motivated by a theory of an attribute, and therefore, still depends upon a realist conception of the attribute. Such a theory dictates which variables should be observed, which instruments should be used to observe them, and how observed data should be processed and presented. For example, even if SES is operationalized as ESCS, ESCS is still designed according to certain implicit theories of how responses to questionnaire items would be influenced by education, occupation, and income. From this perspective, to adopt an operationalist approach to measurement is to use a realist theory of an attribute in the construction of a measure, then subsequently deny the reality of the attribute when the instrument is applied. To believe that one can collect data in a theory-neutral vacuum, and then present the data to researchers who subsequently generate an independent theory of SES ex-post-facto is to misunderstand the scientific method (Tal, 2020). As Kuhn (1961, p. 189) describes this idea of theory-ladenness, "the road from scientific law to scientific measurement can rarely be traveled in the reverse direction." Furthermore, where in a realist measurement paradigm, discrepancies between repeated measures or measures obtained with different instruments could be accounted for as measurement error, this is not possible in the operationalist framework. Indeed, in the PISA assessment, measurement error is not reported alongside estimates of PARED, HISEI and HOMEPOS. Therefore, if separate instances of applying the ESCS instrument to the same subject were to result in different readings, the only possible interpretation is that different attributes were measured. Once again, the notion that distinct versions or instances of ESCS cannot measure the same attribute is limiting because comparisons of "measures" across cycles and countries are no longer possible.

4.2.1 PARED as an operationalization of education

The operationalism of PARED as education is clear-cut: PISA prompts students for the highest degree earned of the most educated parent, and then scales this categorical variable into a numerical variable: years of schooling of the most educated parent. PISA then interprets this number, not as a measure of the parent's education, but as the parent's education itself. While PARED might, in fact, be a reasonable measure of latent education (if that construct indeed exists), it cannot automatically be considered as such. After all, PARED is not the only way that a person's education level can be quantified, and therefore, evidence is required to support its validity as a measure.

4.2.2 HISEI as an operationalization of occupational status

Occupational status is operationally defined as HISEI. Like in the case of PARED, PISA offers no evidence to support the validity of HISEI as a measure of occupational status. Ganzeboom et al. (1992) discuss the validation of ISEI, the base construct from which HISEI is derived (HISEI is the highest of the parents' ISEI scores), but this validation effort is not referenced or cited in PISA documentation, nor do Ganzeboom et al. discuss the application of ISEI to international testing contexts in their validity discussion.

The mechanism of operationalization is not as direct as in PARED. A path model is used to equate occupational status with an estimated parameter that fits observed data while respecting a minimization constraint (see Chapter 2, e.g., Figure 2.4) – a practice that perhaps suits descriptive goals but does not align with the requirement for measurement that the differences between "measured" values be proportional to differences in the quantity of a latent attribute. For example, although the ISEI values are interval-scaled (i.e., the difference between ISEI scores of 20 and 30 is intended to be interpreted as equal to the difference between ISEI scores of 30 and 40), the ISEI "unit" is difficult to meaningfully interpret in relation to an attribute. Likewise, the ISEI model is unable to be validated: the criteria for determining acceptable model fit are necessarily arbitrary and residual variance from the ISEI model cannot be interpreted as measurement error. Rather, it can only be interpreted as the variance in the data which is not explainable by the occupation, income, education, and age variables in the model — it does not refer to the estimated difference between estimated occupational status and actual latent occupational status because the model does not recognize occupational status as a causal attribute.

Moreover, ISEI is not designed to reflect "status" – there is no reference to any sociological "status" construct by Ganzeboom et al. (1992). In fact, the authors even tout the separation of ISEI from the occupational prestige construct utilized in Duncan's (1961) SES model:

> The advantages of our procedure over [Duncan's] older one is simply that (a) the logical relationship with prestige is completely eliminated and (b) it gives a clearer interpretation to SEI. (Ganzeboom et al., 1992, p. 12)

> Although socioeconomic indexes (SEI) of occupational status initially were developed as a way to generalize prestige scores for all occupations (Duncan, 1961), the operations used to derive SEI scales in fact have little to do with prestige scores (Hodge, 1981; Ganzeboom et al., 1992). (Ganzeboom & Treiman, 2003, p. 161)

Rather, ISEI is defined only "as the intervening variable between education and income" (Ganzeboom et al., 1992, p. 11), minimizing the direct effect of education on income and maximizing the indirect effect of education on income through occupation using data from the original 31 ISCO occupation/income datasets. Therefore, if ISEI were a measure, it would be a measure of "occupationally derived" income, rather than of prestige or social status. When Ganzeboom and Treiman (2003) "conceive of ISEI as measuring the attributes of occupations that convert a person's education into income," (p. 171) the only occupational attribute that it can be considered to describe is a profession's propensity to provide more income. To state that "a job's propensity to result in more income" is "the attribute that converts education into income" is tautological. Of course, PISA is also not using HISEI to measure income, as HOMEPOS is already intended as the measure of income. To use HISEI as an income measure would be redundant because ESCS would then measure "education, income, and income related to category of occupation."

Another validity issue is that the ISCO categories and the ISEI scale are, themselves, not rigorously validated. They are supported by evidence of criterion validity alone: ISCO-08 is validated by comparing the structural relationships between the educations of two spouses, their occupations, and their common household income, and then comparing the strength of these relationships with those of ISCO-88 (released in 1988; Ganzeboom and Treiman, 2010, p. 17). The ISCO-88 occupation scale is, in turn, validated in comparison to ISCO-68 (released in 1968). The last link in the chain, ISCO-68, to my knowledge, has never been seriously validated. ISEI is validated by Ganzeboom et al. by comparing its model fit against those of two other "measures" of occupational status, the Standard International Occupational Prestige Scale (SIOPS) and the Erikson, Goldthorpe and Portocarero class schema (EGP; Erikson et al., 1979; Ganzeboom et al., 1992, p. 19). There is significant academic resistance to composite measures of occupational status for other reasons, as well (Boyd, 2008). For example, there is no agreement regarding the extent to which occupational status reflects prestige, as opposed to more traditional economic power. Also, composite scales of occupational status have been shown to be sensitive to gender gaps and to be inadequate for intergenerational comparisons of social inequality (e.g., Hauser and Warren, 1997). IPUMS USA goes so far as to caution that the use of composite measures of occupational status, such as SEI, may be "scientifically obsolete" (IPUMS USA, n.d.).

The preference for data-model fit and statistical power over the ability to conserve consistent relationships between attribute and observation is also evident in how PISA imputes HISEI values.

> In PISA 2018, in order to reduce missing values, an ISEI value of 17 (equivalent to the ISEI value for ISCO code 9000, corresponding to the major group "Elementary Occupations") was attributed to pseudo-ISCO codes 9701, 9702 and 9703 ("Doing housework, bringing up children", "Learning, studying", "Retired, pensioner, on unemployment benefits"). (OECD, 2019c)

The rationale for "reducing missing values" given by Avvisati (2020, p. 17) was to boost the statistical power of the country invariance analysis, stating that, "PISA can eliminate one source of cross-country differences in missing rates and thereby improve cross-country comparability." This procedure of imputing occupational status is not conceptually or empirically justified, however. ISCO documentation refers to "elementary occupations" as jobs that involve "the performance of simple and routine tasks which may require the use of hand-held tools and considerable physical effort" (ILO, 2008). Perhaps PISA is assuming that these jobs are relatively low paying compared to those of the other categories, and therefore, are the most appropriate to be grouped with "non-occupations," such as students, homemakers, and pensioners. However, there are problems with this comparison. First, the ISCO categories are not designed to reflect income, but rather qualitative features of the professions. Accordingly, there is often a large median salary range among the listed occupations in a given category. In fact, there are many lucrative jobs in the elementary occupations category, especially regionally, such as fisherman, miners, and construction workers. For example, in Nevada miners have an average income of almost $98,000 per year, the third highest average paying employment sector in the state (Perry and Visher, 2019). In Alaska, crab fisherman can make $100,000 per year (Alaska Department of Labor and Workforce Development, n.d.). It is not appropriate to universally assign an ISEI value of 17 (out of a range from 11.01 to 88.96) to these professions, nor to group them with other "elementary occupations" that are more uniformly lower paying, such as "food preparation assistants," or with "non-occupations" that have little to no income stream. Additionally, many "non-occupations" may not produce significant direct income but may still be indicative of significant indirect income in the form of savings (e.g., retirees/pensioners), spousal income (e.g., homemaker or stay-at-home parent), or future income (e.g., students). The reality is too complex to assume that these occupations are equal indicators of relative poverty across sociogeographical contexts.

4.2.3 HOMEPOS as an operationalization of income

HOMEPOS values in the 2015 and 2018 cycles are estimated with the GPCM (see Chapter 2), which should not be considered a measurement model because it does not have the property of specific objectivity, in which the "parameters of the subjects in the subgroup can be evaluated without regard to the parameters of the other subjects" (Rasch, 1960):

> Specific Objectivity is the requirement that the measures produced by a measurement model be sample-free for the agents (test items) and test-free for the objects (people). Sample-free measurement means 'item difficulty estimates are as independent as is statistically possible of whichever persons, and whatever distribution of person abilities, happen to be included in the sample.' Test-free measurement means 'person ability estimates are as independent as is statistically possible of

> whichever items, and whatever distribution of item difficulties, happen to be included in the test.' (Wright and Linacre, 1987)

For example, specific objectivity allows for the common interpretation of the "inch" unit, regardless of what object is being measured or whether a tape measure, yardstick, or digital caliper is used to take the measurement.[2] This condition is especially important in the PISA context because students do not receive the same items outside of national samples. Within each PISA cycle, students receive country-specific items, and between PISA cycles the overall item set consistently changes.

Consequently, as opposed to the PCM (used prior to 2015), which does have specific objectivity because it is in the family of Rasch models, estimates derived by the GPCM from different country and cycle samples are only comparable in those local contexts. This is a problem, as the most useful inferences drawn from PISA involve the improvement or decline of national proficiency levels over time for the evaluation of the effectiveness of educational policies applied during that same period. PISA references Glas and Jehangir (2014, p. 98) who recommend using the GPCM model to "identify and account for culture-specific differential item functioning." It is important to recognize, however, that while the GPCM model may be useful for *identifying* DIF as a source of bias in local contexts, this consideration cannot outweigh that of maintaining the fundamental requirements for measurement.

GPCM modeling also presents interpretational issues even within country-by-cycle groups because individuals within these groups are not comparable by a meaningful unit. As a result, under the GPCM, it is rarely possible in practice to map items to different income levels (e.g., using a Wright map [Wilson, 2004; Briggs, 2019]). This also complicates instrument validation because it prevents the researcher from developing

[2] Also, under the condition of specific objectivity, a person's raw score is a sufficient statistic of his ability level.

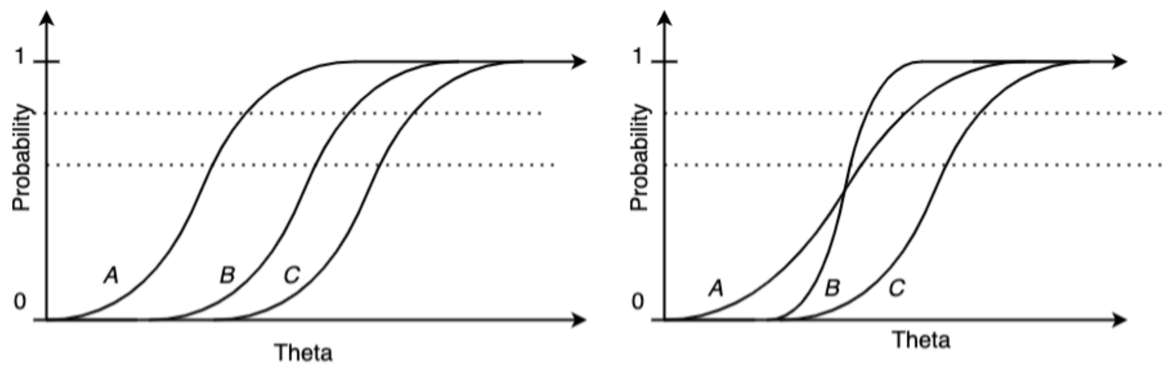
testable a priori hypotheses regarding the relative item difficulties. For example, it is typically impossible in practice to hypothesize the exact point where the ICCs for two items will cross (Figure 4.2). The Rasch model is important for this interpretability because, under it, item characteristic curves (ICC) do not cross because they have equal slopes. For example, if a “refrigerator” item has a greater difficulty parameter than a “washing machine” item, under the Rasch model, the logit distance between the two items will always be the same for all individuals under all contexts. This allows one to infer that the ratio of family income needed to purchase a refrigerator versus a washing machine is always constant. On the other hand, when these two items vary in their slopes in an IRT model, the ICCs cross at one point, inverting the relative item difficulties at a certain income level. Thus, a refrigerator is easier to purchase than a washing machine for some individuals, but not for others. While one might argue that the allowance for flexible slopes is beneficial because it incorporates information regarding the strength of the relationship between the presence of the indicator and the respective attribute, such theories are typically too complex to be generated prior to data collection, and therefore, are impractical to validate through falsification.

While it is true that IRT models conserve interval quantities of latent variables, albeit only in local contexts, I argue that the motivation behind the switch from PCM to GPCM is distinctly operationalist. As with the construction of HISEI, if model fit considerations are placed above measurement criteria, the choice of model becomes arbitrary from a measurement standpoint. Defining measurement, as the application of an arbitrary process is operationalism. Therefore, HOMEPOS must be considered to be an operationalization of income.

There are several other reasons why HOMEPOS values cannot be viewed as meaningfully comparable. First, while invariance is a fundamental assumption of IRT modeling (Hambleton and Rogers, 1989; Mellenbergh, 1982; Meredith, 1993; Millsap, 2012; L.

Figure 4.2

Example item characteristic curves under the Rasch model and the 2PL (or GPCM) IRT model



The slopes of the left pair of ICC curves are equal (as under the Rasch model), whereas the slopes of the right pair are not (as under the GPCM). The instrument on the left has the property of "specific objectivity" because the ratio of the distances between curves A and B, and curves B and C is always constant, ensuring the consistent interpretability of the logit unit across measurement contexts.

Rutkowski and Rutkowski, 2016), the invariance of ESCS (nor that of PARED, HISEI, or HOMEPOS) to other demographic attributes like gender, race, and language group has not properly established by PISA, even though Kreiner and Christensen (2014) demonstrate that nationality has had a meaningful impact on PISA rankings on certain sub-scores of reading achievement, and PISA acknowledges the threat of heterogeneous response bias in its technical report:

> ... measures [of the same construct in different national and cultural contexts] can suffer from various measurement errors, for instance, students are asked to report their behaviour retrospectively. Cultural differences in attitudes towards self-enhancement can influence country-level results in examinees' self-reported beliefs, behaviours and attitudes (Bempechat et al., 2002). The literature consistently shows that response biases, such as social desirability, acquiescence and extreme response choice, are more common in countries with lower socio-economic development, compared with more affluent countries. Within countries, these response styles differ between gender and across socio-economic status levels (Buckley, 2009). (OECD, 2017, p. 295)

While analyses of cross-cycle comparability in individual HOMEPOS items (e.g., Pokropek et al., 2017) have found that HOMEPOS items are largely invariant to time, other research suggests that specific HOMEPOS items are not comparable across country contexts. For example, criticism has highlighted that some countries demonstrate much lower reliability on the HOMEPOS sub-scales: L. Rutkowski and Rutkowski (2010) find that several lower- and middle-income countries have unacceptable levels of reliability on certain items in the HOMEPOS cultural possessions (CULTPOSS) sub-scale. D. Rutkowski and Rutkowski (2013) find low reliability on the 2009 sub-scales, WEALTH, HEDRES, and CULTPOSS and that "reliability estimates across the three sub-scales are highly varied by scale and country" (p. 268).

There is also significant cross-country variation in the degree to which HOMEPOS items adequately represent the HOMEPOS sub-scales. Rutkowski & Rutkowski (2013) conduct a cross-country comparison of the degree to which the HOMEPOS item parameters fit a confirmatory factor analysis in which the three HOMEPOS sub-scales from the 2009 cycle are hypothesized as factors. They find that "few countries" meet three out of four of their minimum fit criteria (CFI/TLI>.90; RMSEA/SRMR<.08; p. 268).

Pokropek et al. (2017) examine which 2012 HOMEPOS items and countries meet an array of acceptable "comparability criteria." The chosen criteria are modification indices (Kaplan, 1989; Whittaker, 2012), expected parameter change (Kaplan, 1989; Whittaker, 2012), and root mean square deviation (RMSD). The countries and items that meet acceptable model fit under more criteria have higher "c-indices" and "i-indices," respectively. The authors conclude that over half of the home possession items should not be considered comparable across-countries.

Lee and von Davier (2020) expand upon the Pokropek et al. (2017) study by analyzing longitudinal invariance of HOMEPOS items across all cycles from 2000 to 2015 and cross-country invariance across a larger body of PISA-participating countries. They find that HOMEPOS items are not invariant across country and sub-national language group contexts.

PISA addresses these comparability concerns by estimating new item parameters for each cycle and estimating unique parameters for country-specific items and other items that do not meet the RMSD fit criteria for certain groups:

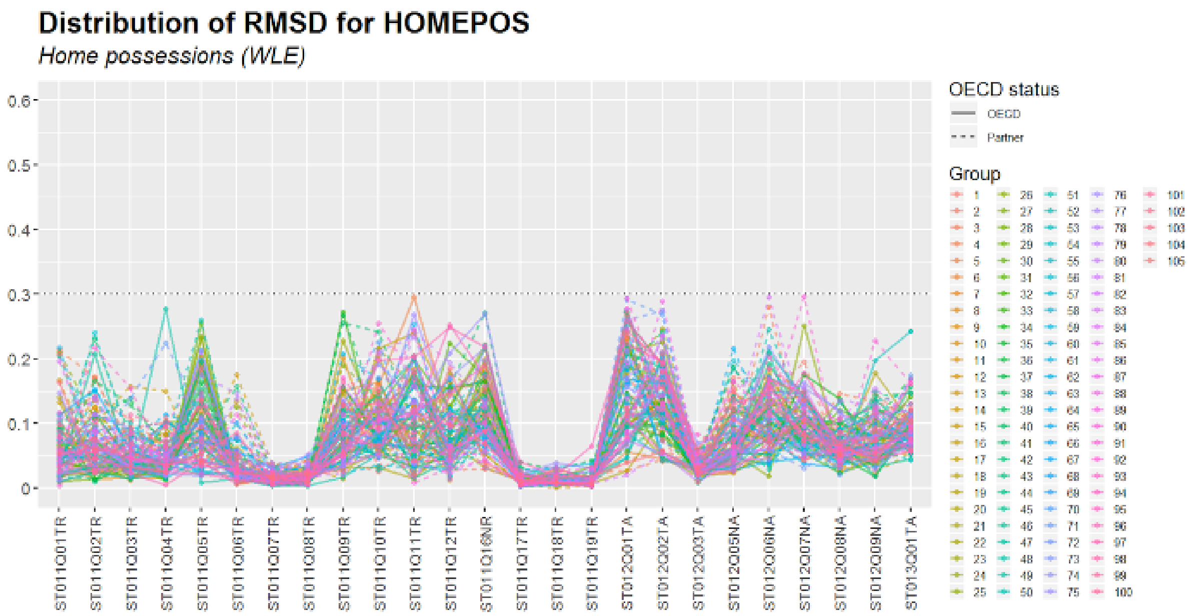
> The comparability of these indicators across countries and over time raises several challenges (Rutkowski and Rutkowski, 2013; Rutkowski and Rutkowski, 2017; Pokropek, Borgonovi and McCormick, 2017). The more serious concerns are related to the items proxied by home possessions, as the meaning and the national examples included in the items may vary significantly across countries, undermining cross-country comparability. In addition, the prevalence of access to technological goods and services, such mobile phones, has increased over time, thus these items convey distinct information at different times. For example, use of a mobile phone shortly after the technology was introduced could be a proxy for high social status; later on, mobile phones may be regarded as a basic resource, accessible to nearly everyone. For this reason, the index summarising home possessions is computed in a different way for all new cycles, and some items may be included in a way specific to each country, in order to take into account distinctive use by countries. (OECD, 2019b, p. 52)

However, due to the theoretical considerations of measurement discussed earlier, this approach is not an appropriate way to address these criticisms because PISA is effectively targeting a different construct each time it re-estimates item parameters under the GPCM. Moreover, the chosen items should not even be established as being indicators of income without some sort of explicit justification.

Finally, the RMSD fit cutoff of 0.3, which serves as the criterion for sufficient item-data fit, is arbitrary. There does not appear to be any justification for the selection of this value. The figures in Annex F of the 2018 Technical Report suggest that this is either a choice of convenience, as most items fall under this cutoff and 11 items have at least one country-by-language group RMSD statistic between 0.25 and 0.3 (Figure 4.3), or that an unknown number of items were culled because they exceeded it.

Figure 4.3

2018 RMSD statistics for HOMEPOS item by country-by-language group



External validation of HOMEPOS

PISA does not investigate the correlation of HOMEPOS with alternative established measures of family income or national income like GDP/GNI per capita in an attempt to establish external validity. While Avvisati (2020) observes that marginal ESCS values from the 2018 cycle moderately correlate with independent estimates of marginal gross national income per capita, the conditional correlations between GDPs per capita and ESCS components across cycles (Table 4.1) and country-groups (Table 4.2) are quite variable. In fact, some OECD member countries (Austria, Belgium, Colombia, Czech Republic, Estonia, Germany, Hungary, Iceland, Italy, Japan, Luxembourg, Mexico, Norway, Slovak Republic, Turkey, and the United States) even exhibit negative correlations in some cases.

While these correlation coefficients are based on few data points (one per cycle in

Table 4.1

Correlation of national estimates of HISEI, HOMEPOS, ESCS with GDP (PPP) per capita

Cycle	HISEI	HOMEPOS	PARED	ESCS
2000	0.36	-	-	-
2003	0.62	0.64	0.58	0.57
2006	0.54	0.64	0.48	0.53
2009	0.56	0.60	0.49	0.62
2012	0.56	0.57	0.44	0.51
2015	0.43	0.59	0.44	0.56
2018	0.50	0.66	0.47	0.56

Note: Official estimates of HOMEPOS, PARED, and ESCS are unavailable for the 2000 cycle (pisadata-explorer.oecd.org).
GDP per capita data from data.worldbank.org.

which the country has participated), this is a threat to the general validity of ESCS, as policymakers in countries which report non-positive correlations may be able to reasonably suggest that "the construct that ESCS measures" is irrelevant in their local context, or even that it measures the opposite of income in the case of a negative correlation, and therefore, could be ignored entirely. Pokropek et al. (2017) and Avvisati (2020) examine the external validity of HOMEPOS, but I consider their results to be inconclusive. Avvisati (p. 19) compares national HOMEPOS estimates against national GNI per capita estimates and finds a correlation of 0.65 with GNI per capital (0.80 if the latter is logarithmically scaled) and a correlation of 0.85 with the percentage of the national population living below the international poverty line. However, these correlations are not convincing validity evidence: while 0.80 may constitute a strong correlation according to common heuristics, the correlation in question relates to two measures that attempt to measure the same thing — the average person's annual income — rather than measures of two related constructs. Given this, it does not seem unreasonable to expect correlations of 0.9 or above. Also, Avvisati's analysis only constitutes a brief section in his study that serves as a broad survey of methodological issues in ESCS.

Pokropek et al., on the other hand, examine the R-squared value of a regression of HOMEPOS on reading comprehension scores. This, however, is not a valid external criterion because one of the main analyses that PISA conducts is to quantify the relationship between SES and academic performance via the same reading comprehension scores. In other words, the authors are treating reading comprehension as an independent variable against which HOMEPOS can be compared for validation purposes, even though PISA treats it as a variable that is dependent upon HOMEPOS in its own analyses. Rather, HOMEPOS should be compared against criteria that are not dependent variables in the PISA analysis, such as GDP or GNI per capita, even if the latter are not individual-level variables.

Selection criteria for HOMEPOS items

No justification is given for the adequate selection of the household possession indicators in HOMEPOS, nor is there testing of item invariance to demographic characteristics such as nationality, gender, or linguistic group. Although RMSD validation is discussed generally in the documentation for the student context questionnaires in the 2015 cycle, neither RMSD nor any other fit statistic was reported for HOMEPOS for the 2015 cycle (only statistics for math, science, reading, and financial literacy attributes were reported; OECD, 2017, Annex H). Conversely, there are reasons to doubt the appropriateness of certain items. For example, there is evidence that even very poor families in the United States often own multiple televisions (Sheffield and Rector, 2011; Kristof, 2016). Also, PISA's own published parameter estimates suggest that the relative difficulty parameters of some outlier items are so extreme that they are of little use as indicators due to the magnitude of the standard errors (see Bond and Fox, 2015, for a review of Rasch standard errors). For example, the *DVD player* item in the 2012 cycle had a mean difficulty of -2.19 logits across OECD countries, implying that only 10% of students would not endorse the item if person ability were normally distributed (and as few as 2.3% in Ireland). Even many of the country-specific items are not well targeted. For example, in the 2018 cycle, the *Smart TV* item in Denmark had an estimated difficulty of -3.34 logits, implying that only 3.4% of the student population (if upon which income is normally distributed) would not endorse the item. The *Smartphone* item in the Russophone Latvian sample had a difficulty of -5.11, implying that only 0.6% of the student population would not endorse the item. Moreover, when the HOMEPOS GPCM displays poor fit to the response pattern generated by a certain item, PISA re-estimates a unique item parameter for that item independently rather than dropping or substituting it:

> A [RMSD] value of 0.3 was set as a cut-off criterion, with larger values indicating that the international item parameters are not appropriate for this group. When the cut-off criterion was exceeded, the group was flagged and a group-specific (unique) item parameter were [sic] calculated for the group. (OECD, 2019c)

If validating HOMEPOS items in respect to how well they indicate an income attribute, qualitative explanations should be proposed for the cause of misfit, and those hypotheses should subsequently be tested. However, PISA re-estimates these problematic items without giving any reason for why the misfit may have occurred.

There also appears to be large item drift across cycles, even after explicitly adjusting the item set over time to account for technological change. This casts doubt on whether person parameter estimates can be compared across cycles (see Table 4.3) and is a serious concern, as the usefulness of PISA is to assess the success of policies within a given country context over time. For example, the dictionary item in the 2012 cycle should not be interpreted as the same item as the dictionary in the 2015 cycle. The fact that it becomes much "easier" to endorse (from +2.37 to -1.75 logits) suggests a shift in the qualitative nature of this item. In the case of other technological items, such as *a link to the internet*, it is evident that, in many national contexts, what was often perceived as a luxury item in 2000 had transformed into a basic household utility by 2018 (as noted by PISA: see page 63).

One of the assumptions of latent variable models is that the indicators are independent, conditional only on the attribute being measured. Currently, several of PISA's HOMEPOS indicators exhibit obvious violations of independence, for example:

- "A room of your own" vs. "A quiet place to study"
- "Books of poetry" vs. "Books to help with your schoolwork" vs. "Books on art, music, or design" vs. "Books"
- "Computers (desktop computer, portable laptop, or notebook)" vs. "A computer you can use for schoolwork" vs. "Educational software" vs. "A link to the internet"

A cursory glance at the selection of country-specific items reveals similarly severe violations of item independence. I am not aware of any analysis, including Rutkowski Rutkowski (2010, 2013) and Pokropek et al. (2017), that analyzes the appropriateness of the choice of country-specific HOMEPOS items. Another issue with the use of country-specific items is that the interpretability of an unbalanced item response matrix depends upon the assumption that the items that are not administered to a given student are missing at random. Usually such a situation arises because testers wish to administer more items than is realistic for a single individual to answer. In other words, it is necessary for the interpretability of HOMEPOS that any given student *could* have received any given item. However, in HOMEPOS, these extra items are administered in a non-random way, tailoring the additional items to specific country contexts. In doing so, PISA is effectively attempting to measure distinct wealth/income constructs for each country (i.e., "Italian wealth", "Brazilian wealth", "Croatian wealth", etc.). Furthermore, these country–specific instantiations cannot be treated as the same item because parameter estimates vary widely across countries and cycles (see Tables 4.4 and 4.5). This is not surprising, as there do not appear to be specific guidelines for the criteria to which each country's PISA coordinators need to abide when selecting items. If such criteria exist, it is not readily apparent in PISA's documentation.

Table 4.2

Correlation of national estimates of HISEI, HOMEPOS, ESCS with GDP (PPP) per capita

Country	HISEI	HOMEPOS	PARED	ESCS
Australia	0.82	0.50	0.94	0.61
Austria	0.65	-0.46	0.89	-0.46
Belgium	0.89	-0.63	0.95	-0.68
Canada	0.88	-0.00	0.97	0.03
Chile***	0.29	0.20	0.87	0.07
Colombia*	-	-	-	-
Czech Republic	-0.19	-0.49	-0.75	-0.84
Denmark	0.95	0.40	0.90	0.73
Estonia	0.96	-0.61	0.93	-0.73
Finland	0.84	-0.60	0.93	0.29
France	0.71	-0.90	0.98	0.27
Germany	0.91	-0.73	0.65	-0.79
Greece	0.47	-0.33	0.35	0.34
Hungary	0.05	-0.74	0.91	-0.38
Iceland	0.92	-0.75	0.90	-0.10
Ireland	0.83	0.52	0.87	0.05
Israel***	0.90	0.99	0.99	0.95
Italy	0.32	-0.58	0.92	-0.56
Japan	0.54	-0.48	0.89	-0.59
Korea	0.88	-0.38	0.83	0.28
Latvia**	1.00	1.00	1.00	1.00
Lithuania*	-	-	-	-
Luxembourg	0.84	-0.68	0.78	-0.80
Mexico	-0.52	-0.72	0.95	-0.44
Netherlands	0.80	0.27	0.94	0.57
New Zealand	0.85	0.34	0.01	0.02
Norway	0.77	-0.21	-0.47	-0.45
Poland	0.66	-0.08	0.76	0.11
Portugal	0.74	-0.48	0.97	0.68
Slovak Republic	-0.18	-0.16	0.79	-0.59
Slovenia***	0.79	-0.80	0.90	0.04
Spain	0.87	-0.66	0.90	0.25
Sweden	0.95	-0.34	0.99	0.79
Switzerland	0.81	-0.19	0.92	0.34
Turkey	-0.74	-0.78	0.79	-0.51
United Kingdom	0.85	0.06	0.55	0.79
United States	0.63	-0.27	0.87	-0.82

Note: Official estimates of HOMEPOS, PARED, and ESCS are unavailable for the 2000 cycle (pisa-dataexplorer.oecd.org).
* Colombia and Lithuania were only OECD members for the 2018 cycle.
** Latvia was only an OECD member country for the 2015 and 2018 PISA cycles.
*** Chile, Israel, and Slovenia were only OECD members since the PISA 2009 cycle.
GDP per capita data from data.worldbank.org.

Table 4.3

Dichotomous item difficulty parameter estimates across cycles

Item	Cycle		
	2012	2015	2018
A desk to study at	-1.54	-1.00	-0.80
A room of your own	-0.80	-0.82	-0.76
A quiet place to study	-1.15	-1.14	-1.10
A computer you can use for school work	-0.81	-0.34	-0.23
Educational software	1.07	0.34	0.30
A link to the Internet	-0.01	-0.42	-0.59
Works of art (e.g., paintings)	0.74	-	0.10
Books to help with your school work	0.98	-1.23	-1.03
Technical reference books	0.80	0.19	0.18
A dictionary	2.37	-1.75	-1.63
Books on art, music, or design	-	-1.03	0.29
Television (at least one)	1.01	1.17	1.25
Car (at least one)	-0.16	1.31	1.27
Room with a bath or shower (at least one)	-1.56	1.79	-
Cellphone with internet access (at least one)	-	-0.09	-0.36
Computer (at least one)	2.20	0.84	0.83
Tablet computer (at least one)	-	1.30	1.20
E-book reader (at least one)	-	1.55	1.46
Musical instrument (at least one)	-	1.01	0.97
Books (more than 25)	0.85	1.67	1.57

Item parameters are not reported for the 2000, 2003, 2006, and 2009 PISA cycles.

Polytomous item responses were dichotomized as endorsing the lowest step or category level (i.e., $\beta_i + \delta_{i1}$). For the 2015 and 2018 cycles, "books" was dichotomoized as the second lowest category level (i.e., $\beta_i + \delta_{i2}$) in order to maintain comparability with the 2012 cycle where 25 was the lowest threshold, rather than 10.

Table 4.4

Items with positive difficulties over 1SD from the country-specific item mean from the last three PISA cycles can be considered outliers

Country	Item	Cycle	Difficulty
Iceland	Security watch or system	2012	2.869
Iceland	Satellite dish	2012	2.783
Netherlands	Piano	2012	2.747
Israel	4x4 vehicle	2012	2.601
Norway	iPhone	2012	2.514
Portugal	Air conditioning	2012	2.446
Sweden	Piano	2012	2.443
Latvia	Scooter	2018	2.328
Norway	iPad	2012	2.295
Latvia	Scooter	2015	2.291
Slovenia	Traveling abroad for one week or more	2012	2.196
Belgium	Alarm system	2012	2.02

Table 4.5

Items with negative difficulties over 1SD from the country-specific item mean from the last three PISA cycles can be considered outliers

Country	Item	Cycle	Difficulty
Latvia	Your own smartphone	2018	-5.107
Denmark	Smart TV	2018	-3.341
Luxembourg	New game console	2018	-3.001
Portugal	Cable TV or satellite dish	2018	-2.569
Japan	Smartphone	2015	-2.544
Ireland	Your own smartphone	2015	-1.646
Ireland	Your own smartphone	2018	-1.535
Denmark	Flat screen TV	2012	-1.402
Germany	Smartphone	2018	-1.345

Chapter 5

The Real Impact

5.1 Replicating 2018 Findings

The previous two chapters explain why ESCS is not qualified to be considered as a measure of SES. In this chapter, I examine the impact, in pragmatic terms, of comparing ESCS values across cycles when estimation methods differ. Specifically, I attempt to replicate the first three "main findings"[1] from the executive summary from the OECD report, "PISA 2018 Results (Volume II): Where All Students Can Succeed" (OECD, 2019b):

> Finding #1: "In 11 countries and economies, including the OECD countries Australia, Canada, Denmark, Estonia, Finland, Japan, Korea, Norway and the United Kingdom, average performance was higher than the OECD average while the relationship between socio-economic status and reading performance was weaker than the OECD average."
>
> Finding #2: "In spite of socio-economic disadvantage, some students attain high levels of academic proficiency. On average across OECD countries, one in ten disadvantaged students[2] was able to score in the top quarter of reading performance

[1] I do not attempt to replicate any other findings, so there was no "cherry-picking" involved in the analysis. The methodology for this replication analysis can be found in Appendix F.

[2] PISA defines a "socio-economically advantaged (disadvantaged) student" as "a student in the top

> in their countries (known as academic resilience), indicating that disadvantage is not destiny. In Australia, Canada, Estonia, Hong Kong (China), Ireland, Macao (China) and the United Kingdom, all of which score above the OECD average, more than 13% of disadvantaged students were academically resilient."
>
> Finding #3: "Disadvantaged students are more or less likely to attend the same schools as high achievers, depending on the school system. In Argentina, Bulgaria, Colombia, the Czech Republic, Hungary, Israel, Luxembourg, Peru, Romania, the Slovak Republic, the United Arab Emirates and Switzerland, a typical disadvantaged student has less than a one-in-eight chance of attending the same school as high achievers[3] (those who scored in the top quarter of reading performance in PISA). By contrast, in Baku (Azerbaijan), Canada, Denmark, Estonia, Finland, Iceland, Ireland, Kosovo, Macao (China), Norway, Portugal, Spain and Sweden, disadvantaged students have at least a one-in-five chance of having high-achieving schoolmates."

To determine the impact of these structural changes on ESCS estimates within each cycle, I re-estimate ESCS values from the 2018 cycle data using the models employed in the 2012 and 2015 cycles. In particular, the 2012 cycle estimated HOMEPOS with the PCM and aggregated ESCS with a PCA, and the 2015 cycles estimated HOMEPOS with a GPCM and aggregated ESCS with a PCA, while the 2018 cycle estimated HOMEPOS with a GPCM and aggregated ESCS as the arithmetic mean of its components (Table 5.1).

I find that differences between PISA's published findings and alternative findings based on the replicated ESCS values are sufficiently large to add or remove certain highlighted countries. From a pragmatic standpoint, these revisions could plausibly result in different policy decisions in those countries and/or generally compromise policymaker confidence in the stability of PISA's findings. The differences in PISA's published con-

(bottom) quarter of ESCS in his or her own country/economy" (OECD, 2019b, p. 17). This distinction is used to make inferences regarding the general distribution of socioeconomic advantage within and across nations (e.g., OECD, 2019b, pp. 50-51), as well as regarding the impact of ESCS on academic achievement.

[3]PISA defines disadvantaged schools as, "those whose average intake of students falls in the bottom quarter of the PISA index of economic, social and cultural status within the relevant country/economy" and advantaged schools as, "those whose average intake of students falls in the top quarter of that index" (OECD, 2019b, p. 106).

Table 5.1

ESCS structure in 2012, 2015, and 2018

Cycle	HOMEPOS estimation	ESCS aggregation
2012	PCM	PCA
2015	GPCM	PCA
2018	GPCM	Mean

clusions and my subsequent replications can be found in columns 2, 6, and 7 of the tables in Appendices B, C, and D. While I attempt to replicate PISA's 2018 findings using the published 2018 methodology detailed in that cycle's technical report, I am unable to replicate them exactly. Therefore, to strictly compare differences in HOMEPOS estimation and ESCS aggregation procedures without the impact of other methodological discrepancies between my analysis and PISA's, my replicated estimates of 2018 ESCS values should be used as a point of comparison, rather than PISA's published 2018 ESCS estimates.

Broadly, we can see that ESCS values often do not strongly correlate when varying the methodologies. In particular, the results obtained using mean aggregation using the 2018 methodology and PCA aggregation using the 2012 and 2015 methodologies only moderately correlate at best (see Tables 5.2, 5.3, and 5.4). In the case of Finding #2, the correlation is even negative.

Table 5.2

Finding #1: Correlation of HOMEPOS estimation and ESCS aggregation methodologies when applied to 2018 data

Procedure	PCA + PCM (2012)	PCA + GPCM (2015)	Mean + GPCM (2018)
PCA + PCM (2012)	1	-	-
PCA + GPCM (2015)	0.944	1	-
Mean + GPCM (2018)	0.513	0.478	1

Table 5.3

Finding #2: Correlation of HOMEPOS estimation and ESCS aggregation methodologies when applied to 2018 data

Procedure	PCA + PCM (2012)	PCA + GPCM (2015)	Mean + GPCM (2018)
PCA + PCM (2012)	1	-	-
PCA + GPCM (2015)	0.966	1	-
Mean + GPCM (2018)	-0.317	-0.351	1

Table 5.4

Finding #3: Correlation of HOMEPOS estimation and ESCS aggregation methodologies when applied to 2018 data

Procedure	PCA + PCM (2012)	PCA + GPCM (2015)	Mean + GPCM (2018)
PCA + PCM (2012)	1	-	-
PCA + GPCM (2015)	0.992	1	-
Mean + GPCM (2018)	0.386	0.326	1

The changes in specific countries' results are particularly striking:

> Finding #1: "In 11 countries and economies, including the OECD countries Australia, Canada, Denmark, Estonia, Finland, Japan, Korea, Norway and the United Kingdom, average performance was higher than the OECD average while the relationship between socio-economic status and reading performance was weaker than the OECD average."

PISA claims that Australia, Canada, Denmark, Estonia, Finland, Japan, Korea, Norway, and the United Kingdom have average reading performance higher than the OECD average, and a lower-than-average relationship between socioeconomic-status and reading performance. However, we can see when the 2018 calculation is replicated with the 2012 and 2015 methodologies, the percentage of reading performance explained by ESCS falls almost to zero for most of these countries (Table 5.5). Even between the 2012 and 2015 methods, when GPCM was substituted for PCM, but the overall aggregation procedure was unchanged, several countries still see shifts of 2-3%. For example, in the case of Indonesia, performance explained by ESCS falls 3.72%, from 8.01% to 4.29%.

Table 5.5

Percentage of reading performance explained by ESCS under distinct estimation/aggregation methods

Country	PCA + PCM (2012)	PCA + GPCM (2015)	Mean + GPCM (2018)
Australia	0.00	0.04	10.36
Canada	0.01	0.05	6.67
Denmark	0.01	0.10	10.47
Estonia	0.35	0.01	6.90
Finland	0.03	0.01	8.38
Japan	0.04	0.00	7.27
Korea	0.44	0.85	7.87
Norway	0.15	0.01	6.26
United Kingdom	0.19	0.88	7.32

Also, the modeling changes are associated with large displacements in the national rankings list. For example, by substituting the 2015 methodology for the 2018 methodology, the rankings of 27 countries change by at least 30 places (out of 79, see Appendix C for full results). Chile has the largest gain of 65 places, while Macao has the largest loss of 50 places. When substituting the 2012 methodology, there is similar variation in rankings: at the extremes, Bulgaria gains 65 places, and Macao loses 54. Even shifting between the 2012 and 2015 methodologies, when PCA was used as the aggregation schema for both cycles, there are large differences. In particular, Poland gains 38 places and Austria loses 31.

> Finding #2: "In spite of socio-economic disadvantage, some students attain high levels of academic proficiency. On average across OECD countries, one in ten disadvantaged students was able to score in the top quarter of reading performance in their countries (known as academic resilience), indicating that disadvantage is not destiny. In Australia, Canada, Estonia, Hong Kong (China), Ireland, Macao (China) and the United Kingdom, all of which score above the OECD average, more than 13% of disadvantaged students were academically resilient."

63 out of 80 countries saw a ranking change of over 10 places when the 2015 aggregation method was used rather than the 2018 methodology. 62 of the countries saw similar changes when the 2018 and 2012 procedures were compared. Again, most of these ranking changes are extreme. For example, an estimated 5.1% of Peruvian disadvantaged students are academically resilient when aggregating ESCS as the mean of PARED, HISEI, and HOMEPOS (as in the 2018 cycle; PISA's published estimate was 5.2%). However, when aggregating the same data using PCA, 36.9% of disadvantaged students qualify as academically resilient. In both the 2018-2015 comparison and the 2018-2012 comparison, only two countries did not change in their ranking (Czech Republic and Austria, and Luxembourg and Qatar, respectively). Even the difference in HOMEPOS estimation between 2012 and 2015 (where ESCS was calculated via PCA for both cycles) resulted

in large ranking differences. Six countries saw changes of 10 places or more (Table 5.6): Saudi Arabia (26), Estonia (20), Malaysia (16), Switzerland (15), United Kingdom (10), Jordan (10).

Table 5.6

Finding #2: Country rankings and changes

Country	PCA + PCM (2012)	PCA + GPCM (2015)	2015 vs. 2012
Estonia	15	35	-20
Switzerland	3	18	-15
United Kingdom	54	64	-10
Jordan	64	54	10
Malaysia	71	55	16
Saudi Arabia	38	12	26

> Finding #3: "Disadvantaged students are more or less likely to attend the same schools as high achievers, depending on the school system. In Argentina, Bulgaria, Colombia, the Czech Republic, Hungary, Israel, Luxembourg, Peru, Romania, the Slovak Republic, the United Arab Emirates and Switzerland, a typical disadvantaged student has less than a one-in-eight chance of attending the same school as high achievers (those who scored in the top quarter of reading performance in PISA). By contrast, in Baku (Azerbaijan), Canada, Denmark, Estonia, Finland, Iceland, Ireland, Kosovo, Macao (China), Norway, Portugal, Spain and Sweden, disadvantaged students have at least a one-in-five chance of having high-achieving schoolmates."

PISA defines the "chance of attending the same school" as a value on Frankel and Volij's (2011) isolation index:

$$I = 1 - \frac{\sum_{j=1}^{J} \frac{n_j^\alpha}{N^\alpha} \frac{(1-n_j^\alpha)}{n_j}}{1 - p^\alpha} \tag{5.1}$$

...where n_j^α represents the number of disadvantaged students in school j, n_j the total number of students in school j, and $p^\alpha = n^\alpha/N$ is the proportion of the disadvantaged students in the country population. The index ranges from 0 to 1, where "the index increases with the concentration of the students of the group in a limited number of schools" (OECD, 2019b, p. 246).

I find this index approach to be problematic because it does not strictly report the chance of attending the same school as high achievers (rather, it describes the degree of clustering of disadvantaged students in the national school system), and because the isolation statistic is not as readily interpretable as other metrics. Instead, I conduct a similar analysis looking at the percentage of disadvantaged students that attend a school in which the majority of students are in the top reading quartile.

Again, methodology differences result in several meaningful differences: 51 out of 80 countries saw a ranking change of over 10 places when the 2015 aggregation methodology was substituted. 49 of the countries saw similar changes when the 2012 methodology was used. Most of these ranking changes are large. For example, an estimated 0.58% of Colombian disadvantaged students attend a school where the majority of students are in the top reading quartile when using the 2018 methodology (PISA's official findings lend an estimate of 0.46%). However, when substituting PCA aggregation for mean aggregation, 23.3% of students are estimated to attend such top performing schools. Also, in the comparison of the 2018 and 2015 methodologies, only two national entities did not change in their ranking (Ireland and the Moscow subregion of Russia). In the comparison of the 2018 and 2012 methodologies, only three country rankings did not change (Luxembourg, Greece, and the Moscow subregion of Russia).

Even the difference in HOMEPOS estimation between 2012 and 2015 (where ESCS was calculated via PCA for both cycles) resulted in large ranking differences. Ten countries saw changes of 5 places or more (Table 5.7): Malaysia (13), Chile (9), UAE (9), Switzerland (9), Jordan (7), Austria (6), Brunei (6), Saudi Arabia (6), and Uruguay (5).

Table 5.7

Finding #3: Country rankings and changes

Country	PCA + PCM (2012)	PCA + GPCM (2015)	2015 vs. 2012
United Arab Emirates	48	57	-9
Switzerland	44	53	-9
Czech Rep.	60	66	-6
Austria	20	26	-6
Uruguay	40	45	-5
Brunei	69	63	6
Saudi Arabia	29	23	6
Jordan	46	39	7
Chile	55	46	9
Malaysia	64	51	13

5.2 CFA as an Alternate Aggregation

The factor score from a one-factor confirmatory factor analysis (CFA) is a more appropriate aggregation of PARED, HISEI, and HOMEPOS because it conserves the interpretation of education, occupational status, and income as indicators of a latent variable or attribute.[4] The difference between CFA and PCA is that CFA analyzes only common variance among the scores, rather than the total variance (which includes variance specific to individual factors, as well as error variance). The common variance is indicative of the influence of one casual factor on income, education, and occupational status. Upon conducting a CFA, roughly equivalent factor loadings would suggest that each of the indicators explains a similar portion of the variance in the underlying attribute. We can see, however, that the proportionality of factor loadings of PARED, HISEI, and HOME-

[4]Note that both CFA and PCA rely on the assumption that the relationships between PARED, HISEI, and HOMEPOS are linear. This assumption is not supported in the PISA technical documentation.

POS on the first principal component of a PCA and a one-factor CFA are quite different (Table 5.8), with the PCA weighting HOMEPOS disproportionately more.

Table 5.8

ESCS component loadings on the first principal component (PCA) vs one-factor CFA

Component	**One-factor CFA**	**First principal component (PCA)**
PARED	1.000 (-)	-0.447
HISEI	1.006 (0.003)	0.442
HOMEPOS	0.882 (0.003)	0.777

Not surprisingly, calculating ESCS values using a CFA as the first common factor of PARED, HISEI, and HOMEPOS presents substantially different estimates than those derived by PISA. Once again, the magnitude of the difference caused by the choice in aggregation technique is such that it could plausibly result in different policy recommendations, therefore impacting the consequential validity of ESCS.

The differences between the 2018 ESCS values replicated from PISA's estimation procedure (the third column of Appendices B, C, and D) and those replicated using PCM for the estimation of HOMEPOS and CFA for the aggregation of PARED, HISEI, and HOMEPOS (the sixth column of Appendices B, C, and D) are of particular interest, as the later combination conserves the interpretation of ESCS as the quantity of a latent variable. As in the previous chapter's analysis, due to discrepancies between my replicated estimates and PISA's original estimates, only the replications (the second column of Appendices B, C, and D) should be compared.

> Finding #1: "In 11 countries and economies, including the OECD countries Australia, Canada, Denmark, Estonia, Finland, Japan, Korea, Norway and the United Kingdom, average performance was higher than the OECD average while the relationship between socio-economic status and reading performance was weaker than the OECD average."

Recalculating ESCS by estimating HOMEPOS with a PCM instead of a GPCM and aggregating PARED, HISEI, and ESCS as the first common factor of HISEI, PARED, and HOMEPOS when using a CFA rather than as the mean, results in a change to this list of countries, as can be seen in Appendix B. First, Portugal (an OECD country with an average reading performance higher than the OECD average[5]) should be added to the list of countries with a lower-than-average relationship between SES and reading performance because, although the point estimate for the average Portuguese R-squared value of 11.3 does not change, the OECD average rises from 11.2 to 11.4. Also noteworthy is that, while not OECD countries with higher-than-average reading scores, Lebanon, Saudi Arabia, and Turkey drop below the OECD average R-squared cut-off.

Also, had ESCS been aggregated using a PCA in 2018, as had been the case in the 2015 cycle and before, the PISA findings change drastically. The estimate of percentage reading scores explained by ESCS falls from at least 10.9% to at most 1.1% overall, and from at least 11.0% to 0.4% in the subset of OECD participant countries. Also, the modeling change is associated with large displacements in the national rankings list for some countries (see Appendix C). The rankings of several countries shift over 10 places (out of 79), a proportion to which, one could imagine, policymaking might be sensitive:

- Lebanon (LBN) gains 19 places, from 57 to 38.
- Kazakhstan (KAZ) gains 16 places, from 48 to 32.
- Dominican Republic (DOM) gains 15 places, from 45 to 30.
- Morocco (MAR) gains 14 places, from 28 to 14.
- Panama (PAN) gains 10 places, from 72 to 62.

[5]It should be noted that PISA's own analysis does not report Portugal to be "significantly" above the OECD average in reading performance, although its point estimate is above the reading performance point estimate.

- Israel (ISR) falls 13 places, from 51 to 64.
- Qatar (QAT) falls 10 places, from 17 to 27.
- United Arab Emirates (ARE) falls 10 places, from 29 to 39.

> Finding #2: "In spite of socio-economic disadvantage, some students attain high levels of academic proficiency. On average across OECD countries, one in ten disadvantaged students was able to score in the top quarter of reading performance in their countries (known as academic resilience), indicating that disadvantage is not destiny. In Australia, Canada, Estonia, Hong Kong (China), Ireland, Macao (China) and the United Kingdom, all of which score above the OECD average, more than 13% of disadvantaged students were academically resilient."

As can be seen in Appendix D, when ESCS is calculated as the first common factor (using CFA) of HISEI, PARED, and HOMEPOS, and HOMEPOS is estimated with a PCM, Bosnia and Herzegovina (BIH), Finland (FIN), Latvia (LVA), Morocco (MAR), and North Macedonia (MKD) cross the 13% threshold of disadvantaged students estimated to be academically resilient.

We can also see changes in the relative position to the OECD average in several countries when the estimation approach changes from GPCM/mean to PCM/CFA: The United Arab Emirates (ARE) fall from above the OECD average to below, and Hungary (HUN), Malaysia (MYS), and Singapore (SGP) rise from below the OECD average to above. Similarly, we can see large differences in rankings in several countries (at least 7 places out of 79):

- Kazakhstan (KAZ) rises 11 places, from 53 to 42.
- Turkey (TUR) rises 9 places, from 15 to 6.
- Montenegro (MNE) rises 8 places, from 16 to 8.
- Belarus (BLR) rises 7 places, from 71 to 64.

- Dominican Republic (DOM) rises 7 places, from 67 to 60.

- United Arab Emirates (ARE) falls 11 places, from 62 to 73.
- New Zealand (NZL) falls 10 places, from 29 to 39.
- Iceland (ISL) falls 9 places, from 24 to 33.
- Poland (POL) falls 8 places, from 39 to 47.
- Switzerland (CHE) falls 7 places, from 52 to 59.

> Finding #3: "Disadvantaged students are more or less likely to attend the same schools as high achievers, depending on the school system. In Argentina, Bulgaria, Colombia, the Czech Republic, Hungary, Israel, Luxembourg, Peru, Romania, the Slovak Republic, the United Arab Emirates and Switzerland, a typical disadvantaged student has less than a one-in-eight chance of attending the same school as high achievers (those who scored in the top quarter of reading performance in PISA). By contrast, in Baku (Azerbaijan), Canada, Denmark, Estonia, Finland, Iceland, Ireland, Kosovo, Macao (China), Norway, Portugal, Spain and Sweden, disadvantaged students have at least a one-in-five chance of having high-achieving schoolmates."

Once again, the estimation change from using a GPCM for the estimation of HOMEPOS and a PCA for the aggregation of PARED, HISEI, and HOMEPOS into ESCS results in several meaningful differences when examining the percentage of disadvantaged students who attend a school in which the majority of students are in the top reading quartile: The United Arab Emirates (ARE) goes from being over the OECD average with the GPCM/PCA approach to under the OECD average with the PCM/CFA change. On the other hand, Albania (ALB), Hungary (HUN), Malaysia (MYS), and Qatar (QAT) go from being equal to or under the OECD average using the GPCM/PCA, to over the OECD average with the PCM/CFA change.

The model substitution also results in changes to the country rankings list. When analyzing the percentage of disadvantaged students who attend an overall "top reading" school, the change does not affect the membership of the top 13 list, but the internal order is quite different. In particular, the rankings of Slovak Republic, Montenegro, and Switzerland change by four places:

- Using PISA's ESCS estimates: Netherlands (1), Turkey (2), Kosovo (3), Germany (4), Morocco (5), Croatia (6), Slovak Republic (7), Switzerland (8), Bosnia and Herzegovina (9), Italy (10), Montenegro (11), Slovenia (12).
- Using PCA/CFA estimation: Turkey (1), Netherlands (2), Kosovo (3), Morocco (4), Germany (5), Croatia (6), Montenegro (7), Bosnia and Herzegovina (8), Slovenia (9), Italy (10), Slovak Republic (11), Switzerland (12).

The shift in the "bottom 12" list is even more pronounced, where New Zealand changes 5 places (64 to 71) and Australia changes 6 places (68 to 62) — enough to remove them from the lower echelon entirely:

- Using PISA's ESCS estimates: Russia (Moscow region; 79), Finland (78), Sweden (77), Colombia (76), Costa Rica (75), Poland (74), Ireland (73), United States (72), Lithuania (71), Spain (70), Estonia (69), Australia (68), Canada (67).
- Using PCA/CFA estimation: Russia (Moscow region; 79), Finland (78), Poland (77), Ireland (76), Costa Rica (75), Estonia (74), Sweden (73), Colombia (72), New Zealand (71), Canada (70), Spain (69), Lithuania (68), United States (67).

Chapter 6

Reconceptualizing ESCS

In the preceding chapters I have demonstrated how an inconsistent conceptual foundation regarding ESCS as a measure of an SES attribute ultimately begets uncertainty in the PISA's public findings. However, as Kane (1992, p. 38) suggests in his discussion of argumentative validity testing, interpretive validation approaches can often be the catalysts for improvement in measurement procedures. In this spirit, I propose several actionable recommendations for how PISA might go about bolstering the validity of ESCS.

Reinterpret ESCS as a Measure of Income

In Chapters 3 and 4, I explained the problems with encouraging the interpretation of ESCS as a measure of SES. Looking forward, I recommend that PISA reconceptualize ESCS as a measure of *income*, of which PARED, HISEI, and HOMEPOS is each an indicator (and rename the instrument accordingly). While this entails a significant conceptual departure from PISA's current SES framework, I think that this would be a practical decision for several reasons.

First, a family's income is more objective. While there may be debate as to its exact method of quantification, there is a common understanding of what family income is and how it causally relates to student achievement on the one hand, and to education, occupation, and household possessions, on the other. It is also a fundamentally quantitative attribute and, therefore, interval differences in units of income are inherently comparable. There is no need to reference webs of oblique constructs like well-being, equity, status, access, nor high-level sociological theories like social, cultural, and human capital, which were intended to generalize the complex social dynamics of resource allocation, rather than as references to quantifiable attributes.

Second, validation of the ESCS instrument will be easier because its relationship with the attribute of income is more readily testable. For example, external validity could be supported by comparing ESCS against other independent measures of income, such as parental tax returns or neighborhood tax revenue estimates. Also, it is useful that income can be directly manipulable by policy interventions, such as a monetary stimulus or a tax credit. Of course, HOMEPOS is already intended to be a measure of income, though it could be improved by incorporating more household possession indicators and dropping the subcategorization of HOMEPOS items (i.e., WEALTH, CULTPOSS, HEDRES). If PISA intends HOMEPOS to be a measure of income, then it must not be divided into separate sub-constructs whose theoretical connection to a traditional conception of income is unclear. The PARED and HISEI descriptions of parental education and occupation are still useful as indicators of income. As discussed in Chapter 4, HISEI is already a function of a measure of income in its current form. Alternatively, parental education and occupation could be quantified separately, apart from income. Policymakers could still make subsequent holistic inferences that consider income, education, and occupation, but that are based on independent measures of the three attributes.

Third, PISA would gain flexibility by being able to report its degree of confidence in its measures (i.e., a margin of error). Currently, PISA only reports standard errors in some appendices (e.g., OECD, 2019b, pp. 252-253) and only gives point estimates of ESCS in its summary reports. The standard errors it does report, however, do not reflect PISA's confidence in how closely its estimates reflect the real quantity of an attribute, but rather strictly the degree to which its ESCS model fits the collected data. Error estimates are also necessary for interpreting PISA's estimates and national rankings of ESCS and ESCS-derived attributes, like educational equity. Without recourse to attributing variance to measurement error, PISA has limited options for dealing with detected data-to-model misfit: just omitting the data points that most contribute to misfit and/or estimating independent parameter estimates for misfitting items by modeling item responses for that item separately. There is nothing inherently wrong with measurement error — even if one assumes that SES can be treated as an attribute, it is to be expected that there would be a significant degree of measurement error, given its generality. To the contrary, overfitting models are undesirable because they may not fit data from future cycles optimally — likely the reason why PISA's model parameters currently change so much each cycle. After all, the aim of measurement is not to provide the closest possible description of observed data, but rather to quantify differences in amounts of an attribute. PISA's primary responsibility is not to produce parameter estimates of well-fitting models — it is to help present the reality of attributes of international education systems. Ultimately, if the truth is not simple, one should not imply that some measure can describe it in a simple way. Keep in mind that PISA claims to be a "yardstick," not a policy analysis group. As long as policymakers are presented with valid measures, it is their job to interpret the resulting measured values (and perhaps conclude that they are not easily interpretable). PISA should not suggest that the data at hand lends to easy interpretation if it does not.

Incorporate country-specific measures

I recommend that PISA measure a distinct income attribute in each country. PISA's stated goals are to help countries assess their own educational policies and to help identify those implemented by other countries that might be worth adopting in the future. To accomplish these goals, PISA does not need to measure attributes that can be compared between countries. For example, a finding that Country A's mathematics ability increases after implementing a certain policy when controlling for a family SES or income attribute unique to that country would still be relevant to Country B if they see their own socioeconomic context as sufficiently similar to that of Country A. By committing to measuring distinct attributes for each country, PISA would make their validation effort significantly easier and would probably reduce measurement error, as indicators of income (as well as education and occupational status) are often not generalizable internationally. A similar recommendation is proposed by Rutkowski & Rutkowski (2013, p. 275), who suggest that "omitting the international scale altogether would encourage researchers and analysts to develop system-specific indicators that are more relevant to a given population." Moreover, the system of reporting national rankings reinforces a mistaken notion that the purpose of PISA is to foster international competition. Rankings are also easily misrepresented. For example, in a 2016 national press conference regarding the release of national PISA results, Mexican Secretary of Public Education, Aurelio Nuño Mayer, highlighted that Mexico's academic performance and educational equity rankings were higher than those of the average Latin American country even though Mexico's scores were "below the OECD average" (Mexico's scores, in reality, were the lowest of any participating OECD country). The Latin American rankings obscure the fact that, since the 2009 PISA cycle, Mexican reading, math, and science scores had not improved. Ultimately, comparing rankings says little about the state of Mexican education in 2015 and is of limited use for policymakers in that country.

Incorporate valid measurement models

If PISA continues to conceptualize ESCS as a common factor of PARED, HISEI, and HOMEPOS, a CFA approach is the only option that maintains the interpretation of SES as an attribute. While this will result in weaker model fit than PCA, PCA does not distinguish between variance attributable to measurement error and variance attributable to differences in attribute quantities. PISA's most recent approach to aggregating PARED, HISEI, and HOMEPOS – taking the arithmetic mean – seems to be the least desirable of the three approaches. Forcing cross-country equivalence of factor loadings sacrifices both the measurement error interpretation of variance, as well as data-to-model fit.

Similarly, returning to estimating HOMEPOS with the PCM will restore the measurement interpretation of this ESCS component. This carries an advantage over generalized IRT models like the GPCM because Rasch-derived person estimates are item-independent, in that distinct indicators can be applied in different country and cycle contexts because items can be swapped in and out if all items are deemed to measure the same attribute. One implication of this is that more "country-specific" items could be chosen to better target the overall income level in the country. Also, items could be randomly sampled from a larger bank of possible items, to some extent alleviating validity concerns regarding invariance and adequate calibration of item difficulty to student ability parameters in the current item set of only two dozen indicators. The PISA sample is large enough that such sub-sampling is a viable approach.

PISA should consider adopting Rasch measures for parental education and occupation, as well. If they are considered to be causal agents of socioeconomic advantage, they should be held to be attributes and, therefore, able to be represented by latent variables.

Also, the current approach of asking only for years of parental education and name of parental occupation is not ideal, because measurement error cannot be estimated from a single indicator. Adding multiple items about parental education and occupation to the student and parental questionnaires will not necessarily contribute concerns over to the overall length of the questionnaire if select students could be sampled to provide these responses and/or a random sample of items could be used. If PISA decides to measure education, occupation, and income as distinct student attributes (rather than solely as income as I have recommended), it could estimate item parameters concurrently by applying a multidimensional Rasch model. This alternative approach allows for correlation between these attributes.

Adopting these recommendations would require refitting models from past cycles using Rasch model and CFA specifications, even though, as demonstrated in Chapter 5, the difference in ESCS estimates due to modeling choice can be quite large. PISA could also apply these measurement models to constructs other than ESCS with the benefit of no longer needing to use "trend scales" to update parameters from past cycles using current cycle data. This type of linking convention would not necessary because, under the Rasch measurement paradigm, the item sets across cycles can vary without sacrificing interpretability, assuming they refer to the same attribute.

Expand validation procedure

I recommend that PISA embrace validation as a cyclical process. PISA documentation describes validation as a distinct phase of the measurement process that takes place after instrument design and the collection of measurements — a "justification" of the completed measurement procedure. Validation, however, is not a procedure to be applied only after measurement — it is an essential component of measurement itself. Once again, the Standards for Educational and Psychological Testing define validity as an "overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations." This "evaluation" is a continuous set of considerations that overlies the entire measurement process, motivating the identification of the attribute of interest, the construction of the measurement instrument, the way the instrument is applied to observe data, and the evaluation of observations to support pragmatic inferences. As such, validation is an ongoing cycle and is inseparable from instrument design, observation, and inference.

Wilson (2004; Figure 6.1) describes a similar cyclical approach in his theory of "constructing measures." The phases include:

1. Construct map: Defining the construct to be measured and identifying varying levels of the construct that might manifest in respondents and be indicated by items.

2. Items design: Determining the specific tasks by which levels of the contract can be observed.

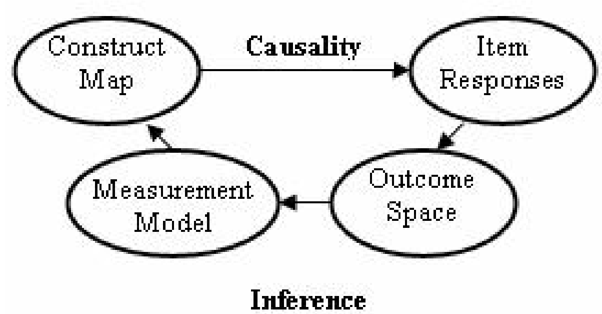
3. Outcome space: Scaling the various observed outcomes to numerical scores.

4. Measurement model: Distilling of the various values in the outcome space to a single value that describes the quantity of the attribute in each person (a justification for using the latent variable measurement model is given in Chapter 3).

This cycle repeats, motivated by the analysis of collected "measures," lending insight as to how the definition of the latent construct can be refined. Ultimately, both construct definition and instrument development are continuously being improved — the cycle never ends and should be revisited at any point when the plausibility of the causal link

Figure 6.1

Wilson's "Construct Modeling" approach to building and validating measures (Wilson, 2004, p. 17)



between the attribute and the observed indicators is threatened (Messick, 1995; AERA, APA & NCME, 2014). The clear documentation of this cycle constitutes a fundamental source of validity evidence (Wilson, 2004, p. 156). Committing to such a simple and practical notion of validity will help PISA define the scope of its validation efforts, as well as boost their transparency and accessibility.

PISA can also improve the reporting of invariance. Avvisati et al. (2019, p. 8) have recognized "a need to better communicate on issues around data comparability in OECD reports," including "more extensive technical documentation on measurement invariance issues" and "items that are crafted in a clearer, more concrete, and less ambiguous manner." PISA should report the contexts in which its measurement models do not fit well or where the assumption of invariance is violated. If a measure appears to not apply to a certain country, linguistic group, etc., this does not mean that the measure is useless; rather, it probably should be interpreted more cautiously.

PISA should also broaden their investigation of invariance to include, not just country-group consistency of factor loadings and Cronbach's alpha statistics, but also the invariance of ESCS values to demographic variables that distinguish minority or underrepresented groups like race, gender, and language group. Explanatory IRT models could be useful for this purpose, as they extend the standard IRT models with a vector of demographic attributes. Explanatory models are not measurement models for the reasons raised in Chapter 4, but the significance of demographic parameters can detect violations of invariance in local contexts. A particularly relevant model for the PISA context is the step difficulty explanatory Linear Partial Credit Model with Multivariate Item-Step Random Error (Kim and Wilson, 2020). This modeling approach considers both person parameters and item category/step parameters as random effects. Considering that the joint estimation of random effects is often computationally taxing for a maximum likelihood approach because an integration is required for each person/item-step pair (De Boeck and Wilson, 2004, p. 195), MCMC estimation methods could be employed. The “blme” (Chung et al., 2013), “brms” (Bürkner, 2017) or “rstan” packages (Stan Development Team, 2023) in R are attractive tools for this purpose.

PISA can also aid validation efforts by identifying external criteria against which they can compare their measures, for example, GDP or GNI per capita data. At the very least, correlations with external criteria could be used as evidence that ESCS is a more valid measure of SES than alternatives. Similarly, PISA could compare measures generated by both the questionnaire responses of students and responses given by parents to the same questions. Rutkowski & Rutkowski (2010) examine the correlation of student responses in the 2006 PIRLS assessment to a *number of books in the home* item that is analogous to the item in HOMEPOS, but with corresponding parental responses, and find only weak-to-moderate correlation between the student-parent response pairs in all 46 participating countries (from r = 0.17 in Indonesia to r = 0.68 in Bulgaria). Currently, such an analysis is not possible with PISA data because parents do not receive the HOMEPOS items in their questionnaire. While the parental questionnaire does include a polytomous response question regarding pre-tax family income, administration of the questionnaire is optional on a country-by-country basis, and it is not clear if this question has been used for validation of the student questionnaire.[1]

Finally, validation of PISA's measures could be improved by boosting the transparency and replicability of PISA's methodology. As it stands, reproducing PISA results is difficult because, while the response data is publicly available and general procedures are given in the PISA technical report, the exact calculations preformed to obtain ESCS are not published. This is especially problematic because the size of the dataset and the number of independent parameters to be estimated require more computational power than is available to many independent researchers. Also, the mdltm software (von Davier, 2005) used by PISA is not widely accessible. Ultimately, PISA is a publicly funded project, which publishes results to the public based on public data. Why should its calculations not be public, as well? PISA could publish these calculations as an appendix to its technical report, as well as use the R environment for its analysis, for which many Rasch/IRT software packages are freely available. This would make it much easier for independent researchers to assist PISA with future validation efforts.

While PISA might be reluctant to change the fundamental structure and theoretical foundation of ESCS because it would require the qualification or correction of many of its findings since the assessment's conception in 2000, it is in PISA's best interest to act sooner rather than later. It is a small price to pay for PISA's ability to defend itself against some of the theoretical criticisms detailed in this paper and others regarding the comparability of data collected across testing cycles, for the breathing room afforded by acknowledging measurement error, and for the conservation of long-term public trust in the assessment. Not only does PISA's public visibility, international government buy-in, and established infrastructure make it worth improving, the cost of improving the validity of an instrument like ESCS is not prohibitive, especially considering the potential price in the erosion of public trust for neglecting to do so. PISA has already prompted real

[1] Parental questionnaire data is not currently available on the online PISA database.

action by many national governments to improve school systems, but this power to foster accountability is only as strong as PISA's own reputation. This reputation, in turn, is dependent upon the trustworthiness of the measures employed in the assessment.

Bibliography

Adams, E. W. (1966). *On the nature and purpose of measurement* (tech. rep. No. 4). Center for Advanced Study in the Behavioral Sciences. Berkeley, CA.

Alaska Department of Labor and Workforce Development. (n.d.). *Seafood and fishing jobs in Alaska*. Retrieved February 9, 2023, from https://jobs.alaska.gov/seafood/

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report.* (tech. rep. No. 4). ERIC.

American Educational Research Association and American Psychological Association and National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. AERA.

Andrews, P., et al. (2014, May 13). Academics call for pause in PISA tests. *The Washington Post*.

Auld, E., & Morris, P. (2016). PISA, policy and persuasion: Translating complex conditions into education 'best practice'. *Comparative Education*, *52*(2), 202–229. https://doi.org/10.1080/03050068.2016.1143278

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

Avvisati, F., Le Donné, N., & Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences*, *1*(1), 8. https://doi.org/10.1186/s42409-019-0010-z

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., Zhang, T., Buehler, E., Lee, S.-j., Sikali, E., Bundsgaard, J., C'deBaca, R., & Christensen, A. A. (2022). *Edsurvey: Analysis of nces education survey and assessment data* [https://CRAN.R-project.org/package=EdSurvey].

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Eds.), *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. ETS.

Bempechat, J., Jimenez, N. V., & Boulay, B. A. (2002). Cultural-cognitive issues in academic achievement: New directions for cross-national research. In *Methodological advances in cross-national surveys of educational achievement* (pp. 117–149). National Academy Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological bulletin*, *110*(2), 305.

Bond, T., & Fox, C. (2015). *Applying the Rasch Model* (3rd). Routledge.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061.

Bourdieu, P. (1983). Okonomisches Kapital, kulturelles Kapital, soziales Kapital. in Soziale Ungleichheiten. *Soziale Welt Gottingen*, 183–198.

Boyd, M. (2008). A Socioeconomic Scale for Canada: Measuring Occupational Status from the Census. *Canadian Review of Sociology/Revue canadienne de sociologie*, *45*(1), 51–91. https://doi.org/10.1111/j.1755-618X.2008.00003.x

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance* (tech. rep. No. 171). OECD.

Bridgman, P. W. (1922). *Dimensional analysis*. Yale University Press.

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch Model. *Measurement*, *146*, 961–971. https://doi.org/10.1016/j.measurement.2019.07.035

Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. *NCES Conference on the Program for International Student Assessment: What We Can Learn from PISA, Washington, DC.*

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Coleman, J. S. (1988). Social capital in the creation of human capital. *American journal of sociology*, *94*, S95–S120.

Cowan, Hauser, Levin, Spencer, B., & Chapman. (2012). *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation* (tech. rep.). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test Validation. In R. Thorndike (Ed.), *Educational Measurement* (Second, p. 443). American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281–302.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

de Leeuw, J. (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, *53*(4), 437–454. https://doi.org/10.1007/BF02294399

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 471–503. https://doi.org/10.1007/BF02296971

de Ayala, R. (2009). *The theory and practice of item response theory*. Guilford Publications. https://books.google.com/books?id=NchXAQAAQBAJ

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). Socioeconomic background and achievement.

Duncan, O. D. (1961). A socioeconomic index for all occupations, and properties and characteristics of the socioeconomic index. *Occupations and Social Status, A. Reiss, Ed.(Free Press of Glencoe, New York, 1961)*, 109–161.

Edwards, J. R., & Bagozzi, R. P. (2000). On the Nature and Direction of Relationships Between Constructs and Measures. *Psychological Methods*, *5*(2), 155–174.

Engel, L. C., & Rutkowski, D. (2020). Pay to play: What does PISA participation cost in the US? *Discourse: Studies in the Cultural Politics of Education*, *41*(3), 484–496. https://doi.org/10.1080/01596306.2018.1503591

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational Class Mobility in Three Western European Societies: England, France and Sweden. *The British Journal of Sociology*, *30*(4), 415–441. https://doi.org/10.2307/589632

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., & Tucker, W. S. (1940). Quantitative estimates of sensory events. *Advancement of Science*, *2*, 331–349.

Frankel, D. M., & Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory*, *146*(1), 1–38. https://doi.org/10.1016/j.jet.2010.10.008

Ganzeboom, H., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Ganzeboom, H., & 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: A european working book for demographic and socio-economic variables* (pp. 159–193). Springer US. https://doi.org/10.1007/978-1-4419-9186-7_9

Ganzeboom, H., & Treiman, D. J. (2010). Occupational status measures for the new International Standard Classification of Occupations ISCO-08; with a discussion of the new classification. *Annual Conference of International Social Survey Programme*.

Gaylord, R. (n.d.). *Conceptual consistency and criterion equivalence: A dual approach to criterion analysis*. Unpublished Manuscript.

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

Goldstein, D. (2019, December 3). 'It Just Isn't Working': PISA Test Scores Cast Doubt on U.S. Education Efforts. *The New York Times*.

Gottfried, A. W. (1985). Measures of socioeconomic status in child development research: Data and recommendations. *Merrill-Palmer Quarterly*, *311*(1), 85–92.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x

Hauser, R. M. (1994). Measuring Socioeconomic Status in Studies of Child Development. *Child Development*, *65*(6), 1541–1545. https://doi.org/10.2307/1131279

Hauser, R. M., & Warren, J. R. (1997). Socioeconomic Indexes for Occupations: A Review, Update, and Critique. *Sociological Methodology*, *27*(1), 177–298. https://doi.org/10.1111/1467-9531.271028

Heim, J. (2016, December 12). What's behind Finland's Pisa slide, Parenting & Education News & Top Stories - The Straits Times. *Washington Post*.

Hodge, R. W. (1981). The measurement of occupational status. *Social Science Research*, *10*(4), 396–415.

HP. (2018, March 17). Luxembourg reduces PISA participation to every six years. *Luxembourg Times*. https://www.luxtimes.lu/en/luxembourg/education-luxembourg-reduces-pisa-participation-to-every-six-years-602e35d2de135b92361e5122.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424.

International Labour Office. (1988). *ISCO-88: International standard classification of occupations* (tech. rep.).

International Labour Office. (2008). *ISCO-08: International standard classification of occupations* (tech. rep.).

IPUMS USA. (n.d.). User note on composite measure of occupational standing. Retrieved February 2, 2023, from https://usa.ipums.org/usa/chapter4/sei_note.shtml

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Kaplan, D. (1989). Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic. *Multivariate Behavioral Research*, *24*(3), 285–305. https://doi.org/10.1207/s15327906mbr2403_2

Kim, J., & Wilson, M. (2020). Polytomous item explanatory IRT models with random item effects: Concepts and an application. *Measurement*, *151*, 107062. https://doi.org/10.1016/j.measurement.2019.107062

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. Academic Press.

Kreiner, S., & Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, *79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

Kristof, N. (2016, October 28). Opinion — 3 TVs and No Food: Growing Up Poor in America. *The New York Times*.

Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis*, *52*(2), 161–193.

Lazarsfeld, P. F. (1959). Latent Structure Analysis. In S. Koch (Ed.), *Psychology: A study of a science*. McGraw-Hill.

Lee, 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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Educational Testing Service.

Lumsden, J. (1976). Test theory. *Annual review of psychology*, *27*(1), 251–280.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 technical report* (tech. rep.). Chestnut Hill, MA, International Study Center.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowsi, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258). CRC Press.

McGrane, J. A. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, *6*, 431.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of educational statistics*, *7*(2), 105–118.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*(4), 285–294. https://doi.org/10.1016/j.measurement.2005.09.004

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Mueller, C. W., & Parcel, T. L. (1981). Measures of Socioeconomic Status: Alternatives and Recommendations. *Child Development*, *52*(1), 13–30. https://doi.org/10.2307/1129211

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–76.

Nuño Mayer, A. (2016). *Mensaje del secretario de Educación Pública, Aurelio Nuño Mayer, durante el informe de resultados México en PISA 2015*. Retrieved February 9, 2023, from http://www.gob.mx/sep/prensa/mensaje-del-secretario-de-educacion-publica-aurelio-nuno-mayer-durante-el-informe-de-resultados-mexico-en-pisa-2015-86084?idiom=es

O'Connell, M. (2019). Is the impact of SES on educational performance overestimated? Evidence from the PISA survey. *Intelligence*, *75*, 41–47. https://doi.org/10.1016/j.intell.2019.04.005

OECD. (n.d.-a). *Data*. Retrieved September 30, 2022, from https://www.oecd.org/pisa/data/

OECD. (n.d.-b). *Education GPS*. Retrieved September 30, 2022, from http://gpseducation.oecd.org

OECD. (n.d.-c). *Germany's PISA Shock - OECD*. Retrieved February 9, 2023, from https://www.oecd.org/about/impact/germany-pisa-shock.htm

OECD. (2012a). *Strong Performers and Successful Reformers In education: Lessons from PISA for Japan*. Retrieved February 9, 2023, from https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/49802616.pdf

OECD. (2012b). *PISA 2009 Technical Report* (tech. rep.). https://doi.org/10.1787/9789264167872-en

OECD. (2013). *PISA as a Yardstick for Educational Success*. https://doi.org/10.1787/9789264207585-3-en

OECD. (2014). *PISA 2012 Technical Report* (tech. rep.). Retrieved February 9, 2023, from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. https://doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 Technical Report* (tech. rep.). Retrieved February 9, 2023, from https://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2018a). *PISA 2015: Results in Focus*. Retrieved February 9, 2023, from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

OECD. (2018b). *Equity in Education: Breaking Down Barriers to Social Mobility*. https://doi.org/10.1787/9789264073234-en

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. Retrieved February 9, 2023, from https://www.oecd-ilibrary.org/education/pisa-2018-assessment-and-analytical-framework_b25efab8-en

OECD. (2019b). *PISA 2018 Results (Volume II): Where All Students Can Succeed*. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2019c). *PISA 2018 Technical Report* (tech. rep.). Retrieved February 9, 2023, from https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2019d). *PISA 2018: Insights and Interpretations* (tech. rep.). Retrieved February 9, 2023, from https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Oliveri, M., & Von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling*, *53*, 315–333.

Perry, R., & Visher, M. (2019). *Major Mines of Nevada 2018 Mineral Industries in Nevada's Economy* (tech. rep. Special Publication P-30). The Nevada Division of Minerals. Univerisity of Reno.

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

Porter, E. (2015, November 3). America's Students Are Lagging. Maybe It's Not the Schools. *New York Times*.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rhodan, M. (2013, December 5). *What Does it Mean That American Students are Barely Average?* Retrieved February 9, 2023, from https://nation.time.com/2013/12/05/what-does-it-mean-that-american-students-are-barely-average/

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

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430. https://doi.org/10.1080/00220272.2010.487546

Rutkowski, L., & Rutkowski, D. (2016). A Call for a More Measured Approach to Reporting and Interpreting PISA Results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Sahlberg, P., & Hargreaves, A. (2015, March 24). The tower of PISA is badly leaning. An argument for why it should be saved. *Washington Post*.

Schuller, T. (2001). The complementary roles of human and social capital. *Canadian Journal of Policy Research*, *2*(1), 18–24.

Sheffield, R., & Rector, R. (2011). Air Conditioning, Cable TV, and an Xbox: What is Poverty in the United States Today? *The Heritage Foundation*. Retrieved February 2, 2023, from https://www.heritage.org/poverty-and-inequality/report/air-conditioning-cable-tv-and-xbox-what-poverty-the-united-states

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

Spiezia, V. (2011). Does computer use increase educational achievements? Student-level evidence from PISA. *OECD Journal: Economic Studies*, *2010*(1), 1–22.

Stan Development Team. (2023). RStan: The R interface to Stan [R package version 2.21.8]. https://mc-stan.org/

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173–180.

Stenner, J. A., Stone, M. H., & Burdick, D. S. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions*, *21*(1), 1152–3.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.

The Economist. (2010, December 9). No longer bottom of the class. *The Economist*.

The Economist. (2019, December 5). PISA results can lead policymakers astray. *The Economist*.

Torres Irribarra, D. (2021). A Pragmatic Perspective of Measurement. In *A Pragmatic Perspective of Measurement* (pp. 43–62). Springer International Publishing. https://doi.org/10.1007/978-3-030-74025-2_4

Traynor, A., & Raykov, T. (2013). Household Possessions Indices as Wealth Measures: A Validity Evaluation. *Comparative Education Review*, *57*(4), 662–688. https://doi.org/10.1086/671423

von Davier, M. (2005). Mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models. Retrieved February 9, 2023, from http://www.von-davier.com/

von Davier, M. (2007). Hierarchical General Diagnostic Models. *ETS Research Report Series*, *2007*(1), i–19. https://doi.org/10.1002/j.2333-8504.2007.tb02061.x

Von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models (L. Rutkowsi, M. Von Davier, & D. Rutkowski, Eds.). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 155–174.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481. https://doi.org/10.1037/0033-2909.91.3.461

Whittaker, T. A. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *The Journal of Experimental Education*, *80*(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Wickham, H., Miller, E., & Smith, D. (2022). *Haven: Import and export 'spss', 'stata' and 'sas' files* [https://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat].

Willms, D. J., & Tramonte, L. (2015). *Towards the development of contextual questionnaires for the PISA for development study* (OECD Education Working Papers No. 118). https://doi.org/10.1787/5js1kv8crsjf-en

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Wright, B., & Linacre, J. M. (1987). Dichotomous Rasch model derived from specific objectivity. *Rasch Measurement Transactions*, *1*(1), 5–6.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Mesa Press.

Wuttke, J. (2007). *Uncertainty and Bias in PISA* (SSRN Scholarly Paper No. ID 1159042). Social Science Research Network. Rochester, NY.

Zhao, Y. (2020, December 3). How PISA created an illusion of education quality and marketed it to the world. *Washington Post*.

Appendix A: ISO Country Codes

Country	ISO Code	Country	ISO Code
ALB	Albania	LBN	Lebanon
ARE	United Arab Emirates	LTU	Lithuania
ARG	Argentina	LUX	Luxembourg
AUS	Australia	LVA	Latvia
AUT	Austria	MAC	Macao
BEL	Belgium	MAR	Morocco
BGR	Bulgaria	MDA	Moldova
BIH	Bosnia and Herzegovina	MEX	Mexico
BLR	Belarus	MKD	North Macedonia
BRA	Brazil	MLT	Malta
BRN	Brunei Darussalam	MNE	Montenegro
CAN	Canada	MYS	Malaysia
CHE	Switzerland	NLD	Netherlands
CHL	Chile	NOR	Norway
COL	Colombia	NZL	New Zealand
CRI	Costa Rica	PAN	Panama
CZE	Czech Republic	PER	Peru
DEU	Germany	PHL	Philippines
DNK	Denmark	POL	Poland
DOM	Dominican Republic	PRT	Portugal
ESP	Spain	QAT	Qatar
EST	Estonia	QAZ	Baku (Azerbaijan)
FIN	Finland	QCI	B-S-J-Z (China)*
FRA	France	QMR	Moscow Region (RUS)
GBR	United Kingdom	QRT	Tatarstan (RUS)
GEO	Georgia	ROU	Romania
GRC	Greece	RUS	Russian Federation
HKG	Hong Kong	SAU	Saudi Arabia
HRV	Croatia	SGP	Singapore
HUN	Hungary	SRB	Serbia
IDN	Indonesia	SVK	Slovak Republic
IRL	Ireland	SVN	Slovenia
ISL	Iceland	SWE	Sweden
ISR	Israel	TAP	Chinese Taipei
ITA	Italy	THA	Thailand
JOR	Jordan	TUR	Turkey
JPN	Japan	UKR	Ukraine
KAZ	Kazakhstan	URY	Uruguay
KOR	South Korea	USA	United States of America
KSV	Kosovo	VNM	Vietnam

Note: Chinese participation in the 2018 cycle was limited to the provinces of Beijing, Shanghai, Jiangsu, and Zhejiang.

Appendix B: Replicated ESCS Values

Country	PISA	Mean & GPCM	Mean & PCM	CFA & GPCM	CFA & PCM	PCA & GPCM	PCA & PCM
ALB	-0.832	-0.828	-0.792	-0.700	-0.685	0.862	0.793
ARE	0.254	0.307	0.240	0.330	0.283	0.162	0.315
ARG	-0.716	-0.601	-0.593	-0.503	-0.504	0.671	0.660
AUS	0.357	0.303	0.256	0.182	0.153	-0.600	-0.506
AUT	0.067	0.136	0.139	0.072	0.073	-0.282	-0.291
BEL	0.139	0.212	0.176	0.196	0.170	-0.044	0.037
BGR	-0.166	-0.125	-0.131	-0.075	-0.083	0.354	0.372
BIH	-0.542	-0.431	-0.433	-0.394	-0.400	0.315	0.322
BLR	-0.088	-0.133	-0.110	-0.020	-0.012	0.687	0.649
BRA	-1.072	-1.066	-1.045	-0.917	-0.911	1.007	0.970
BRN	-0.249	-0.076	-0.048	-0.041	-0.027	0.271	0.216
CAN	0.408	0.462	0.451	0.445	0.436	-0.078	-0.049
CHE	0.033	0.124	0.100	0.105	0.087	-0.042	0.011
CHL	-0.262	-0.190	-0.173	-0.147	-0.141	0.325	0.294
COL	-1.040	-1.156	-1.102	-1.040	-1.012	0.846	0.731
CRI	-0.966	-0.895	-0.909	-0.783	-0.799	0.783	0.821
CZE	-0.061	-0.013	-0.021	-0.057	-0.064	-0.154	-0.140
DEU	-0.009	0.174	0.187	0.134	0.141	-0.157	-0.187
DNK	0.423	0.699	0.603	0.706	0.642	0.011	0.227
DOM	-1.043	-1.136	-1.080	-0.894	-0.869	1.509	1.405
ESP	-0.025	-0.060	-0.026	-0.115	-0.094	-0.206	-0.285
EST	0.111	0.088	0.075	0.084	0.073	0.047	0.076
FIN	0.325	0.296	0.276	0.305	0.289	0.084	0.132
FRA	-0.052	-0.163	-0.172	-0.167	-0.175	0.076	0.095
GBR	0.302	0.272	0.262	0.215	0.208	-0.253	-0.235
GEO	-0.347	-0.486	-0.419	-0.388	-0.351	0.654	0.516
GRC	-0.068	-0.049	0.014	-0.012	0.024	0.274	0.142
HKG	-0.492	-0.586	-0.591	-0.557	-0.564	0.313	0.323
HRV	-0.230	-0.214	-0.227	-0.171	-0.184	0.320	0.353
HUN	-0.035	-0.074	-0.054	-0.072	-0.061	0.100	0.057
IDN	-1.370	-1.486	-1.355	-1.287	-1.214	1.322	1.048
IRL	0.146	0.088	0.132	0.051	0.077	-0.134	-0.231
ISL	0.570	0.767	0.754	0.763	0.752	-0.045	-0.009
ISR	0.338	0.225	0.214	0.211	0.202	-0.017	0.010
ITA	-0.203	-0.092	-0.049	-0.143	-0.117	-0.189	-0.287
JOR	-0.710	-0.756	-0.775	-0.559	-0.581	1.218	1.276
JPN	-0.082	-0.238	-0.136	-0.148	-0.090	0.581	0.369
KAZ	-0.313	-0.514	-0.489	-0.360	-0.352	0.965	0.924
KOR	0.101	-0.036	0.030	0.025	0.062	0.402	0.267
KSV	-0.439	-0.280	-0.265	-0.117	-0.116	0.969	0.952
LBN	-0.496	-0.472	-0.438	-0.392	-0.376	0.577	0.508
LTU	0.071	-0.003	-0.011	0.016	0.007	0.176	0.196
LUX	0.058	0.203	0.207	0.130	0.132	-0.341	-0.355
LVA	0.016	-0.095	-0.105	-0.085	-0.094	0.148	0.171
MAC	-0.508	-0.488	-0.519	-0.500	-0.522	0.083	0.147
MAR	-1.914	-1.688	-1.649	-1.526	-1.511	1.148	1.069

Continued on next page

MDA	-0.531	-0.598	-0.603	-0.479	-0.489	0.782	0.800
MEX	-1.060	-1.107	-1.073	-0.975	-0.961	0.919	0.852
MKD	-0.283	-0.186	-0.208	-0.110	-0.130	0.496	0.553
MLT	0.107	0.235	0.216	0.133	0.122	-0.490	-0.456
MNE	-0.162	-0.225	-0.198	-0.173	-0.160	0.376	0.322
MYS	-0.748	-0.650	-0.544	-0.498	-0.439	0.959	0.742
NLD	0.304	0.312	0.253	0.244	0.206	-0.326	-0.200
NOR	0.564	0.527	0.479	0.428	0.398	-0.509	-0.409
NZL	0.217	0.187	0.155	0.087	0.067	-0.473	-0.410
PAN	-0.948	-0.938	-0.873	-0.761	-0.728	1.138	1.009
PER	-1.095	-1.217	-1.125	-1.015	-0.967	1.297	1.110
PHL	-1.414	-1.388	-1.298	-1.159	-1.113	1.458	1.276
POL	-0.128	-0.158	-0.156	-0.210	-0.210	-0.178	-0.188
PRT	-0.340	-0.238	-0.226	-0.294	-0.287	-0.184	-0.218
QAT	0.304	0.405	0.338	0.452	0.405	0.286	0.441
QAZ	-0.534	-0.725	-0.665	-0.603	-0.573	0.821	0.700
QCI	-0.332	-0.329	-0.256	-0.310	-0.268	0.224	0.067
QMR	0.354	0.209	0.238	0.271	0.284	0.378	0.326
QRT	0.180	0.011	0.038	0.105	0.116	0.576	0.529
ROU	-0.433	-0.450	-0.429	-0.431	-0.422	0.224	0.179
RUS	0.174	0.008	0.032	0.095	0.104	0.538	0.497
SAU	-0.635	-0.548	-0.638	-0.523	-0.583	0.288	0.483
SGP	0.185	-0.045	-0.058	-0.006	-0.019	0.299	0.333
SRB	-0.221	-0.258	-0.229	-0.200	-0.187	0.413	0.356
SVK	-0.114	0.045	0.033	0.071	0.060	0.201	0.233
SVN	0.003	-0.037	-0.075	-0.086	-0.112	-0.183	-0.103
SWE	0.408	0.389	0.340	0.343	0.310	-0.212	-0.104
TAP	-0.304	-0.364	-0.349	-0.330	-0.325	0.302	0.272
THA	-1.021	-1.026	-0.973	-0.978	-0.949	0.457	0.340
TUR	-1.165	-1.249	-1.177	-1.211	-1.169	0.440	0.277
UKR	-0.151	-0.197	-0.181	-0.070	-0.067	0.779	0.757
URY	-0.919	-0.842	-0.835	-0.803	-0.803	0.390	0.374
USA	0.097	0.159	0.129	0.111	0.090	-0.197	-0.133
VNM	-1.717	-1.779	-1.724	-1.667	-1.639	0.905	0.784

Appendix C: Reading Explained by ESCS

Estimates of percentage of national reading performance explained by ESCS

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	8.6	8.2	8.6	7.5	7.8	1.6	1.9
ARE	11.3	8.7	9.2	10.6	10.8	0.2	0.1
ARG	17.7	17.0	17.4	15.4	15.7	0.8	1.0
AUS*	10.2	9.6	10.4	9.9	10.4	0.0	0.0
AUT	11.7	11.0	12.2	10.8	11.5	0.0	0.3
BEL*	16.1	14.5	15.0	14.3	14.6	0.0	0.0
BGR	17.4	16.8	16.9	16.6	16.7	0.1	0.0
BIH	7.5	6.6	6.8	6.4	6.6	0.1	0.0
BLR	18.6	18.3	18.2	17.6	17.7	0.0	0.2
BRA	14.0	13.9	13.5	12.2	12.1	1.4	1.0
BRN	16.8	16.7	15.0	14.8	13.7	4.0	2.0
CAN*	6.6	6.1	6.7	6.5	6.8	0.0	0.1
CHE	14.7	13.2	14.2	12.7	13.4	0.1	0.0
CHL	17.6	16.9	17.1	17.0	17.1	0.0	0.0
COL	16.2	16.5	16.2	14.9	14.7	4.5	3.5
CRI	13.3	13.9	13.5	12.3	12.1	2.4	1.9
CZE*	19.9	19.3	20.6	19.3	20.2	0.1	0.1
DEU*	16.7	16.3	17.1	15.1	15.6	0.1	0.5
DNK*	10.9	10.4	10.5	9.9	10.0	0.0	0.1
DOM	10.9	12.0	11.4	10.0	9.7	4.4	3.6
ESP	8.7	8.1	8.6	7.7	8.0	0.1	0.0
EST*	6.8	6.0	6.9	6.6	7.2	0.4	0.0
FIN*	8.5	7.9	8.4	7.9	8.2	0.0	0.0
FRA*	20.3	20.7	21.5	19.7	20.2	1.8	2.8
GBR*	6.7	6.1	7.3	5.9	6.7	0.2	0.9
GEO	11.4	11.2	11.3	10.8	11.0	0.4	0.7
GRC	10.1	9.6	10.0	9.7	9.9	0.1	0.0
HKG*	3.6	3.4	3.6	3.0	3.2	0.3	0.6
HRV	8.7	8.1	8.2	8.0	8.1	0.1	0.0
HUN	18.6	18.4	18.5	17.6	17.8	0.1	0.4
IDN	14.1	15.3	13.6	13.3	12.3	8.0	4.3
IRL*	9.9	9.1	10.5	9.1	9.9	0.0	0.3
ISL	6.2	5.9	6.5	6.5	6.9	0.5	0.2
ISR	14.9	13.3	13.4	14.9	14.9	0.1	0.1
ITA	7.7	7.9	8.4	7.0	7.3	0.1	0.3
JOR	8.7	7.9	7.3	7.6	7.3	0.9	0.5
JPN*	8.0	6.8	7.3	7.5	7.8	0.0	0.0
KAZ	9.2	12.6	11.7	10.2	9.8	4.1	3.3
KOR*	7.4	7.6	7.9	7.3	7.5	0.4	0.9
KSV	4.7	4.5	4.5	3.7	3.7	0.7	0.8
LBN	13.6	14.3	13.2	11.2	10.6	6.5	4.8
LTU	11.8	11.2	11.6	10.8	11.2	0.0	0.0
LUX	18.5	17.0	17.6	16.6	17.0	0.1	0.0

Continued on next page

LVA	5.8	4.8	5.2	5.1	5.3	0.4	0.1
MAC*	1.6	1.8	2.1	1.5	1.6	0.6	1.0
MAR	9.7	8.4	7.9	7.2	7.0	2.1	1.3
MDA	16.4	16.0	16.8	14.5	15.1	1.1	2.1
MEX	13.6	13.8	13.3	13.2	12.9	1.8	1.1
MKD	9.8	9.3	9.4	8.3	8.5	0.4	0.6
MLT	7.7	7.0	7.2	7.2	7.4	0.6	0.5
MNE	6.0	5.1	5.6	5.1	5.5	0.0	0.1
MYS	15.8	15.5	14.7	14.1	13.6	2.5	1.0
NLD	10.0	9.8	10.7	9.0	9.7	0.2	0.9
NOR*	6.7	5.6	6.3	6.4	6.8	0.1	0.0
NZL*	12.6	12.6	13.2	12.6	12.9	0.2	0.4
PAN	17.5	17.5	16.8	15.2	14.8	5.6	3.5
PER	21.8	21.7	20.6	19.9	19.3	6.9	4.8
PHL	18.3	17.9	18.2	15.9	16.1	7.2	6.8
POL*	12.2	11.7	12.9	12.0	12.8	0.4	0.0
PRT*	12.2	11.3	11.8	11.1	11.3	0.8	0.5
QAT	9.1	7.5	7.6	8.4	8.4	0.0	0.0
QAZ	4.5	4.4	4.6	4.4	4.5	0.0	0.1
QCI*	14.0	14.0	14.8	13.9	14.4	0.0	0.0
QMR	3.3	2.8	2.9	3.5	3.6	0.3	0.1
QRT	5.3	4.6	4.7	5.2	5.2	0.3	0.1
ROU	15.9	16.0	16.4	14.9	15.3	1.2	1.9
RUS	8.0	7.5	7.3	7.5	7.5	0.1	0.1
SAU	13.2	13.1	12.0	11.7	11.1	0.4	0.1
SGP*	14.4	14.1	14.2	13.2	13.4	1.2	1.6
SRB	9.3	8.4	8.6	8.2	8.4	0.3	0.1
SVK	15.4	14.9	16.0	14.2	14.9	0.0	0.5
SVN	11.1	10.4	10.7	10.0	10.2	0.0	0.0
SWE*	9.4	8.1	8.9	8.0	8.5	0.1	0.4
TAP*	11.8	11.5	12.0	11.4	11.7	0.3	0.8
THA	25.6	26.4	26.5	24.7	24.9	5.0	4.6
TUR	12.0	12.1	12.4	10.8	11.0	0.7	1.0
UKR	12.6	11.9	12.4	11.3	11.7	0.1	0.4
URY	16.5	15.7	16.1	14.9	15.2	0.6	1.0
USA*	11.4	11.2	11.7	10.6	11.0	0.8	1.1
VNM	-	-	-	-	-	-	-
Overall average	11.8	11.4	11.7	10.9	11.1	1.1	1.0
OECD average	11.8	11.2	11.8	11.0	11.4	0.4	0.4

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

$*$ Countries marked with an asterisk have a mean 2018 reading score higher than the OECD average.

Estimates of national reading performance explained by ESCS: National ranking

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	21	26	25	20	22	64	66
ARE	38	29	29	38	39	36	31
ARG	71	71	71	69	69	56	60
AUS*	34	32	32	34	37	1	18
AUT	41	37	45	41	46	4	35
BEL*	62	58	60	60	60	12	2
BGR	68	68	68	71	71	30	2
BIH	15	14	12	11	9	22	8
BLR	74	74	73	74	74	16	33
BRA	54	53	53	50	50	63	60
BRN	67	67	61	62	58	70	68
CAN*	10	13	11	12	11	4	21
CHE	57	50	56	53	56	32	15
CHL	70	69	69	73	73	4	8
COL	63	66	64	66	61	73	72
CRI	50	54	54	51	50	68	67
CZE*	76	76	76	76	77	22	26
DEU*	66	65	70	67	68	22	42
DNK*	35	36	33	33	35	4	27
DOM	36	45	38	35	30	72	74
ESP	24	23	26	23	23	18	2
EST*	13	11	13	14	15	44	8
FIN*	20	22	24	24	25	12	8
FRA*	77	77	78	77	78	65	70
GBR*	12	12	18	9	10	34	53
GEO	40	40	37	42	40	48	49
GRC	33	33	31	32	34	28	8
HKG*	3	3	3	2	2	40	47
HRV	23	24	22	26	24	25	15
HUN	75	75	75	75	75	31	41
IDN	55	60	55	56	51	79	75
IRL*	31	30	34	31	33	8	36
ISL	9	10	10	13	13	50	34
ISR	58	51	52	63	64	25	23
ITA	17	21	24	15	17	28	37
JOR	22	20	16	22	16	59	46
JPN*	18	15	15	19	21	16	2
KAZ	26	48	40	37	32	71	71
KOR*	14	19	21	18	20	49	52
KSV	5	5	4	4	4	54	50
LBN	52	57	49	45	38	76	78
LTU	43	39	39	43	44	8	18
LUX	73	70	72	72	72	25	15
LVA	7	7	7	6	7	44	25
MAC*	1	1	1	1	1	52	55
MAR	29	28	20	17	14	67	63
MDA	64	63	67	61	65	60	69
MEX	51	52	51	54	53	66	61
MKD	30	31	30	28	29	44	48
MLT	16	16	14	16	18	52	45
MNE	8	8	8	7	8	8	31

Continued on next page

MYS	60	61	58	58	57	69	56
NLD	32	34	36	30	31	34	54
NOR*	11	9	9	10	12	32	8
NZL*	47	47	50	52	54	37	38
PAN	69	72	66	68	62	75	73
PER	78	78	77	78	76	77	78
PHL	72	73	74	70	70	78	79
POL*	46	43	48	49	52	47	8
PRT*	45	41	42	44	45	57	44
QAT	25	17	19	29	27	8	15
QAZ	4	4	5	5	5	12	24
QCI*	53	55	59	57	59	12	15
QMR	2	2	2	3	3	40	31
QRT	6	6	6	8	6	40	29
ROU	61	64	65	66	67	62	66
RUS	18	18	18	21	19	22	28
SAU	49	49	43	48	43	46	21
SGP*	56	56	57	55	56	61	64
SRB	27	28	27	27	26	40	21
SVK	59	59	62	59	64	16	43
SVN	37	36	35	36	36	12	8
SWE*	28	25	28	26	28	18	39
TAP*	42	42	44	47	47	42	50
THA	79	79	79	79	79	74	76
TUR	44	46	46	40	42	55	58
UKR	48	44	47	46	48	27	40
URY	65	62	63	64	66	53	58
USA*	39	38	41	39	41	58	62
VNM	-	-	-	-	-	-	-

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

$*$ Countries marked with an asterisk have a mean 2018 reading score higher than the OECD average.

Appendix D: Disadvantaged Top Readers

Estimates of percentage disadvantaged students in the top reading quartile

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	11.5	11.5	11.6	11.8	12.2	28.6	29.5
ARE	6.6	8.2	7.8	7.7	7.5	18.3	19.6
ARG	6.1	6.0	5.9	6.5	6.7	25.3	26.0
AUS	12.1	11.7	11.6	12.1	12.3	21.9	23.4
AUT	9.6	9.9	9.1	9.8	9.6	22.8	24.1
BEL	9.2	9.3	9.0	10.0	9.8	21.4	22.1
BGR	6.0	6.1	6.1	6.2	6.2	20.5	20.5
BIH	12.4	13.2	13.3	12.9	13.1	23.3	24.1
BLR	8.3	7.6	8.2	8.5	8.8	24.6	24.9
BRA	8.7	8.8	9.3	9.6	9.8	28.6	27.6
BRN	8.1	7.8	8.8	8.6	8.8	32.3	28.8
CAN	13.4	13.7	13.4	13.5	13.2	22.5	22.9
CHE	9.1	9.2	9.1	9.3	9.1	18.8	22.2
CHL	7.4	7.6	7.5	8.1	8.2	24.3	23.9
COL	8.0	7.6	8.0	8.5	8.7	31.7	31.0
CRI	10.4	10.2	10.4	11.0	11.3	30.1	29.8
CZE	6.7	7.0	6.7	7.1	6.5	19.7	22.0
DEU	8.9	8.5	8.8	9.4	9.1	20.0	21.2
DNK	10.8	11.4	11.2	11.6	11.4	22.0	23.6
DOM	9.0	7.9	8.7	8.7	9.0	35.5	33.5
ESP	13.8	13.7	13.4	14.2	13.9	22.1	22.8
EST	13.9	13.7	13.1	14.3	14.1	21.0	23.8
FIN	12.9	12.9	13.1	13.4	13.3	22.6	23.2
FRA	8.1	7.2	6.8	8.0	8.0	26.4	28.0
GBR	13.8	13.8	13.2	14.7	14.1	25.2	27.5
GEO	10.6	10.4	10.5	11.2	11.2	25.6	26.5
GRC	12.0	12.6	12.3	12.8	12.9	20.6	21.0
HKG	15.9	16.1	16.1	17.2	16.7	26.3	27.3
HRV	14.7	14.2	14.0	16.2	15.9	23.0	23.4
HUN	7.5	6.8	6.6	8.2	7.8	23.6	24.1
IDN	8.4	8.4	9.0	9.1	9.2	40.2	35.3
IRL	12.4	11.9	11.9	12.8	12.4	23.3	24.4
ISL	11.5	12.0	11.8	11.4	11.4	19.2	19.6
ISR	7.4	7.9	7.7	7.7	7.8	21.4	22.6
ITA	12.4	12.8	12.6	13.0	12.8	23.4	23.9
JOR	10.6	11.9	11.9	11.8	12.0	27.8	26.2
JPN	10.6	11.0	10.9	11.1	11.5	20.3	20.6
KAZ	11.0	9.0	9.7	10.2	10.5	32.4	32.0
KOR	12.8	12.3	12.4	13.1	12.8	24.8	24.9
KSV	15.6	15.5	15.6	16.4	16.5	26.6	27.2
LBN	7.8	8.1	8.6	9.0	9.3	35.0	33.0
LTU	11.5	11.5	11.4	12.0	12.0	22.6	23.7
LUX	6.7	6.9	6.7	7.6	7.4	20.0	20.4
LVA	12.9	13.5	12.9	13.7	13.6	20.6	21.7
MAC	18.4	17.4	17.0	18.2	17.5	25.7	26.9
MAR	10.8	12.0	12.7	12.8	13.2	30.2	27.3
MDA	7.6	8.4	8.1	8.7	8.6	25.7	27.2
MEX	9.8	10.0	10.3	10.1	10.0	29.2	28.8
MKD	12.4	12.9	12.3	13.7	13.0	23.5	24.1

Continued on next page

MLT	14.7	14.3	13.9	14.2	13.9	18.9	18.5
MNE	13.2	13.1	13.2	14.7	14.3	23.4	24.0
MYS	8.8	8.5	9.5	9.3	9.6	30.2	26.4
NLD	11.7	11.9	10.6	12.4	11.8	25.1	26.7
NOR	12.2	13.2	12.6	12.9	12.9	22.5	23.5
NZL	11.3	11.7	11.2	10.7	10.8	22.1	23.2
PAN	7.6	7.9	7.5	8.7	8.9	33.1	29.5
PER	5.2	5.1	6.2	5.3	6.0	36.9	35.0
PHL	7.2	8.3	7.8	8.4	8.2	40.0	39.6
POL	9.9	10.6	9.6	10.2	9.8	21.6	23.0
PRT	9.4	10.5	10.3	10.3	10.0	17.5	17.9
QAT	8.6	9.3	9.5	9.3	9.4	22.7	23.5
QAZ	14.0	14.4	14.6	14.3	15.1	24.4	24.4
QCI	9.8	9.9	9.4	9.9	9.7	21.0	21.6
QMR	13.5	14.0	14.6	14.9	14.6	19.8	21.0
QRT	12.7	12.9	13.0	12.6	12.7	22.0	22.8
ROU	8.1	7.9	8.1	8.2	8.1	26.6	27.3
RUS	10.7	11.0	11.1	11.2	11.3	23.8	24.1
SAU	9.4	9.4	9.8	10.0	10.2	23.3	21.2
SGP	8.7	8.6	8.6	9.3	9.2	27.3	28.3
SRB	11.7	11.9	11.8	12.4	12.0	19.5	19.7
SVK	9.7	9.7	8.7	9.8	9.4	22.5	23.8
SVN	10.5	10.7	10.9	11.6	11.3	21.5	21.6
SWE	9.9	10.0	9.7	10.8	10.6	24.1	24.8
TAP	10.4	10.8	10.5	10.5	10.7	24.9	24.7
THA	6.1	5.5	5.6	6.4	6.4	34.2	32.9
TUR	13.3	13.1	13.1	14.6	14.6	23.5	23.7
UKR	11.3	11.2	11.5	12.1	11.6	21.7	23.4
URY	8.3	8.5	8.2	8.4	8.8	24.4	24.8
USA	10.2	9.8	9.9	10.3	10.3	28.1	28.4
VNM	-	-	-	-	-	-	-
Overall average	10.43	10.53	10.48	10.98	10.92	24.90	25.16
OECD average	10.63	10.75	10.49	11.11	10.95	22.70	23.62

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

Estimates of percentage disadvantaged students in the top reading quartile: National ranking

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	29	32	29	31	26	16	11
ARE	75	62	68	73	73	78	77
ARG	77	77	78	76	75	27	29
AUS	23	30	28	28	25	58	54
AUT	48	46	53	51	52	47	41
BEL	51	51	55	48	49	63	63
BGR	78	76	77	78	78	68	73
BIH	21	14	10	20	17	45	40
BLR	61	71	63	65	64	32	31
BRA	57	54	51	52	48	15	17
BRN	64	68	57	63	63	9	13
CAN	11	11	9	15	16	53	58
CHE	52	52	52	57	59	77	62
CHL	71	70	71	70	68	35	44
COL	65	69	66	64	65	10	8
CRI	41	42	40	38	37	13	9
CZE	74	73	74	75	76	73	64
DEU	54	58	56	53	58	71	69
DNK	34	33	33	33	34	57	49
DOM	53	67	59	62	60	4	4
ESP	9	10	8	12	12	55	60
EST	7	9	15	10	10	65	46
FIN	15	19	14	16	14	50	56
FRA	63	72	72	71	70	22	16
GBR	8	8	12	7	9	28	18
GEO	38	41	39	36	38	26	26
GRC	24	21	23	23	20	67	71
HKG	2	2	2	2	2	23	21
HRV	5	6	6	4	4	46	53
HUN	69	75	75	69	72	38	39
IDN	59	60	54	58	57	1	2
IRL	20	28	25	22	24	44	36
ISL	28	24	27	34	33	75	76
ISR	70	66	69	72	71	62	61
ITA	19	20	20	18	22	42	43
JOR	37	27	24	30	29	18	28
JPN	36	36	36	37	32	69	72
KAZ	32	53	46	45	42	8	7
KOR	16	22	21	17	21	31	30
KSV	3	3	3	3	3	21	23
LBN	66	63	61	59	55	5	5
LTU	27	31	31	29	28	49	48
LUX	73	74	73	74	74	70	74
LVA	14	12	17	14	13	66	65
MAC	1	1	1	1	1	25	24
MAR	33	23	18	21	15	12	20
MDA	68	59	65	61	66	24	22
MEX	46	44	42	46	46	14	12
MKD	18	18	22	13	18	40	38
MLT	4	5	7	11	11	76	78

Continued on next page

MNE	13	16	11	6	8	41	42
MYS	55	57	49	56	51	11	27
NLD	26	26	37	26	30	29	25
NOR	22	13	19	19	19	52	51
NZL	31	29	32	40	39	54	55
PAN	67	65	70	60	61	7	10
PER	79	79	76	79	79	3	3
PHL	72	61	67	67	67	2	1
POL	44	39	47	44	47	60	57
PRT	50	40	41	43	45	79	79
QAT	58	50	48	55	54	48	50
QAZ	6	4	5	9	5	34	35
QCI	45	45	50	49	50	64	67
QMR	10	7	4	5	7	72	70
QRT	17	17	16	24	23	56	59
ROU	62	64	64	68	69	20	19
RUS	35	35	34	35	36	37	37
SAU	49	49	44	47	44	43	68
SGP	56	55	60	54	56	19	15
SRB	25	25	26	25	27	74	75
SVK	47	48	58	50	53	51	45
SVN	39	38	35	32	35	61	66
SWE	43	43	45	39	41	36	33
TAP	40	37	38	41	40	30	34
THA	76	78	79	77	77	6	6
TUR	12	15	13	8	6	39	47
UKR	30	34	30	27	31	59	52
URY	60	56	62	66	62	33	32
USA	42	47	43	42	43	17	14
VNM	-	-	-	-	-	-	-

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

Appendix E: Reading Environment of Disadvantaged Students

Estimates of percentage of disadvantaged students who attend a top reading quartile school

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	2.3	2.0	2.0	2.2	2.5	12.0	12.2
ARE	1.7	2.8	2.4	2.6	2.4	13.4	15.4
ARG	1.0	0.9	1.0	1.1	1.1	12.6	13.1
AUS	0.8	1.0	1.0	0.9	1.0	4.6	4.9
AUT	3.6	3.8	3.6	3.5	3.3	8.0	8.9
BEL	4.4	4.3	4.2	4.8	4.6	13.7	13.4
BGR	2.9	2.7	2.7	2.9	2.6	14.1	14.1
BIH	5.5	5.9	6.0	5.5	5.7	12.2	12.6
BLR	1.0	1.3	1.5	1.2	1.3	11.7	12.1
BRA	2.0	1.7	1.7	2.1	2.2	17.2	16.6
BRN	3.0	2.9	3.3	3.2	3.6	19.1	16.8
CAN	0.9	0.8	0.9	0.8	0.8	3.4	3.2
CHE	5.5	5.4	5.3	5.7	5.3	12.8	14.2
CHL	0.9	0.8	0.7	1.1	1.0	14.3	13.2
COL	0.5	0.6	0.5	0.6	0.7	23.2	21.8
CRI	0.5	0.4	0.5	0.6	0.6	17.9	17.2
CZE	4.1	4.3	3.9	4.1	3.5	16.0	17.4
DEU	6.5	6.2	6.5	6.5	6.7	12.1	12.8
DNK	1.3	1.3	1.3	1.7	1.7	3.7	3.5
DOM	1.4	1.2	1.2	1.3	1.4	26.2	23.8
ESP	0.8	0.8	0.8	0.8	0.8	3.0	2.8
EST	0.8	0.6	0.5	0.7	0.7	5.0	5.5
FIN	0.3	0.3	0.3	0.3	0.3	1.8	1.7
FRA	2.8	2.5	2.0	2.5	2.5	13.8	13.8
GBR	1.9	1.9	1.9	2.0	1.9	8.6	9.1
GEO	2.8	2.9	3.1	2.9	3.0	11.2	11.6
GRC	0.9	0.9	0.9	0.8	0.9	5.2	5.2
HKG	3.9	4.1	3.9	4.4	4.2	10.9	11.4
HRV	5.7	5.7	5.5	6.8	6.4	14.6	15.3
HUN	2.7	2.1	1.8	2.9	2.6	14.1	14.5
IDN	3.2	3.3	3.5	3.7	3.8	40.4	35.2
IRL	0.6	0.6	0.5	0.6	0.6	2.2	2.4
ISL	0.9	0.9	0.9	0.9	0.9	1.3	1.4
ISR	1.8	1.8	2.0	1.6	1.8	8.2	8.2
ITA	5.4	5.1	5.2	5.5	5.5	12.5	12.4
JOR	2.0	2.0	2.1	2.1	2.0	13.1	12.6
JPN	3.4	4.0	3.5	3.5	3.6	13.4	13.4
KAZ	4.2	3.6	4.1	4.2	4.4	26.8	25.8
KOR	1.3	1.1	1.2	1.4	1.4	8.9	8.9
KSV	8.0	7.2	7.2	8.4	8.3	16.3	16.4
LBN	3.5	3.3	3.5	3.9	3.9	31.8	28.8
LTU	0.7	0.7	0.7	0.8	0.8	8.4	8.9
LUX	1.1	1.1	1.0	1.1	1.1	9.3	8.7
LVA	1.1	1.3	1.3	1.3	1.2	6.0	6.3
MAC	1.4	1.2	1.3	1.6	1.5	13.9	13.5

Continued on next page

MAR	5.8	6.4	6.5	7.3	7.5	25.2	23.0
MDA	2.6	2.9	2.7	3.2	2.9	12.4	13.1
MEX	1.3	1.4	1.5	1.4	1.6	24.2	22.9
MKD	2.4	2.6	2.5	2.5	2.6	8.0	8.5
MLT	0.9	0.9	0.9	0.9	0.9	2.9	2.6
MNE	5.2	5.0	5.4	6.0	5.9	15.4	16.1
MYS	2.5	2.1	2.8	2.5	2.9	17.4	14.0
NLD	9.7	9.8	8.7	10.4	9.9	18.8	19.8
NOR	1.4	1.2	1.2	1.8	1.7	3.5	3.2
NZL	0.9	0.8	0.8	0.8	0.7	2.0	1.9
PAN	1.1	1.1	1.5	1.2	1.3	25.0	21.6
PER	0.9	0.9	1.1	0.9	1.2	22.6	21.0
PHL	1.0	1.1	0.9	0.9	0.9	28.8	27.5
POL	0.5	0.7	0.5	0.5	0.5	5.0	4.9
PRT	1.2	1.0	1.3	1.2	1.2	2.8	2.8
QAT	1.6	1.9	1.8	2.5	2.5	18.9	19.1
QAZ	1.2	1.1	1.1	1.3	1.4	9.4	9.3
QCI	4.0	3.9	3.9	4.2	4.2	14.2	14.2
QMR	0.0	0.0	0.0	0.0	0.0	0.0	0.0
QRT	1.9	1.7	1.6	1.9	1.9	11.7	12.0
ROU	4.2	3.8	4.1	4.4	4.2	18.1	18.2
RUS	1.2	1.2	1.1	1.3	1.4	9.7	9.8
SAU	2.4	2.3	2.4	2.4	2.4	9.6	8.8
SGP	2.3	2.3	2.3	2.7	2.7	17.0	17.2
SRB	4.0	3.9	3.7	4.4	4.3	11.3	11.7
SVK	5.5	5.4	5.1	5.6	5.4	14.6	15.6
SVN	5.0	5.5	4.9	5.5	5.5	15.3	15.8
SWE	0.4	0.5	0.6	0.6	0.7	2.4	2.2
TAP	3.1	3.4	3.0	3.1	2.9	11.0	11.7
THA	1.6	1.2	1.2	1.7	1.8	32.5	31.3
TUR	9.4	9.1	8.8	10.7	10.6	20.6	21.2
UKR	1.6	1.5	1.5	1.8	1.7	8.2	8.6
URY	1.4	1.5	1.8	1.8	2.1	12.3	13.2
USA	0.6	0.7	0.9	0.7	0.8	5.5	5.8
VNM	-	-	-	-	-	-	-
Overall average	2.5	2.5	2.5	2.7	2.7	13.0	12.8
OECD average	2.5	2.4	2.3	2.6	2.5	9.4	9.5

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

Estimates of percentage of disadvantaged students who attend a top reading quartile school: National ranking

Country	Mean (PISA)	Mean		CFA		PCA	
	GPCM (PISA)	GPCM	PCM	GPCM	PCM	GPCM	PCM
ALB	35	37	37	37	34	44	44
ARE	41	28	32	31	37	34	25
ARG	59	61	59	60	60	38	39
AUS	68	60	60	64	62	67	67
AUT	20	19	20	22	24	60	54
BEL	13	14	13	13	13	33	35
BGR	26	29	28	27	31	30	30
BIH	9	6	6	10	8	42	41
BLR	58	46	47	57	54	46	45
BRA	37	41	42	39	38	19	20
BRN	25	25	24	24	21	13	19
CAN	67	69	67	70	70	70	70
CHE	8	10	9	8	12	37	29
CHL	65	67	71	58	61	27	36
COL	76	74	74	74	72	10	9
CRI	75	77	77	76	75	17	18
CZE	16	13	16	19	23	22	16
DEU	4	5	4	6	5	43	40
DNK	49	48	48	46	47	68	68
DOM	46	53	55	54	52	6	6
ESP	70	66	69	66	69	71	72
EST	69	73	75	71	74	65	64
FIN	78	78	78	78	78	77	77
FRA	28	31	35	33	33	32	32
GBR	38	38	38	40	41	56	53
GEO	27	26	25	28	25	48	49
GRC	64	65	66	68	66	64	65
HKG	19	15	17	14	18	50	50
HRV	6	7	7	5	6	26	26
HUN	29	35	41	29	32	29	27
IDN	23	23	22	21	20	1	1
IRL	73	75	76	73	76	75	74
ISL	63	63	64	63	65	78	78
ISR	40	40	36	48	43	58	61
ITA	10	11	10	12	10	39	43
JOR	36	36	34	38	40	36	42
JPN	22	16	23	23	22	35	34
KAZ	14	21	14	17	14	5	5
KOR	50	54	54	49	50	55	56
KSV	3	3	3	3	3	21	21
LBN	21	24	21	20	19	3	3
LTU	71	72	70	67	68	57	55
LUX	56	58	61	59	59	54	58
LVA	57	47	50	53	57	62	62
MAC	47	51	49	47	49	31	33
MAR	5	4	5	4	4	7	7
MDA	30	27	29	25	26	40	38
MEX	51	45	46	50	48	9	8
MKD	32	30	30	32	30	61	60
MLT	62	62	63	62	64	72	73

Continued on next page

MNE	11	12	8	7	7	23	22
MYS	31	34	27	35	27	18	31
NLD	1	1	2	2	2	15	13
NOR	48	50	53	44	46	69	69
NZL	66	68	68	69	71	76	76
PAN	55	56	45	55	55	8	10
PER	61	64	56	65	56	11	12
PHL	60	57	62	61	63	4	4
POL	74	71	73	77	77	66	66
PRT	54	59	51	56	58	73	71
QAT	42	39	39	34	35	14	14
QAZ	53	55	58	51	53	53	52
QCI	18	17	18	18	17	28	28
QMR	79	79	79	79	79	79	79
QRT	39	42	43	41	42	45	46
ROU	15	20	15	16	16	16	15
RUS	52	52	57	52	51	51	51
SAU	33	32	31	36	36	52	57
SGP	34	33	33	30	29	20	17
SRB	17	18	19	15	15	47	48
SVK	7	9	11	9	11	25	24
SVN	12	8	12	11	9	24	23
SWE	77	76	72	75	73	74	75
TAP	24	22	26	26	28	49	47
THA	44	49	52	45	44	2	2
TUR	2	2	1	1	1	12	11
UKR	43	44	44	43	45	59	59
URY	45	43	40	42	39	41	37
USA	72	70	65	72	67	63	63
VNM	-	-	-	-	-	-	-

Note: Estimates are differentiated by aggregation type (Mean, Confirmatory Factor Analysis [CFA], Principal Component Analysis [PCA]) and HOMEPOS model (Generalized Partial Credit Model [GPCM] and Partial Credit Model [PCM]). In 2018, PISA calculated ESCS as the arithmetic mean of HOMEPOS, PARED, and HISEI. HOMEPOS was estimated with a GPCM.

Appendix F: Replication Methodology

This appendix details the methodology for replicating the three findings from the executive summary of Volume II of the 2018 PISA report, as discussed in Chapter 5.

First, 2018 PISA data was downloaded from the official PISA website (OECD, n.d.-a). As analyses for this project were conducted in R, and the official data is only available in SPSS and SAS formats, the "EdSurvey" (Bailey et al., 2022) and "haven" (Wickham et al., 2022) R packages were used to download data into an R data frame and convert SPSS data types into standard R numeric and string data types. If one has access to SPSS, data can also be accessed in R by first downloading in SPSS, exporting data in .csv format, and then importing into the R environment.

The next step was to reduce PISA's dataset into a usable data format for analysis. Student response patterns with less than three valid responses for the home possession questions were removed listwise. Five types of variables were singled out: 1) HOMEPOS item responses, 2) student demographic variables, 3) person- and country-senate weights, 4) student reading achievement plausible values, and 5) PISA's estimates of ESCS and its constituent constructs, PARED, HISEI, HOMEPOS. Responses coded as Not Applicable, Invalid, No Response, or Missing were recoded as NA. Certain HOMEPOS item response variables were divided into country-by-language groups, as specified in the PISA technical report. These include the three "national indicators of home possessions" (ST011D17TA, ST011D18TA, and ST011D19TA), as well as three items that are "considered to have country-specific meaning" (OECD, 2019c; ST011Q07TA - *Classic literature*, ST011Q08TA - *Books of poetry*, and ST012Q03TA - *Rooms with a bath or shower*).

I next attempted to replicate PISA's concurrent WLE estimation of HOMEPOS person parameters for the full 2018 sample using a Generalized Partial Credit Model

(GPCM) and PISA's published person weights. I was unable to conduct this estimation, even when using 1.4 TB of RAM (estimating unique country-by-group parameters for the six country-specific items implies 723 distinct item parameters and 612,004 person parameters to be estimated). I, therefore, used maximum a posteriori estimates instead.

After estimating HOMEPOS person parameters, I attempted to replicate the ESCS values reported by PISA. Then, the replicated HOMEPOS values and PISA's reported HISEI and PARED values were standardized using "senate-weighted" means and standard deviations, and then averaged to obtain the final replicated ESCS values. The purpose of senate weights is to force all countries to contribute equally to the standardization, regardless of the number of data points contributed by each. The replicated HOMEPOS and ESCS values differed from PISA's reported HOMEPOS and ESCS values to the extent that the statistics obtained from the replication effort often differed significantly, as can be seen in the tables in Appendices B, C, D, and E, where the replications using PISA's reported ESCS values are designated as "Mean (PISA)" and "GPCM (PISA)." As PISA does not publish their exact calculation procedure, it is difficult to ascertain to what exactly the difference can be attributed. I was also unable to replicate PISA's reported ESCS values when using WLE estimation with bootstrapped subsamples of the full data set. One possible explanation for the discrepancy is that PISA may have excluded some data to compensate for oversampling of certain national subpopulations.

Subsequently, new HOMEPOS and ESCS values were estimated using the alternate methodologies described in Chapter 5 (i.e., substituting the PCM model for the GPCM and a CFA of HOMEPOS, HISEI, and PARED rather than a simple averaging or PCA). Finally, the resulting ESCS values generated using the PCM and CFA models were compared with the reading achievement plausible values reported by PISA to generate findings analogous to those of the PISA executive summary.