Sempé *Large-scale Assess Educ* (2021) 9:9
https://doi.org/10.1186/s40536-021-00103-7

Large-scale Assessments in Education

RESEARCH **Open Access**

School-level inequality measurement based categorical data: a novel approach applied to PISA

Lucas Sempé*

*Correspondence: l.sempe@uea.ac.uk
School of International Development, University of East Anglia, Norwich, UK

Abstract

This paper introduces a new method to measure school-level inequality based on Item Response Theory (IRT) models. Categorical data collected by large-scale assessments poses diverse methodological challenges hinder measuring inequality due to data truncation and asymmetric intervals between categories. I use family possessions data from PISA 2015 to exemplify the process of computing the measurement and develop a set of country-level mixed-effects linear regression models comparing the predictive performance of the novel inequality measure with school-level Gini coefficients. I find school-level inequality is negatively associated with learning outcomes across many non-European countries.

Keywords: PISA, Item Response Theory, Inequality, Ordinal data, School inequality, HOMEPOS

Introduction

Although the relevance of socioeconomic factors as predictors of children's cognitive learning attainment is a highly disputed issue in terms of causality (Mayer, 1997), there is extensive and long-standing research recognising their important role in explaining educational disparities in terms of access and outcomes (Coleman, 1966; Del Bello et al., 2015). Furthermore, research from a range of disciplines has highlighted a negative association between socioeconomic disparity and individual outcomes, offering various explanations for a detrimental role of inequality on domains such as health and subjective well-being (Deaton, 2003; Schneider, 2016; Wilkinson & Pickett, 2006).

Socioeconomic variables play also an important role in Large-Scale Assessments to explain or control for differences among groups in terms of learning outcomes and other variables of interest (Hopfenbeck et al., 2018). However, the possible interplay between school-level inequality and educational outcomes has been less addressed. Although previous research has developed alternatives to address the measurement of inequality based dichotomous or ordinal data, there has not been to my knowledge an alternative that computes inequality in the same statistical framework used in Large-Scale Assessments by using Item Response Theory (IRT). In this paper, I develop a novel method to measure school-level assets inequality utilising IRT models based on the discrimination parameter α. The proposed inequality measure computes the dispersion of the data at a certain aggregated level—such as schools or countries. The measure allows both to rank observations in terms of inequality, and to compare the average of inequality across the schools. I exemplify this case computing inequality based on PISA in 2015 home possessions index (HOMEPOS).

The remainder of the paper is structured as follows. "Socioeconomic measurement in PISA" section discusses the role and limitations of socioeconomic variables in PISA and International Large-Scale Assessments (ILSAs), while "The complexity of measuring inequality based on categorical data" section reviews the relevant literature regarding the measurement of inequality using categorical data, discussing the main methods previously developed in recent literature. "Alpha inequality: inequality based on an Item Response Theory paradigm" section briefly introduces IRT and summarises the methodological construction process of the inequality measure, named as Alpha Inequality. "Methods" section introduces the criteria used to analyse Alpha Inequality and the data used in the empirical section. "Results and discussion" section presents the findings of the construction process of Alpha Inequality and a comparative analysis of results with a Gini coefficient in terms of descriptive and inferential parameters, while "Conclusion" section concludes the study.

Socioeconomic measurement in PISA

The relevance of socioeconomic background questions in PISA as well as in ILSAs is twofold. First, socioeconomic variables are constantly used as control regressors as well as in the analysis of equality of opportunities of educational systems. For instance, PISA reports differences among scores within quintiles of wealth and report gaps explained by less privileged socioeconomic backgrounds (OECD, 2016). Second, due to the nature of PISA and other ILSAs, where there is limited time to cover diverse aspects of knowledge, students are exposed only to a portion of cognitive tests. Subsequently, socio-economic information is used as auxiliary information to impute final learning scores, through a technique called plausible values, which are "drawn from a posteriori (data) distribution by combining the IRT scaling of the test items with a latent regression model using information from the student context questionnaire in a population model" (OECD, 2017, p. 128).

Extensive research has been done analysing background questionnaires in PISA, showing diverse limitations on socioeconomic indicators. For instance, there is evidence of cross-country comparability deficiencies within and between PISA cycles (Lee & Von Davier, 2020; Sandoval-Hernandez et al., 2019) and poor model fit (Rutkowski & Rutkowski, 2013). One of the main consequences is the distortion of achievement estimates—see, for example, Rutkowski (2011, 2014) and also Rutkowski and Zhou (2015). Additionally, prior research also reports deficiencies regarding the cultural validity of some questions. For instance, there is a particular bias towards describing better contexts of developed countries, such as the number of questions that reflect a certain type of cultural possession (Rutkowski & Rutkowski, 2010, 2013), The greater access to electronic goods or internet in current days does not necessarily differentiate among higher and lower classes as could happen in a recent past (Avvisati, 2020).

Turning specifically to HOMEPOS in PISA 2015, I observe questions' wording that raises concerns regarding their weight in the index computation. For instance, 6 of the common 22 questions (27%) refers to the possession of different books, while 4 questions (18%) refer to electronic possessions. In that dimension, two questions present similar topics ('Computers [desktop computer, portable laptop, or notebook]' and 'A computer you can use for school work', which presents a strong polychoric correlation, $r(492{,}640) = 0.739$, $p < 0.001$). Additionally, there is one general question that does not seem to reflect socioeconomic status ('a quiet place to study'), but an educational or academic environment. Finally, the question asking about the possession of 'works of art' at home is open to diverse interpretations, which may confuse respondents. This last question parameter is not included in official reports, although it was not formally excluded from the index (OECD, 2016, 2017).

Another relevant topic relates to the national items—three questions used by each country, which has been praised as a step forward in terms of each country better contextualisation of socioeconomic status (Rutkowski & Rutkowski, 2013). However, diverse points can be raised about those questions: first, they do not necessarily discriminate socioeconomic status but household choices (e.g., expresso machine in France or cultural television programs with payment in Albania). Second, they may refer to outdated technology ('BluRay player' in Mexico) or are biased towards specific sensitivities ('Violin/Cello' in Hong Kong, 'Piano or violin' in Taipei and Macao, or a 'piano' in the Netherlands). Third, only in a few cases, they relate to the possessions of luxury goods ('summer residence' and 'swimming pool' in Malta), which produce extreme parameters. It is also possible to detect redundancy of those national questions with the common questions. For instance, many questions regarding electronics are repeated (e.g. 'laptop' in Moldova and Finland or 'tablet' in Norway, Spain, Switzerland and UK; 'musical instruments' in the United States; an 'encyclopaedia' in Colombia), while local dependencies and inconsistencies among answers are not explicitly assessed by PISA (Avvisati, 2020). Finally, it is possible to find important differences in terms of factor loadings among countries (OECD, 2017), which suggests room for improvement in terms of capturing wealth in families. Additionally, one of the trade-offs of extending national items in HOMEPOS is the difficulty to address cross-country comparability issues using fewer common items across countries. While many criticisms can be made to HOMEPOS highlighting limitations and challenges, there still are a relevant source to be used with caution to shed light on the role of socioeconomic differences in schools.

The complexity of measuring inequality based on categorical data

Measuring inequality based on ordinal or binomial data—or a mixture of both, portrays a set of methodological challenges. First, certain distributional statistics such as the mean or variance or standard deviation cannot be properly drawn (Cowell & Flachaire, 2017; Zheng, 2008). Proportions and modes will be appropriate tools to analyse this type of data. Second, in many cases, ordinal data depict an arbitrary scale or asymmetric intervals in their response alternatives, which may also bias the analysis. For instance, a 5-point Likert scale question does not necessarily represent the same difference between pairs of options. I could either choose the category to 'agree' or 'strongly agree'—both options are closer in my mind in this case—with an opinion regarding certain policy addressing inequality within schools, although I will never choose the middle-point category—'neither agree nor disagree'—because I understand as very far from the 'agree' I might have chosen.

One of the consequences of dealing with categorical data is that traditional inequality measures, such as the Gini coefficient and generalised entropy indexes—for example, Theil or Atkinson indexes, which refer to inequality as a deviation from the mean or are mean-normalised, cannot be suitably employed to measure inequality using categorical raw data (Cowell & Flachaire, 2017; Zheng, 2011).

Recent research has been developing alternatives to develop inequality measurements based on categorical data. Allison and Foster (2004) suggest comparing one-variable cumulative distributions of Likert-type questions by ordering the data and identifying the distance from the median as an inequality measure. As they mention, their method only applies when each case's median coincides among them. Additionally, this method does not meet a desirable characteristic of any inequality index—the normalization axiom, where a distribution of identical observations, where there is total equality, desirably portrays a zero value. Based on that seminal idea, Abul Naga and Yalcin (2008) introduce a family of inequality indices based on the analysis of one variable normalising different questions' scales. Under their method, different Likert-scale questions—portraying 3, 5 or 7 alternatives—can be compared in terms of inequality. Zheng (2011) extends the approach to measuring inequality based on two variables. However, if the median does not provide an adequate reference for inequality—for example, when there is skewness on data, all previously measures may not capture the extent of the inequality.

A second approach developed to address this limitation is proposed by Cowell and Flachaire (2012, 2017). Instead of using the median as a reference, they compute inequality relative to a reference status. They suggest counting ranking positions of all observations and expressing them as proportions of the population. The measure could be either 'downwards' or 'upwards' in terms of relative position on a scale. Although very suggestive, this method does not seem adequate for measuring assets inequality due to the multivariate nature of a continuous wealth trait. However, the idea of maintaining the ordinality of the scales and ranking them rather than measuring inequality remain concepts in my proposed approach.

A third approach that addresses multiple variables consists of computing inequality based on latent variable methods. For instance, Mckenzie suggests a relative inequality measure towards identifying subpopulations' disparity based on a polychoric Principal Component Analysis index data (2005). His method computes each subpopulation's standard deviations divided by the variance explained by the first principal component, which additionally allows the comparisons of subgroups to the overall population inequality. The idea of ratios and comparing to the overall inequality average are kept in my proposal. In this case, IRT is chosen over polychoric PCA as a specific approach to model categorical data.

Finally, at least three caveats can be drawn when assessing school-level inequality based on HOMEPOS. First, HOMEPOS is derived through a posterior weighted maximum likelihood estimation (WLE), which assumes a normal distribution (Warm, 1989). In the case of PISA 2015, significant differences between countries occur in terms of the mean of HOMEPOS while there are fewer variations in the distribution across countries (see

Fig. 5 in Annex 2). Second, simulations show that WLE tends to overestimate within-school variance (OECD, 2009). This is relevant for our case as school-level inequality is relative to the variance of school HOMEPOS. Third, WLE is sensitive to ceiling and floor effects if items are too easy or difficult, respectively. This contradicts another desired property of any inequality measure—scale invariance, where proportion changes to answers should not modify inequality. For example, if we add 10% of wealth to everyone, inequality remains the same as previous. Finally, as WLE are only a single possible realization of the estimation it does not addresses the uncertainty of the model, which could be adapted by using plausible values as independent variables (Pokropek, 2015). However, to address current limitations with measuring inequality based on WLE, I compute inequality based on the raw answers of family possessions rather than using the derived-index HOMEPOS.

Alpha inequality: inequality based on an Item Response Theory paradigm

Item Response Theory models

The proposed inequality measure—hereafter, Alpha Inequality—builds upon the discrimination parameter from IRT models. IRT is a statistical family of latent construct analysis that focuses on categorical data and is mainly used in educational and psychological fields. IRT assumes that each person has a certain level—called individual trait—of an unobservable continuous construct (e.g., knowledge, competences, attitudes) that predict the probability of answering correctly or endorsing an observable item (e.g., cognitive questions). In this case, the higher the possession of the construct—family wealth, the higher the probability of answering the possession of an item—electronic good.

It is based on the notion that the probability of a correct response or endorsement to an item is a function of both the person's trait and certain item parameters—such as difficulty, discrimination or pseudo guessing (Embretson & Yang, 2006). The item parameters determine the information offered by each item to any person's trait level.

The simplest IRT model is often called the Rasch model (Rasch, 1960). According to the Rasch model, an individual's response to a binary item (i.e., right/wrong, agree/disagree) is determined by the individual's trait level and one item parameter—the difficulty of the item. Because this model uses the logistic density function and uses a single item parameter, it is called the one-parameter logistic model (1-PL) (Fischer, 1995)—although there are some conceptual differences between Rasch and 1-PL. Other IRT models have been developed covering ordinal and nominal data; adding parameters to the logistic function such as the discrimination or guessing parameters (Embretson & Yang, 2006); and also using distinct methods towards dichotomising data for the analytical modelling process.

For instance, in 2015, PISA uses two IRT models: the generalised partial credit model (GPCM) (Muraki, 1992) for multi-item questions and the two-parameter logistic model for dichotomous items. In both cases, it adds the item discrimination parameter α_i to the function, which will be explained later. The GPCM presents the following notation:

$$P\left(X_{ij} = k|\theta_j, \beta_k, \alpha_i\right) = \frac{e^{\left[ak_{k-1}\left(\alpha_{2i}\theta_{2j}+\alpha_{2i}\theta_{2j}+\cdots+\alpha_{ni}\theta_{nj}\right)+\beta_{k-1}\right)\right]}}{\sum_{k=1}^{K} e^{\left[ak_{k-1}\left(\alpha_{2i}\theta_{2j}+\alpha_{2i}\theta_{2j}+\cdots+\alpha_{ni}\theta_{nj}\right)+\beta_{k-1}\right)\right]}}, \tag{1}$$

which expresses the probability of an individual i correct response (or endorsement) X_i to an item j for the total number of categories K of each question. θ_j represents the individual's trait level, while β_k refers to the item difficulty or location. The parameter ak_k indicates the ordering of the categories from 0 to $k-1$ (Chalmers, 2012).

The discrimination parameter α_i represents the degree to which an item differentiates between respondents in different regions of the measured latent trait θ_j (in this case, household possessions). The parameter defines the steepness of the slope when $P(\theta) = 0.5$, where higher values suggest a better separation between individuals with higher and lower latent traits. Therefore, if $\alpha_i \to \infty$, the item represents a perfect separation between those who respond correctly, in this case, have a specific possession, and those who do not have ownership of it. Figure 1 is a simulated example of item characteristics curve (ICC) for three items, where item 3 has a higher discrimination parameter than the other two items because it 3 shows a steeper curve than items 1 and 2. The item discrimination parameter α_i reflects the sensitivity of the response probability to trait levels changes (Embretson & Yang, 2006) and gives information on the importance of the item to the individual trait—in this case, how relevant possessing certain good reflects family wealth.

Now I depart from the usual IRT parameter interpretation to turn into the consideration of inequalities. First, let us remember that inequality is an aggregated measure and not an individual condition. Therefore, we can think the latent trait as a continuum of equality (or inequality) of wealth for all respondents. In the hypothetical case that all respondents fall into the same value of θ, then the item represents an egalitarian condition—irrespective of the location in the x-axis of $P(\theta) = 0.5$, where values in the left of the axis would represent poverty while in the right would represent richness). If the same occurred for all items, then there will be a status of full egalitarianism. Additionally, as the parameter defines the steepness of the ICC, larger item discrimination also means that the gap between those that are below the 50% probability of endorsing the item and those over that threshold has greater weight in terms of splitting individuals

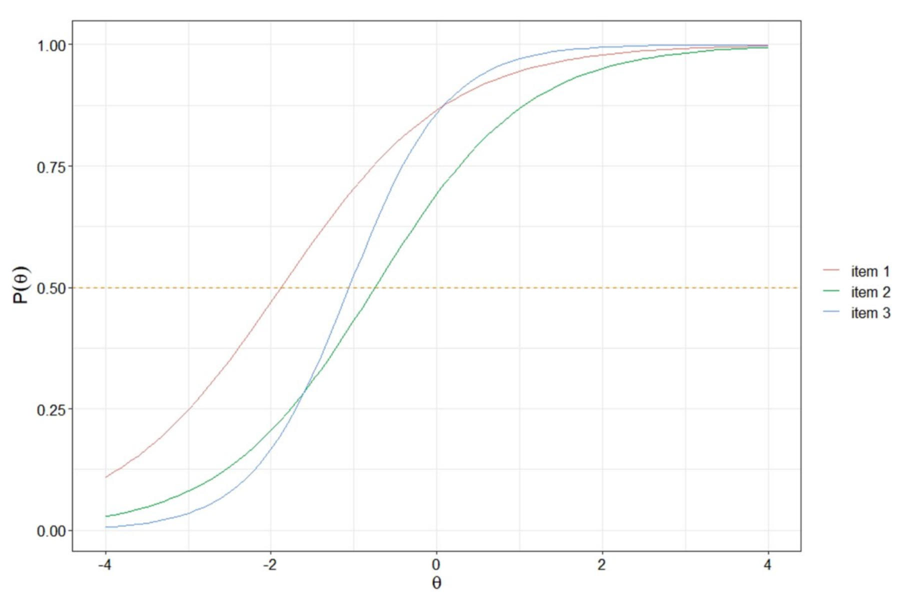

Fig. 1 Example of 3-items characteristic curve

in the trait. The Alpha Inequality is based on this interpretation of the discrimination parameter.

Developing Alpha Inequality

The building process of Alpha Inequality, $I_j(x)$, of any economic variable of interest—in this case, household assets possession—implies the following steps. First, the method involves modelling any IRT or latent variable model that considers the binary or ordinary nature of the responses—such as the graded response model, continuation ratio model, among many—and assumes the existence of a discrimination parameter that differs between items—which is not the case of a 1-PL model. In this example, I use GPMC for polytomous questions and 2-PL for binary items to coincide with the PISA 2015 modelling strategy.

The first step involves computing the IRT models for each item used in building the index and extracting the α_i parameters. The second step consists of normalising all answers alternatives, ς_i, into the same range of values, in this case, from 0 to 1. This is done to give the same importance to polytomous and binomial questions in terms of a similar contribution to the inequality measure. The third step involves the sum of the product of each parameter α_i and the observation score ς_{ij} for each observation (person), j, of the dataset. This is noted as follows:

$$\xi_j = \sum_{j=1}^{n} \alpha_i \varsigma_{ij}. \tag{2}$$

In the case of missing data, I weight each observation j according to the number of questions answered, q_j to differentiate questions not answered from the absence of possession of an item, such as in:

$$\omega\xi_j = \frac{\xi_j}{q_j}. \tag{3}$$

The final step implies computing the inequality measure for each school, I_φ, which allows comparing between school, as well as assessing if schools reach an egalitarian status, where $I_\varphi = 0$. The inequality measure for each school φ is computed as the ratio between the standard deviation of $\omega\xi_j$ by the standard deviation of the entire population c, in this case, each country, ξ_c, which can be expressed as:

$$\mathrm{I}_\varphi = \frac{\sigma\left(\omega\xi_j\right)}{\sigma(\xi_c)}. \tag{4}$$

Following McKenzie (2005), this provides additional information such as if I_φ is greater that one, the school displays more inequality than the country average inequality.

Every inequality measure has some properties to fulfil to provide reliable information regarding the distribution of any variable, in this case, wealth: scale and anonymity invariance, population independence, and binding the Pigou–Dalton transfer principle (Cowell, 2016). The Lemma containing how I_φ fulfills all main axioms and its proof can be found in Annex 1.

Methods

Data

I use the wealth index, HOMEPOS from PISA 2015 to exemplify and evaluate the performance of Alpha Inequality. PISA 2015 collects data from dichotomous and ordinal questions on 25 household indicators across 73 countries and economies. The target population and sampling strategy aimed to represent the universe of 15-year-old students enrolled in each educational system. Students are sampled following a stratified design, where a minimum of 150 schools with proportional probabilities to the student population is initially selected. The minimum sample expected by a school is 20 students to ensure adequate accuracy in estimating between and within schools variance (OECD, 2017).

HOMEPOS is computed based on data collected from three student's questions (ST011, ST012, ST013), with 25 questions covering different household assets and characteristics. Question ST011 displays two sets of dichotomic questions (possible answers: 'yes', 'no'): thirteen that are common to all countries and three questions which differ by each country (called national items). Question ST012 displays eight 4-response option questions (possible answers are: 'none', 'one', 'two' and 'three or more'), common to all countries, while Question ST013 present one questions with six scales (with the following possible answers: '0–10 books', '11–25 books', '26–100 books', '101–200 books', '201–500 books', and 'More than 500 books').

Following PISA's criteria (OECD, 2017), I subset those observations with at least 3 answers on the HOMEPOS scale and no missing values for the computed HOMEPOS scale. I exclude observations from schools with less than 20 observations. Additionally, data from two USA states and Puerto Rico, which did not provide identification of schools, are also excluded. The sample was reduced from 519,334 to 454,734 observations belonging to 69 countries, administrative regions, and economies and 13,387 schools. Descriptive statistics per country used in this study are in Tables 1 and 7 in Annex 2 shows the frequency of observations per country.

PISA's modelling strategy for HOMEPOS is a two-step process. First, a multiple group IRT two-parameter model is estimated (GPCM for ordinal questions and 2PL for dichotomous questions). Subsequently, HOMEPOS is computed based on the posterior weighted maximum likelihood estimation (WLE) (OECD, 2017). As HOMEPOS published parameters by PISA are estimated from a sample and do not reflect the observations used in this study (OECD, 2017), I replicate the first step of PISA's modelling strategy to extract the α discrimination parameters for each country and items. Following PISA, I estimate 22 common questions with equal parameters while 3 questions had parameters freely estimated per country. Correlations between PISA's HOMEPOS and the replicated index are over 0.939 for each country (see Table 8 in Annex 2).

Great variability is seen in terms of discrimination across items (Table 2), where, for instance, the questions 'book of poetry' and 'classic literature' present lower values, and in the opposite side, 'internet access' and 'computers' present the highest values among the common parameters.

There is also large variability in the parameters of the national-specific items, shown in Table 3. For instance, some countries present higher values in all three items, such as the case of Thailand, while the opposite also occurs, such as in the case of the United

Table 1 Descriptive statistics HOMEPOS items. Source: OECD (2017)

Statistic	n	M	SD	Min	Median	Max	Description
ST011Q01TA	448,112	0.886	0.318	0	1	1	A desk to study at
ST011Q02TA	443,628	0.818	0.386	0	1	1	A room of your own
ST011Q03TA	447,922	0.878	0.328	0	1	1	A quiet place to study
ST011Q04TA	448,642	0.858	0.349	0	1	1	A computer you can use for school work
ST011Q05TA	439,751	0.522	0.500	0	1	1	Educational software
ST011Q06TA	448,498	0.897	0.304	0	1	1	A link to the Internet
ST011Q07TA	442,540	0.522	0.500	0	1	1	Classic literature
ST011Q08TA	442,974	0.483	0.500	0	0	1	Books of poetry
ST011Q09TA	443,413	0.599	0.490	0	1	1	Works of art
ST011Q10TA	446,376	0.824	0.381	0	1	1	Books to help you with your school work
ST011Q11TA	440,147	0.588	0.492	0	1	1	Technical reference books
ST011Q12TA	447,453	0.926	0.262	0	1	1	A dictionary
ST011Q16NA	441,943	0.561	0.496	0	1	1	Books on art, music or design
ST011D17TA	444,699	0.631	0.482	0	1	1	*Country-specific item 1*
ST011D18TA	429,510	0.612	0.487	0	1	1	*Country-specific item 2*
ST011D19TA	408,365	0.534	0.499	0	1	1	*Country-specific item 3*
ST012Q01TA	450,081	3.156	0.829	1	3	4	Televisions
ST012Q02TA	442,555	2.419	0.967	1	2	4	Cars
ST012Q03TA	439,415	2.506	0.839	1	2	4	Rooms with a bath or shower
ST012Q05NA	448,358	3.499	0.870	1	4	4	Cell phones with internet access
ST012Q06NA	448,500	2.847	1.003	1	3	4	Computers (desktop computer, portable laptop)
ST012Q07NA	443,428	2.116	1.040	1	2	4	Tablet computers
ST012Q08NA	442,489	1.326	0.685	1	1	4	E-books
ST012Q09NA	448,337	2.063	1.108	1	2	4	Musical instruments
ST013Q01TA	450,608	2.978	1.460	1	3	6	Number of books in your house
HOMEPOS	454,734	−0.338	1.199	−9.481	−0.248	5.994	Home possessions index

Italics denote specific items for each country

Kingdom. Germany is the only case that presents a negative discrimination parameter for the question 'A TV in your own room.' A negative discrimination parameter suggests the latent trait diminishes with the ownership of the good.

As the objective of the study is to exemplify the construction of the inequality measure, I do not address and evaluate model fit and invariance analysis. I rely on PISA's item invariance analysis—named root mean square deviance (RMSD), which states that invariance of HOMEPOS items across countries was analysed and 'unique parameters were assigned if necessary' (OECD, 2017, p. 342). However, as I was previously mentioned, prior research reports dispute the reliability and validity of socioeconomic scales in PISA. I acknowledge those limitations and focus, on the present study, only on the methodological contribution of building an inequality measure.

Criteria to assess Alpha Inequality validity

The strategy chose to examine Alpha Inequality assessing its validity in comparison to prior evidence and comparing results to a well-known inequality index based on HOMEPOS such as the Gini coefficient. The Gini coefficient is computed based on HOMEPOS

Table 2 Item Alpha parameter—common items. Source: author's calculations based on OECD (2017)

Item	Alpha
ST011Q01TA	1.116
ST011Q02TA	0.803
ST011Q03TA	0.931
ST011Q04TA	2.18
ST011Q05TA	0.914
ST011Q06TA	2.536
ST011Q07TA	0.643
ST011Q08TA	0.509
ST011Q09TA	0.906
ST011Q10TA	0.623
ST011Q11TA	0.903
ST011Q12TA	0.788
ST011Q16NA	0.827
ST011D17TA	1.288
ST011D18TA	1.215
ST011D19TA	1.247
ST012Q01TA	0.632
ST012Q02TA	0.923
ST012Q03TA	0.918
ST012Q05NA	0.901
ST012Q06NA	1.612
ST012Q07NA	0.796
ST012Q08NA	0.671
ST012Q09NA	0.594
ST013Q01TA	0.475

applying a correction for finite populations (Nygärd & Sandström, 1985). HOMEPOS was transformed into a range of positive values [0, 15.457] to address a requirement of the Gini coefficient computation.

First, I compare cross-countries rankings statistics from both measures and exemplify the relevance of inequality on learning scores in the case of the USA by comparing schools at both extremes of the inequality continuum.

Second, I model a set of textbook regressions to examine how Alpha Inequality and the Gini coefficient are associated with Mathematics scores. For each country, I fit two sets of two-level mixed-effects linear models, allowing random intercepts to vary at school-levels. This addresses the hierarchical structure of PISA, where students are nested in schools. Formally, the equation of two-level random intercept model reads as:

$$\mathrm{Y_{ij}} = \beta_{0j} + \beta_1 homepos_{1ij} + \beta_2 \mathrm{x_{1ij}} + \mathrm{u_j} + \epsilon_{ij}, \tag{5}$$

where Y_{ij} denotes the outcome variable for the i-th observation (student) of group j (School), β_{0j} the school intercepts (which are random variables enabling the quantification of the differences between groups). β's are regression parameters invariant across groups. The different inequality measures are denoted by x_{1ij}, while u_j is the group-dependent deviation from the intercept mean and ϵ_{ij} represents the error term.

Table 3 Item Alpha parameter—country-specific items. Source: author's calculations based on OECD (2017)

CNT	ST011D17TA	ST011D18TA	ST011D19TA
ARE	1.288	1.215	1.247
AUS	0.655	0.610	1.015
AUT	1.132	0.539	1.106
BEL	0.522	0.749	1.365
BGR	1.575	1.734	1.110
BRA	1.331	1.149	1.448
CAN	1.090	0.745	0.704
CHE	1.644	0.499	1.107
CHL	0.764	1.599	1.058
COL	1.819	0.851	0.907
CRI	1.189	1.552	1.177
CZE	0.851	0.851	0.851
DEU	0.212	−0.197	1.484
DNK	1.883	0.270	0.851
DOM	1.318	1.694	1.158
DZA	0.851	0.851	0.851
ESP	1.311	0.648	0.763
EST	1.228	1.435	1.314
FIN	1.727	0.655	0.851
FRA	0.785	1.109	1.171
GBR	0.347	0.835	0.570
GEO	1.013	1.119	1.195
GRC	1.441	0.993	1.227
HKG	1.036	1.569	0.653
HRV	0.806	0.921	1.157
HUN	0.750	1.114	1.858
IDN	1.631	1.036	2.162
IRL	0.956	0.741	0.746
ISL	1.038	0.976	0.622
ISR	0.800	1.207	0.914
ITA	0.881	0.952	0.742
JOR	0.953	1.238	1.589
JPN	1.360	0.619	0.738
KOR	1.077	1.128	1.291
KSV	1.221	1.199	1.536
LBN	1.456	1.365	0.749
LTU	1.781	0.929	1.511
LUX	0.614	1.218	0.353
LVA	1.340	1.193	0.679
MAC	1.587	1.860	1.482
MDA	1.712	0.851	0.851
MEX	1.155	1.381	1.374
MKD	0.969	0.757	0.851
MLT	0.969	0.683	0.991
MNE	1.557	1.761	1.900
NLD	0.631	1.724	0.767
NOR	1.747	0.811	0.851
NZL	0.784	0.739	1.098

Table 3 (continued)

CNT	ST011D17TA	ST011D18TA	ST011D19TA
PER	1.510	1.922	2.132
POL	1.474	1.853	1.789
PRT	0.448	1.063	1.071
QAR	0.383	1.495	1.115
QAT	0.837	1.348	1.094
QCH	1.899	2.501	1.315
QES	1.271	0.623	0.771
ROU	1.018	0.720	1.725
RUS	1.694	1.286	1.246
SGP	1.615	1.283	0.851
SVK	1.177	1.994	0.851
SVN	0.709	1.186	1.176
SWE	1.092	0.725	0.785
TAP	1.724	1.148	1.465
THA	2.377	1.155	1.932
TTO	1.156	0.810	0.566
TUN	1.572	1.824	1.098
TUR	1.075	1.659	1.502
URY	0.499	1.251	2.294
USA	0.863	1.375	1.142
VNM	2.577	0.794	2.052

HOMEPOS was included in the model due to the influence of the difficulty parameter on the posterior estimation of HOMEPOS, which may allow a better understanding of the role of an inequality measure independent from the wealth possessions.

There are three key methodological considerations which should be considered when modelling data from PISA. First, it is important to consider that PISA is based upon a two-stage stratified sampling strategy to select schools and students. I address this using sampling weights to account for differences in the probabilities of students, classes and schools being selected in the sample (Rutkowski et al., 2010). Considering a multilevel analysis setting, I follow current PISA's practice since 2012 (OECD, 2017) using weights both at the student and school levels in the regression analysis. For the student level, I scale student weights following Rabe-Hesketh and Skrondal (2006), which adjusts students' weights by the ratio of the school size and the sum of students' weighs, as follows:

$$\mathrm{W\_FSTUWT}_{RH-S} = \mathrm{W\_FSTUWT} * \left(\sum students\ in\ school \Big/ \sum_{s=1}^{n} \mathrm{W\_FSTUWT} \right). \tag{6}$$

School level weights correspond to the sum of $\mathrm{W\_FSTUWT}_{RH-S}$ for each school.

Secondly, due to PISA's design, tests scores are estimated as plausible values, where each student has 10 different marks. To address this uncertainty, I apply Rubin's rules for handling multiple imputations (Rubin, 1987) both in terms of computing schools averages and modelling regressions for each plausible value, where I compute adjusted sets of coefficients and standard error estimates and join them in a final estimate. Finally, due to the stratified multistage sampling design mentioned earlier, I estimate the uncertainty associated with the sampling using PISA's approach—Fay's modification of the balanced repeated replication (BRR) method, which allows computing the sampling variance.

Item parameters are estimated through an iterative marginal maximum likelihood approach (Bock & Aitkin, 1981), using the expectation–maximization algorithm provided by *mirt* package (Chalmers, 2012) in statistical software R (R Core Team, 2020) and statistical analysis was performed using package *BIFIEsurvey* (Robitzsch & Oberwimmer, 2015).

Results and discussion

Comparison between two school-level inequality measurements

Comparisons between countries are only feasible if we assume the existence of measurement invariance across countries, which allows further inferential analysis in the same metric. Conditionally to the assumption of measurement invariance claimed by PISA (OECD, 2017, p. 342). Table 3 presents the average inequality per country and the inequality coefficient of variation (CV) for both inequality measurements. While Alpha Inequality/Gini aims to assess the level of school-level inequality per country, CV provides a sense of the variability of inequality within the educational system.

Looking at the Alpha Inequality values, countries from Latin America and South Asia such as Peru, China (4 cities), Indonesia, Thailand and Colombia present the lowest values of Alpha Inequality and, at the same time, high values of CV. The opposite occurs with countries such as Iceland, Finland, Estonia, Poland, and Norway, which present Alpha Inequality close to 1 while having low values of CV. This suggests important differences between the two groups of countries. The first group of countries are characterised by educational systems with socioeconomically more homogeneous schools and larger degrees of segregation between schools, dividing poor and rich in different schools. The second group presents relatively smaller socioeconomic differences between schools while having larger within schools' economic diversity. This coincides with recent research focused on the analysis of segregation on different waves of PISA (Gutiérrez et al., 2019). Additionally, Alpha Inequality allows comparisons between countries (Table 4). For instance, Iceland, Kosovo, Moldova, Montenegro, Iceland, New Zealand, and Qatar present more than 35% schools with school-inequality above their national average, while Indonesia, Israel, Peru, China (4 cities) and Thailand only present less than 5% schools above the national average of inequality (see Table 9 in Annex 2).

Figure 2 shows the distribution of Alpha Inequality for each school by countries. Alpha Inequality presents different distributions across countries, as could be expected based on prior cross country analysis (Thomas et al., 2001). In some cases, they approximate to Gaussian functions, such as the case of Brazil, Indonesia, and Australia, while in other cases there are bimodal distributions such as in the case of Malta, Macedonia, and Trinidad and Tobago. In many cases, kurtosis and skewness are relevant features to be observed on the distributions and inferential analysis.

On the other hand, the Gini inequality presents, in general, very low coefficients across countries and schools. National averages are in a range between 0.003 and 0.006, and countries such as The Netherlands, Denmark and the Slovak Republic appear with the

Table 4 Mean and coefficient of variation per country—Alpha inequality and Gini coefficient. Source: author's calculations based on OECD (2017)

Country	Alpha inequality		Gini	
	Mean	CV	Mean	CV
ARE	0.875	0.279	0.005	0.263
AUS	0.881	0.288	0.005	0.239
AUT	0.888	0.249	0.004	0.189
BEL	0.884	0.229	0.004	0.193
BGR	0.856	0.343	0.004	0.258
BRA	0.832	0.168	0.005	0.179
CAN	0.922	0.248	0.005	0.190
CHE	0.915	0.215	0.003	0.198
CHL	0.785	0.269	0.004	0.175
COL	0.770	0.207	0.005	0.210
CRI	0.827	0.202	0.005	0.181
CZE	0.861	0.334	0.003	0.192
DEU	0.882	0.267	0.003	0.213
DNK	0.920	0.246	0.003	0.239
DOM	0.845	0.165	0.005	0.173
DZA	0.924	0.131	0.006	0.218
ESP	0.904	0.221	0.004	0.172
EST	0.951	0.184	0.004	0.172
FIN	0.961	0.191	0.003	0.185
FRA	0.873	0.259	0.003	0.198
GBR	0.906	0.266	0.005	0.185
GEO	0.862	0.256	0.004	0.201
GRC	0.902	0.245	0.004	0.208
HKG	0.921	0.157	0.004	0.174
HRV	0.911	0.237	0.003	0.192
HUN	0.861	0.252	0.004	0.191
IDN	0.729	0.201	0.005	0.209
IRL	0.933	0.228	0.004	0.157
ISL	0.963	0.218	0.003	0.193
ISR	0.581	0.336	0.005	0.270
ITA	0.919	0.212	0.003	0.210
JOR	0.870	0.194	0.006	0.246
JPN	0.931	0.170	0.004	0.163
KOR	0.912	0.191	0.003	0.171
KSV	0.918	0.238	0.004	0.180
LBN	0.792	0.278	0.005	0.229
LTU	0.911	0.300	0.003	0.248
LUX	0.933	0.141	0.005	0.102
LVA	0.935	0.249	0.003	0.190
MAC	0.900	0.115	0.004	0.116
MDA	0.873	0.272	0.004	0.277
MEX	0.808	0.167	0.005	0.164
MKD	0.908	0.233	0.004	0.189
MLT	0.894	0.247	0.004	0.132
MNE	0.930	0.183	0.004	0.141
NLD	0.897	0.213	0.003	0.180
NOR	0.942	0.285	0.004	0.214

Table 4 (continued)

Country	Alpha inequality		Gini	
	Mean	CV	Mean	CV
NZL	0.925	0.249	0.004	0.191
PER	0.714	0.225	0.005	0.204
POL	0.943	0.222	0.004	0.176
PRT	0.906	0.210	0.004	0.155
QAR	0.782	0.310	0.004	0.182
QAT	0.883	0.303	0.005	0.264
QCH	0.719	0.208	0.004	0.184
QES	0.926	0.207	0.004	0.168
ROU	0.828	0.303	0.004	0.188
RUS	0.920	0.250	0.004	0.225
SGP	0.877	0.244	0.004	0.190
SVK	0.788	0.340	0.004	0.326
SVN	0.908	0.268	0.003	0.218
SWE	0.933	0.232	0.004	0.209
TAP	0.906	0.161	0.004	0.163
THA	0.744	0.221	0.005	0.163
TTO	0.884	0.245	0.006	0.166
TUN	0.834	0.170	0.005	0.178
TUR	0.840	0.148	0.005	0.164
URY	0.846	0.267	0.004	0.188
USA	0.911	0.211	0.005	0.171
VNM	0.825	0.159	0.004	0.224

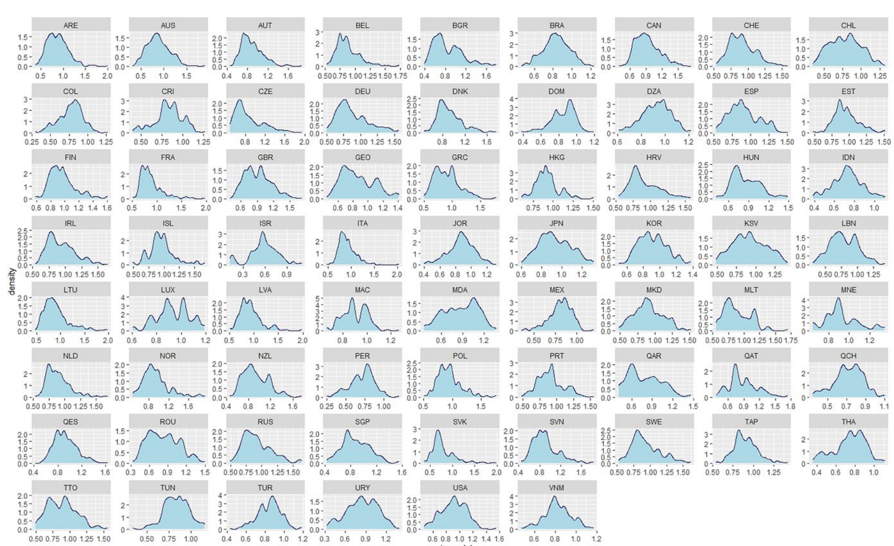

Fig. 2 Density plots per country: school Alpha Inequality

smallest values while countries such as Trinidad and Tobago, Qatar and Algeria display the largest values. However, countries like Denmark and the Slovak Republic present high coefficients of variation, which contradicts previous empirical evidence in terms

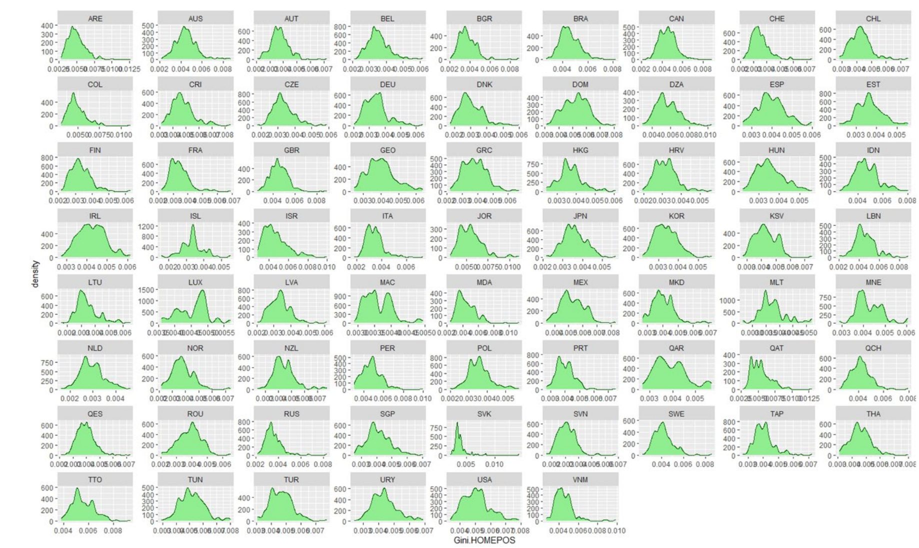

Fig. 3 Density plots per country: school Gini coefficient

of segregation in schooling systems (Gutiérrez et al., 2019). Figure 3 shows school Gini density functions for each country, where in general, they present heavy-tailed distributions. Exceptions of bimodal distributions are Macedonia and Montenegro.

Country-level correlations of both inequality measurements present an overall mean of 0.612(*SD* 0.131) ranging from 0.105 (Israel) to 0.846 (Qatar) (Table 10 in Annex 2).

To examine the impact of differences between both measurements, I turn to the case of the USA, which has more prior empirical analysis on segregation and inequality. The Gini coefficient does not provide a hint of difference between schools in the top 20% and the bottom 20% of the Gini index in terms of the average of mathematics learning scores per school. This contradicts prior estimations (Rutkowski et al., 2018) as well as cross-country studies that focus on the segregation levels of USA schools and educational scores (Benito et al., 2014; OECD, 2018). Contrarily, Fig. 4 shows how schools with lower Alpha inequality outperform in terms of Mathematics Average by 0.57 standard deviations schools with the largest share of inequality with statistically significant differences between groups, $t(60.36) = -7.01$, $p < 0.001$. This represents about 2 more years of schooling according to PISA (2009).

Models' coefficients

Results from country-level mixed-effects regressions models can be seen in Table 5 with Alpha Inequality as a predictor of Mathematics score. I find that in 67 out of 69 countries show statistically significant negative parameters, while in the case of Indonesia and Vietnam the null hypothesis of a parameter different from 0 cannot be rejected under a standard cut-off of $p < 0.05$.

On the other hand, Table 6 presents the estimations of regression parameters using the Gini coefficient for each county. In this case, the number of cases not showing a statistically significant association raises to 5, being Estonia, Iceland, Latvia, United Kingdom, and the United States of America. The case of the United States, as previously discussed,

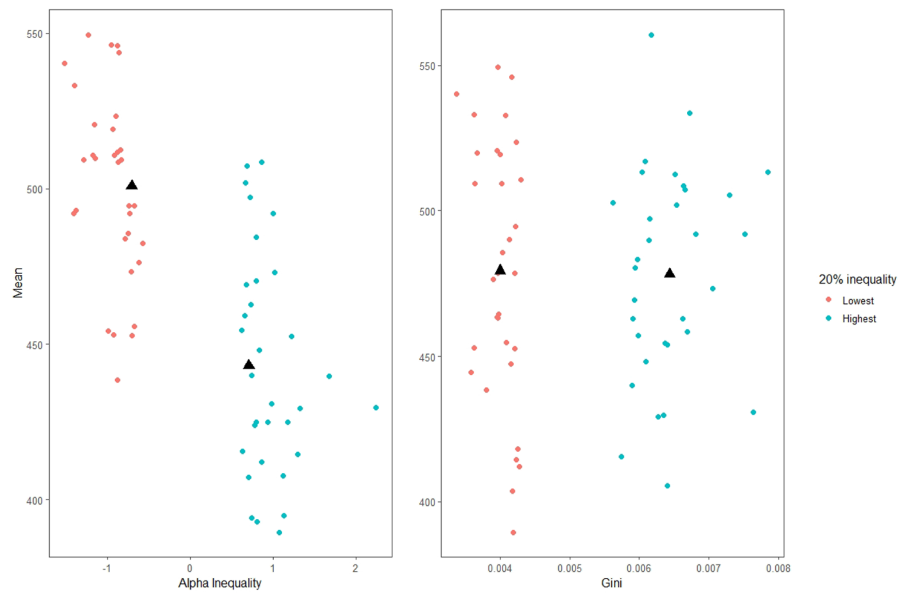

Fig. 4 20% schools with higher and lower inequality (Alpha Inequality and Gini) and mean Maths scores—USA

raises concerns in terms of estimation reliability of the Gini parameter due to the lack of ability to find statistical significance given the previous empirical evidence found in the literature. Additionally, Luxembourg is the only case portraying a positive coefficient for the slope of school inequality and mathematics scores.

Conclusion

This paper has found that a set of multivariate household possessions collected as categorical data can be used to provide a novel measure of inequality. The proposed inequality measure is independent of the scale of wealth and fulfils the main properties of inequality measures. Additionally, Alpha Inequality also allows for comparisons between and within countries.

Computing school-level inequality using data of PISA 2015, I find a consistent significant negative association of school-level inequality and Mathematics scores across countries—the great exemption being a majority of European countries. It is also suggested that the inequality measure outperforms the Gini coefficient in terms of assessing the association of school-level inequality and learning outcomes. This is consistent with previous research on the topic that identifies different levels of inequality within and cross-countries. In the case of known negative effects of inequality, Alpha Inequality is shown to better grasp the relevance of socioeconomic disparities between schools in terms of learning scores.

There are important limitations to be acknowledged. While the improvement of socio-economic scales such as HOMEPOS focusing on the need of updating items to represent wealth in current times, cross-compatibility and model fit becomes a requisite to apply and study thoroughly the effects of school-inequality, further research could point to different directions such as the assessment of inequality on cognitive and non-cognitive

Table 5 Regression coefficients per country—Alpha inequality. Source: author's calculations based on OECD (2017)

CNT	HOMEPOS			Alpha inequality		
	Est	SE	p-value	Est	SE	p-value
ARE	2.97	0.854	0.006	− 39.9	1.5	0
AUS	22.5	1.29	0	− 19.3	1.61	0
AUT	27.4	1.37	0	− 25	1.37	0
BEL	28.7	1.1	0	− 33.9	0.94	0
BGR	13	1.55	0	− 41.1	1.76	0
BRA	24.2	0.976	0	− 23.2	1.14	0
CAN	14.6	1.19	0	− 16.5	0.589	0
CHE	25.4	1.33	0	− 31.2	1.29	0
CHL	17.5	1.12	0	− 35.8	1.34	0
COL	16.8	0.998	0	− 16.3	1.09	0
CRI	16.8	0.571	0	− 14.7	1.16	0
CZE	25.6	1.2	0	− 39	2.32	0
DEU	30.4	1.36	0	− 29.6	1.07	0
DNK	34.1	1.59	0	− 13	0.87	0
DOM	14.1	1.29	0	− 15.7	1.23	0
DZA	4.65	0.906	0	− 14.5	2.18	0
ESP	29.2	0.861	0	− 11.6	0.785	0
EST	17.5	1.52	0	− 13.4	1.36	0
FIN	24.3	1.47	0	− 8.51	1.04	0
FRA	31.6	1	0	− 33.8	0.989	0
GBR	20.9	0.836	0	− 20.6	1.06	0
GEO	22.4	1.68	0	− 27.2	2.02	0
GRC	14.4	1.56	0	− 23.7	1.41	0
HKG	19.7	1.46	0	− 20.4	1.54	0
HRV	24.1	1.91	0	− 34	1.15	0
HUN	23	1.28	0	− 36.9	1.85	0
IDN	26.6	0.884	0	− 2.31	1.23	0.062
IRL	22.6	1.09	0	− 14.7	0.968	0
ISL	22.1	1.51	0	− 4.25	1.4	0.008
ISR	12	1.33	0	− 32.4	1.37	0
ITA	22.4	1.62	0	− 20.2	1.3	0
JOR	13.8	1.3	0	− 18.1	1.4	0
JPN	23.4	1.65	0	− 24.8	2.04	0
KOR	35	2	0	− 23.6	1.18	0
KSV	14.1	1.24	0	− 21.3	1.08	0
LBN	23.9	1.65	0	− 32	1.41	0
LTU	18.9	1.67	0	− 22.1	1.2	0
LUX	25.4	0.846	0	− 27.2	0.648	0
LVA	20.1	1.69	0	− 8.92	0.992	0
MAC	7.58	1.29	0	− 22.1	0.802	0
MDA	22.4	1.2	0	− 21.2	1.34	0
MEX	13.8	0.753	0	− 6.59	1.31	0
MKD	20.4	1.17	0	− 29.6	1.24	0
MLT	30.7	1.41	0	− 41.1	0.915	0
MNE	7.03	0.981	0	− 30.8	0.972	0
NLD	26.3	2.51	0	− 35.2	1.82	0
NOR	26.1	1.39	0	− 7.08	1.3	0

Table 5 (continued)

CNT	HOMEPOS			Alpha inequality		
	Est	SE	p-value	Est	SE	p-value
NZL	21.3	1.23	0	− 16.7	1.51	0
PER	23.6	1.02	0	− 13.4	1.25	0
POL	30.6	1.25	0	− 10.9	1.04	0
PRT	30.3	1.77	0	− 22.5	1.08	0
QAR	21.4	1.95	0	− 32.4	3.01	0
QAT	0.094	0.569	0.868	− 52.5	1.03	0
QCH	31.8	1.4	0	− 18.6	1.56	0
QES	27.2	1.15	0	− 11.3	0.475	0
ROU	22.4	1.1	0	− 29.3	1.09	0
RUS	9.82	1.74	0	− 16.2	1.7	0
SGP	28.4	1.03	0	− 26.8	0.87	0
SVK	19	1.72	0	− 35.2	1.69	0
SVN	17.6	1.54	0	− 26.3	0.981	0
SWE	23.8	1.09	0	− 16.7	0.958	0
TAP	26.4	0.875	0	− 34.2	1.17	0
THA	19.2	1.12	0	− 27.2	1.5	0
TTO	12.8	1.37	0	− 43.1	0.801	0
TUN	19.6	1.15	0	− 10.9	1.99	0
TUR	20.8	1.24	0	− 12.9	1.77	0
URY	18.5	1.52	0	− 33.2	0.968	0
USA	21.2	0.639	0	− 14.5	1.19	0
VNM	22.2	1.62	0	− 2.39	1.74	0.17

educational outcomes across different waves of PISA as well the interplay between inequality, segregation and educational outcomes.

Second, there is a methodological debate regarding the inclusion of survey weights design into IRT scoring procedures to take account of the complex sampling designs and nested structure of item response data of PISA and other ILSAs. This uses multilevel item response methods and different weighting strategies (Zheng & Yang, 2016).

Third, alternative sampling weights scaling methods at both levels were explored (Mang et al., 2021) addressing the complexity of using within and between weights in multilevel clustered analysis. Although the number of statistically significant models varied, similar negative coefficients were found in all cases, and, in all cases, models with Alpha Inequality predictors were more sensitive than Gini. However, in some weighting configurations, large standard errors were found suggesting model identification or convergence issues.

Table 6 Regression coefficients per country—Gini coefficient. Source: author's calculations based on OECD (2017)

CNT	HOMEPOS			Gini		
	Est	SE	p-value	Est	SE	p-value
ARE	7.806	0.919	0	−38.54	1.448	0
AUS	25.86	1.354	0	−11.87	1.41	0
AUT	33.9	1.251	0	−5.152	1.361	0
BEL	32.9	1.15	0	−23.09	1.293	0
BGR	22.15	1.587	0	−23.78	1.219	0
BRA	26.57	1.078	0	−13.07	1.233	0
CAN	16.89	1.143	0	−10.24	0.739	0
CHE	32.2	1.323	0	−16.6	1.422	0
CHL	33.64	1.122	0	−13.12	0.988	0
COL	19.76	1.012	0	−9.5	0.913	0
CRI	20.28	0.638	0	−6.613	1.258	0
CZE	41.04	1.345	0	−12.21	1.402	0
DEU	38.26	1.365	0	−8.616	1.218	0
DNK	35.5	1.612	0	−10.1	0.946	0
DOM	14.71	1.356	0	−11.44	1.439	0
DZA	3.332	0.856	0.001	−15.3	1.829	0
ESP	32.59	0.768	0	−4.624	0.766	0
EST	20.16	1.545	0	−2.324	1.583	0.143
FIN	25.32	1.486	0	−4.941	0.981	0
FRA	40.24	1.123	0	−19.72	1.152	0
GBR	24.79	0.913	0	−0.239	0.969	0.803
GEO	28.2	1.692	0	−17.42	1.966	0
GRC	19.09	1.568	0	−8.292	1.674	0
HKG	23.35	1.408	0	−7.728	1.31	0
HRV	32.71	1.849	0	−16.78	1.018	0
HUN	36.97	0.957	0	−11.31	1.713	0
IDN	24.22	0.966	0	−11.79	1.25	0
IRL	25.34	1.019	0	−4.903	0.785	0
ISL	22.17	1.531	0	1.03	0.967	0.3
ISR	16.04	1.404	0	−33.42	1.565	0
ITA	26.79	1.641	0	−7.458	1.496	0
JOR	14.89	1.304	0	−14.56	1.204	0
JPN	27.44	1.419	0	−14.31	2.228	0
KOR	40.99	2.105	0	−14.57	1.299	0
KSV	18.48	1.128	0	−8.201	1.338	0
LBN	32.87	1.511	0	−11.9	1.917	0
LTU	22.81	1.606	0	−13.67	1.231	0
LUX	31.25	0.971	0	4.094	0.77	0
LVA	22.19	1.642	0	−1.699	0.959	0.083
MAC	16.36	1.258	0	−12.88	0.793	0
MDA	25.61	1.482	0	−11.58	1.665	0
MEX	13.7	0.715	0	−7.866	1.205	0
MKD	25.14	1.187	0	−17.75	1.249	0
MLT	41.74	1.322	0	−16.23	0.793	0
MNE	13.91	1.007	0	−10.79	1.128	0
NLD	32.84	2.349	0	−18.22	2.519	0
NOR	27.06	1.386	0	−4.043	1.131	0.001

Table 6 (continued)

CNT	HOMEPOS			Gini		
	Est	SE	p-value	Est	SE	p-value
NZL	23.49	1.259	0	−9.83	1.592	0
PER	21.76	1.065	0	−14.65	1.379	0
POL	32.85	1.16	0	−3.204	1.126	0.005
PRT	36.02	1.731	0	−12.37	1.713	0
QAR	33.41	1.796	0	−18.13	2.734	0
QAT	3.285	0.593	0	−47.95	0.966	0
QCH	37.11	1.386	0	−11.16	1.317	0
QES	30.08	1.122	0	−5.165	0.498	0
ROU	32	1.415	0	−14.79	1.759	0
RUS	11.37	1.655	0	−4.531	1.666	0.015
SGP	34.19	1.021	0	−18.75	0.758	0
SVK	27.7	1.911	0	−16.98	2.016	0
SVN	25.59	1.516	0	−12.04	0.86	0
SWE	25.39	1.104	0	−11.42	1.229	0
TAP	35.24	1.083	0	−18.9	1.043	0
THA	28.82	1.315	0	−3.797	1.003	0
TTO	22.19	1.387	0	−12.99	0.826	0
TUN	20.44	1.115	0	−3.963	1.675	0.022
TUR	20.98	1.286	0	−9.549	1.432	0
URY	27.59	1.5	0	−16.99	1.232	0
USA	24.46	0.61	0	−0.655	1.57	0.675
VNM	19.47	1.344	0	−10.62	1.804	0

This is relevant as sample design in PISA is informed by school socioeconomic attributes and the estimation of parameters—discrimination, among them—could be affected by the lack of weights. Further research could point the relevance of weighting IRT models to address socioeconomic sampling variances. In this case, I mimic IRT modelling single-level strategy and address the stratified complex sampling design in the multilevel regression model regression analysis including replicate and scale weights.

Annexes

Annex 1

Lemma 1 *I_φ satisfies the main properties of an inequality measure.*

- I_φ is continuous on the domain of distributions I.

- I_φ is invariant to permutations of the measure among students in the same population (anonymity invariance).
- I_φ is invariant to any multiplication of each student score observation by any positive integer constant. The inequality measure is, therefore, independent of the aggregate level of income (scale invariance).
- I_φ remains invariant to the size of the population, and therefore, to the replication of observation of the original population (population independence).
- Redistributing benefits from richer to poorer individuals (without individuals' reranking) reduces I_φ, as the standard deviation at the numerator decreases while the denominator remains unchanged (Pigou–Dalton transfer).
- I_φ takes a zero value when all individuals rank their health status identically (normalisation).

Proof of Lemma 1 (Continuity) $I_{\varphi 1}$ and $I_{\varphi 1}$ represent two inequality measures. If $I_{\varphi 1} \approx I_{\varphi 2}$, then they will have very similar inequality values.

(Anonymity) Let x denote any distribution of assets with elements $\{x_1, x_2, \ldots\}$. As $I_\varphi(x)$ depends only on the set $\{x_1, x_2, \ldots\}$, any permutation of elements of x does not produce changes in I_φ, so $I_\varphi(P(x)) = I_\varphi(x)$.

(Scale invariance) For any $I_\varphi(x)$, multiplying a constant $\gamma > 0$ to all elements of the set $\{x_1, x_2, \ldots\}$ produces $I'_\varphi(x\gamma) = I_\varphi(x)$.

(Population invariance) For any x, replicating the population would produce $\xi'_l = \alpha_{l1}\varsigma_{l1} + \alpha'_{l1}\varsigma_{l1} + \alpha_{l2}\varsigma_{l2} + \alpha'_{l2}\varsigma_{l2} + \cdots + \alpha_{ln}\varsigma_{ln} + \alpha'_{ln}\varsigma_{ln}$. Then $I'_\varphi = \frac{\xi'_l}{\xi'_p} = I_\varphi(x \cup x) = I_\varphi$.

(Pigou–Dalton transfer property) Let ξ denote a wealth score of individuals l and m, where $\xi_l > \xi_{\mathrm{m}}$. Let $\widehat{\xi_l} = \xi_l - \delta$ and $\hat{\xi}_m = \xi_m + \delta$, when $\delta > 0$ transferred from l to m. Let I_φ and $\widehat{I_\varphi}$ represent the initial and transformed inequality measure. As $\sigma_j > \widehat{\sigma_j}$, then $I_\varphi > \widehat{I_\varphi}$.

(Normalisation) For any x where $\{x_1 = x_2, \ldots\}$, $\sigma(\xi) = 0$, then $I_\varphi = 0$.

This section suggests the inequality measure fulfils main properties customarily deemed desirable for an inequality measure, and therefore, can be accepted as a desirable measurement of inequality.

Annex 2

See Tables 7, 8, 9, and 10; Fig. 5.

Table 7 Frequency per country. Source: OECD (2017)

Country	Freq.
ARE	12,695
AUS	6398
AUT	6494
BEL	9273
BGR	5393
BRA	19,440
CAN	17,791
CHE	5241
CHL	6634
COL	10,896
CRI	6471
CZE	4828
DEU	4949
DNK	5346
DOM	4167
DZA	5207
ESP	6384
EST	4819
FIN	5735
FRA	5245
GBR	13,214
GEO	3884
GRC	5117
HKG	5270
HRV	5679
HUN	5124
IDN	5966
IRL	5654
ISL	2819
ISR	6413
ITA	10,030
JOR	6928
JPN	6614
KOR	5411
KSV	4013
LBN	3374
LTU	5170
LUX	5204
LVA	3660
MAC	4419
MDA	4293
MEX	6667
MKD	5214
MLT	3524
MNE	5400
NLD	5172
NOR	4940
NZL	3431
PER	6112

Table 7 (continued)

Country	Freq.
POL	4248
PRT	6680
QAR	1520
QAT	11,719
QCH	9632
QES	31,066
ROU	4610
RUS	5240
SGP	6084
SVK	4722
SVN	5338
SWE	5019
TAP	7617
THA	7542
TTO	4381
TUN	4939
TUR	5664
URY	5428
USA	5539
VNM	5624

Table 8 Correlation between HOMEPOS and replication per country. Source: author's calculations based on OECD (2017)

CNT	Correlation
ARE	0.993
AUS	0.989
AUT	0.982
BEL	0.946
BGR	0.977
BRA	0.995
CAN	0.991
CHE	0.975
CHL	0.999
COL	0.998
CRI	0.998
CZE	0.962
DEU	0.964
DNK	0.941
DOM	0.992
DZA	0.999
ESP	0.988
EST	0.984
FIN	0.943
FRA	0.974
GBR	0.986

Table 8 (continued)

CNT	Correlation
GEO	0.992
GRC	0.987
HKG	0.986
HRV	0.950
HUN	0.979
IDN	0.998
IRL	0.987
ISL	0.947
ISR	0.961
ITA	0.971
JOR	0.996
JPN	0.981
KOR	0.973
KSV	0.985
LBN	0.996
LTU	0.970
LUX	0.991
LVA	0.963
MAC	0.994
MDA	0.996
MEX	0.998
MKD	0.989
MLT	0.981
MNE	0.983
NLD	0.953
NOR	0.955
NZL	0.994
PER	0.994
POL	0.980
PRT	0.984
QAR	0.998
QAT	0.988
QCH	0.997
QES	0.987
ROU	0.996
RUS	0.980
SGP	0.990
SVK	0.985
SVN	0.940
SWE	0.981
TAP	0.983
THA	0.998
TTO	0.995
TUN	0.993
TUR	0.996
URY	0.988
USA	0.996
VNM	0.998

Table 9 Percentage of schools above mean country Inequality. Source: author's calculations based on OECD (2017)

CNT	# schools > national	Average # schools < national	Average proportion schools > national average
ARE	96	255	27.35
AUS	68	187	26.67
AUT	55	151	26.7
BEL	61	190	24.3
BGR	43	101	29.86
BRA	83	517	13.83
CAN	190	398	32.31
CHE	52	121	30.06
CHL	32	150	17.58
COL	21	288	6.8
CRI	31	154	16.76
CZE	46	116	28.4
DEU	54	137	28.27
DNK	62	155	28.57
DOM	13	123	9.56
DZA	38	104	26.76
ESP	54	130	29.35
EST	45	91	33.09
FIN	52	101	33.99
FRA	41	149	21.58
GBR	162	342	32.14
GEO	37	93	28.46
GRC	51	111	31.48
HKG	31	106	22.63
HRV	50	104	32.47
HUN	45	122	26.95
IDN	9	181	4.74
IRL	56	106	34.57
ISL	25	44	36.23
ISR	3	161	1.83
ITA	98	255	27.76
JOR	51	173	22.77
JPN	67	129	34.18
KOR	47	111	29.75
KSV	46	85	35.11
LBN	23	120	16.08
LTU	49	130	27.37
LUX	13	26	33.33
LVA	43	98	30.5
MAC	9	30	23.08
MDA	55	88	38.46
MEX	10	176	5.38
MKD	33	70	32.04
MLT	16	33	32.65
MNE	15	26	36.59
NLD	42	132	24.14
NOR	62	134	31.63

Table 9 (continued)

CNT	# schools > national	Average # schools < national	Average proportion schools > national average
NZL	41	74	35.65
PER	4	191	2.05
POL	42	102	29.17
PRT	52	138	27.37
QAR	11	39	22
QAT	54	83	39.42
QCH	3	252	1.18
QES	291	604	32.51
ROU	35	102	25.55
RUS	47	103	31.33
SGP	52	123	29.71
SVK	29	116	20
SVN	56	147	27.59
SWE	51	114	30.91
TAP	50	158	24.04
THA	9	194	4.43
TTO	36	93	27.91
TUN	16	130	10.96
TUR	15	139	9.74
URY	51	123	29.31
USA	56	106	34.57
VNM	20	146	12.05

Table 10 Correlation between GINI HOMEPOS and Ineq. Alpha in each country. Source: author's calculations based on OECD (2017)

CNT	Correlation
ARE	0.771
AUS	0.761
AUT	0.603
BEL	0.757
BGR	0.665
BRA	0.576
CAN	0.632
CHE	0.649
CHL	0.611
COL	0.722
CRI	0.771
CZE	0.303
DEU	0.618
DNK	0.626
DOM	0.567
DZA	0.61
ESP	0.638
EST	0.559
FIN	0.58
FRA	0.654

Table 10 (continued)

CNT	Correlation
GBR	0.337
GEO	0.76
GRC	0.544
HKG	0.706
HRV	0.627
HUN	0.533
IDN	0.39
IRL	0.533
ISL	0.479
ISR	0.11
ITA	0.678
JOR	0.766
JPN	0.768
KOR	0.671
KSV	0.479
LBN	0.525
LTU	0.633
LUX	0.555
LVA	0.276
MAC	0.563
MDA	0.691
MEX	0.626
MKD	0.687
MLT	0.611
MNE	0.583
NLD	0.687
NOR	0.584
NZL	0.733
PER	0.635
POL	0.609
PRT	0.658
QAR	0.759
QAT	0.847
QCH	0.531
QES	0.615
ROU	0.684
RUS	0.54
SGP	0.84
SVK	0.757
SVN	0.387
SWE	0.664
TAP	0.692
THA	0.6
TTO	0.541
TUN	0.71
TUR	0.604
URY	0.634
USA	0.511
VNM	0.533

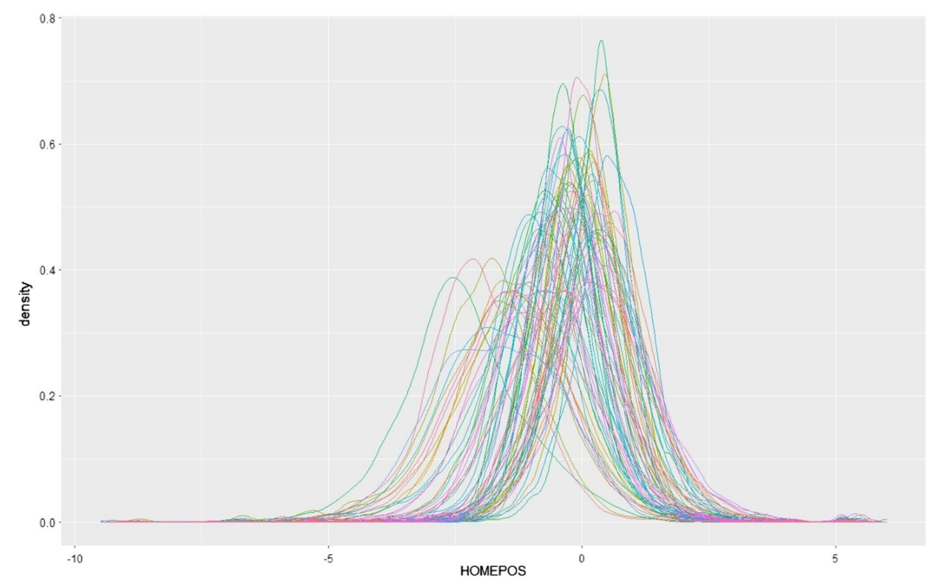

Fig. 5 Density distribution HOMEPOS per country

Authors' contribution
As the sole author, LS conducted the analysis and wrote the whole manuscript. The authorread and approved the final manuscript.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The author gave consent for this article's publication.

Competing interests
No potential conflict of interest is reported by the author.

Received: 4 March 2020 Accepted: 17 April 2021
Published online: 03 May 2021

References
Abul Naga, R. H., & Yalcin, T. (2008). Inequality measurement for ordered response health data. *Journal of Health Economics, 27*(6), 1614–1625
Allison, R. A., & Foster, J. E. (2004). Measuring health inequality using qualitative data. *Journal of Health Economics, 23*(3), 505–524
Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-scale Assessments in Education*. https://doi.org/10.1186/s40536-020-00086-x
Benito, R., Alegre, M. À., & Gonzàlez-Balletbò, I. (2014). School segregation and its effects on educational equality and efficiency in 16 OECD comprehensive school systems. *Comparative Education Review, 58*(1), 104–134
Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29
Coleman, J. (1966). *Equality of educational opportunity*.
Cowell, F. (2016). Inequality and poverty measures. In M. Adler & M. Fleurbaey (Eds.), *The Oxford handbook of well-being and public policy*.Oxford University Press.
Cowell, F. A. & Flachaire, E. (2012). *Inequality with ordinal data*. In European economic association and econometric society conference, Málaga. https://doi.org/10.1111/ecca.12232/epdf
Cowell, F. A., & Flachaire, E. (2017). Inequality with ordinal data. *Economica, 84*(334), 290–321
Deaton, A. (2003). Health, inequality, and economic development. *Journal of Economic Literature, 41*(1), 113–158
Del Bello, C. L., Patacchini, E. & Zenou, Y. (2015). *Neighborhood effects in education*. IZA Discussion Papers.
Embretson, S. E., & Yang, X. (2006). Item response theory. In J. Green, G. Camilli, P. Elmore, A. Skukauskaiti, & E. Grace (Eds.), *Handbook of complementary methods in education research.* (pp. 385–409). Routledge.
Fischer, G. H. (1995). The linear logistic test model. *Rasch models.* (pp. 131–155). Springer.
Gutiérrez, G., Jerrim, J., & Torres, R. (2019). School segregation across the world: Has any progress been made in reducing the separation of the rich from the poor? *Journal of Economic Inequality, 18*(2), 157–179
Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353
Lee, S. S., & Von Davier, M. (2020). Improving measurement properties of the PISA home possessions scale through partial invariance modeling 1. *Psychological Test and Assessment Modeling, 62*(1), 55–83
Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education, 9*(1), 1–39
Mayer, S. E. (1997). *What money can't buy: Family income and children's life chances*. Harvard University Press.
McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics, 18*, 229–260
Muraki, E. (1992). *A generalized partial credit model: Application of an EM algorithm*. Princeton.
Nygärd, F., & Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics, 1*(4), 399–426
OECD. (2009). *PISA data analysis manual*. (2nd ed.). OECD.
OECD. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. OECD.
OECD. (2017). *PISA 2015 technical report*. OECD Publishing.
OECD. (2018). *Equity in education: Breaking down barriers to social mobility*. OECD Publishing. https://doi.org/10.1787/9789264073234-en%0A
Pokropek, A. (2015). Phantom effects in multilevel compositional analysis: Problems and solutions. *Sociological Methods & Research, 44*(4), 677–705
R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.
Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 169*(4), 805–827
Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Danmarks Paedag-ogiske Institut.
Robitzsch, A. & Oberwimmer, K. (2015). BIFIEsurvey: Tools for survey statistics in educational assessment. *R package version*, pp. 1–2.
Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. Wiley.
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
Rutkowski, D., Rutkowski, L., Wild, J., & Burroughs, N. (2018). Poverty and educational achievement in the US: A less-biased estimate using PISA 2012 data. *Journal of Children and Poverty, 24*(1), 47–67
Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement, 48*(3), 293–312. https://doi.org/10.1111/j.1745-3984.2011.00144.x
Rutkowski, L. (2014). Applied measurement in education sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education, 27*(2), 115–132
Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170
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
Rutkowski, L., & Zhou, Y. (2015). The impact of missing and error-prone auxiliary information on sparse-matrix sub-population parameter estimates. *Methodology, 11*(3), 89–99. https://doi.org/10.1027/1614-2241/a000095
Sandoval-Hernandez, A., Rutkowski, D., Matta, T., & Miranda, D. (2019). Back to the drawing board: Can we compare background scales? *Revista de Educación, 383*, 37–61
Schneider, S. M. (2016). Income inequality and subjective wellbeing: Trends, challenges, and research directions. *Journal of Happiness Studies, 17*(4), 1719–1739
Thomas, V., Wang, Y., & Fan, X. (2001). *Meaasuring education inequality. Gini coefficients of education*. World Bank Publications.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450
Wilkinson, R. G., & Pickett, K. E. (2006). Income inequality and population health: A review and explanation of the evidence. *Social Science and Medicine, 62*(7), 1768–1784

Zheng, B. (2008). Measuring inequality with ordinal data: A note. *Research on Economic Inequality*. https://doi.org/10.1016/S1049-2585(08)16008-2

Zheng, B. (2011). A new approach to measure socioeconomic inequality in health. *Journal of Economic Inequality, 9*(4), 555–577

Zheng, X. & Yang, J. S. (2016). Using sample weights in item response data analysis under complex sample designs. In *Springer proceedings in mathematics and statistics* (pp. 123–137). Springer. https://doi.org/10.1007/978-3-319-38759-8_10.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.