Mittal *et al.*
Large-scale Assessments in Education (2022) 10:13
https://doi.org/10.1186/s40536-022-00132-w

Large-scale Assessments in Education

RESEARCH Open Access

Assessing the evidence for the comparability of socioeconomic status between students with and without immigrant background in Norway and Sweden

Oleksandra Mittal[1*], Ronny Scherer[2] and Trude Nilsen[1]

*Correspondence:
oleksandra.mittal@ils.uio.no

[1] Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Blindern, Postbox 1099, 0317 Oslo, Norway
[2] Centre for Educational Measurement at the University of Oslo, Faculty of Educational Sciences, University of Oslo, Oslo, Norway

Abstract

The prerequisite for meaningful comparisons of educational inequality indicators across immigration status is the comparability of socioeconomic status (SES) measures. The Programme for International Student Assessment (PISA) uses its index of economic, social, and cultural status (ESCS) to provide insights into the problems of inequality across students' socioeconomic and immigration statuses. However, missing evidence regarding the comparability of the ESCS index or its components across students with and without immigrant background challenges the accuracy of empirical inferences. Our study sheds light on the comparability of the index of household possessions (HOMEPOS) across immigration status in Norway and Sweden—two countries that continue to be two largest recipients of immigration flows among their Nordic neighbours. We tested the PISA 2018 HOMEPOS scale for the overall measurement invariance and possible differential item functioning (DIF) across three student groups with first-generation, second-generation, or no immigrant background. Several HOMEPOS items exhibited DIF within these countries. Moreover, we examined how four strategies to deal with DIF items may affect the inferences regarding educational inequalities across immigration status. The strength of the HOMEPOS–achievement association was sensitive to the choice of approach for 15-year-old immigrant students, while it remained stable and moderate for native students. Our findings encourage researchers using the HOMEPOS scale to consider the invariance testing to avoid measurement bias and provide robust evidence characterizing immigrant achievement gaps.

Keywords: Socioeconomic status, Household possessions, PISA, Measurement invariance, Differential item functioning, Immigrant background

Introduction

The empirical evidence from the Nordic countries suggests that immigrant students quite often have poor educational outcomes (e.g., Bakken & Elstad, 2012; Kilpi-Jakonen, 2014; Rangvid, 2007; Skolverket, 2016b). In PISA 2018, two Nordic countries with the largest immigrant populations—Norway and Sweden—were among the 11 participants where more than 45% of the 15-year-old immigrant students were found to be

socioeconomically disadvantaged, with a large achievement gap existing between them and native students after accounting for socioeconomic profiles (Organisation for Economic Co-operation and Development [OECD], 2019a). These findings are worrisome since the Nordic region has strong integration policies for immigrant, school-aged children and adults that primarily target those arriving on humanitarian grounds (Breidahl, 2017; Bunar, 2010; Hernes et al., 2019; OECD, 2019a; Skolverket, 2016a). Furthermore, the lower explanatory power of PISA's ESCS index—a multidimensional SES construct—for a large immigrant achievement gap calls for investigating how well this index captures the SES of immigrant students and whether it is comparable between students with and without immigrant background. Unless addressed, this knowledge gap will continue to challenge the validity of findings used to inform policymakers on facilitating solutions within the student populations and subpopulations.

Valid inferences about educational inequalities that exist among immigrant and non-immigrant populations require the comparability of their SES (Braveman et al., 2005; Hansson & Gustafsson, 2013; Lenkeit et al., 2015). In practice, an immigrant and a native student with equivalent SES scores should not differ in their actual SES, i.e., their SES scores should not be conditioned by the characteristics of a group they belong to. Since SES indicators reflect "the social standing" (American Psychological Association [APA], 2018) based on the society's value judgement, they are better adapted to capture economic, cultural, and social realities of the non-immigrant group rather than of the minority group with immigrant background (Lenkeit et al., 2015; Modood, 2012). PISA uses the ESCS index to operationalize SES through three common indicators: household possessions (HOMEPOS; a proxy for family wealth or income), parental education, and parental occupation (Willms & Tramonte, 2015). This index is an essential tool for comparisons between native and immigrant students in the OECD reports that inform policymakers about the status of educational inequalities for these groups (OECD, 2019a; Schleicher, 2006). Despite multiple reports and secondary analyses on immigrant students using the ESCS index, evidence about the comparability of its indicators across immigration status is missing. This may result in the misinterpretation of educational inequalities existing for immigrant students and limited nature of cross-immigration status comparisons.

One example of different social realities that an immigrant student may experience and that may not be captured by common SES indicators centred towards non-immigrant students is downward social mobility of immigrant student's parents (Modood, 2005). This factor may lead to a potential non-comparability between parents' social class and educational capital in the country of origin and occupational status in the country of destination (Modood, 2012). Furthermore, the educational levels in the country of origin and destination may not be equivalent challenging their non-biased estimation for immigrant parents (Dronkers et al., 2014; Hansson & Gustafsson, 2013). The household possessions, a proxy for "long-term economic well-being" (Hannum et al., 2017, p. 85), can be a reliable data source on family wealth that is not influenced by sudden changes in, e.g., occupation (Andersen et al., 2008; Currie et al., 2008; Wardle et al., 2002; Yang & Gustafsson, 2004). In PISA, this widely used construct is represented with 22 international and three country-specific items that may be a robust alternative to measuring SES across immigration statuses. PISA's household possessions (HOMEPOS) index,

however, is a broad indicator of economic status (Hannum et al., 2017), which may not truly capture family wealth across students with and without immigrant background within countries-participants. For example, HOMEPOS includes 'number of books at home' that was previously found to be potentially biased against immigrant student groups in another cross-country education survey—Trends in International Mathematics and Science Study (TIMSS) 2003 (Hansson & Gustafsson, 2013). Multiple factors may affect the item response patterns on a household possessions scale, e.g., culture, urban or rural place of residence, consumption preferences (Currie et al., 2008; May, 2006). These differences in an item ownership are natural as long as they do not become systematic, i.e., the item ownership largely reflects the belonging of a student to an immigrant group rather than the actual socioeconomic status. The more HOMEPOS items exhibit such trend, the less information on actual variability in family wealth may be derived and the lower may be the explanatory power of HOMEPOS for an immigrant achievement gap.

We approached the problem by investigating the comparability of the HOMEPOS scale across native and non-native students in Sweden and Norway. Of the three ESCS indicators, the HOMEPOS index has the strongest predictive power for reading achievement (Lee et al., 2019). The measurement and scaling procedures of HOMEPOS are continuously updated (Avvisati, 2020) with researchers focusing on the cross-country and cross-cycle comparability of HOMEPOS in the last decade (e.g., Lee & von Davier, 2020; Pokropek et al., 2017; Rolfe, 2021; Rutkowski & Rutkowski, 2013). We go a step further in the comparability analysis and unravel the complexities of HOMEPOS at the item level to understand whether each of the 22 international items works equally well for the native, first-generation, and second-generation immigrant students in Norway and Sweden. Our findings on the comparability of HOMEPOS items are then used to test how four approaches to handle non-comparable items (Cho et al., 2016; Liu & Rogers, 2021) influence the strength of HOMEPOS–reading achievement relationship for three student groups. This may guide future research in finding adequate SES measures to identify educational inequalities in diverse student subpopulations and circumvent measurement bias.

Theoretical background

Immigrant achievement gap, PISA, and immigration trend

The seventh cycle of the Programme for International Student Assessment (PISA) provided insights into the problems of inequality in reading literacy across students' socioeconomic and immigration statuses. Denmark, Finland, Iceland, Norway, and Sweden were among the 11 participating countries for which this problem was the most pronounced for immigrant students. In Norway and Sweden, the largest receivers of immigrant families, the score-point differences in reading performance associated with immigrant background (after accounting for gender and students' and schools' socioeconomic profiles) were higher than the OECD average (Table 1; OECD, 2019a). Furthermore, an immigrant achievement gap persists in many OECD countries (Andon et al., 2014), with immigrant students' low academic achievement usually explained by the low SES of their foreign-born parents (Ammermüller, 2007; Marks, 2005; Shapira, 2012). Nevertheless, several studies highlight a weaker association between SES and achievement for students with immigrant background compared to non-immigrant students

Table 1 The snapshot of immigrant students in Norway and Sweden

Country	Proportion of immigrant students, %	Proportion of disadvantaged immigrant students, %	Performance in reading			Gap in average reading performance between immigrant and non-immigrant students	Score-point difference in reading performance associated with immigrant background*
			Non-immigrant students	2nd-gen. immigrant students	1st-gen. immigrant students		
Norway	12.4	46.9	509	463	451	52 points	− 33
Sweden	20.5	45.6	525	471	410	82 points	− 54

Derived from: OECD, PISA 2018 Database, Tables II.B1.9.1 and II.B1.9.3; https://doi.org/10.1787/888934037051). OECD average: − 24

*After accounting for gender, and students' and schools' socioeconomic status

(Elmeroth, 2006; Kingdon & Cassen, 2010; Strand, 2014). The shortcomings of common SES indicators and their potential non-equivalence when capturing SES across the heterogeneous body of children and adolescents with and without immigrant background have been discussed elsewhere (Braveman et al., 2005; Fekjær, 2007; Modood, 2012; Rothon, 2007). However, few studies in general evaluated measurement invariance of SES across immigration status (Hansson & Gustafsson, 2013; Lenkeit et al., 2015), with no study addressing this problem with the PISA data.

Many secondary analyses of the PISA data use the ESCS indicators as control or predictor variables to investigate factors associated with achievement gaps between immigrant and native students (e.g., Areepattamannil et al., 2013; Gramaṭki, 2017; Martin et al., 2012; Marx et al., 2012; Schnepf, 2007). Additionally, the ESCS index has been central to the construct of academic resilience (Agasisti et al., 2017; Cerna et al., 2021; Cheung et al., 2014; Gabrielli et al., 2021; OECD, 2018). The unawareness of how equally well the ESCS index or its components capture the SES of native and non-native students may impair the validity of findings and the effectiveness of policy recommendations. For instance, the non-comparability of the HOMEPOS index, one of the three ESCS components, may prove not useful in locating and explaining immigrant achievement gaps within countries, potentially compromising a just distribution of educational resources among schools with larger and smaller shares of immigrant students, or, e.g., inhibiting appropriate school budget allocations that are driven by findings of educational inequalities existing across immigration status.

This study is thus relevant for the OECD countries due to the need to understand causes behind a persistent immigrant achievement gap (Andon et al., 2014), and people's increasing global mobility (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2018) which may challenge the validity of using identical SES measures across immigration status to capture educational inequalities. It is further relevant for the Nordic countries in the face of the refugee crisis of 2015, with Sweden and Norway having received the largest proportion of asylum seekers in the Nordic region (Byström & Frohnert, 2017; Hagelund, 2020). By the end of 2015, Sweden had registered approximately 163,000 refugees (Adan & Antara, 2018), whereas Norway had accepted about 31,000 (Parveen, 2020). These are substantial numbers considering that in 2015 Sweden's population was 9.8 million and Norway's 5.2 million. The latest refugee crisis caused by Russia's invasion of Ukraine, with 5 million Ukrainians registered in Europe

at the time of writing (United Nations High Commissioner for Refugees [UNHCR], 2022), suggests that the comparability testing of SES indicators across immigration status should be of a systematic nature. Our study is, hence, an attempt to facilitate such investigations which may in turn improve our understanding of challenges and successes that immigrant students experience in schools.

Comparability of the PISA SES measures across immigrant status

Meaningful group comparisons are prerequisites for the validity of findings in cross-cultural studies (e.g., Rutkowski & Svetina, 2014; Van de Vijver, 2018). Such comparisons are valid if sufficient evidence that a scale and its items operate in the same way across populations exists (Bauer, 2017). For example, if we are to compare students' SES across different immigration statuses, the first step is to test for the equivalence or measurement invariance (MI) of this construct. We want to make sure that students' item responses are dependent solely on the level of SES they have and not on the effects of a group they belong to. The measurement invariance (MI) of PISA's ESCS index is therefore of great interest because it is one of the student background characteristics used to derive estimates of student achievement (Rutkowski et al., 2014; von Davier et al., 2009). Hence, a systematic lack of invariance of the ESCS index, its subscales, or items across, for instance, students with or without an immigrant background may bias the proficiency scores and thus the subsequent policy decisions (Rutkowski & Rutkowski, 2010). Furthermore, MI is a question of fairness and equity (Meredith, 1993). If the differences in the ESCS index, subscales, or items depend on certain students' characteristics and not on the differences in the students' level of SES, then the measure is biased against one group of students (Bauer, 2017; He & Van de Vijver, 2013).

To our best knowledge, two empirical studies have investigated the invariance of SES measures across immigration status (Hansson & Gustafsson, 2013; Lenkeit et al., 2015). For instance, using TIMSS 2003 data, Hansson and Gustafsson (2013) operationalised SES by the mother's and father's educational level, the number of books at home, and the student's study aspirations. They concluded that the reflective latent variable SES had the same meaning across the eighth-grade students with Swedish and non-Swedish backgrounds. Conversely, large group differences in the probability of endorsing 'number of books at home' and 'mother's education' item categories indicated a potential bias against first- and second-generation immigrant students. The authors further suggested testing the comparability of family income and parental occupation 'to obtain a valid measurement model' of SES for the diverse groups of students (Hansson & Gustafsson, 2013, p. 163). The second known study used data from the Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) in England and found that SES measures were 'not equivalent representations of family SES across different groups' with and without immigrant background (Lenkeit et al., 2015, p. 77). The authors warned researchers to compare SES and its associations with educational attainment across immigration status with caution.

Since PISA 2015, the scaling procedures for HOMEPOS, one of the three ESCS subscales, have partially addressed the problem of cross-cultural comparability (OECD, 2017). These procedures included the performance of country-by-language invariance testing for the countries that administered the PISA test in more than two languages

and with the weighted sample size of each language group above 300 (OECD, PISA 2018 Technical Report, Chapter 16, https://www.oecd.org/pisa/data/pisa2018technicalreport/). The invariance tests were conducted across language groups for the Norwegian student sample (language groups: Nynorsk and Bokmål), yet not the Swedish one. However, the two language groups in Norway do not represent the immigrant backgrounds in the Norwegian student population. Evidence on the comparability of the HOMEPOS index across native and non-native student subpopulations in Norway and Sweden is still lacking.

PISA's household possessions index and its comparability

The household possessions-based aspect of SES is a reliable data source on family wealth that is less prone to error due to a high parent-student agreement and response rate (e.g., Andersen et al., 2008). This asset index captures "the presence or absence of various consumer durable goods and home construction features in their [students'] primary dwelling" (Traynor & Raykov, 2013, p. 664). This aspect may be particularly important for students with immigrant background since their parents' occupational status may not be indicative of their educational level in the country of origin (Lenkeit et al., 2015). Moreover, the household possessions may capture contributions to the family income by older employed sons and daughters of, e.g., South Asian immigrant background (Basit, 1997). In PISA, the HOMEPOS index, one of the three indicators used in the computation of the ESCS index, represents such aspect of SES that is also known to be a strong predictor of academic achievement (Lee et al., 2019). The measurement and scaling procedures of HOMEPOS have been continuously updated, and there was a low percentage of missing item responses, compared to other ESCS components (Avvisati, 2020). Previously, researchers argued that individual SES components are stronger predictors of inequalities than the composite SES indices (Watermann et al., 2016; White, 1982). As a subscale of ESCS, the HOMEPOS index may hence reveal the complex nature of equity mechanisms better than the composite ESCS index that keeps implicit the equity profiles of the countries (Keskpaik & Rocher, 2011). Besides, HOMEPOS is the only ESCS scale that manifests the trends within social, technical, and economic contexts across participating countries (OECD, 2017). Despite that, the challenge of PISA to keep the HOMEPOS items comparable across countries remains with several studies having established full or partial cross-cultural non-invariance of the construct (e.g., Lee & von Davier, 2020; Pokropek et al., 2017; Rutkowski & Rutkowski, 2013). This may happen because possessions do not have the same meaning across developed and developing countries (e.g., Kim et al., 2019). For example, Keskpaik and Rocher (2011) used PISA 2009 data and concluded that most HOMEPOS items were strong predictors of achievement for many non-OECD countries, whereas only two items regarding the availability of 'books of poetry' and 'classic literature' had a higher-than-average correlation with achievement across OECD countries. This finding points to possible cross-cultural differences or construct biases (e.g., He et al., 2019). Rapidly changing immigration patterns with at least one out of five 15-year-old students in the OECD having an immigrant background (UNESCO, 2018) may further hamper the detection of inequalities across immigrant populations within OECD countries. Despite or possibly due to this trend, the household possessions scale may still have a great potential to detect inequalities within and across

Table 2 Sample characteristics

Country	Age (yrs)	Gender (%)		Immigration status			Sample size	
	M (*SD*)	Boys	Girls	Native	Second-generation	First-generation	Total	Missing
Norway	15.8 (0.29)	50.5	49.5	4883	347	345	5575	236
Sweden	15.7 (0.28)	49.8	50.2	4244	581	511	5336	166

The samples are based on the available HOMEPOS data

diverse immigrant student sub-populations, compared to the indicators of occupational and educational status that may not be equivalent across immigration statuses (Lenkeit et al., 2015; Modood, 2012; Rothon, 2007). However, this potential can be tapped only when we learn to fully utilize the scale. By full utilization we mean examining the scale comparability at the item-level across the groups of interest prior to any further analysis. This is especially essential when the aim is to detect inequity or inequalities in schools.

The present study

The present study examines the comparability of the PISA 2018 HOMEPOS scale—an indicator of SES—across immigration status in Norway and Sweden. Specifically, we evaluate (a) the overall measurement invariance of the HOMEPOS scale; (b) the differential functioning of the HOMEPOS items; and (c) the relationship between HOMEPOS and reading achievement across immigration status. We further provide recommendations for the use of the HOMEPOS scale when comparisons across immigration status are of interest. Specifically, our study addresses the following research questions (RQs):

RQ1 To what extent does the measure of students' HOMEPOS demonstrate overall invariance across three student groups, namely native students and first- and second-generation immigrant students, in Norway and Sweden? *(Comparability of the overall scale)*

RQ2 To what extent do the 22 items of the HOMEPOS scale exhibit DIF across immigration status? *(Comparability of specific items)*

RQ3 To what extent are HOMEPOS and reading achievement related across immigration status, and does the strength of this relationship depend on the strategy of handling the possible non-comparability of the HOMEPOS items? (*Relations to reading achievement*)

Addressing the last question, we examine the following strategies: (a) Ignoring the existence of non-comparable items; (b) deleting non-comparable items; (c) deleting only non-uniform DIF items; and (d) accounting for non-comparable items in the HOMEPOS measurement model.

Methods

Sample

The present study draws on the PISA 2018 data from nationally representative samples of 15-year-old students in Norway and Sweden (see Table 2). PISA 2018 followed a two-stage stratified sampling design with a sample of 35 or 42 students per sampled school

Table 3 Stratification variables used for Norway and Sweden

Country	Explicit stratification variables	Number explicit strata	Implicit stratification variables
Norway	School level (2)	2	None
Sweden	Funding (2); ISCED level (2); Urbanisation for lower secondary (3)	8	Geographic LAN—for upper secondary (21); Responsible authority—for upper secondary (3); Level of immigrants (3); Income Quartiles—for lower secondary/mixed (4)

PISA 2018 Technical Report, Chapter 4, https://www.oecd.org/pisa/data/pisa2018technicalreport/

(OECD, PISA 2018 Technical Report, Chapter 16, https://www.oecd.org/pisa/data/pisa2018technicalreport/). Each country had a unique list of stratification variables that indicated school characteristics and was used to aggregate schools into mutually exclusive groups prior to the school sampling (see Table 3). This information is essential for understanding differences in the mechanism of sampling students with an immigrant background. Given these unique country features, we report the *within*-country findings.

Measures

Household possessions scale

The HOMEPOS scale is one of the three components of the PISA 2018 Index of ESCS. It captures four aspects of family wealth, cultural possessions, home educational resources, information and communication technology (ICT) resources, and the number of books at home. The scale includes 22 indicators common across the participating countries and economies and up to three country-specific items. In Norway, the three national indicators on the availability of tablet computers, smart telephones, and e-book readers were added; in Sweden, students indicated the availability of a piano, cleaning services, and an espresso machine (OECD, PISA 2018 Technical Report, Annex E, https://www.oecd.org/pisa/data/pisa2018technicalreport/). In our study, we used the 22 international indicators that the PISA team used to compute the HOMEPOS index. By using only international indicators we aimed at showing the same phenomenon for the two countries albeit we are not comparing them. The corresponding internal consistencies were $\alpha = 0.76$ for Norway and $\alpha = 0.75$ for Sweden (OECD, PISA 2018 Technical Report, Chapter 16, https://www.oecd.org/pisa/data/pisa2018technicalreport/).

Of the 22 items, 13 were scored dichotomously (0 = *no*, 1 = *yes*) and indicated the possession of 'a desk', 'a room of one's own', 'a quiet place to study', 'a computer one can use for school work', 'educational software', 'a link to the Internet', 'classic literature', 'books of poetry', 'works of art', 'books to help with school work', 'technical reference books', 'a dictionary', and 'books on art, music or design'. Eight polytomous items indicated the number of 'televisions', 'cars', 'rooms with a bath or shower', 'cell phones with Internet access', 'computers', 'tablet computers', 'e-book readers', and 'musical instruments' (0 = *none*, 1 = *one*, 2 = *two*, 3 = *three or more*). The item 'books' had six categories: 0 to 10, 11 to 25, 26 to 100, 101 to 200, 201 to 500, and more than 500 books.

To illustrate the properties of items composing the HOMEPOS scale, we provide item parameter estimates and response distributions for the three immigration status groups in Norway and Sweden in Additional file 1: Appendix A.

Immigration status

We used the index of immigrant background (IMMIG) provided in the PISA 2018 dataset to indicate the three groups of students within each country. This index distinguishes between (a) native students (i.e., students with at least one parent born in the country of assessment), (b) second-generation immigrant students (i.e., students born in the country of assessment with both parent(s) born in another country), and (c) first-generation immigrant students (i.e., they and their parents were born outside the country of assessment; see Table 2; OECD, PISA 2018 Technical Report, Chapter 16, https://www.oecd.org/pisa/data/pisa2018technicalreport/). In our analyses, we refer to these categories as 'native', '2ndGEN', and '1stGEN' students, respectively.

Reading achievement

Reading literacy was the focal domain in PISA 2018 and was defined as 'understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society' (OECD, 2019b, p. 28). This concept involves cognitive and metacognitive processes of navigating the plural realm of reading, effectively synthesising and integrating information from multiple sources, and being 'active, purposeful, and functional' in one's application of reading strategies in any given life scenario (see OECD, 2019b, p. 28). Three major components of reading literacy were further defined: texts (classified according to their source, structure, format, and type), cognitive processes (i.e., locating information, understanding, evaluating, reflecting, and reading fluently), and scenarios (see OECD, 2019b).

The computer-based reading literacy assessment contained 245 items (45 units) that were delivered to students in three adaptive stages. The response formats included selected-response, short-constructed, and open-response items. Each student responded to 33 to 40 items in 7 units within 60 min. Sixty-five reading fluency items were administered prior to the main test to better capture students' reading proficiency at the lower level of achievement. The multistage adaptive testing design was a new feature used for the reading domain in PISA 2018. Test reliabilities were 0.94 for both Norway and Sweden. The proficiency distribution in reading literacy is represented by 10 plausible values that account for the measurement uncertainty and ensure reliable achievement estimates in the population. In our analyses, we used all 10 plausible values (Rutkowski et al., 2010).

Analytic strategy

Testing for measurement invariance and differential item functioning

The scaling procedures for the HOMEPOS items were based on the two-parameter logistic model (2PLM) for dichotomously scored responses and the generalised partial credit model (GPCM) for polytomous responses (OECD, 2017). Both models belong to the item response theory tradition of estimating the item response probability as a nonlinear relationship between categorical item responses and the latent trait *theta*, with the probability bounded between 0 and 1 (De Ayala, 2009). The 2PLM describes the probability that a student v responds in category 1 (e.g., checking the

specific home possession) to an item i as a function of the student's trait level θ_v, the item difficulty b_i, and the item discrimination a_i (with a scaling constant $D = 1.7$; e.g., Desjardins & Bulut, 2018):

$$P(X_{vi} = 1|\theta_v, b_i, a_i) = \frac{\exp[Da_i(\theta_v - b_i)]}{1 + \exp[Da_i(\theta_v - b_i)]}$$

This model extends the popular Rasch 1PL model by relaxing the equality constraint on the item discriminations a_i, that is, allowing for item-specific relations between the item and the latent trait. In polytomously scored items, students can respond in several categories $k = 0, \ldots, m_i$. The GPCM describes the probability of responding in category k as a function of the student's trait level θ_v, the item difficulty b_i, the item discrimination a_i, and the item threshold parameters d_i between categories (with a scaling constant $D = 1.7$ and a zero sum of all threshold parameters for each item; see the OECD's PISA 2018 Technical Report, Chapter 9, https://www.oecd.org/pisa/data/pisa2018technicalreport/):

$$P(X_{vi} = k|\theta_v, b_i, a_i, d_i) = \frac{\exp\left[\sum_{r=0}^{k} Da_i(\theta_v - b_i + d_{ir})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{r=0}^{h} Da_i(\theta_v - b_i + d_{ir})\right]}$$

Similar to the 2PLM, this model allows for item-specific discriminations and is therefore more flexible than the PCM, in which these parameters are equal across items. In the present study, we adhered to the PISA procedure and implemented these models as reflective measurement models of HOMEPOS.

We took two approaches to test the equivalence of the HOMEPOS scale across immigration status: MI testing via multigroup item response theory (MG-IRT) modelling and testing for item-specific differential item functioning (DIF; see Bauer, 2017; Millsap, 2011). MG-IRT invariance testing allows for testing the scale's overall comparability ('scale functioning') and has limited sensitivity to identify non-invariant items; conversely, DIF testing identifies such non-invariant items (Bauer, 2017). Both approaches were implemented with the IRT treatment of the HOMEPOS scale in the framework of confirmatory factor analysis (CFA) using M*plus* (Muthén & Muthén, 1998–2017). Several researchers compared the IRT- and CFA-based MI testing and DIF detection (Kim & Yoon, 2011; Stark et al., 2006), and proposed an integrated IRT- and CFA-based approach (see, for instance, Dimitrov, 2017). Please find the respective M*plus* input files in the Additional file 4: Appendix D.

Multigroup Item Response Theory Invariance Testing We estimated and compared three MG-IRT invariance models: the configural, metric, and scalar invariance models (Millsap, 2011). The *configural* invariance model estimates the cross-group equivalence of the setup of the factor structure, assuming the same number of factors and item-factor patterns yet freely estimating model parameters. The *metric or weak* invariance model constrains the item discriminations across groups. This model establishes that the relationships between the latent variable and the manifest item responses are the same. Deviations from metric invariance indicate the presence of non-uniform DIF items. The test for *scalar or strong* invariance is a prerequisite for factor mean comparisons across groups. It builds upon metric invariance and constrains the item difficulties/thresholds

to be equal across groups (Bialosiewicz et al., 2013). The absence of scalar invariance indicates the presence of uniform DIF items. As a final step, the metric and scalar invariance models are compared to the configural model via likelihood-ratio tests, differences in information criteria, or other fit indices to examine the extent to which additional model constraints may deteriorate the model fit (Putnick & Bornstein, 2016).

Multiple-Indicators-Multiple-Causes Differential Item Functioning Testing With their tests of global model fit, multi-group models have a low sensitivity to detect item-specific DIF across groups (Bauer, 2017). DIF occurs when the probability of endorsing an item varies for respondents with the same amount of latent trait depending on the group to which respondents belong (Stark et al., 2006). Two types of measurement non-invariance can be identified at the item level: uniform and non-uniform DIF (De Ayala, 2009). *Uniform DIF* is associated with group differences in item difficulties/thresholds (Stark et al., 2006). It occurs when the probability of answering an item correctly or selecting a higher response category is different for one subgroup over the entire range of its latent trait (Fig. 1a; Woods, 2009). *Non-uniform DIF* is associated with situations in which item discriminations (factor loadings) and possibly item difficulties differ between groups. With regard to the HOMEPOS scale, this means that, for instance, the probability of endorsing the item 'books of poetry' may be equal across the subgroups with a HOMEPOS score of '0' on the latent continuum but may be systematically higher or lower for one subgroup with a HOMEPOS score of '1' (Fig. 1b). Hence, an expected item response is a function of both group membership and the level of the HOMEPOS latent trait.

In the present study, we tested for uniform and non-uniform DIF via Multiple-Indicators-Multiple-Causes (MIMIC)-DIF modelling (e.g., Chun et al., 2016; Woods et al., 2009). MIMIC-DIF models introduce the grouping variable as an endogenous variable to the measurement model (Bauer, 2017). To test for DIF with the MIMIC approach, we implemented the *constrained baseline* method. This method begins by estimating a baseline model in which the two dummy-coded grouping variables 2ndGen and 1stGen are related only to the latent variable HOMEPOS; all other possible effects on items are constrained to 0 (Chun et al., 2016; see Fig. 1a). The following steps include tests for uniform and non-uniform DIF. To detect *uniform DIF*, the baseline model is extended by two paths connecting the grouping variables with an individual item ($\gamma_i^{b_1}$ and $\gamma_i^{b_2}$ for an item i; see Fig. 1b). If the model fit improves significantly relative to the baseline, then the item is flagged with uniform DIF. This procedure is then repeated for all other items. To test for *non-uniform DIF*, we added two variables that represented the interactions between the latent variable HOMEPOS and the two grouping variables (specified via the 'XWITH' command in Mplus; see Fig. 1c). In the subsequent testing, both the grouping and interaction variables were regressed on one item at a time, with the latter reflecting the moderating effect on the latent variable HOMEPOS (paths $\gamma_i^{a_1}$ and $\gamma_i^{a_2}$ for an item i; see Chun et al., 2016). We compared models assuming non-uniform DIF to the corresponding uniform DIF models to see potential between-group differences in item discriminations in addition to item difficulties. The constrained baseline approach implements an all-other-item method in which all other items except the one studied are constrained to have equal parameters across groups and are assumed to be DIF-free (Wang et al., 2009). Given that

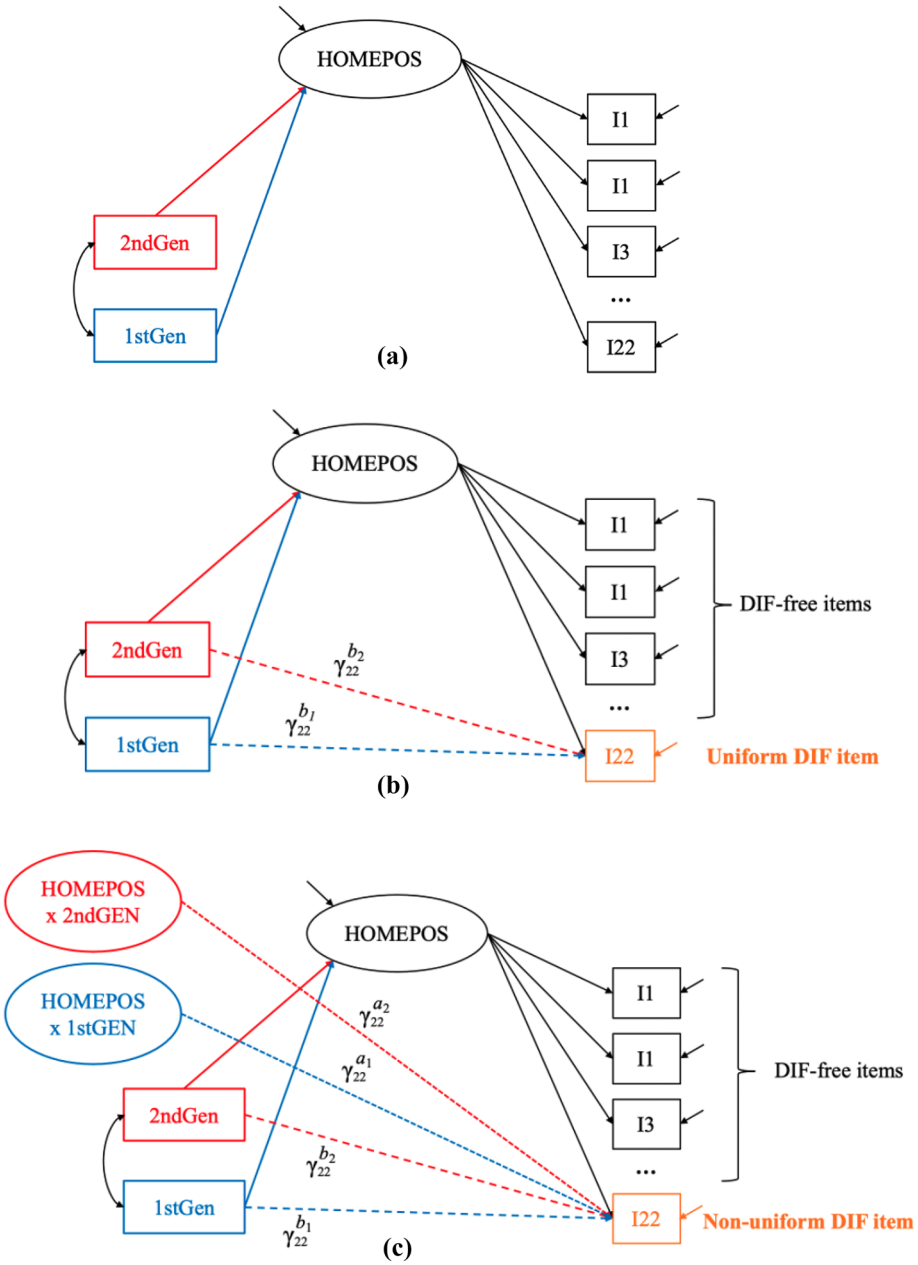

Fig. 1 Constrained baseline approach for DIF detection: **a** the baseline model with two covariates, 2ndGEN and 1stGEN (second-generation and first-generation immigrant student group variables); **b** the augmented model to test for uniform (threshold) DIF in the HOMEPOS items; **c** the augmented model to test for non-uniform (loading and threshold) DIF with two additional variables "HOMEPOS × 2ndGEN" and "HOMEPOS × 1stGEN" that represent the interactions between the latent variable HOMEPOS and two covariates

this approach included multiple significance tests, we adjusted the *p*-values using the Benjamini–Hochberg procedure for the 1% significance level. As part of our sensitivity analyses, we present the results obtained from an alternative method, namely the

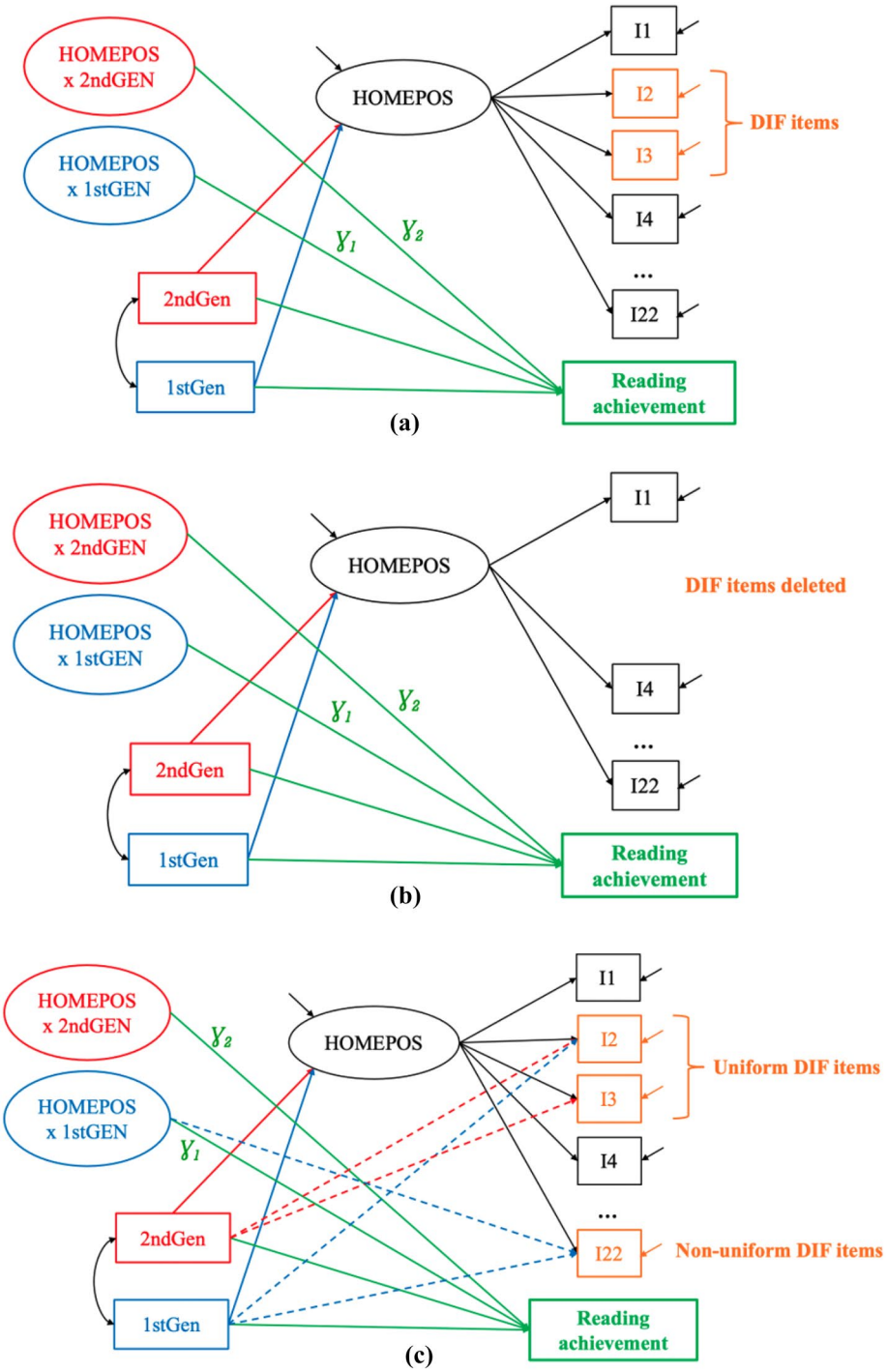

Fig. 2 DIF treatment approaches when quantifying the HOMEPOS–reading achievement relationship: **a** ignoring DIF; **b** deleting DIF items; **c** accounting for DIF. In this example, item *I2* exhibited uniform DIF across 2ndGEN, 1stGEN and the reference groups; item *I3* exhibited uniform DIF only between 2ndGEN and the reference group; item *I22* exhibited non-uniform DIF between 1stGEN and the reference group. $\gamma_{1,2}$—the moderation effect of "HOMEPOS × 2ndGEN/1stGEN". This allows us to report the HOMEPOS–reading achievement relationship for each group

sequential-free baseline method, in Additional file 3: Appendix C (for details, please refer to Chun et al., 2016).

Quantifying the household possessions–achievement relation via different strategies

To address the problem of DIF items in the HOMEPOS measurement model, we examined how four different approaches of treating DIF items affected the relation between HOMEPOS and reading achievement across immigration status (Fig. 2):

Ignoring DIF In this approach, the latent variable HOMEPOS was represented by all items, irrespective of the evidence on DIF. Ignoring DIF can be a feasible approach when the interpretation is made on the population level (Cho et al., 2016). However, the parameter estimates may not be accurate if many items show DIF (Liu & Rogers, 2021).

Deleting DIF items In this approach, the latent variable HOMEPOS is represented only by items that exhibited neither uniform nor non-uniform DIF. In a recent simulation study by Liu and Rogers (2021), this strategy resulted in the largest average standard error and performed the worst under most conditions. Deleting DIF items may also reduce scale reliability and content validity due to the loss of information (Liu & Rogers, 2021).

Deleting non-uniform DIF items In this approach, only non-uniform DIF items are deleted from the HOMEPOS scale. This ensures the validity of group comparisons of the HOMEPOS–reading achievement relation, because non-uniform DIF items have different item discriminations (factor loadings) across groups.

Accounting for DIF In this approach, we accounted for uniform and non-uniform DIF items by allowing that the corresponding item parameters could vary between the reference and focal groups, with other items constrained to be equal across groups.

Analytic setup

The PISA 2018 data have a clustered structure with students nested in schools, which may have been purposefully over or under sampled in a specific region and may vary in size and non-response rates. This leads to unequal selection probabilities of students. To minimise this potential source of bias, we incorporated the final student weight (W_FSTUWT) in our analyses (OECD, 2017).

Addressing RQ3, we included the 10 plausible values by estimating each model with achievement 10 times and combining the resultant model parameters via Rubin's combination rules (Rutkowski et al., 2010). This procedure can be accessed in the software *Mplus* via the TYPE = IMPUTATION command. We performed all analyses using the software *Mplus* 8.5 (see Additional file 4: Appendix D for the inputs). All models were based on maximum-likelihood estimation with robust standard errors (MLR estimator) with a built-in expectation–maximization algorithm to handle missing data.

Results

Testing for the measurement invariance of the overall scale (RQ1)

Prior to testing for the invariance of the overall HOMEPOS scale, we fit the single-factor GPCM for HOMEPOS to the data of the total student samples and in each of the three subsamples. Next, we estimated the configural invariance models for each country as baseline models (see Tables 4 and 5).

Table 4 Likelihood-ratio tests and information criteria of the multi-group invariance models for Norway

Model	LL	SCF	Npar	AIC	BIC	aBIC	Model comparisons		
							cLRT	Δ Npar	*p*
Model 1. Configural	−74,628	1.099	194	149,644	150,929	150,313	–	–	–
Model 2. Metric	−74,694	1.061	152	149,692	150,699	150,216	107.2	42	< .001
Model 3. Scalar	−81,155	1.229	56	162,421	162,792	162,615	12,483.8	138	< .001
Measurement model for the total sample ($N = 5575$)	−73,860	1.106	64	147,849	148,273	148,070	–	–	–
Measurement model for the native sample ($n = 4883$)	−62,309	1.113	64	124,747	125,162	124,959	–	–	–
Measurement model for the 2ndGEN sample ($n = 347$)	−4706	1.094	64	9541	9787	9584	–	–	–
Measurement model for the 1stGEN sample ($n = 345$)	−5047	1.101	64	10,223	10,469	10,266	–	–	–

LL: Log-likelihood value; Npar: Number of free parameters; SCF: Scaling correction factor; AIC: Akaike's information criterion; BIC: Bayesian information criterion, aBIC: sample size adjusted BIC; cLRT: corrected Likelihood-ratio test statistic

Table 5 Likelihood-ratio tests and information criteria of the multi-group invariance models for Sweden

Model	LL	SCF	Npar	AIC	BIC	aBIC	Model comparisons		
							cLRT	Δ Npar	*p*
Model 1. Configural	−78,166	1.158	194	156,720	157,997	157,380	–	–	–
Model 2. Metric	−78,345	1.130	152	156,994	157,994	157,511	284.6	42	< .001
Model 3. Scalar	−85,243	1.237	56	170,597	170,966	170,788	12,568.9	138	< .001
Measurement model for the total sample ($N = 5336$)	−77,659	1.145	64	155,446	155,868	155,665	–	–	–
Measurement model for the native sample ($n = 4244$)	−59,277	1.142	64	118,683	119,090	118,886	–	–	–
Measurement model for the 2ndGEN sample ($n = 581$)	−7947	1.115	63	16,021	16,293	16,093	–	–	–
Measurement model for the 1stGEN sample ($n = 511$)	−7457	1.166	64	15,042	15,311	15,108	–	–	–

LL: Log-likelihood value; Npar: Number of free parameters; SCF: Scaling correction factor; AIC: Akaike's information criterion; BIC: Bayesian information criterion; aBIC: sample size adjusted BIC; cLRT: corrected Likelihood-ratio test statistic

Constraining further the item discriminations across groups in Norway (metric invariance) resulted in a significant loss of model fit, $\Delta\chi^2(42) = 107.2$, $p < 0.001$. Conversely, the BIC and aBIC information criteria parameters indicated an improvement in the model fit while the AIC parameter suggested a deteriorated fit (see Table 4). This potentially indicates a partial metric invariance which could be checked with further item-level DIF testing to identify items with non-uniform DIF. Similarly, the scalar invariance model indicated a significantly deteriorated fit compared to the configural model, $\Delta\chi^2(138) = 12{,}483.8$, $p < 0.001$. All three information criteria parameters also suggested a

deteriorated model fit. Hence, we did not have evidence for scalar invariance of the overall HOMEPOS scale across immigration status.

For the Swedish data, the metric invariance model fit significantly worse than the baseline model, $\Delta\chi^2(42) = 284.6$, $p < 0.001$. This result was supported by the AIC and aBIC information criteria parameters and contradicted by the BIC parameter that indicated a minor improvement in the model fit. Hence, the poor metric invariance could set the stage for further item-level DIF detection to flag items with non-uniform DIF. Analogous to the metric model, constraining the thresholds (scalar invariance) resulted in a substantial loss of model fit, $\Delta\chi^2(138) = 12{,}568.9$, $p < 0.001$. This time all three information criteria parameters indicated a deteriorated model fit compared to the configural model. Consequently, we proceeded with identifying potential DIF items in the scale.

DIF testing (RQ2)

Uniform DIF

As noted earlier, we compared the baseline model for uniform DIF detection to models with direct paths from the two grouping variables 2ndGEN and 1stGEN to one item. Significant likelihood-ratio tests and uniform DIF effects for either of the focal groups would point to significant between-group differences in item thresholds and hence the presence of uniform DIF. Possible negative values of uniform DIF effects on certain items indicate that the reference native group had a higher expected score for those items after controlling for the level of the HOMEPOS latent trait. Conversely, positive values indicated that 2ndGEN, 1stGEN, or both focal groups had a higher probability of endorsing the items flagged for significant uniform DIF effects. In both cases, the difference in item response probabilities is assumed constant over the entire latent continuum.

For the Norwegian data, 14 HOMEPOS items demonstrated uniform DIF (Additional file 2: Appendix B: Table B1). Seven of these items had a significant difference in item thresholds in favour of the reference group and seven items in favour of the focal groups. Both 2ndGEN and 1stGEN reported on the availability of 'books to help with school work', 'a dictionary', and 'e-book readers' at a significantly higher frequency than the reference group did. Having the same HOMEPOS score, first-generation immigrant students consistently reported more often than native students that they have 'classic literature', 'books of poetry', and 'books on art, music or design'. Furthermore, it was more likely for 2ndGEN students than native students to answer that they have their own desk. Conversely, two focal groups with the same amount of the HOMEPOS latent trait as the reference group were less likely to endorse the items indicating the availability of 'a room of one's own' and the number of 'televisions', 'cars', 'musical instruments' and 'books' at home. In addition, the 1stGEN group had a significantly lower probability than the native group of endorsing the items regarding the number of 'rooms with a bath or shower' or 'tablet computers'.

In Sweden, 19 HOMEPOS items exhibited uniform DIF (Appendix B: Table B3); 10 of these were biased towards the two focal groups, and nine indicated that these groups had a significantly higher probability of endorsing the items after controlling for the HOMEPOS score. Both 2ndGEN and 1stGEN groups responded positively regarding the availability of 'books to help with school work', 'technical reference books', 'a dictionary', 'educational software', 'desk', 'books of poetry', and 'e-book readers' at a significantly

higher frequency rate. Additionally, the 1stGEN group had a significantly higher response probability for the 'books on art, music or design' item, and the 2ndGEN group had a consistently higher expected value on the item indicating the availability of 'a computer one can use for school work' than the native group did. The pattern of uniform DIF showed that the focal groups endorsed the items indicating educational resources and cultural possessions significantly more often than the reference group with the same level on the HOMEPOS trait did. Counter to that, the native group had a significantly higher expected value on eight items indicating family wealth, one cultural possession item ('musical instruments'), and the number of books at home item.

Non-uniform DIF

To test for non-uniform DIF, we compared the models with interaction effects to the corresponding uniform DIF models. Significant likelihood-ratio test statistics and interaction effects of *HOMEPOS x 2ndGEN* or *HOMEPOS x 1stGEN* would indicate the presence of non-uniform DIF. A positive interaction effect indicates that an item is less discriminating for the reference group. Different item discrimination parameters for each group imply that the between-group difference in endorsing an item is not constant over the latent continuum.

For the Norwegian data, only one item (i.e., the availability of classic literature at home) exhibited non-uniform DIF between the 1stGEN and native groups (Additional file 2: Appendix B: Table B2). A significant negative interaction effect indicated that first-generation immigrant students who are average on the HOMEPOS latent trait were more likely to endorse the item than the native students were. For the Swedish data, eight items were flagged for the differences in discrimination parameters (Additional file 2: Appendix B: Table B4), two of which (i.e., number of 'televisions' and 'e-book readers') exhibited non-uniform DIF between the reference and both focal groups. The other six items had significant differences in their ability to discriminate between the native and 1stGEN groups. The first-generation immigrant students with the average level of the HOMEPOS latent trait were more likely to endorse the items on the availability of 'books of poetry', 'books to help with school work', and 'a dictionary'. The native students who were average on the HOMEPOS latent trait were more likely to select a higher category for the items regarding the number of 'televisions', 'cars', 'tablet computers', 'e-book readers', and 'books'.

Relations to reading achievement (RQ3)

To address RQ3, we investigated how four approaches to handle DIF items influenced the strength of the relationship between reading achievement and HOMEPOS across immigration status. This influence was compared across groups within each approach and across approaches for each group separately (see Table 6; Fig. 3). We conducted pairwise significance testing using the slope, standard error, and sample size.

For the native student subpopulation in Norway, the second approach of deleting 14 DIF items was distinct from others when comparing it both across the groups and across the approaches within one group. For example, compared to ignoring DIF in the regression analysis, deleting 14 DIF items (see Fig. 4a; Additional file 2: Appendix B: Table B1—for DIF item names) increased the strength of the relationship between HOMEPOS and

Table 6 Regression coefficients reflecting the relationship between HOMEPOS and reading achievement across approaches and groups

	Approach 1: Ignore DIF β (*SE*)	Approach 2: Delete DIF items β (*SE*)	Approach 3: Delete non-uniform DIF items β (*SE*)	Approach 4: Account for DIF β (*SE*)
Norway				
Native	0.253 (0.021)	0.239 (0.027)	0.235 (0.023)	0.257 (0.021)
2ndGen	0.227 (0.073)	0.378 (0.075)	0.243 (0.074)	0.213 (0.069)
1stGen	0.313 (0.067)	0.384 (0.074)	0.310 (0.069)	0.303 (0.070)
Sweden				
Native	0.319 (0.021)	0.369 (0.022)	0.322 (0.022)	0.325 (0.021)
2ndGen	0.246 (0.059)	0.259 (0.077)	0.256 (0.066)	0.209 (0.058)
1stGen	0.259 (0.069)	0.247 (0.141)[#]	0.342 (0.077)	0.251 (0.069)

[#] The regression coefficient of the 1stGen group was found to be significantly different from that of the native group within this approach ($p < .05$)

reading achievement for 2ndGEN by 0.151 and for 1stGEN by 0.071 points. None of the four DIF treatment approaches made a significant difference for the strength of the relationship in the native group.

For the Swedish data, the correlations were stronger for the native group than for the 2ndGEN and 1stGEN groups across the four approaches except for the 1stGEN group in Approach 3. When eight non-uniform items were deleted (see Additional file 2: Appendix B: Table B4 for DIF item names), the relationship between HOMEPOS and achievement for 1stGEN increased by 0.083 compared to the 'ignore DIF' approach. In Approach 2, we had only three comparable items that we used for the HOMEPOS latent variable (StuPlace, ClassLit, ArtWorks; see Fig. 4b; Additional file 2: Appendix B: Table B3). This did not change the correlations compared to other approaches for all the groups. Conversely, the correlation slightly increased for the native group that differed significantly from the 1stGEN group within the approach.

Discussion

Measuring household possessions across immigration status

Previous research focused on the comparability of the HOMEPOS *scale* across countries and cycles (e.g., Lee & von Davier, 2020; Pokropek et al., 2017). Our study took a step further and unfolded the (non-)comparability of the HOMEPOS scale and its consequences across immigration status within Norway and Sweden in three steps (see Additional file 4: Appendix D for M*plus* inputs).

First, we examined the overall invariance of the HOMEPOS measurement model scaled according to the PISA procedure (OECD, 2017) and could not find support for full metric invariance across immigration status within the Norwegian and Swedish PISA samples. Similar challenges were identified in earlier studies (e.g., Rutkowski & Rutkowski, 2013, 2018; Sandoval-Hernandez et al., 2019). This finding may imply (1) a potential difference in the sociocultural value for certain items for students with and without immigrant background (Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004); (2) a systematic failure to capture actual differences in SES across the student groups. The latter means that the item ownership systematically depends on culture, geography

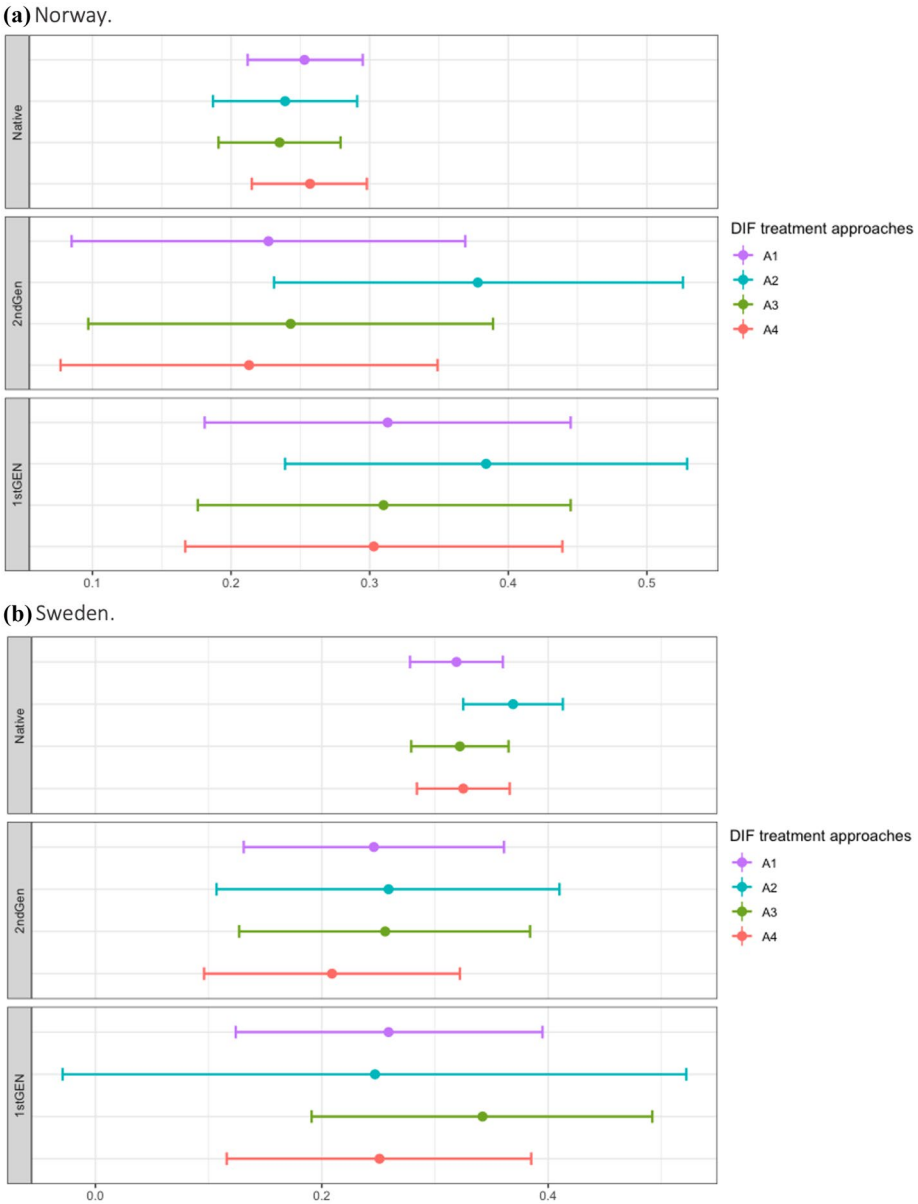

Fig. 3 The change in the strength of the HOMEPOS–Achievement relationship with different DIF treatment approaches. The bars reflect the 95% confidence interval

(May, 2006), i.e., the part of the country one lives in, or consumption preferences (Currie et al., 2008). The item response patterns will certainly vary due to these factors motivating the relevance of specific item ownership; however, this variation should reflect true variability in wealth rather than belonging to an immigrant or non-immigrant student group. In practice, full or partial metric non-invariance suggests that two students with the same actual level of SES but different immigration statuses will have different SES scores or vice versa (Lenkeit et al., 2015). This questions the valid use of the HOMEPOS scale scores for cross-immigrant group comparisons. Additionally, the lack of invariance for HOMEPOS may constrain meaningful comparisons across immigration status with the ESCS index that comprises three indicators, namely, household possessions, highest

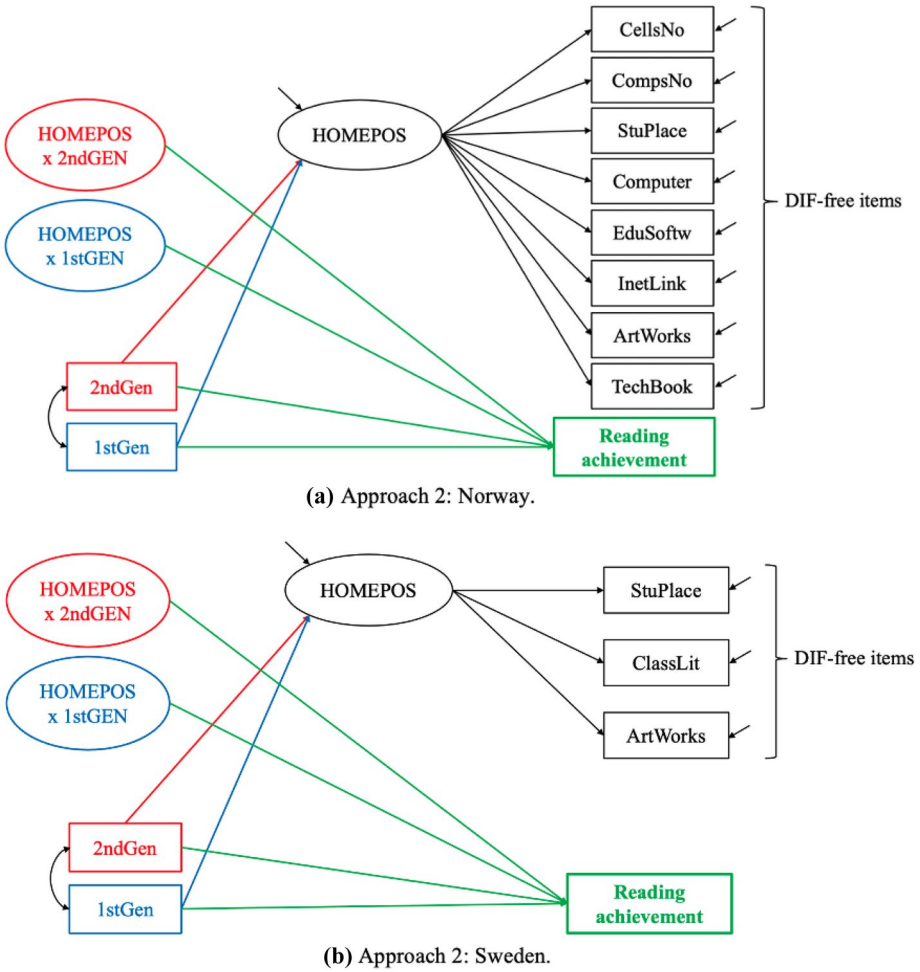

Fig. 4 Quantifying the HOMEPOS–reading achievement relationship after deleting non-comparable items: **a** HOMEPOS measurement model in the Norwegian sample is represented by comparable items—items that do not exhibit uniform or non-uniform DIF; **b** HOMEPOS measurement model in the Swedish sample is represented by comparable items—items that do not exhibit uniform or non-uniform DIF

parental education, and occupation. In PISA 2018, the ESCS index was constructed as the arithmetic mean of these three indicators that were given equal arbitrary factor loadings (see the OECD's PISA 2018 Technical Report, Chapter 16, https://www.oecd.org/pisa/data/pisa2018technicalreport/). As of this step, we retrieved no sufficient evidence for comparability of HOMEPOS. Hence, transferring the inferences drawn from the HOMEPOS scale to subgroups of students with different immigration statuses can create bias and misinform policymaking (Hansson & Gustafsson, 2013; Lenkeit et al., 2015).

For the second RQ, we identified items that functioned differently across immigration status and found several items exhibiting DIF. However, the findings varied in terms of the number and type of non-comparable items (i.e., uniform and non-uniform DIF) in the Norwegian and Swedish samples, the student group (i.e., certain items exhibited DIF only for the first- or second-generation immigrant students), and the relation of an item

to the HOMEPOS scale (i.e., which group the item was biased against) (for details, see Additional file 2: Appendix B). A possible explanation to our finding may be the variation in ethnicities that second- and first-generation immigrant groups belong to in Norway and Sweden (Fekjær, 2007; Heath et al., 2008; Lundahl & Lindblad, 2018). Furthermore, non-uniform DIF[1] was mainly observed between the first-generation immigrant and native students, although to a varying degree in two countries (Additional file 2: Appendix B). Since a potentially greater assimilation level is observed among the second-generation immigrant students (Alba et al., 2011; Drouhot & Nee, 2019; Heath et al., 2008; Hermansen, 2016; Jonsson & Rudolphi, 2011), our finding may imply that possessions hold a more equivalent value and relevance to the household circumstances of the native and second-generation immigrant students compared to the first-generation immigrant group. Hence, treating these two immigrant groups as one against the native group when comparing educational inequalities is of questionable value.

Overall, we found a tendency for wealth possession items (e.g., number of televisions, cars, bathrooms) and the number of books at home to be biased against students with immigrant background. Previously, Rutkowski and Rutkowski (2018) found the PISA 2012 wealth possessions scale overall non-comparability across Nordic countries and low parent-student agreement on the number of books at home. Besides, bringing books from the country of origin may be challenging and impractical, even if they were collected over generations (Elmeroth, 2006; Hansson & Gustafsson, 2013; Lenkeit et al., 2015). Conversely, immigrant students were more likely to endorse e-readers and home educational resources regardless of their HOMEPOS level. Since immigrant parents commonly have high aspirations for their children (e.g., Basit, 2012; Drouhot & Nee, 2019; Fekjær & Leirvik, 2011; Lauglo, 1999), their priority may be mobilizing capital for providing a study-motivating environment (Modood, 2005).

Finally, we examined how four approaches to adjust for DIF items influenced the strength of the HOMEPOS–reading achievement relationship to provide recommendations for the use of the scale. Neither of approaches had any effect on the HOMEPOS–achievement association for the native students in both countries[2] and for the second-generation immigrant students in Sweden. The HOMEPOS–achievement relationship remained moderate and stable even after deleting as many as 14 and 19 non-comparable items in Norway and Sweden, respectively. Conversely, two approaches for deleting non-comparable items considerably strengthened the HOMEPOS–achievement relationship for two immigrant student groups in Norway (after deleting all DIF items), and for the first-generation immigrant students in Sweden (after deleting non-uniform DIF items; see Figs. 3, 4, & Table 6; see Additional file 2: Appendix B for the full list of uniform and non-uniform DIF items).

Several practical implications arise from our findings for the use of the HOMEPOS scale. First, the non-comparability of multiple items suggests the limited nature of

[1] We remind that uniform DIF implies the non-equivalence of item response probability for students of different immigration status after controlling for their HOMEPOS latent trait. This hampers HOMEPOS factor mean cross-group comparisons. Non-uniform DIF is more peculiar since it points to a potential systematic difference in the value of HOMEPOS item for different groups with a HOMEPOS score of, e.g., "1" but not "0". This challenges the valid use of HOMEPOS to compare its relationship with achievement across groups.

[2] In Sweden, deleting DIF item approach increased the strength of HOMEPOS–achievement association for the native subsample to a small extent.

inferences we may draw about immigrant and non-immigrant student sub-populations with regard to their success or failure in schools. This necessitates invariance testing of the HOMEPOS measurement model to ensure that it reflects true variability in family wealth across all three student groups. Second implication concerns the deletion of non-comparable items that did not affect the strength of HOMEPOS–achievement association for native students in both countries, and 2ndGEN students in Sweden. The household possessions indices usually have a strong predictive power for academic achievement (Hannum et al., 2017; Lee et al., 2019). Hence, two conclusions may arise from our finding, (1) the HOMEPOS scale truly has a lower explanatory power for the achievement of groups specified above; (2) the deleted DIF items are potentially non-effective for capturing the SES of those groups. The latter is a common problem among higher-income countries (Avvisati, 2020), or countries with higher levels of wealth equality, since it is difficult to develop items that adequately discriminate among groups with different SES levels (Traynor & Raykov, 2013). Further analysis of the HOMEPOS item properties may give an insight of how well each item discriminates among advantaged and disadvantaged students across immigration status.[3] Third, ignoring non-comparable items by using the HOMEPOS index potentially masks high importance of SES for immigrant student achievement. Certain items (e.g., wealth possessions) may be negatively associated with reading achievement (Brese & Mirazchiyski, 2013; Traynor & Raykov, 2013), hence concealing SES effects. Fourth, an approach to account for DIF items which is usually preferred to deleting or ignoring DIF (Cho et al., 2016; Liu & Rogers, 2021) had no effect on the strength of HOMEPOS-achievement relationship for any group. The effectiveness of the approach is thus questionable. Fifth, we suggest caution in using the ESCS index to capture SES or to interpret educational inequalities across immigration status since HOMEPOS may compromise its adequate functioning. We further advise invariance testing for parental occupational status and educational level, since several researchers indicated potential problems with the equivalence of these socioeconomic status indicators (Lenkeit et al., 2015; Modood, 2005; Rothon, 2007). To conclude, several studies suggested the redundancy of the idea that all items function in the same way across different countries or groups (Rutkowski & Rutkowski, 2018), further introducing new methods, such as implementing partial invariance constraints to improve cross-country comparability (Lee & von Davier, 2020). Our study, however, took a more conservative approach and illustrated that, even after accounting for non-comparable items, we risk misinterpreting the SES – achievement relationship for immigrant student groups.

Limitations

Our study has some limitations that suggest future research directions. First, the PISA's IMMIG does not allow us to generalise our findings to specific ethnic groups due to the vague distinction between the native and second-generation immigrant categories (Basarkod et al., 2022), which may have assigned students who had one or two parents of the same ethnicity born outside the country of assessment to the different categories.

[3] Additional file 1: *Appendix A* contains item parameter estimates and response distributions for the 22 international items of HOMEPOS for students with first-generation, second-generation, and no immigrant background.

Second, the sampling criteria at the student level may have impacted the representativeness of the immigrant student samples. Over the years, Norway and Sweden have documented an increase in the proportion of 15-year-old students excluded from the PISA surveys. Compared to the average 4% for the OECD countries, 11.1% of the target population in Sweden and 7.9% in Norway were excluded from participating in PISA 2018 (Aursand & Rutkowski, 2021; Skolverket, 2019). Third, different sample characteristics due to different stratification variables used in Norway and Sweden do not allow for using the information on DIF items from the Swedish sample to construct a scale for the Norwegian sample and vice versa. The sample characteristics may not be important for international comparisons. However, when selecting comparable items within a country, using DIF test results of other country's sample is not advisable. Therefore, our findings cannot be applied to the construction of the HOMEPOS measure for countries other than Norway or Sweden. Fourth, by using only 22 international HOMEPOS items with the exclusion of three country-specific items we aimed at illustrating the same phenomenon for two countries albeit we are not comparing the findings due to the above-mentioned argument. Fifth, our invariance testing highlighted the immigrant background of a student as a possible cause of the non-comparability of HOMEPOS items. However, there may be more implicit groups that can potentially affect the comparability of HOMEPOS, such as the country of student origin (Dronkers et al., 2014).

Conclusions

Our study contributes to the discussion of the PISA HOMEPOS scale's comparability by investigating immigration status as a potential source of non-equivalence within two Nordic countries. The comparability of SES indicators is needed to accurately capture inequalities among immigrant and native children and adolescents (e.g., Braveman et al., 2005; Lenkeit et al., 2015). Immigrant students usually represent a heterogeneous group with a nuanced background depending on their country of origin, the reasons for migration, and the length of stay (Basarkod et al., 2022; Elmeroth, 2006). Thus, SES indicators likely have differential meanings and values for this group compared to a native one which may in turn influence the validity of our interpretations regarding the success or failure of students across immigration status (Fekjær, 2007; Modood, 2005; Rothon, 2007).

Our study revealed that deleting uniform DIF items from the HOMEPOS measurement model increased its effect on reading achievement for both immigrant groups in Norway, and deleting non-uniform DIF items increased this effect for the first-generation immigrant group in Sweden (see Additional file 2: Appendix B for the list of items). This finding suggested the dependency between the methodological approach researchers choose to deal with DIF items and the relation to a key educational outcome. Despite the small size of this dependency, there is a risk to draw conclusions based on evidence that may be over- or under-emphasizing achievement gaps for the three immigration status groups. Hence, we encourage researchers using the HOMEPOS scale to consider the invariance testing to avoid implicit methodological bias of the scale against immigrant groups. We further suggest an in-depth analysis of the HOMEPOS item properties to fully understand how well the items discriminate among students of various SES levels with and without immigrant background. This type of analysis will add value to

immigrant-related research that utilizes the PISA data to investigate disadvantage in schools, achievement gaps, or academic resilience.

Acknowledging the effort PISA teams around the world have taken to develop the HOMEPOS index, we aimed at initiating a productive discussion of this measure to improve its effectiveness in capturing educational inequalities across heterogeneous student subpopulations. Our study also highlighted the limitations of the HOMEPOS index as a source of evidence on immigration gaps in educational achievement in Norway and Sweden.

Abbreviations

1stGEN	First-generation immigrant students (a category of the index of immigrant background)
2ndGEN	Second-generation immigrant students (a category of the index of immigrant background)
2PLM	Two-parameter logistic model
CILS4EU	Children of Immigrants Longitudinal Survey in Four European Countries
DIF	Differential item functioning
ESCS	Economic, social and cultural status
GPCM	Generalised partial credit model
HOMEPOS	Household possessions (a composite variable in the PISA Database)
ICT	Information and communication technology
IMMIG	Index of immigrant background
MG-IRT	Multigroup item response theory
MI	Measurement invariance
MIMIC-DIF	Multiple-indicators-multiple-causes differential item functioning
MIMIC	Multiple-indicators-multiple-causes
OECD	Organisation for Economic Co-Operation and Development
PISA	Programme for International Student Assessment
SES	Socioeconomic status
TIMSS	Trends in International Mathematics and Science Study
UNESCO	United Nations Educational, Scientific and Cultural Organization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-022-00132-w.

Additional file 1: Appendix A. Table A1–A4: Item parameter estimates & item response distributions for the native, second-generation (2ndGEN), and first-generation (1stGEN) immigrant student sub-populations in Norway and Sweden.

Additional file 2: Appendix B. Likelihood-ratio tests, information criteria, uniform, and non-uniform DIF effects for the Norwegian sample (Table B1–B2) and the Swedish sample (Table B3–B4).

Additional file 3: Appendix C. Sensitivity analyses: Sequential-free baseline approach.

Additional file 4: Appendix D. Mplus input commands.

Acknowledgements
Not applicable.

Author contributions
OM, RS, and TN conceptualized the study together. OM conducted the literature review, ran the analyses, interpreted the findings, and prepared the manuscript draft. RS provided guidance concerning the methodology, analyses, and findings. RS also contributed to the different parts of the written manuscript. TN provided comments on the theory and manuscript overall. All authors read and approved the final manuscript.

Author information
Oleksandra Mittal is a doctoral research fellow and Trude Nilsen is a research professor at the Department of Teacher Education and School Research at the University of Oslo. Ronny Scherer is a professor at the Centre for Educational Measurement at the University of Oslo.

Funding
Not applicable.

Availability of data and materials
The PISA 2018 student-questionnaire data file analysed in this study is publicly available in the OECD PISA database https://www.oecd.org/pisa/data/2018database/.

Declarations

Ethics approval and consent to participate
Not applicable since the secondary data provided by PISA does not contain any sensitive confidential information. Pseudonymisation is applied to data with ID numbers being assigned to all schools and students participating in the survey. As a result, the assessment data is completely anonymous in the published data files. Therefore, neither consent to participate nor ethics approval were required for the analyses.

Consent for publication
The authors consent to the publication of the manuscript in ***Large-scale Assessments in Education***.

Competing interests
The authors declare that they have no competing interests.

Received: 13 May 2022 Accepted: 5 September 2022
Published online: 13 September 2022

References

Adan, T., & Antara, L. (2018). *Political participation of refugees. The case of Syrian and Somali refugees in Sweden*. International Institute for Democracy and Electoral Assistance. https://www.idea.int/sites/default/files/publications/political-participation-of-refugees-the-case-of-syrian-and-somali-refugees-in-sweden.pdf

Agasisti, T., & Longobardi, S. (2017). Equality of educational opportunities, schools' characteristics and resilient students: An empirical study of EU-15 countries using OECD-PISA 2009 data. *Social Indicators Research, 134*(3), 917–953. https://doi.org/10.1007/s11205-016-1464-5

Alba, R., Kasinitz, P., & Waters, M. C. (2011). The kids are (mostly) alright: Second-generation assimilation: Comments on Haller, Portes and Lynch. *Social Forces, 89*(3), 763–773. https://doi.org/10.1353/sof.2011.0024

American Psychological Association. (n.d.). *Socioeconomic status*. Retrieved 5 Aug 2022 from http://www.apa.org/topics/socioeconomic-status/

Ammermüller, A. (2007). Poor background or low returns? Why Immigrant Students in Germany Perform so Poorly in PISA. *Education Economics, 15*(2), 215–230. https://doi.org/10.2139/ssrn.686722

Andersen, A., Krølner, R., Currie, C., Dallago, L., Due, P., Richter, M., Örkényi, A., & Holstein, B. E. (2008). High agreement on family affluence between children's and parents' reports: International study of 11-year-old children. *Journal of Epidemiology & Community Health, 62*(12), 1092–1094. https://doi.org/10.1136/jech.2007.065169

Andon, A., Thompson, C. G., & Becker, B. J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-Scale Assessments in Education, 2*(1), 1–20. https://doi.org/10.1186/s40536-014-0007-2

Areepattamannil, S., & Kaur, B. (2013). Factors predicting science achievement of immigrant and non-immigrant students: A multilevel analysis. *International Journal of Science and Mathematics Education, 11*(5), 1183–1207. https://doi.org/10.1007/s10763-012-9369-5

Aursand, L., & Rutkowski, D. (2021). Exemption or exclusion? A study of student exclusion in PISA in Norway. *Nordic Journal of Studies in Educational Policy, 7*(1), 16–29. https://doi.org/10.1080/20020317.2020.1856314

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

Bakken, A., & J. I. Elstad. (2012). *For store forventninger? Kunnskapsløftet og ulikheten i skolekarakterer* [Too high expectations? Knowledge promotion and inequality in school grades, in Norwegian]. *NOVA Report, 7*(12). NOVA. https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/2012/nova_slutt.pdf

Basarkod, G., Marsh, H. W., Parker, P. D., Dicke, T., & Guo, J. (2022). The immigrant paradox and math self-concept: An SES-of-origin-country hypothesis. *Learning and Instruction, 77*, 101539. https://doi.org/10.1016/j.learninstruc.2021.101539

Basit, T. N. (1997). 'I want more freedom, but not too much': British Muslim girls and the dynamism of family values. *Gender and Education, 9*, 425–440. https://doi.org/10.1080/09540259721178

Basit, T. N. (2012). 'My parents have stressed that since I was a kid': Young minority ethnic British citizens and the phenomenon of aspirational capital. *Education, Citizenship and Social Justice, 7*(2), 129–143. https://doi.org/10.1177/1746197912440857

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–526. https://doi.org/10.1037/met0000077

Bialosiewicz, S., Murphy, K., & Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants* [Demonstration Session]. American Evaluation Association, Washington, DC, 1–37. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Braveman, P., Cubbin, C., Egerter, S., Chideya, S., Marchi, K., Metzler, M., & Posner, S. (2005). Socioeconomic status in health research: One size does not fit all. *JAMA the Journal of the American Medical Association, 294*(22), 2879–2888. https://doi.org/10.1001/jama.294.22.2879

Breidahl, K. N. (2017). Scandinavian exceptionalism? Civic integration and labour market activation for newly arrived immigrants. *Comparative Migration Studies, 5*(1), 1–19. https://doi.org/10.1186/s40878-016-0045-8

Brese, F., & Mirazchiyski, P. (2013). *Issues and Methodologies in Large-Scale Assessments. Special Issue 2: Measuring Students' Family Background in Large-Scale International Education Studies. IERI Monograph Series. International Association for the Evaluation of Educational Achievement*. https://files.eric.ed.gov/fulltext/ED561898.pdf, Appendices: https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/Special_Issue_2/08_IERI_Special_Issue_2_Appendix.pdf

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*. https://doi.org/10.3389/feduc.2017.00051

Bunar, N. (2010). *Nyanlända och lärande - En forskningsöversikt om nyanlända elever i den svenska skolan [Newly arrived students and their learning. A review of studies on newly arrived students in the Swedish school]*. Vetenskapsrådets rapportserie 6. Vetenskapsrådet. https://www.vr.se/download/18.2412c5311624176023d25b5f/1529480533281/Nyanlaenda-och-laerande_VR_2010.pdf

Byström, M., & Frohnert, P. (2017). *Invandringens historia: från "folkhemmet" til dagens Sverige. [History of immigration: From the "People's Home" to present-day Sweden]*. Report 5. Delegationen för migrationsstudier: Delmi. Derived from: https://www.delmi.se/publikationer/kunskapsoversikt-2017-5-invandringens-historia-fran-folkhemmet-till-dagens-sverige/

Cerna, L., Brussino, O., & Mezzanotte, C. (2021). The resilience of students with an immigrant background: An update with PISA 2018, *OECD Education Working Papers*, No. 261, OECD Publishing. https://doi.org/10.1787/e119e91a-en

Cheung, K. C., Sit, P. S., Soh, K. C., Ieong, M. K., & Mak, S. K. (2014). Predicting academic resilience with reading engagement and demographic variables: Comparing Shanghai, Hong Kong, Korea, and Singapore from the PISA perspective. *The Asia-Pacific Education Researcher, 23*(4), 895–909.

Cho, S. J., Suh, Y., & Lee, W. Y. (2016). After differential item functioning is detected: IRT item calibration and scoring in the presence of DIF. *Applied Psychological Measurement, 40*(8), 573–591. https://doi.org/10.1177/0146621616664304

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement, 40*(7), 486–499. https://doi.org/10.1177/0146621616659738

Currie, C., Molcho, M., Boyce, W., Holstein, B., Torsheim, T., & Richter, M. (2008). Researching health inequalities in adolescents: The development of the Health Behaviour in School-Aged Children (HBSC) family affluence scale. *Social Science & Medicine, 66*(6), 1429–1436. https://doi.org/10.1016/j.socscimed.2007.11.024

De Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

Desjardins, C. D., & Bulut, O. (2018). Handbook of educational measurement and psychometrics using R. *CRC Press*. https://doi.org/10.1201/b20498

Dimitrov, D. M. (2017). Examining differential item functioning: IRT-based detection in the framework of confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 50*(3), 183–220. https://doi.org/10.1080/07481756.2017.1320946

Dronkers, J., Levels, M., & de Heus, M. (2014). Migrant pupils' scientific performance: The influence of educational system features of origin and destination countries. *Large-Scale Assessments in Education, 2*(3), 1–28. https://doi.org/10.1186/2196-0739-2-3

Drouhot, L. G., & Nee, V. (2019). Assimilation and the second generation in Europe and America: Blending and segregating social dynamics between immigrants and natives. *Annual Review of Sociology, 45*, 177–199. https://doi.org/10.1146/annurev-soc-073117-041335

Elmeroth, E. (2006). Monokulturella studier av multikulturella elever. Att mäta och förklara skolresultat [Mono-cultural studies of multicultural students. Measuring and explaining school performance]. *Pedagogisk Forskning i Sverige, 11*(3), 177–194.

Fekjær, S. N. (2007). New differences, old explanations: Can educational differences between ethnic groups in Norway be explained by social background? *Ethnicities, 7*(3), 367–389. https://doi.org/10.1177/1468796807080234

Fekjær, S., & Leirvik, M. (2011). Silent gratitude: Education among second-generation Vietnamese in Norway. *Journal of Ethnic and Migration Studies, 37*(1), 117–134. https://doi.org/10.1080/1369183X.2011.521365

Gabrielli, G., Longobardi, S., & Strozza, S. (2021). The academic resilience of native and immigrant-origin students in selected European countries. *Journal of Ethnic and Migration Studies*. https://doi.org/10.1080/1369183X.2021.1935657

Gramaţki, I. (2017). A comparison of financial literacy between native and immigrant school students. *Education Economics, 25*(3), 304–322. https://doi.org/10.1080/09645292.2016.1266301

Hagelund, A. (2020). After the refugee crisis: Public discourse and policy change in Denmark, Norway and Sweden. *Comparative Migration Studies, 8*(13), 1–17. https://doi.org/10.1186/s40878-019-0169-8

Hannum, E., Liu, R., & Alvarado-Urbina, A. (2017). Evolving approaches to the study of childhood poverty and education. *Comparative Education, 53*(1), 81–114. https://doi.org/10.1080/03050068.2017.1254955

Hansson, Å., & Gustafsson, J. E. (2013). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research, 57*(2), 148–166. https://doi.org/10.1080/00313831.2011.625570

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39–56). Information Age Publishing.

Heath, A. F., Rothon, C., & Kilpi, E. (2008). The second generation in Western Europe: Education, unemployment, and occupational attainment. *Annual Revision of Sociology, 34*, 211–235. https://doi.org/10.1146/annurev.soc.34.040507.134728

Hermansen, A. S. (2016). Moving up or falling behind? Intergenerational socioeconomic transmission among children of immigrants in Norway. *European Sociological Review, 32*(5), 675–689. https://doi.org/10.1093/esr/jcw024

Hernes, V., Arendt, J. N., Joona, P. A., & Tronstad, K. R. (2019). Nordic integration and settlement policies for refugees: A comparative analysis of labour market integration outcomes. *Nordic Council of Ministers*. https://doi.org/10.6027/TN2019-529

Jonsson, J. O., & Rudolphi, F. (2011). Weak performance—strong determination: School achievement and educational choice among children of immigrants in Sweden. *European Sociological Review, 27*(4), 487–508. https://doi.org/10.1093/esr/jcq021

Keskpaik, S., & Rocher, T. (2011). La mesure de l'équité dans PISA: pour une décomposition des indices statistiques. *Éducation et formations, 80*, 69–78. http://media.education.gouv.fr/file/revue_80/30/4/Depp-EetF-2011-80-mesure-equite-pisa-indices-statistiques_203304.pdf
Kilpi-Jakonen, E. (2014). Citizenship and educational attainment amongst the second generation: An analysis of children of immigrants in Finland. *Journal of Ethnic and Migration Studies, 40*(7), 1079–1096. https://doi.org/10.1080/1369183X.2013.831543
Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337
Kim, S. W., Cho, H., & Kim, L. Y. (2019). Socioeconomic status and academic outcomes in developing countries: A meta-analysis. *Review of Educational Research, 89*(6), 875–916. https://doi.org/10.3102/0034654319877155
Kingdon, G., & Cassen, R. (2010). Ethnicity and low achievement in English schools. *British Educational Research Journal, 36*(3), 403–431. https://doi.org/10.1080/01411920902989185
Lauglo, J. (1999). Working harder to make the grade: Immigrant youth in Norwegian schools. *Journal of Youth Studies, 2*(1), 77–100. https://doi.org/10.1080/13676261.1999.10593025
Lee, J., Zhang, Y., & Stankov, L. (2019). Predictive validity of SES measures for student achievement. *Educational Assessment, 24*(4), 305–326. https://doi.org/10.1080/10627197.2019.1645590
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.
Lenkeit, J., Caro, D. H., & Strand, S. (2015). Tackling the remaining attainment gap between students with and without immigrant background: An investigation into the equivalence of SES constructs. *Educational Research and Evaluation, 21*(1), 60–83. https://doi.org/10.1080/13803611.2015.1009915
Liu, X., & Jane Rogers, H. (2021). Treatments of differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211012050
Lundahl, L., & Lindblad, M. (2018). Immigrant student achievement and education policy in Sweden. In L. Volante, D. Klinger, & O. Bilgili (Eds.), *Immigrant student achievement and education policy. Policy implications of research in education* (Vol. 9, pp. 69–85). Springer: Springer. https://doi.org/10.1007/978-3-319-74063-8_5
Marks, G. N. (2005). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. *Ethnic and Racial Studies, 28*(5), 925–946. https://doi.org/10.1080/01419870500158943
Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology, 104*(4), 1054–1073. https://doi.org/10.1037/a0029152
Marx, A. E., & Stanat, P. (2012). Reading comprehension of immigrant students in Germany: Research evidence on determinants and target points for intervention. *Reading and Writing, 25*(8), 1929–1945. https://doi.org/10.1007/s11145-011-9307-x
May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics, 31*(1), 63–79. https://doi.org/10.3102/10769986031001063
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
Modood, T. (2005). The educational attainments of ethnic minorities in Britain. In G. C. Loury, T. Modood, & S. M. Teles (Eds.), *Ethnicity, social mobility, and public policy. Comparing the US and UK* (pp. 288–308). Cambridge University Press.
Modood, T. (2012). Capitals, ethnicity and higher education. In T. N. Basit & S. Tomlinson (Eds.), *Social inclusion and higher education* (pp. 17–40). The Policy Press.
Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement, 44*(2), 118–136. https://doi.org/10.1177/0146621619835496
Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). Muthén & Muthén. https://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r6_web.pdf
Organisation for Economic Co-operation and Development (2017). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf
Organisation for Economic Co-operation and Development (forthcoming). *PISA 2018t technical report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/
Organisation for Economic Co-operation and Development. (2018). *The resilience of students with an immigrant background: Factors that shape well-being OECD Reviews of Migrant Education*. OECD Publishing. https://doi.org/10.1787/9789264292093-en
Organisation for Economic Co-operation and Development. (2019a). *PISA 2018 results (Volume II): Where all students can succeed*. OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en
Organisation for Economic Co-operation and Development. (2019b). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en
Parveen, S. (2020). Norwegian asylum policy and response to the 2015 refugee crisis. *International Studies, 57*(4), 391–406. https://doi.org/10.1177/0020881720965050
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
Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004
Rangvid, B. S. (2007). Sources of immigrants' underachievement: Results from PISA—Copenhagen. *Education Economics, 15*(3), 293–326. https://doi.org/10.1080/09645290701273558
Rolfe, V. (2021). Tailoring a measurement model of socioeconomic status: Applying the alignment optimization method to 15 years of PISA. *International Journal of Educational Research, 106*, 101723. https://doi.org/10.1016/j.ijer.2020.101723
Rothon, C. (2007). Can achievement differentials be explained by social class alone? An examination of minority ethnic educational performance in England and Wales at the end of compulsory schooling. *Ethnicities, 7*(3), 306–322. https://doi.org/10.1177/1468796807080231

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., Gonzales, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Chapman & Hall/CRC Press.
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. https://doi.org/10.1080/00220272.2010.487546
Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354–367. https://doi.org/10.1080/00313831.2016.1261044
Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257
Sandoval-Hernandez, A., Rutkowski, D., Matta, T., & Miranda, D. (2019). Back to the drawing board: Can we compare socioeconomic background scales? [Pensémoslo de nuevo:¿ Podemos comparar las escalas de antecedentes socioeconómicos?] *Revista de Educación,* 383, 37–61. https://doi.org/10.4438/1988-592X-RE-2019-383-400
Schleicher, A. (2006). Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003: © OECD 2006. *Intercultural Education, 17*(5), 507–516. https://doi.org/10.1080/14675980601063900
Schnepf, S. V. (2007). Immigrants' educational disadvantage: An examination across ten countries and three surveys. *Journal of Population Economics, 20*(3), 527–545. https://doi.org/10.1007/s00148-006-0102-y
Shapira, M. (2012). An exploration of differences in mathematics attainment among immigrant pupils in 18 OECD countries. *European Educational Research Journal, 11*(1), 68–95. https://doi.org/10.2304/eerj.2012.11.1.68
Skolverket. (2016a). *Utbildning för nyanlända elever. Skolverkets allmänna råd med kommentarer*. Skolverket. https://www.skolverket.se/download/18.6bfaca41169863e6a65bceb/1553966475563/pdf3576.pdf
Skolverket. (2016b). *Invandringens betydelse för skolresultaten [The importance of immigration for school performance]*. Skolverket. https://www.skolverket.se/publikationer?id=3604
Skolverket. (2019). *PISA 2018: 15-åringars kunskaper i läsförståelse, matematik och naturvetenskap*. Skolverket. https://www.skolverket.se/download/18.75bdbbb116e7434ebf8595/1575624399449/pdf5347.pdf
Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292
Strand, S. (2014). Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and 'getting it' for the white working class. *Research Papers in Education, 29*(2), 131–171. https://doi.org/10.1080/02671522.2013.767370
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
United Nations Educational, Scientific and Cultural Organization (2018). *Global education monitoring report 2019: Migration, displacement and education: Building bridges, not walls* (2nd ed.). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000366946
United Nations High Commissioner for Refugees Operational Data portal (2022). *Ukraine Refugee Situation*. UNHCR, Retrieved 5 Aug 2022 from https://data2.unhcr.org/en/situations/ukraine
Van de Vijver, F. J. (2018). Towards an integrated framework of bias in noncognitive assessment in international large-scale studies: Challenges and prospects. *Educational Measurement: Issues and Practice, 37*(4), 49–56. https://doi.org/10.1111/emip.12227
Von Davier, M., Gonzalez, E., & Mislevey, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*(1), 9–36.
Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*(5), 713–731. https://doi.org/10.1177/0013164409332228
Wardle, J., Robb, K., & Johnson, F. (2002). Assessing socioeconomic status in adolescents: The validity of a home affluence scale. *Journal of Epidemiology and Community Health, 56*(8), 595–599. https://doi.org/10.1136/jech.56.8.595
Watermann, R., Maaz, K., Bayer, S., & Roczen, N. (2016). Social background. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 117–145). Springer International Publishing.
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
Willms, J., & Tramonte, L. (2015). Towards the development of contextual questionnaires for the PISA for development study. *OECD Education Working Papers*, No. 118. OECD Publishing. https://doi.org/10.1787/5js1kv8crsjf-en
Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1–27. https://doi.org/10.1080/00273170802620121
Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339–361.
Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *Journal of Psychopathology and Behavioral Assessment, 31*, 320–330. https://doi.org/10.1007/s10862-008-9118-9
Yang, Y., & Gustafsson, J.-E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation, 10*(3), 259–288. https://doi.org/10.1076/edre.10.3.259.30268

Publisher's Note