Measurement 66 (2015) 263–271

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Correcting for differential item functioning in multi-level regression models in cross-national surveys

Khurrem Jehangir *, Stephanie M. van den Berg [1], Cees A.W. Glas [2]

Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form 9 December 2014
Accepted 6 February 2015
Available online 14 February 2015

Keywords:
Differential item functioning
PISA 2009
Country specific item parameters
Partial credit model
Generalised partial credit model

ABSTRACT

Results of the PISA project have shown that the school average socio-economic status is an important background variable that explains a lot of variance in the student results. However, if the socio-economic variable which is measured at the student level is biased across countries due to a cultural bias, then the aggregated variable (at school level) is also subject to error. In this article, DIF (Differential Item Functioning, i.e., item bias) is mitigated using country specific item parameters. The effect of using this approach is studied on the results from multilevel regression for different measurement models and person parameter estimation procedures. Results showed that for countries affected by DIF the impact on the regression coefficients cannot be ignored. The effect is shown to be more for the PCM than the GPCM and generally more for the EAP estimates than the WML estimates.

1. Introduction

The PISA (Program for International Student Assessment) educational survey of the OECD aims to evaluate students' skills and aptitude for lifelong learning in the areas of reading literacy, mathematics, science and problem solving. The target population of PISA consists of 15 year-old students. The first PISA cycle started in the year 2000 and is repeated every three years with one of the three mentioned domains being the focus in each cycle. Besides collecting responses on cognitive tests, PISA also collects data on background characteristics of the students through so-called context questionnaires. The information provided by these background questionnaires is key to the ongoing success of PISA, since these questionnaires provide valuable information about the factors which affect students' test performance. This helps the participating countries to frame effective educational policies to tackle the shortcomings in their educational school systems that are revealed by PISA. It also gives the participating countries an opportunity to observe how background variables may relate to student performance in other countries.

One of the key background variables that explains much variation in student performance is the socio-economic status variable. However, the questionnaire items administered to the students are international and there is always the question of whether the items used as a proxy for socio-economic status function consistent across countries. For example, an item like the number of cars owned by a family may be a proxy for socio-economic status in New York or in a suburban town, but the importance given to the number of cars might be different in these two places. The presence of such a bias in the measurement scale can lead to measurement error in the variable across

* Corresponding author at: Witbreuksweg 397-109, 7522 ZA Enschede, The Netherlands. Tel.: +31 621273773.

E-mail addresses: k_jehangir@hotmail.com (K. Jehangir), Stephanie.vandenberg@utwente.nl (S.M. van den Berg), C.A.W.Glas@utwente.nl (C.A.W. Glas).

[1] Address: University of Twente, Faculty of Behavioral Sciences, Department OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands. Tel.: +31 53 4892422.

[2] Address: University of Twente, Faculty of Behavioral Sciences, Department OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands. Tel.: +31 53 4893565.

http://dx.doi.org/10.1016/j.measurement.2015.02.019
0263-2241/

respondents from diverse cultural or geographical groups. This measurement bias where an item is behaving differently across distinct groups is called Differential Item Functioning (DIF). This can have a bearing on the measurement precision of a key variable like socio economic status and subsequently on the amount of variance it explains in the regression model. Since the background variables are consequential for educational policy making in the various participating countries, it is important to take steps to ensure that cross country bias is minimised in the measurement of the variables of interest.

PISA currently uses a composite index for the socio-economic status of students. This index is composed of three attributes of a student's background, that is, possessions at home (HOMEPOS), highest parental education (PARED) and highest parental occupation (HISEI). PISA also makes use of these three attributes of socio-economic status separately in the regression models. HOMEPOS is a latent variable, which means that it is not directly observed but inferred using a measurement model. PARED and HISEI on the other hand, are directly observed scores based on a comparative framework across countries. To form a composite index from these three sub-components a principal component analysis is done to compute a composite scale score. The method to assess DIF is applicable to multi-item scales, so we use the HOMEPOS scale in our analysis. The HOMEPOS scale consists of a common set of 20 items that are administered across all participating countries. The items are both dichotomous and polytomous. The dichotomous items ask the examinee if he or she has a certain household possession at home, like an own room or a computer or a dishwasher etcetera. The polytomous items ask the examinee how many televisions or cars or bathrooms are present in the household. There has been a long debate in PISA whether to use national or international item parameters to measure the HOMEPOS scale. The item parameters of the measurement model used to scale HOMEPOS represent one or more characteristics of the test item. National item parameters are those which are obtained by using within country scaling and international item parameters are those which are obtained by scaling a sample of students drawn from a group of countries. Both methods have their advantages and disadvantages and so far there is no consensus on which is better. Using international item parameters is impacted by DIF as already explained. On the other hand using national item parameters avoids DIF but the consequence is that meaningful comparisons between countries become harder to make.

The aim of this article is to study the impact of cross-country DIF on multi-level regression of student performance on HOMEPOS and Mean HOMEPOS (average HOMEPOS of all individuals in a school), that is, to assess how the regression coefficients of HOMEPOS and Mean HOMEPOS vary with and without compensating for DIF. We propose to assess the impact of DIF in the latent scale HOMEPOS by using country specific item parameters for any items displaying large DIF. This approach enables the use of international calibration for the item parameters and for making meaningful international comparisons while tackling the cultural bias inherent in international measurement instruments. The effect of using country specific item parameters for DIF items in the HOMEPOS scale is shown for different measurement models and different parameter estimation procedures to show how well this approach functions in different settings.

2. Method

This section gives an overview of how the analyses were conducted and discusses the theory they are based on. Firstly, we describe the Item Response Theory (IRT) models which are used as measurement models in our analyses. Secondly, we present the estimation process for our analyses starting with the calibration phase of the estimation process, that is, how item parameters are estimated. Thirdly, we describe how DIF is identified and compensated for with country specific items parameters. Then the scoring methods for person parameters, that is, the latent trait scores of persons are described. Finally the multi-level regression model is presented.

2.1. Item response theory

IRT has been gaining popularity amongst survey researchers as a tool for analysing survey data, especially in the field of education. IRT relates the item responses on a test to a latent variable, for example reading achievement (see, for instance, [13]). The use of IRT is especially relevant for the analyses of the PISA scales. The term scales refers to groups of questions/items which are clustered together and aimed at measuring certain constructs of interest like for instance 'attitude towards reading'. Most scales in PISA consist of polytomously scored items. There are several reasons for using IRT-based scores rather than conventional sum scores or weighted sum scores for measuring latent constructs. As will become clear in the example given below, one of the main advantages is that IRT-based latent scale scores come with standard errors which reflect the measurement error of the instruments used. Further, IRT-based IRT scale scores remain comparable when the responses are collected in a so-called incomplete design, that is, when respondents are presented different sets of items.

IRT models describe the relationship between an examinee's standing on a latent variable, say educational achievement or an attitude, and item responses, based on the characteristics of the items of the test. The dependence of the observed responses on the dichotomously or polytomously scored items on the latent person variable is fully specified by the item characteristic function, which is the regression of an item score on latent variable. The item characteristic function allows inference about latent variable to be made from the observed item responses. The item characteristic functions cannot be directly observed because the ability parameter is not observed. But under certain assumptions it is possible to infer the information of interest from the examinee's responses to the test items (e.g., [13]).

An example of an item characteristic function for a dichotomous item is the one parameter model. The probability that an examinee answers an item correctly

depends on his ability and the difficulty of the item. The difficulty parameter is the point on the ability scale where the probability of a correct response is 50%. So the larger the value of the difficulty parameter, the higher the ability that is required to have a 50% chance of getting the item correct. In Fig. 1, two different item characteristic functions are plotted for a one parameter model. The functions differ by location on the ability scale. Item 2 is more difficult and shifted to the right on the difficulty scale.

The one-parameter model can be extended to a two-parameter model where the probability that an examinee answers an item correctly depends not only on his or her ability and the difficulty of the item but also on the discriminating behaviour of the item. The difficulty parameter is the point on the ability scale where the probability of a correct response is 50%. The discrimination parameter is proportional to the slope of the item response function at the point of the difficulty parameter on the ability scale.

Items with high discrimination parameter values are useful for separating examinees into different ability levels. In Fig. 2, two different item characteristic functions are plotted. The functions differ by location on the ability scale and by the slopes. Item 2 is more difficult and shifted to the right on the ability scale. The item characteristic curve of Item 2 corresponds with a higher discrimination parameter. As a result, a small increase in ability leads to a higher increase in probability of scoring correct compared to item 1.

An IRT model may provide an adequate description of the test data. It is essential to test the fit of the model to the data. The examinees' abilities are unobservable but can be estimated. An IRT model provides a framework for the uncertainty regarding the estimate of the ability. So, an IRT model can be used to measure the abilities of the examinees, and quantifies the uncertainty regarding the estimate. For an overview of different IRT models, see, for example, Hambleton and Swaminathan (1985).

An IRT model for polytomous items is the Generalised Partial Credit model GPCM by Muraki [18]. If an item is dichotomous the GPCM reduces to the two-parameter model. If one of the parameters in the GPCM is fixed to 1, the GPCM reduces to the PCM [14]. The GPCM is an example of an adjacent category model. Others models have also been proposed that fall in the category of continuation-ratio models [21] or cumulative probability models [20]. Though the rationales underlying the models are very different, the practical implications are often negligible, because their item-category curves are so close that they can hardly be distinguished on the basis of empirical data [21]. The PCM (or the version of the GPCM with the discrimination parameter constrained to 1) is the model that has been selected by large scale surveys like PISA, TIMMS and PIRLS to calibrate the response data. Therefore we also selected the GPCM as the measurement models for our analysis.

In the GPCM, the probability of a student n scoring in category j on item i (denoted by $X_{nij} = 1$) is given by

$$P(X_{nij} = 1|\theta_n) = P_{ij}(\theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_{ij})}{1 + \sum_{j=1}^{M_i} \exp(j\alpha_i\theta_n - \beta_{ij})}, \qquad (1)$$

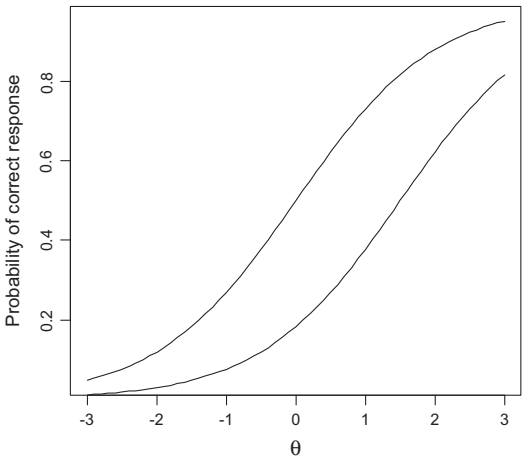

Fig. 1. Item characteristic function for a one parameter model.

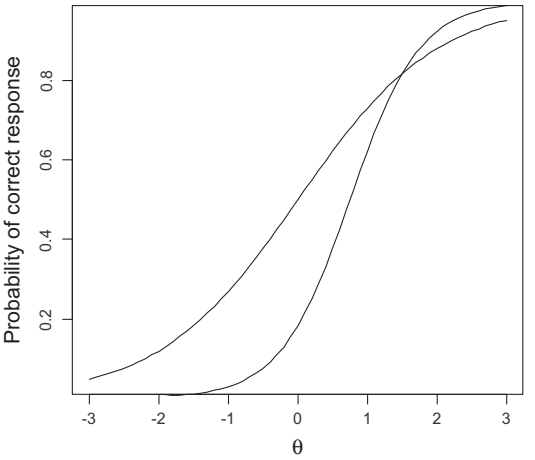

Fig. 2. Item characteristic function for a two parameter model.

for $j = 1,\ldots, M_i$. As the latent trait level of the student increases the probability of scoring in higher categories increases. For every latent trait level θ (theta) there is a certain category where the probability of responding is the highest. An example of the category response functions $P_{ij}(\theta_n)$ for an item with four ordered response categories is given in Fig. 3. Further, the graph also shows the expected item-total score

$$E(T_i|\theta) = \sum_{j=1}^{M_i} jE(X_{ij}|\theta) = \sum_{j=1}^{M_i} jP_{ij}(\theta), \qquad (2)$$

where the item-total score is defined as $T_i = \sum_{j=1}^{M_i} jX_{ij}$. Note that the expected item-total score increases as a function of θ.

If the discrimination parameter α_i is constrained to one, the GPCM reduces to the PCM. The models for the analysis using the PCM and the GPCM were identified as follows. The item parameters were estimated over all countries using the cross-country data-set. When using the PCM the model was identified by fixing the ability distribution

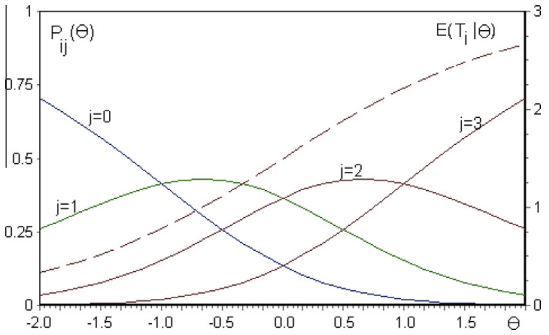

Fig. 3. Response functions and expected item-total score under the GPCM.

of one group (or one country in our case). In this study the ability distribution of the last country in the analysis, the United States was fixed to standard normal when estimating the model parameters. For the analysis using the GPCM, again the ability distribution of United States was fixed at standard normal when estimating the item parameters. Furthermore, for the GPCM, the product of the discrimination parameters for all items was fixed at 1.

2.2. Estimation process

The estimation process described below is for the independent variable HOMEPOS that is used in the regression analyses. The dependent variable in the regression analyses is the reading ability of the student. The reading ability, which is a latent trait, is not estimated here, but directly used from the PISA 2009 student questionnaire database. A brief description of how the reading ability was estimated in PISA 2009 is given later in the Multi-level model section.

2.2.1. Item calibration

The estimation process for the independent variable HOMEPOS begins with a calibration run to estimate the item parameters. This is done using the marginal maximum likelihood (MML) method [1]. In this approach, the person parameters θ are considered as nuisance parameters. They are assumed to be samples from one or more normal distributions and intenerated out of the likelihood function. Thus, the likelihood function no longer depends on the person parameters, but only on the item parameters and the means and standard deviations of the population distributions. In the application presented here, every country has its own distribution.

2.2.2. Country-specific item parameters

The next step after obtaining the item parameters was to investigate the presence of DIF or item bias across countries. In this study we focussed on uniform DIF across countries, which mean that only changes in the item difficulty due to DIF are targeted and changes in the item discrimination parameter are ignored. Several options are available in the presence of DIF. One extreme option is to eliminate the DIF items from the measurement instrument. If the number of eliminated items is large, this has the drawback that the measurement precision decreases and the construct validity is threatened. An alternative is to model DIF using country-specific item parameters. The items without DIF identify the latent scale and it is assumed that the items with DIF still load on this latent scale but with specific item parameters for the concerning subgroup(s) [6,8]. Thus, these items can be used to estimate the value on the latent variable and contribute to the precision of the estimates. The data of the set of countries in the study is analysed simultaneously to identify items with country-specific DIF and an item which is identified as a DIF item is modelled with country-specific item parameters. This is done in an iterative process where the item with the worst misfit is replaced with country specific item parameters, followed by a new overall analysis. So the DIF items were treated one-at-a-time until eight items had been replaced. The number of items that are calibrated using country specific item parameters in a scale is an arbitrary choice. The only thing that has to be ensured is that there is a sufficient number of anchor items remaining in the scale. The scale consisted of a total of twenty items, so replacing eight items with country specific item parameters meant that there were twelve items that were used to anchor the scale.

In order to identify DIF in the HOMEPOS scale we proceeded as follows. Though several techniques for detecting DIF have been proposed, most of them are based on evaluating differences in response probabilities between groups, conditional on some measure of ability. The most generally used technique is based on the Mantel–Haenszel statistic [10], others are based on log-linear models [12], on IRT models [9], or on log-linear IRT models [11]. In the Mantel–Haenszel, log-linear and log-linear IRT approaches, the difficulty level of the item is evaluated conditionally on the respondents' un-weighted sum scores. However, adopting the assumption that the un-weighted sum score is a sufficient statistic for ability (together with some technical assumptions, which will seldom be inappropriate) necessarily leads to the adoption of the Rasch model. However, with the exception of the log-linear IRT approach, the validity of the Rasch model is rarely explicitly tested. Therefore, Glas and Verhelst [8] suggested a procedure consisting of two steps, (1) searching for an IRT model for fitting the data of the sample from the reference population, and as far as possible, the sample from the focal population, (2) evaluating the differences in response probabilities between the two samples in homogeneous ability groups. The tests like Mantel–Haenszel mentioned earlier cannot be used for performing the second step of the above approach for the two-parameter logistic and nominal response model while the Lagrange Multiplier (LM) test can. Therefore, we employ the LM test which can handle the two-parameter logistic model that we use in our study.

However because of a large sample set the power of the test was large and even minor differences in model fit became significant. Though the size of the LM statistic is an indicator of the size of the DIF, a more useful measure of DIF for our immediate purposes was the residual analysis of the differences between observed and expected scores of the examinees. A set of observed and expected

responses for an item in the scale was constructed over the entire student population in a country, while controlling for the different marginal distribution in each country. The observed responses on an item in a country are the sum of responses of all students on that item in that country. The expected responses for an item in a country are the sum of the expected response for all students on that item in that country. The expected response on an item is computed as a posterior expected response given the student's response pattern, the item parameters and the population parameters of the country from which the student is sampled. The difference between the sum of the observed and expected responses on an item in a country is an indicator of item bias in that country. Summing up this difference for all the countries and then dividing it by the number of countries is an indicator of the size of the global DIF affecting that item across the set of countries in the analysis. The larger the value of this difference, the larger is the item bias or DIF across the countries. Likewise a statistic for the difference between observed and expected responses can be obtained for all the items in a scale like HOMEPOS. Next it can be seen which items have a comparatively higher difference between the observed and the expected values. Those are the items which are labelled as DIF items in our analyses. The item difficulty parameters of these DIF items are allowed to vary across countries in the measurement model to fit the data. This variation is country specific and thus the item parameters of the DIF item become country specific.

2.2.3. Scoring procedures

After obtaining the item parameters and correcting for DIF, the students can be assigned latent trait scores given the item parameters and their response patterns. In this section the theory behind the scoring procedures is discussed and in the next section the scoring process is described.

Point estimates of the students' latent trait scores are obtained given the estimates item parameters and the response pattern. However this point estimate of the latent trait score has uncertainty associated with it and using it directly in subsequent statistical analyses leads to too narrow confidence intervals. Therefore, both for the dependent and the independent variables, plausible values rather than point estimates are used to take the uncertainty of the estimates into account. Plausible values were first developed for the analyses of NAEP (National Assessment of Educational Progress) data, by Ref. [15], based on Rubin's work on multiple imputations Rubin [19]. They are draws from the posterior distribution of a student's ability given his response pattern. Therefore, plausible values provide not only information about the estimate of a variable, but also the uncertainty associated with this estimate.

The two most widely used procedures to estimate person parameters are Maximum Likelihood (ML) estimation and Expected A Posteriori (EAP) estimation. These two estimation procedures resulted out of respectively a frequentist and a Bayesian approach to estimation.

2.2.4. WML estimation

A frequentist approach to estimating the parameters of an IRT model is given by the maximum likelihood (ML) method. The likelihood function models the likelihood of a certain response pattern. If we maximising this function with respect to the person parameter we obtain the Maximum Likelihood (ML) estimate. However this ML estimate has a bias. This bias can be corrected by attaching a weight to the ML estimates of each person. Correcting for the bias by attaching a weight to the ML estimates results in unbiased weighted maximum likelihood or WML estimates [22].

2.2.5. EAP estimation

The Bayesian approach makes use of prior distributions and inferences are based on the posterior distribution. The prior distribution is a beforehand notion about the parameters, for instance about the mean and variance of the population, often based on some theoretical assumption. The posterior distribution incorporates both this prior information and the information from the data. An often noted disadvantage of Bayesian statistics is that the choice of the prior in the parameter estimation procedure is in some way subjective. However, as the sample size increases the weight of the data far outweighs that of the prior [5].

A point estimate of the latent score can be obtained by using the mode or the expectation of the posterior distribution. We use the latter, which is known as the expected-a posteriori or EAP estimate [16]. Ref. [17] recommend this approach for IRT models. The technique capitalises on an insight of Thomas Bayes which enables us to find the conditional probability of an event theta (e.g., ability) given the conditional probability of event X (e.g., response pattern of student) and the unconditional probabilities of events theta and X. The unconditional probability of theta is the prior information about a person's ability estimate. The EAP estimates are the expected value of the conditional probability distribution of theta.

2.2.6. Estimation process

The MML, WLM and EAP estimates were computed using the MIRT software [7]. Plausible values for person parameters were then drawn for the WML and the EAP approaches. The WML estimates have an asymptotic normal distribution with expectation equal to the WML estimate and variance equal to the square of the standard error. The plausible values were drawn from this normal distribution. For the EAP based plausible value, the draws are made directly from the posterior distribution without asymptotic assumptions.

2.3. The multilevel regression model

The dataset used for the analyses was the PISA 2009 International Student dataset. All the analyses were done for equal-sized samples of students from 10 culturally diverse countries. From each of the 10 countries, 1500 students were sampled. The list of these countries for which the analyses were conducted are presented in the tables. The regression analyses were done separately for each country. The regression model used for the analyses was a 2-level random intercepts model [2]. The multi-level model consisted of a student and school level. The

independent variables were the home possession variable at the student level (HOMEPOS) and the Mean of the home possessions variable at the school level (Mean MHOMEPOS). Details about the item composition of the PISA home possessions scale can be found in Appendix 1. The mean of the HOMEPOS variable was obtained by summing the HOMEPOS indices of individuals in a school and then dividing the sum over the number of students in the school. PISA studies have shown that Mean HOMEPOS explains more variance than the individual level HOMEPOS index (as is the case for other facets of socio-economic status variables) and therefore we used a multilevel model that includes the Mean HOMEPOS variable. The model that was run for the multilevel analyses is presented below:

$$\theta_{ij} = \beta_{0j} + \beta_1(\text{HOMEPOS}_{ij}) + \varepsilon_{ij} \quad (3)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{10}(\text{Mean\_HOMEPOS}_j) + \mu_j \quad (4)$$

This is a random intercepts model. The random intercept component of the model is given by Eq. (4). The subscript i represents a student and subscript j denotes the school. Thus the left hand side of Eq. (3) represents a plausible value of the reading ability of a student i in school j. These plausible values for reading ability were based on EAP estimates and the measurement model was the PCM. These plausible value draws were fully conditional draws that are conditioned on the background variables (see PISA 2009 Technical Report). Five Plausible values for reading ability were drawn from the posterior distribution based on this fully conditional model. For our analyses, these five plausible values of reading ability and five plausible values for the HOMEPOS variable (which we estimated) were used. Thus five regression analyses were done for five sets of plausible values of the dependent and independent variables, where each of the five values for the HOMEPOS variable is regressed on one of the five values for the dependent variable. The ranges of the plausible values for the dependent variable (ability estimate) were between 300 and 700 units on the PISA scale approximately. The range of plausible values for the independent variable HOMEPOS ranged between −3 and 3. The multi-level regression analyses were done five times using a different set of plausible values of the dependent and independent variable. The results for the effect size of the independent variables were then averaged by summing the effect size from the five runs and then dividing by five. The standard errors and the significance tests were then adjusted for the variation between the five sets of results. The variance of an estimate is the average of the variances of the estimate from five runs plus the between-runs variance of the five estimates [19]. The standard errors were simply the square root of the estimated variance.

For the HOMEPOS variable, different measurement estimates are used in the analyses. The measurement models used to calibrate the Home Possessions scale are the PCM and the GPCM and the scales scores used for home possessions variable are the WML and EAP based plausible values. Afterwards these estimation exercises are repeated using country specific item parameters to model DIF. The investigated procedures and their labels are listed in Tables 1–4.

3. Results

The results of the analyses are presented in Tables 1–4. Tables 1 and 2 present the results of regression analyses for the PCM and the GPCM estimates of the HOMEPOS variable without using any country specific item parameters. These different estimates are computed using WML and EAP based plausible values for HOMEPOS. The level 1 results are presented under the heading HOMEPOS and the level 2 effects under the heading Mean HOMEPOS.

The analyses presented in Tables 1 and 2 are repeated using country specific item parameters in Tables 3 and 4, which present the results for the case where eight country specific item parameters have been used to compensate for DIF items in the HOMEPOS scale.

Table 1 compares the results obtained using the PCM for the WML and EAP estimation procedures listed in the tables. Table 2 presents the same results for the GPCM. From the tables it can be seen that the level 1 effects are significant for about half of the countries. The level 2 effects are significant for all countries except Finland. The effects are systematically larger for the PCM. However the effect size for the PCM and the GPCM cannot be compared because they are on different scales. However, the ordering of the countries with respect to the size of the regression coefficients and the significances can be compared and they are very similar using the PCM and the GPCM. Another result that becomes evident is that the size of the regression effects is larger for the EAP based

Table 1
Regression coefficients and standard errors (S.E.) for HOMEPOS using the PCM with regular item parameters. Plausible values are used in this table.

Country	HOMEPOS				Mean HOMEPOS			
	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	8.2	6.8	**17.8**	8.1	**25.4**	10.3	**23.0**	10.3
Germany	**11.6**	4.2	**17.6**	6.3	**44.0**	8.9	**67.0**	13.4
Finland	**18.2**	6.4	**27.6**	12.9	4.2	13.7	14.4	18.2
Indonesia	−4.4	3.2	−5.5	5.0	**26.2**	7.4	**37.8**	8.1
Japan	4.6	7.1	4.3	8.7	**31.4**	13.5	**53.4**	20.3
Mexico	2.4	5.6	1.1	7.4	**17.2**	6.6	**25.2**	8.1
Netherlands	−2.6	6.0	−2.5	9.1	**62.4**	15.1	**118.8**	24.8
Shanghai	6.2	4.6	8.9	7.5	**40.2**	8.9	**57.4**	11.0
Thailand	4.7	4.5	3.9	5.6	**29.0**	6.1	**33.2**	6.3
United States	**13.3**	5.6	**16.0**	7.2	**31.8**	12.4	**46.4**	11.2

The significant effects (at 5%) are highlighted in bold.

Table 2
Regression coefficients and standard errors (S.E.) for HOMEPOS using the GPCM with regular item parameters. Plausible values are used in this table.

Country	HOMEPOS				Mean HOMEPOS			
	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	9.3	4.9	**11.8**	5.6	**10.8**	5.4	**16.0**	6.8
Germany	**7.0**	3.1	**14.2**	4.9	**30.8**	7.6	**46.4**	10.7
Finland	**17.0**	4.9	**25.3**	9.1	0.6	8.4	−1.7	15.7
Indonesia	−1.1	2.1	−2.8	2.9	**14.4**	5.7	**23.7**	5.8
Japan	1.4	5.3	2.8	8.4	**25.3**	10.5	**36.4**	16.7
Mexico	1.8	5.0	2.4	5.6	**13.9**	5.4	**16.8**	6.3
Netherlands	−1.8	4.2	−7.4	9.6	**35.2**	11.5	**86.1**	22.8
Shanghai	4.1	3.3	5.7	4.5	**28.9**	6.8	**40.9**	8.1
Thailand	2.3	2.8	3.4	3.1	**22.5**	4.2	**24.1**	4.2
United States	**7.6**	3.7	**12.7**	5.5	**23.1**	6.5	**35.1**	9.1

The significant effects (at 5%) are highlighted in bold.

Table 3
Regression coefficients and standard errors (S.E.) for HOMEPOS using the PCM with eight country specific (C.S.) item parameters. Plausible values are used in this table.

Country	HOMEPOS				Mean HOMEPOS			
	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	**13.0**	6.0	15.2	8.5	**17.2**	7.3	**24.0**	10.3
Germany	**11.6**	3.8	**15.6**	5.3	**35.0**	8.1	**48.0**	10.2
Finland	**19.2**	5.7	**31.4**	8.3	−1.4	8.9	−7.8	14.4
Indonesia	−2.7	2.7	−3.8	3.8	**22.0**	6.3	**30.4**	7.6
Japan	1.6	4.6	7.0	7.8	**33.6**	11.4	**48.4**	17.5
Mexico	0.2	5.8	3.2	7.2	**21.9**	6.8	**23.6**	8.3
Netherlands	1.7	3.8	0.5	6.6	**40.9**	11.7	**74.0**	13.2
Shanghai	6.6	3.4	**10.0**	4.9	**33.5**	7.0	**45.4**	9.5
Thailand	3.2	3.7	5.9	4.5	**30.9**	5.7	**34.4**	5.9
United States	**10.6**	4.4	**18.1**	6.7	**25.9**	8.6	**36.8**	10.7

The significant effects (at 5%) are highlighted in bold.

Table 4
Regression coefficients and standard errors (S.E.) for HOMEPOS using the GPCM with eight country specific (C.S.) item parameters. Plausible values are used in this table.

Country	HOMEPOS				Mean HOMEPOS			
	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	**11.6**	5.4	12.9	5.9	**13.4**	6.5	**18.0**	7.0
Germany	**9.4**	3.5	**16.2**	4.2	**34.2**	7.2	**47.4**	9.7
Finland	**19.3**	4.8	**28.2**	7.4	−0.8	8.0	0.8	13.4
Indonesia	−2.2	2.2	−2.1	2.8	**18.5**	5.7	**21.1**	5.7
Japan	5.2	4.1	7.4	5.9	**29.2**	10.8	**46.9**	13.3
Mexico	2.1	5.2	1.6	5.4	**16.1**	6.1	**19.2**	6.1
Netherlands	0.5	2.9	0.8	4.9	**33.5**	7.3	**58.3**	11.4
Shanghai	5.2	3.3	**8.9**	4.2	**29.3**	6.3	**39.6**	7.8
Thailand	3.2	3.3	2.5	3.2	**23.3**	4.7	**27.7**	4.5
United States	**7.8**	3.5	**15.0**	5.3	**23.2**	6.6	**30.9**	8.6

The significant effects (at 5%) are highlighted in bold.

estimates than the WML based estimates for both the PCM and the GPCM.

Tables 3 and 4 present the results obtained using country specific item parameters for the PCM and the GPCM.

For the PCM in Table 3, introducing country specific item parameters causes the effect sizes to change compared to Table 1. For the GPCM in Table 4, introducing country specific item parameters causes a smaller change in the effect sizes (when comparing with Table 2) than the change in effect sizes for the PCM. Figs. 4 and 5 show this for the Mean HOMEPOS variable for all the countries (except Finland for which the result was not significant).

The results showed that change for the Netherlands was especially large when using country specific item parameters for EAP estimates with the PCM or the GPCM. This is because the results for DIF showed that the Netherlands was most impacted by DIF in the HOMEPOS scale. A large number of items in the scale displayed sizeable DIF for the Netherlands during item calibration.

These results suggest that the PCM is more affected by DIF than the GPCM as country specific item parameters

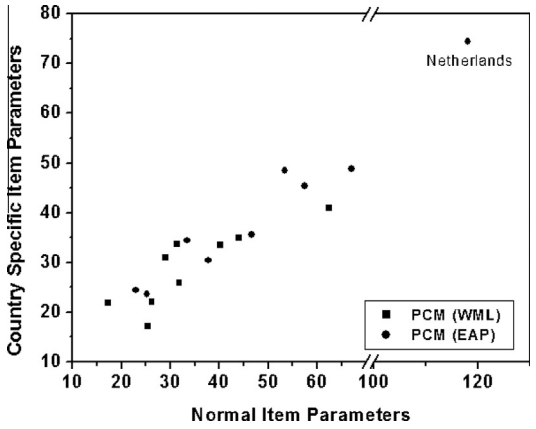

Fig. 4. Effect on regression parameters when introducing country-specific item parameters for the PCM.

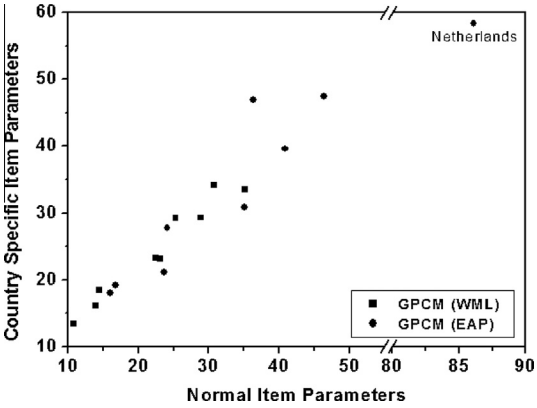

Fig. 5. Effect on regression parameters when introducing country-specific item parameters for the GPCM.

which mitigate the impact of DIF affect the PCM results more than the GPCM results.

4. Conclusions

The validity of cross national surveys is threatened by the presence of cultural bias especially for a survey like PISA which is conducted across a diverse group of countries around the world. The cultural bias can compromise comparisons across countries which are essential to cross-national surveys. According to some recent studies on cultural bias in measurement scales [3,4] response styles amongst the participating countries can be very different and researchers need to be cognizant of the possibility that the data they collect when using multi-country surveys might be influenced by the responding styles of the participating countries. According to one study, people from Brazil and China, for example, often gave extreme responses, while the Japanese leaned towards midpoint answers. That showed that researchers must not assume that countries in the same region respond in the same way; data must be studied per country or researchers must correct the data accordingly by using calibrated scales. Though the variable in question in our study is not an attitudinal variable that is influenced by the response styles of the respondents, yet there exists a cultural bias in the wealth variable being measured. For instance taking the ownership of a car or two in New York as a proxy for wealth is very different from making a similar comparison with ownership of a similar number of cars in a Mid-west US town. Likewise the number of bathrooms in the home as a proxy for wealth is not applicable when comparing a Tokyo apartment with a home in a suburban Australia. Due to this inherent bias in the measurement of wealth scale, it becomes essential to use a calibrated scale that can compensate for the cultural bias as recommended by previous studies also. In this study we investigated the impact of cultural bias in the framework of IRT modelling, which is increasingly used for measurement in survey research. We selected an important background scale from PISA 2009 to study the impact of cultural bias or DIF in the framework of IRT. We wished to see the impact of DIF on the regression model as ultimately the inferences from the background scales are made on the regression outcomes. For our analyses we investigated the impact of DIF using different measurement models and scoring methods. We did this for the PCM and the GPCM as the measurement models and the WLM and EAP scoring methods. We modeled DIF in the framework of these different approaches using country specific item parameters and analysed the results from the subsequent regression analyses.

The first thing that is evident from the analyses is that the country specific item parameters affect the results for the PCM more than the GPCM across the spectrum of analyses. Though the effect sizes for the GPCM and the PCM cannot be directly compared with each other because the two models are estimated on a different scale, however the differences produced when using regular or country specific item parameters for each of the two models can be compared. After using country specific item parameters the change in the effect sizes for the PCM was in most cases larger than the GPCM. As country specific item parameters mitigate the impact of DIF, it can be concluded that the PCM is more vulnerable to DIF than the GPCM.

Another result that becomes evident is that the change in regression coefficients when using normal versus country specific item parameters was larger when using EAP estimates than when using WML estimates of the variable Home Possessions, especially for the PCM. The difference in effect sizes was especially prominent for countries with larger DIF in the Home Possessions variable like the Netherlands.

Concluding the results, we see that DIF in the latent scale under investigation impacts the results of the regression analyses and the use of country specific item parameters helps in mitigating the impact of DIF. The PCM is more affected by DIF than the GPCM especially when using EAP estimates of person parameters, so the use of country specific item parameters has greater need if the PCM is the measurement model. For the GPCM, overall the change in effect sizes was not large when using country specific item parameters but the precision of the estimates is still

enhanced. For a country severely affected by DIF, using country specific item parameters with the GPCM using the EAP method also made a substantial difference (see Netherlands in Fig. 5).

Thus for countries affected by DIF the impact on the regression coefficients cannot be ignored. The effect has been shown to be more for the PCM than the GPCM and generally more for the EAP estimates than the WML estimates. However as the ordering of the countries and the significances (in terms of effect sizes) were quite similar for the PCM and the GPCM using both the WML and EAP estimates which is often of interest in cross national surveys, either method may be used for future analyses provided DIF is adequately accounted for with country specific item parameters.

Acknowledgements

This research was funded by the University of Twente, the Netherlands, as part of the faculty research Grants.

Appendix 1. Items in home possession scale

(Q20) Which of the following are in your home?
(Please tick one box in each row) Yes/No

(a) A desk to study at
(b) A room of your own
(c) A quiet place to study
(d) A computer you can use for school work
(e) Educational software
(f) A link to the Internet
(g) Classic literature (e.g., <Shakespeare>)
(h) Books of poetry
(i) Works of art (e.g., paintings)
(j) Books to help with your school work
(k) <Technical reference books>
(l) A dictionary
(m) A dishwasher
(n) A <DVD> player
(o) <Country-specific wealth item 1>
(p) <Country-specific wealth item 2>
(q) <Country-specific wealth item 3>

(Q21) How many of these are there at your home?
(Please tick only one box in each row) None, One, Two, Three or more

(a) Cellular phones
(b) Televisions
(c) Computers
(d) Cars
(e) Rooms with a bath or shower 4

(Q22) How many books are there in your home?

(a) 0–10 books
(b) 11–25 books
(c) 26–100 books
(d) 101–200 books
(e) 201–500 books
(f) More than 500 books

References

[1] R.D. Bock, M. Aitkin, Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, Psychometrika 46 (1981) 443–459.
[2] A.S. Bryk, S.W. Raudenbush, Hierarchical linear models: applications and data analysis, Sage, Newbury Park, CA, 2001.
[3] C. Chen, S. Lee, H.W. Stevenson, Response style and cross-cultural comparisons of rating scales among East Asian and North American students, Psychol. Sci. 6 (1995) 170–175.
[4] M.M. Chiu, B.W.-Y. Chow, C. McBride-Chang, Universals and specifics in learning strategies: explaining adolescent achievement in mathematics, science, and reading across 34 countries, Learn. Individ. Differ. 17 (2007) 344–365.
[5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, London, Chapman & Hall/CRC, Boca Raton, FL, 2004.
[6] C.A.W. Glas, Detection of differential item functioning using Lagrange multiplier tests, Stat. Sin. 8 (1998) 647–667.
[7] C.A.W. Glas, MIRT, Software program and Manual, University of Twente, Enschede, The Netherlands, 2010. http://www.utwente.nl/gw/omd/en/employees/employees/glas/.
[8] C.A.W. Glas, N.D. Verhelst, Tests of fit for polytomous Rasch models, in: G.H. Fischer, I.W. Molenaar (Eds.), Rasch models. Their Foundation, Recent Developments and Applications, Springer, New York, 1995, pp. 325–352.
[9] R.K. Hambleton, H.J. Rogers, Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods, Applied Measurement in Education 2 (1989) 313–334.
[10] P.W. Holland, D.T. Thayer, Differential item functioning and the Mantel-Haenszel procedure, in: H. Wainer, H.I. Braun (Eds.), Test Validity, Lawrence Erlbaum Associates Inc, Hillsdale, NJ, 1988.
[11] H. Kelderman, Item bias detection using loglinear IRT, Psychometrika 54 (1989) 681–697.
[12] F.G. Kok, G.J. Mellenbergh, H. van der Flier, Detecting experimentally induced item bias using the iterative logit method, Journal of Educational Measurement 22 (1985) 295–303.
[13] F.M. Lord, Applications of Item Response Theory to Practical Testing Problems, Lawrence Erlbaum, Hillsdale, 1980.
[14] G.N. Masters, A Rasch model for partial credit scoring, Psychometrika 47 (1982) 149–174.
[15] R.J. Mislevy, Estimating latent distributions, Psychometrika 49 (1984) 359–381.
[16] R.J. Mislevy, Bayes modal estimation in item response models, Psychometrika 51 (1986) 177–195.
[17] R.J. Mislevy, R.D. Bock, A hierarchical item-response model for educational testing, in: R.D. Bock (Ed.), Multilevel Analysis for Educational Data, Academic Press, San Diego, CA, 1989, pp. 57–74.
[18] E. Muraki, A generalized partial credit model: application of an EM algorithm, Appl. Psychol. Meas. 16 (1992) 159–176.
[19] D.B. Rubin, Multiple Imputations in sample surveys—a phenomenological Bayesian approach to nonresponse, Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, 1978, pp. 20–34.
[20] Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. Psychometrika, Monograph Supplement, No 17.
[21] N.D. Verhelst, C.A.W. Glas, H.H. de Vries, A steps model to analyze partial credit, in: W.J. van der Linden, R.K. Hambleton (Eds.), Handbook of modern item response theory, Springer, New York, NJ, 1997, pp. 123–138.
[22] T.A. Warm, Weighted likelihood estimation of ability in item response models, Psychometrika 54 (1989) 427–450.