Wendt *et al. Large-scale Assess Educ (2017) 5:10*
DOI 10.1186/s40536-017-0043-9

Large-scale Assessments in Education

RESEARCH Open Access

Assuming measurement invariance of background indicators in international comparative educational achievement studies: a challenge for the interpretation of achievement differences

Heike Wendt[1], Daniel Kasper[1*] and Matthias Trendtel[2]

*Correspondence: daniel.kasper@tu-dortmund.de
[1] Institute for School Development Research, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany
Full list of author information is available at the end of the article

Abstract

Background: Large-scale cross-national studies designed to measure student achievement use different social, cultural, economic and other background variables to explain observed differences in that achievement. Prior to their inclusion into a prediction model, these variables are commonly scaled into latent background indices. To allow cross-national comparisons of the latent indices, measurement invariance is assumed. However, it is unclear whether the assumption of measurement invariance has some influence on the results of the prediction model, thus challenging the reliability and validity of cross-national comparisons of predicted results.

Methods: To establish the effect size attributed to different degrees of measurement invariance, we rescaled the 'home resource for learning index' (HRL) for the 37 countries ($n = 166,709$ students) that participated in the IEA's combined 'Progress in International Reading Literacy Study' (PIRLS) and 'Trends in International Mathematics and Science Study' (TIMSS) assessments of 2011. We used (a) two different measurement models [one-parameter model (1PL) and two-parameter model (2PL)] with (b) two different degrees of measurement invariance, resulting in four different models. We introduced the different HRL indices as predictors in a generalized linear mixed model (GLMM) with mathematics achievement as the dependent variable. We then compared three outcomes across countries and by scaling model: (1) the differing fit-values of the measurement models, (2) the estimated discrimination parameters, and (3) the estimated regression coefficients.

Results: The least restrictive measurement model fitted the data best, and the degree of assumed measurement invariance of the HRL indices influenced the random effects of the GLMM in all but one country. For one-third of the countries, the fixed effects of the GLMM also related to the degree of assumed measurement invariance.

Conclusion: The results support the use of country-specific measurement models for scaling the HRL index. In general, equating procedures could be used for cross-national comparisons of the latent indices when country-specific measurement models are fitted. Cross-national comparisons of the coefficients of the GLMM should take into

account the applied measurement model for scaling the HRL indices. This process could be achieved by, for example, adjusting the standard errors of the coefficients.

Keywords: PIRLS/TIMSS combined, Invariance background models, Measurement and prediction invariance, Generalized linear mixed model, Sensitivity analyses for variance components

Background

Introduction

In order to report international trends in educational achievement over time and to compare achievement results across countries, the International Association for the Evaluation of Educational Achievement (IEA) conducts, among other studies, regular iterations of the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS). PIRLS has assessed the reading comprehension achievement of fourth-grade students every 5 years since 2001 (Mullis et al. 2012a), while TIMSS has assessed the mathematics and science achievement of fourth- and eighth-grade students every 4 years since 1995 (Martin et al. 2012). In 2011, IEA conducted both studies jointly for the first time. Thirty-four countries and three benchmark participants collected data on Grade 4 students' educational achievement in three competence domains: reading comprehension, mathematics, and science (Martin and Mullis 2013).

In their efforts to explain observed achievement differences in the data from large-scale assessment studies, researchers have increasingly combined different background indicators (Bos et al. 2012; Martin et al. 2008; Mullis et al. 2007, 2008; OECD 2014a) by scaling them into latent background variables. Scaling these variables usually requires application of an item response theory (IRT) model (Martin and Mullis 2012; OECD 2014b). The approach has several advantages, among which is the ability to control the measurement errors in the manifest variables. Controlling for measurement error is especially important in educational research studies because the multilevel prediction models commonly used in this area are very sensitive to these errors (Lüdtke et al. 2011).

Although using IRT models to scale latent background variables before including them in a prediction model works very well in large-scale assessment studies, the method presents several challenges (van den Heuvel-Panhuizen et al. 2009). First researchers wanting to use latent indices instead of manifest indicators need to develop a coherent theoretical framework for the construct they intend to measure. Second, they need to define the assessment's desired target population and the sampling procedure. Third, they need to choose not only a suitable measurement model for the construct but also a statistical model that will allow them to scale the latent indices according to this model. Finally, they must specify a useful and appropriate prediction model.

These tasks also need to be considered within the context of two central challenges that researchers face when conducting cross-national studies of educational achievement. The first centers on the need to ensure that the indices used for international comparison are comparable across the countries participating in each study (Nagengast and Marsh 2013), and the second concerns the need to ensure that the latent variables are comparable across the participating countries. Researchers conducting these large-scale

assessment studies usually endeavor to meet these challenges by assuming measurement invariance across countries when they scale the latent indices. However, as work by Millsap (1995, 1997, 1998, 2007) shows, this approach leads to inconsistent measurement invariance and predictive invariance. Thus, when researchers assume that there will be measurement invariance across countries and then, during data analysis, use the scaled latent indices as predictors in the country-specific prediction models, the prediction coefficients across countries will only be the same under very restricted conditions. However, researchers are unlikely to deem these conditions reasonable in practice. What is obvious here is that the different decisions that those designing large-scale assessment studies must make before latent indices can be used, will influence the results of these studies. Generalizability theory calls these sources of influence facets or dimensions, and emphasizes that researchers must take the variance in those research results that can be traced back to these dimensions into account before they attempt to generalize the results (Brennan 2001).

The aim of the study presented in this paper was to investigate the extent to which the assumption of cross-national measurement invariance of latent background variables affected the results of prediction models that use these indices as predictors in large-scale assessment studies. To achieve this aim, we reanalyzed the PIRLS/TIMSS 2011 data that Martin et al. (2013) used in their study on effective school environment. We considered this study especially useful for the desired purpose because Martin and colleagues used latent indices scaled under the assumption of cross-national measurement invariance as predictors in their country-specific hierarchical linear models and then compared the results of these models across the countries. We considered that reanalyzing these data sets by allowing different degrees of cross-national measurement invariance could help to answer the question of whether this assumption has an influence on (1) the cross-national comparisons performed by Martin et al. (2013) in particular, and (2) the results of large-scale assessment studies that use a design comparable to the one Martin and his colleagues employed in general. We begin by providing a summary of the study by Martin and his colleagues (2013). We then describe how we conducted our study, before presenting the results from that study and a discussion of those findings.

Assessment of Martin et al.'s study

Overview

Martin et al. (2013) performed a "school effectiveness" analysis of data from the 37 countries that participated in PIRLS/TIMSS 2011. According to Martin and his colleagues (2013, p. 111) "School effectiveness analyses seek to improve educational practice by studying what makes for a successful school beyond having a student body where most of the students are from advantaged socioeconomic backgrounds." In their analysis, Martin et al. used five school effectiveness variables and two student home background variables as predictors in the country-specific hierarchical linear models. They used students' achievement scores (reading comprehension, mathematics achievement, science achievement) as dependent variables. Because the goal of the study was to "present an analytic framework that could provide an overview of how these relationships vary across countries", (Martin et al. 2013, p. 110) the results from the hierarchical linear modeling could be assumed to be comparable across the participating countries.

One of the major findings of the study by Martin et al. (2013) was that the strength of the relationships between the school effectiveness variables and the student achievement scores decreased substantially in nearly all 37 countries when Martin et al. included the home background control variables in their models; country-specific effects were also apparent. For example, in 15 countries, only one out of the five effectiveness indicators still presented a statistically significant prediction coefficient after Martin and his colleagues had controlled for students' home background. In four countries, three prediction coefficients remained significant. If the results of these analyses were, in fact, comparable across countries, in most countries the strength of the relationships between school effectiveness variables and student achievement should be relatively weak after controlling for student home background.

However, by scaling the school effectiveness variables and the home background variables as latent variables, Martin et al. (2013) assumed measurement invariance across countries (see the next section). Thus, it is also possible that the cross-national variation of the prediction coefficients of the school effectiveness variables and the home background variables was at least partially a methodological artifact due to the general inconsistency of measurement invariance and predictive invariance. Studying the relationship between assumed measurement invariance and the observed prediction coefficients more closely therefore seems worthwhile. We accordingly decided that reanalyzing one of the data sets that Martin et al. (2013) used would be a useful exercise. We determined we could rescale one of the home background control variables (the "home resources for learning scale", hereafter HRL) while assuming different degrees of cross-national measurement invariance. We could then, in an effort to explain students' mathematics achievement, introduce the rescaled variable as a predictor in a generalized linear mixed model (GLMM).

We considered the reduction in our reanalysis to only one independent variable out of the eight and one dependent variable out of the three that Martin et al. (2013) used would lead to a valuable reduction of complexity, particularly given that no other study has yet analyzed the relationship between measurement invariance and predictive invariance in large-scale assessment study data. Therefore, nothing is known about possible interaction or compensatory effects in situations where the relationship between measurement invariance and predictive invariance affects more than one latent variable. We believed a reduced model would consequently increase the likelihood of finding such effects in the PIRLS/TIMSS data sets.

Also, because the selection of the HRL indices is somewhat arbitrary, we decided it would make sense to concentrate on the HRL variable. Many large-scale assessment studies have shown that the cross-national assumption of measurement invariance is unlikely to hold for social background variables see, for example, (Caro and Sandoval-Hernandez 2012; Hansson and Gustafsson 2013; Lakin 2012). Therefore, rescaling the HRL in a way that assumes measurement non-invariance would be consistent with the findings of this prior research. In addition, it is plausible to assume that indicators of the HRL indices will show country-specific characteristics. For example, the indicator "students have own room at home" could, in some countries, be a very important indicator with respect to differentiating students with many home resources from students with only a few home resources. However, in most of the countries participating in PIRLS/TIMSS 2011, this indicator was unlikely to be a strong one because nearly all of

the students had their own room at home. In terms of the IRT approach, this indicator should therefore show cross-national variation in the discrimination parameter.

It is useful at this point to outline the procedures on which the study of Martin et al. (2013) was based, especially those used to scale the HRL indices. This explanation may seem unnecessary given the wealth of literature on IRT models, but we consider it necessary for two reasons. First, Martin et al. did not explicitly use the term measurement invariance in their report. We are therefore left with the notion that they simply assumed there was measurement invariance. Second, a clear description of the scaling model they used is required to illustrate why we deemed it necessary to use a modified version of this model in our reanalysis. We also considered it necessary to introduce the prediction model.

Scaling procedures used to develop the HRL index

Martin et al. (2013) used, as indictors for the HRL index, three items from the PIRLS 2011 home questionnaire (the "Learning to Read Survey") given to the parents of the students who participated in the study, and two items from the PIRLS 2011 student questionnaire. The home questionnaire items were "number of children's books in the home," "highest level of education of either parent," and "highest level of occupation of either parent." The student questionnaire items were "number of books in the home" and "number of home study supports" (see Table 1). The PIRLS and TIMSS studies use these items as indicators of the economic and cultural capital of students' families (Mullis et al. 2012b). The positive association between these indicators and student achievement are evident in many of the reported findings from large-scale studies of educational achievement (see, for example, Martin et al. 2008, 2012; Mullis et al. 2007, 2008, 2012a; OECD 2014a). In line with Bourdieu 's (1986) work on cultural capital, the HRL index can thus be interpreted as a measure of students' socioeconomic and cultural home learning environments (Smith et al. 2016).

Determining an appropriate scaling procedure for the HRL index presented Martin et al. (2013) with a statistical challenge. They decided to use the partial credit model (Masters 1982; Wright and Masters 1982) to derive the index. That is, assuming $i = 1, \ldots, p$ are p items with $k_i = 0, \ldots, m_i$ response levels, then

$$\Pr_g(X_{gij} = k_i | \theta_{gj}, \boldsymbol{\xi_{gi}}) = \frac{\exp \sum_{t=0}^{k_i} (\theta_{gj} - \tau_{gti})}{\sum_{a=0}^{m_i} \exp \sum_{t=0}^{a} (\theta_{gj} - \tau_{gti})}$$

gives the probability of a response in category k_i of item i for a person j ($j = 1, \ldots, n_g$) in group g ($g = 1, \ldots, G$) with the latent value θ_{gj} and an item i with a group-specific item parameter vector $\boldsymbol{\xi}'_{\boldsymbol{gi}} = \begin{pmatrix} \tau_{g0i} & \ldots & \tau_{gm_i i} \end{pmatrix}$, where τ_{gti} is the t-th threshold location of item i in group g on a latent continuum. For identification purposes, it is usually assumed that $\tau_{g0i} = 0$ for all g and i. In addition, applications of the partial credit model frequently assume local independence. Accordingly, given the value of θ_{gj}, the item responses should be conditionally independent. This means that

$$\Pr_g(\boldsymbol{x}'_{\boldsymbol{gj}} | \theta_{gj}, \boldsymbol{\Xi_g}) = \prod_{i=1}^{p} \frac{\exp \sum_{t=0}^{k_i} (\theta_{gj} - \tau_{gti})}{\sum_{a=0}^{m_i} \exp \sum_{t=0}^{a} (\theta_{gj} - \tau_{gti})},$$

Table 1 Items of the home resources for learning scale (fourth grade) and percentage of yes responses overall countries (n = 138,103)

Item	Response option	% yes	*SE*
Number of books in the home (students)	0–10	17.1	0.24
	11–25	25.9	0.21
	26–100	31.5	0.20
	101–200	14.0	0.15
	More than 200	11.5	0.17
Number of home study supports (students)	None	12.8	0.18
	Internet connection or own room	36.9	0.19
	Both	50.3	0.23
Number of children's books in the home (parents)	0–10	24.2	0.20
	11–25	21.9	0.17
	26–50	24.7	0.16
	51–100	17.4	0.14
	More than 100	11.9	0.14
Highest level of education of either parent (parents)	Finished some primary or lower secondary or did not go to school	7.9	0.18
	Finished lower secondary	13.6	0.17
	Finished upper secondary	31.4	0.22
	Finished post-secondary education	19.1	0.16
	Finished university or higher	28.0	0.28
Highest level of occupation of either parent (parents)	Has never worked outside home for pay, general laborer, or semi-professional (skilled agricultural or fishery worker, craft or trade worker, plant or machine operator)	27.7	0.21
	Clerical (clerk or service or sales worker)	25.8	0.16
	Small business owner	13.1	0.14
	Professional (corporate manager or senior official, professional, or technician or associate professional)	33.4	0.23

where $\boldsymbol{x'_{gj}} = \left(x_{g1j} \cdots x_{gpj}\right)$ is the response vector of person j in group g and $\boldsymbol{\Xi_g}$ is a $i \times p$ block-diagonal matrix, with the item parameter vectors $\boldsymbol{\xi'_{g1}} \cdots \boldsymbol{\xi'_{gp}}$ in the diagonal.

Different procedures exist for the estimation of $\boldsymbol{\Xi_g}$ and θ_{gj}, given the observed data $\boldsymbol{X_g} = (x_{ij})_g$ (Fischer and Molenaar 1995). In order to estimate the item parameters (a procedure also know as item calibration), Martin et al. (2013) used the marginal maximum likelihood approach. According to this approach, the marginal likelihood of $\boldsymbol{X}$ in group g is

$$L_g(\boldsymbol{X}) = \prod_{j=1}^{n_g} \int_{-\infty}^{+\infty} \Pr_g(\boldsymbol{x'_j}|\theta)\phi_g(\theta)d\theta = \prod_{j=1}^{n_g}\prod_{i=1}^{p} \int_{-\infty}^{+\infty} \Pr_g(x_{ij}|\theta)\phi_g(\theta)d\theta.$$

This likelihood is maximized with respect to $\boldsymbol{\Xi}_g$, where $\phi_g(\theta)$ is the population density function for θ in group g (in the case of the HRL index, it was assumed that $\theta \sim \mathrm{N}_g(0,1)$).

For the calibration of the item parameters, Martin et al. (2013) used the combined data from the 37 countries participating in both TIMSS and PIRLS 2011, with each country contributing equally to the calibration. This was achieved by weighting each country's student data to sum up to 500. The item parameters across groups were therefore

fixed $\boldsymbol{\Xi_1} = \cdots = \boldsymbol{\Xi_{37}}$, which also meant that $\Pr_1(\boldsymbol{x'}|\theta) = \cdots = \Pr_{37}(\boldsymbol{x'}|\theta)$. Hence, Martin et al. assumed, with respect to the HRL index, measurement invariance across participating countries. If we assume that the item responses are conditionally independent across groups, then the marginal likelihood of $\boldsymbol{X}$ would be

$$L(\boldsymbol{X}) = \prod_{g}^{37} \prod_{j=1}^{n_g} \int_{-\infty}^{+\infty} w_j \Pr{}_g(\boldsymbol{x'_j}|\theta)\phi_g(\theta)d\theta, \tag{1}$$

where w_j is a person-specific weighting factor so that each country's student data sums up to 500.

Once the items have been calibrated by maximizing Eq. (1) with respect to $\boldsymbol{\Xi}$, estimators of θ can be observed by maximizing the weighted likelihood function,

$$g(\theta)L(\boldsymbol{x'}|\theta, \boldsymbol{\Xi}) = g(\theta) \prod_{i=1}^{p} \Pr(x_i|\theta, \boldsymbol{\Xi}),$$

where $g(\theta)$ is a function of the first and second partial derivatives of $L(\boldsymbol{x'}|\theta, \boldsymbol{\Xi})$ with respect to θ. The aforementioned equation is known as weighted likelihood estimation, and the resulting estimator is called Warm's likelihood estimator (WLE; Warm 1989), which has been shown to produce less bias than the unweighted maximum likelihood estimator of θ.

The prediction model used to explain student achievement

To explain the achievement differences among the fourth-grade students participating in PIRLS/TIMSS 2011, Martin et al. (2013) used the WLE estimators $\hat{\boldsymbol{\theta}}'_{\boldsymbol{g}} = \left(\hat{\theta}_{g1} \ \cdots \ \hat{\theta}_{gn_g}\right)$ of the HRL index and the average of two other indices—"early literacy tasks" and "early numeracy tasks"—as predictors in their country-specific hierarchical linear models (which they called the Home Background Control Model). For example, for a given country g, let y_{us} be the achievement value of student u in school s ($s = 1, \cdots, N_g$), $\hat{\theta}_{Hus}$ be the corresponding value on the WLE estimate of the HRL index, and $\hat{\theta}_{Eus}$ be the average value of the early literacy tasks and the early numeracy tasks indices. The combined model for explaining achievement is therefore

$$\begin{aligned} y_{us} \ &= \gamma_{00} + \gamma_{10}\hat{\theta}^*_{Hus} + \gamma_{20}\hat{\theta}^*_{Eus} + \gamma_{01}\hat{\bar{\theta}}_{Hus} + \gamma_{02}\hat{\bar{\theta}}_{Eus} \\ &\quad + \alpha_{1s}\hat{\theta}^*_{Hus} + \alpha_{2s}\hat{\theta}^*_{Eus} + u_{0s} + r_{us}. \end{aligned} \tag{2}$$

In this equation, γ represents the intercept and the fixed effects of the predictors, α are random effects representing variation in the fixed effects across schools, $\hat{\theta}^*$ are the school mean-centered WLEs, and $\hat{\bar{\theta}}$ are the school average of the respective WLEs. Note that the u and r are error terms associated with the school and the individual. Note also the assumption that y and r are normally distributed (Raudenbush and Bryk 2002).

We should mention, however, that Martin et al. (2013) only includes the random effects when there was significant variation in the relationship between the WLEs and achievement across schools and only when they could estimate this relationship reliably. Furthermore, they usually used the variance components $\sigma^2_\alpha = \text{var}(\alpha)$ and not the

coefficients for α to estimate these effects. In addition, because Martin et al. used plausible values for y, they performed all analyses five times and averaged the results according to Rubin's formulas (Rubin 1987).

Comments on Martin and colleagues' procedures

In order to address the challenges identified above, the construct underlying the HRL index needed to be based on a coherent and robust theoretical framework. Such a framework can indeed be derived by drawing on various conceptualizations of capital (Bourdieu 1986; Coleman 1988). However, because the HRL index drew on only five indicators (from the many available), it was very narrowly defined. We consider that the index would have particularly benefited from inclusion of the more reliable and valid indicators of social reproduction (Caro et al. 2014). Martin et al.'s (2013) assumption of measurement invariance also merits consideration for two reasons. First, because cross-national and comparative research in various disciplines challenges the validity of this assumption (Çetin 2010; Caro et al. 2014; Hansson and Gustafsson 2013; Schulte et al. 2013; Schulz 2005; Segeritz and Pant 2013). We assumed that at least some of the HRL indicators would show differential item functioning across the participating countries. For example, having an internet connection and/or a room of one's own may be more discriminating indicators of social status among students in southern or eastern European countries than among students in central European countries. Also, it seems prudent to conceptualize highest level of occupation of either parent in terms of the characteristics of each country. For example, a small business ownership might represent high social status in some countries but denote a broader category representing both lower and middle social status in other countries. These considerations suggest that the apparent lack of research studies on the invariance of the HRL index across countries needs to be remedied.

The second reason why critiquing the assumption of measurement invariance is critical relates to the general inconsistency of measurement invariance and predictive invariance shown in the work by Millsap (1995, 1997, 1998, 2007). Assuming that the HRL index presents no measurement invariance across countries, then the implication of that assumption is that the variance of the coefficients of the hierarchical linear model across countries is a purely methodological artifact. In addition, where this methodological variance does exist, then, according to generalizability theory (Brennan 2001) it should be added to the actual variance of the coefficients across countries (by, for example, increasing the standard errors of the coefficients). However, enacting this proviso is difficult because the size of the effect between measurement invariance and predictive invariance is presently unclear. The same can be said of the relationships between different degrees of measurement invariance, different measurement models, and other (more general) prediction models.

The concerns we have expressed here led to the following research questions:

- To what extent can measurement invariance across participating countries be assumed for the HRL index of Grade 4 students assessed in the combined PIRLS and TIMSS studies of 2011?

- If the assumption of measurement invariance does not hold, to what extent do country-specific measurement models differ?
- Is there an effect of different degrees of measurement invariance on the parameter estimates of the prediction model?
- If there is an effect, how large is it?

In an effort to answer these questions, and as already indicated, we reanalyzed some of the data that Martin et al. (2013) used in their school effectiveness study.

We began by addressing the first research question. Here, we fitted two different measurement models with two different degrees of measurement invariance to the combined data and then used well-established fit criteria to compare the resulting models. To answer the second research question, we compared the discrimination parameters of the measurement models across countries, a procedure that allowed us to derive the country-specific measurement validity of the indicators. In order to answer the third research question, we introduced the different HRL indices as predictors in generalized linear mixed models (GLMMs) where mathematics achievement was the dependent variable. By comparing the regression coefficients across countries and across different measurement models, we were able to observe both the overall effect of different degrees of measurement invariance on the prediction coefficients and the country-specific effect on the coefficients. We also analyzed the variance component, that is, the random part of the hierarchical linear model, in the same manner as we analyzed the regression coefficients. A fuller explanation of how we conducted our analyses follows.

Methods

Data

We used the combined international data sets for all countries participating in PIRLS/TIMSS 2011.[1] We then drew from these data sets, the country-specific data files named ASG***B1 and ASH***B1: *** stands for a country-specific code, ASG are the fourth-grade student background data sets and ASH are the corresponding home background data sets.[2] Our next step was to merge the different data sets, first according to countries and then according to data resources. This process resulted in a data-set that included the student background data and the home background data for $n = 166,709$ Grade 4 students across 37 participating countries.

Scaling procedure

We used Muraki's (1992) generalized partial credit model to scale the HRL index. We decided to apply this model instead of the partial credit model used by Martin et al. (2013) because it allows for modeling the different discrimination parameters of the indicators. Opportunity to model different discrimination parameters seemed to us especially important given the number of studies that show that the discrimination parameters of different social indicators vary across countries (see section "Comments

[1] The data sets are freely available under http://timss.bc.edu/timsspirls2011/international-database.html.

[2] These data sets contained all necessary variables for the analysis. For a detailed description of the data sets, see Foy (2013).

on Martin and colleagues' procedures" section). According to the generalized partial credit model, the probability of a response in category k_i ($k_i = 0, \ldots, m_i$) of item i ($i = 1, \ldots, p$) for a person j ($j = 1, \ldots, n_g$) of group g ($g = 1, \ldots, G$) is

$$\Pr_g(X_{gij} = k_i|\theta_{gj}, \boldsymbol{\xi_{gi}}) = \frac{\exp\sum_{t=0}^{k_i}\alpha_{gi}(\theta_{gj} - \tau_{gti})}{\sum_{a=0}^{m_i}\exp\sum_{t=0}^{a}\alpha_{gi}(\theta_{gj} - \tau_{gti})}, \tag{3}$$

with the latent value θ_{gj} and the item parameter vector $\boldsymbol{\xi'_{gi}} = \left(\alpha_{gi}\ \tau_{g0i}\ \ldots\ \tau_{gm_ii}\right)$, τ_{gti} is the t-th threshold location of item i in group g, and α_{gi} are the group-specific discrimination parameters of item i on a latent continuum. For identification purposes, it is usually assumed that $\tau_{g0i} = 0$ for all g and i. The partial credit model that Martin et al. used can be seen as a special case of the generalized partial credit model, with $\alpha_{gi} = c$ for all i and g (normally $c = 1$). However, the generalized partial credit model we used allowed different discrimination parameters between the items i and between the groups g.

We used the item response function (3) to estimate four different measurement models, each with different degrees of measurement invariance for the HRL index.

1. Model 1: In this model, all discrimination parameters $\alpha_{gi} = c$ and all $\tau_{gki} = \tau_{ki}$ were held constant both between the items and across the countries whereas the threshold parameters were allowed to vary between items but remain constant across countries. This model was the same as the one used by Martin et al. (2013).
2. Model 2: In contrast to Model 1, the discrimination parameters $\alpha_{gi} = c_g$ were held constant between the items but allowed to vary across the countries. However, the assumptions for the threshold structure were the same as those for Model 1.
3. Model 3: Here, discrimination parameters $\alpha_{gi} = c_i$ were allowed to vary between the items but were held constant across the countries. Again, the threshold structure remained unchanged.
4. Model 4: All discrimination parameters $\alpha_{gi} = c_{gi}$ were allowed to vary both between the items and across the countries. As before, the threshold structure remained unchanged.

According to this design, Model 1 was the most restrictive model because it assumed strict measurement invariance across the countries. Model 4 was the least restrictive model because it allowed for country-specific measurement models (at least with respect to the item discrimination parameters α_{gi}).

Table 1 depicts the items i, with their corresponding names in the international data sets, that were used for scaling the HRL index. We used the marginal maximum likelihood approach to calibrate the item parameters of the four models. After estimating the parameters, we used the maximum a posterior probability (MAP) estimate to generate the scores for θ. The following formula describes the corresponding posterior distribution of θ:

$$p_g(\theta|\boldsymbol{x'_{gj}}, \boldsymbol{\Xi_g}) \propto \Pr_g(\boldsymbol{x'_{gj}}|\theta_{gj}, \boldsymbol{\Xi_g})\phi_g(\theta),$$

Generally, this procedure results in more efficient estimates of θ than the WLE approach, especially when there are only a few items to scale ($p \leq 10$; Wang and Wang 2001).

However, the MAP bias seems slightly greater than the bias of the WLEs (at least under some circumstances). Overall, this procedure made it possible to derive four estimates of θ for every student.

Prediction model

We used the generalized linear mixed model (GLMM) as the prediction model (Zeger and Karim 1991; Karim and Zeger 1992). We chose the GLMM as the framework rather then the hierarchical linear model applied by Martin et al. (2013) because cross-national comparisons of the fixed effects from the GLMM require use of a test statistic. However, the statistic we needed was not yet available, so we developed one as part of this study. Provision of the mathematical proof of this statistic, which we based on the GLMM, is beyond the scope of this paper. We have therefore covered this matter in a separate paper (see Kasper 2017). We also selected the GLMM because the hierarchical linear model is a special case of it, which means that nothing is lost when this framework is used. Use of the GLMM framework furthermore makes it easier for readers to follow the development and proof of the test statistic in Kasper (2017), and thus check the validity of our application of this test statistic in our current study. In order to use this very general prediction model [for a detailed description of it, see, McCulloch and Searle (2001)] in our study, we needed to simplify some aspects of it. For example, because we used the plausible values of the Grade 4 students' mathematics achievement as the dependent variable and assumed the random effects were normally distributed, we could also assume that the dependent variable y_g was approximately normally distributed in accordance with the assumptions made during generation of these plausible values (Martin and Mullis 2012). This approach led to a GLMM with identity link function $g(\cdot)$, which meant that $\boldsymbol{\eta_g} = g(\mathrm{E}(\boldsymbol{y}_g)) = \mathrm{E}(\boldsymbol{y}_g)$ and

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{Z}_g\boldsymbol{\alpha}_g + e_g. \tag{4}$$

Here, $\boldsymbol{y_g}$ is a $n_g \times 1$ vector with the plausible values on mathematics achievement as the dependent variable; $\boldsymbol{X_g}$ is a $n_g \times 5$ matrix with the school mean-centered values and the school average values of θ_{Hg} and θ_{Eg} in the columns (plus a constant vector of 1s for the intercept); $\boldsymbol{\beta_g}$ is a 5×1 vector with the corresponding fixed effects; $\boldsymbol{Z_g}$ is a $n_g \times 2s$ block matrix with two block-diagonal matrices each of size $n_g \times s$ in the columns representing the random predictors; $\boldsymbol{\alpha_g}$ is a $2s \times 1$ vector with the corresponding random effects; and $\boldsymbol{e_g}$ is a $n_g \times 1$ vector of residuals.

Estimation of the coefficients of this model requires use of the pseudo-likelihood approach. However, due to the distributional assumptions about the dependent variable, $\boldsymbol{y_g}$ can be used in the pseudo-likelihood approach instead of the working variate $\boldsymbol{t_g}$. This alternative use results in a real objective function $l(\boldsymbol{\theta}_g, \boldsymbol{y}_g)$. The derived pseudo-likelihood estimates in our study were therefore formally equivalent to the restricted likelihood estimates of the fixed and random effects that Martin et al. (2013) derived in their study. Also, because we wanted to analyze the influence of different scaling procedures for the HRL index on the GLMM results, we introduced only the intercept and the slope of the HRL in the model as random effects. This meant that, unlike the study by Martin and colleagues, our study did not include a random slope for the early literacy/numeracy

task indicator. However, the random effects could still be correlated and, given the random effects, it could then be assumed that the schools were independent, resulting in

$$\begin{aligned}\boldsymbol{D_g} &= \boldsymbol{G_g} \otimes \boldsymbol{I_{gs}},\\ &= \begin{pmatrix} \sigma^2_{\alpha_0} & \sigma^2_{\alpha_0,\alpha_1} \\ \sigma^2_{\alpha_1,\alpha_0} & \sigma^2_{\alpha_1} \end{pmatrix} \otimes \boldsymbol{I_{gs}},\end{aligned}$$

where $\otimes$ is the Kronecker product and $\boldsymbol{I}_{gs}$ is a identity matrix of order s.

Outcomes

Scaling models

In order to compare the scaling models, we calculated the log-likelihood, the Bayesian information criterion (BIC; Schwarz 1978) and Akaike's information criterion (AIC; Akaike 1974) for each of the four models. We also calculated the variance of c_g across countries, the variance of c_i across items, the variance of c_{gi} across items (given country g), and the variance of c_{gi} across countries (given item i):

$$\begin{aligned} s^2_{c_g} &= \frac{\sum_{g=1}^{G}(c_g - \bar{c}_g)^2}{G-1}, & \bar{c}_g &= \frac{\sum_{g=1}^{G} c_g}{G},\\ s^2_{c_i} &= \frac{\sum_{i=1}^{p}(c_i - \bar{c}_i)^2}{p-1}, & \bar{c}_i &= \frac{\sum_{i=1}^{p} c_i}{p},\\ s^2_{c_{gi|g}} &= \frac{\sum_{i=1}^{p}(c_{gi|g} - \bar{c}_{gi|g})^2}{p-1}, & \bar{c}_{gi|g} &= \frac{\sum_{i=1}^{p} c_{gi|g}}{p},\\ s^2_{c_{gi|i}} &= \frac{\sum_{g=1}^{G}(c_{gi|i} - \bar{c}_{gi|i})^2}{G-1}, & \bar{c}_{gi|i} &= \frac{\sum_{g=1}^{G} c_{gi|i}}{G}. \end{aligned}$$

To test the hypotheses that these variances would be equal to zero, we used the χ^2-test. We also calculated the asymmetric confidence intervals for the different variance estimations. Thus, if $H_0: \sigma^2_k = t$ and s^2_k is an estimate of σ^2_k, then

$$\chi^2_\nu = \frac{\nu s^2_k}{\sigma^2_k} \quad \text{and} \quad \frac{\nu s^2_k}{\chi^2_{\alpha/2}} \leq \sigma^2_k \leq \frac{\nu s^2_k}{\chi^2_{1-\alpha/2}},$$

with ν degrees of freedom. However, because $t = 0$ is not a testable assumption, it was necessary to choose small values $t > 0.000$ for the respective χ^2-calculations.

Comparison of the conditional variances $s^2_{c_{gi|g}}$ and $s^2_{c_{gi|i}}$ required use of two further approaches. The first involved calculation of the overall variances

$$\begin{aligned} s^2_{c_{gi.g}} &= \frac{\sum_{g=1}^{G}(s^2_{c_{gi|g}} - \bar{s}^2_{gi.g})^2}{G-1}, & \bar{s}^2_{gi.g} &= \frac{\sum_{g=1}^{G} s^2_{c_{gi|g}}}{G},\\ s^2_{c_{gi.i}} &= \frac{\sum_{i=1}^{p}(s^2_{c_{gi|i}} - \bar{s}^2_{gi.i})^2}{p-1}, & \bar{s}^2_{gi.i} &= \frac{\sum_{i=1}^{p} s^2_{c_{gi|i}}}{p}, \end{aligned}$$

and then (by using the above-mentioned χ^2-test and confidence intervals) testing of the hypothesis $H_0: \sigma^2_{c_{gi.g}} = \sigma^2_{c_{gi.i}} = 0$. The second approach, used whenever the results of

these overall tests were significant, required multiple comparisons of $s^2_{c_{gi|g}}$ across countries and of $s^2_{c_{gi|i}}$ across items. We performed these comparisons by using $[G!/(G-2)!2!]$ -times and $[p!/(p-2)!2!]$-times the F-ratio:

$$F_{v1,v2} = \frac{s^2_{c_{gi|x}}}{s^2_{c_{gi|y}}}, \quad |\forall_{x<y} \in K \vee \forall_{x<y} \in L,$$

with $K := \{1, \ldots, G\}$ and $L := \{1, \ldots, p\}$, assuming that the variances are ordered by decreasing size.

Prediction model

To obtain an indication of the effect that the different scaling models had on the fixed and random effect coefficients of the GLMM, we performed different analyses. We based the analyses for the fixed effects on F and χ^2 tests. Thus, if $\hat{\boldsymbol{\beta}}_{\boldsymbol{gz}}$ are the estimated fixed effects for country g and scaling model $z (z = 1, \ldots, 4)$, then the hypothesis that a linear combination of the difference of the fixed effects between two scaling models w and q $(w \neq q)$ equals a constant value $\boldsymbol{m}$, that is $\mathrm{H}_0 : \boldsymbol{L_g}(\boldsymbol{\beta_{gw}} - \boldsymbol{\beta_{gq}}) = \boldsymbol{m}$, can be tested with

$$F = \frac{\left[\boldsymbol{L}\hat{\beta}_{\boldsymbol{diff}} - \boldsymbol{m}\right]' \left[\boldsymbol{L_g}\left(\Sigma_{\hat{\beta}_{gw}} + \Sigma_{\hat{\beta}_{gq}}\right)L_g'\right]^{-1} \left[\boldsymbol{L}\hat{\boldsymbol{\beta}}_{\boldsymbol{diff}} - \boldsymbol{m}\right]}{\mathbf{r}(\boldsymbol{L_g})\hat{\sigma}_g^2},$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{diff}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{gw}} - \hat{\boldsymbol{\beta}}_{gq}$ and $\hat{\sigma_g}^2 = (\hat{\sigma_g}^2_w + \hat{\sigma_g}^2_q)/2$ is the pooled residual variance estimate for the separate GLMM models w and q. Under the null hypothesis, the test statistic is noncentral F-distributed with r $(\boldsymbol{L_g})$ and n_g degrees of freedom [the proof is given in Kasper (2017)]. The F-statistic is calculated for each country separately under the assumption that the difference of the fixed effects between each non-redundant pair of scaling models is zero, that is, $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{m} = \mathbf{0}$.

In addition to analyzing the global tests of significant difference between the fixed effects, we analyzed the variances of the respective fixed effects across scaling models (given a country) and the variance of the fixed effects across countries (given a fixed effect). Thus, if $\hat{\beta}_{jgz}$ is the estimated fixed effect for predictor j $(j = 1, \ldots, 5)$ in country g given scaling model z, then the variance

$$s^2_{\hat{\beta}_{jgz.g}} = \frac{\sum_{z=1}^{4}(\hat{\beta}_{jgz} - \bar{\beta}_{jg.g})^2}{4-1}, \quad \bar{\beta}_{jg.g} = \frac{\sum_{z=1}^{4}\hat{\beta}_{jgz}}{4},$$

is calculated for every combination of j and g. The hypotheses $\mathrm{H}_0 : \sigma^2_{\hat{\beta}_{jgz.g}} = t$ are then tested with $\chi^2_v = v s^2_{\hat{\beta}_{jgz.g}}/t$, where $v = 4 - 1$ are the respective degrees of freedom for this test. We next calculated the variance of the fixed effects across countries (given scaling model z). Here, the variances

$$s^2_{\hat{\beta}_{jgz.z}} = \frac{\sum_{g=1}^{G}(\hat{\beta}_{jgz} - \bar{\beta}_{jgz.z})^2}{G-1}, \quad \bar{\beta}_{jgz.z} = \frac{\sum_{g=1}^{G}\hat{\beta}_{jgz}}{G},$$

are separately calculated for every combination of j and z, and then the hypotheses $H_0 : \sigma^2_{\hat{\beta}_{jgz.z}} = t$ are tested with $\chi^2_v = v s^2_{\hat{\beta}_{jgz.z}}/t$, where $v = G - 1$ are the respective degrees of freedom for this test.

As with the analysis of the slope coefficients, whenever significant results emerged from these overall tests, we performed multiple comparisons of $s^2_{\hat{\beta}_{jgz.g}}$ across countries and of $s^2_{\hat{\beta}_{jgz.z}}$ across fixed effects by using $[G!/(G-2)!2!]$-times and $[j!/(j-2)!2!]$-times the F-ratio

$$F_{v1,v2} = \frac{s^2_{\hat{\beta}_{jgz.x}}}{s^2_{\hat{\beta}_{jgz.y}}}, \quad |\forall_{x<y} \in K \vee \forall_{x<y} \in L,$$

with $K := \{1, \ldots, G\}$ and $L := \{1, \ldots, p\}$, assuming that the variances are ordered by decreasing size.

We used structural equation models to analyze the random effect coefficients. Here, the hypothesis that the covariance matrices of the random effect coefficients $\boldsymbol{D_g} = \boldsymbol{G_g} \otimes \boldsymbol{I_{gs}}$, given a country g is equal across scaling models, that is, $H_0 : \boldsymbol{G_{g1}} = \cdots = \boldsymbol{G_{g4}}$, can be tested by calculating the overall discrepancy function value

$$\begin{aligned} F_g(\boldsymbol{\theta}) &= \sum_{z=1}^{4} t_{gz} F_{gz}(\boldsymbol{\theta}) \\ &= \frac{t_{g1}}{2} Tr\left[\boldsymbol{G_{g1}^{-1}}\left(\boldsymbol{G_{g1}} - \boldsymbol{\Sigma_{g1}}\right)\right]^2 + \cdots + \frac{t_{g4}}{2} \text{Tr}\left[\boldsymbol{G_{g4}^{-1}}\left(\boldsymbol{G_{g4}} - \boldsymbol{\Sigma_{g4}}\right)\right]^2, \end{aligned}$$

with the restriction $\boldsymbol{\Sigma_{g1}} = \cdots = \boldsymbol{\Sigma_{g4}}$ and $t_{gz} = (n_g - 1)/(4n_g - 4)$. Under the null hypothesis, the overall discrepancy function value is approximately chi-square distributed $\chi^2_F \approx v F_g(\boldsymbol{\theta})$ with v degrees of freedom. Significant χ^2_F statistics therefore lead to rejection of the hypothesis that the scaling procedure has no influence on the random effect coefficients.

Dealing with missing values, weighting and software

We used a Markov chain Monte Carlo (MCMC) method to impute missing values in the indicators of the HRL indices. The imputation model included all indicators of the HRL indices and the plausible values of mathematics achievement, and so produced five complete data sets. Of course, a fully nested imputation strategy would have resulted in 25 imputed data sets (e.g., for each plausible value, five imputed data sets). However, because Martin et al. (2013) applied only a single imputation strategy (which seemed to us an inaccurate approach of conducting an analysis involving analysis of the variance), an increase from 1 to 25 imputations would have made it impossible to compare the results of this current paper with Martin and colleagues' results. Every analysis in our study was performed once for every completed data-set, and then the results were averaged according to Rubin's (1987) formula. *Senwgt* was used as the weighting variable for the scaling models. Senwgt summed up to a total sample size of students $n_g = 500$ for every country and so led to the equal weighting of the countries in the scaling process. The GLMM analysis, however, uses *houwgt*, which sums up to the observed sample size of students for every country. Unless we state otherwise in this paper, all the analyses in

our study were generated by way of Statistical Analysis System (SAS) software, Version 9.4 (TS1M1) of the SAS System for Windows.[3] We used the procedure MI to carry out the multiple imputations, the procedure IRT to scale the HRL index, the procedure GLIMMIX for the GLMM analysis, and the procedure CALIS for the structural equation models. We used the IML-module insight of SAS to implement the derived test statistics.

Results

Descriptive statistics

Table 1 shows the percentage of *yes* responses on the HRL scale items for the total sample of Grade 4 students $n = 138,103$.[4] Overall, the responses of the students were equiproportionally distributed across the response category of the items. However, a highly skewed distribution was evident for the indicator "number of home study supports": over 50% of the students had both an internet connection and their own room at home. Thus, for the majority of the students, this indicator provided no useful information. Also noteworthy is the relatively low percentage (7.9%) of parents who had completed only some primary or lower-secondary education or who had not attend school.

In order to verify that we had correctly implemented the scaling models, we replicated the original HRL index that Martin et al. (2013) used. Table 2 presents the descriptive statistics for these replicated values together with the newly created HRL indices, average student mathematics achievement scores, student sample sizes, and school sample sizes. The correlation between the original HRL index and the replicated HRL index (RP) was $r = 0.97$, suggesting that the scaling models were correctly implemented in this study (Table 3 shows the correlations between the other indices).[5]

When we compare the average values on the different HRL indices across the scaling models, we observed, on average, only small changes between the different indices per country. However, some noteworthy exceptions were apparent. These included changes of around 0.3 points for Germany, Honduras, Hungary, and Poland. Hence, for these countries, the influence of the scaling model on the average HRL indices was approximately one-third of a standard deviation of this index. For Malta, the influence of the scaling model on the average HRL indices was even more pronounced, at approximately two-thirds of a standard deviation of the HRL index.

Scaling models

We based our assessment of the accuracy of the four different measurement models used to scale the HRL index on three criteria: the log-likelihood (the higher the value, the better the fit), the AIC, and the BIC (the smaller the value, the better the fit).

[3] Copyright © 2002-2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

[4] Due to iteration problems, the GLMM could not be fitted to nine countries: Botswana, Dubai (UAE), Hong Kong SAR, Northern Ireland, Norway, Quebec (Canada), Russian Federation, and United Arab Emirates. The student samples from these countries were therefore not used in this study.

[5] Note that the newly created HRL indices were not, as was the case with the original HRL index, transformed to an $N \sim (10.03, 1.82)$ metric. Instead, we left the scaling metric $N \sim (0, 1)$ unchanged. We chose to do this because the transformation that Martin et al. (2013) applied made sense when the latent variable was measured on the same scale, that is, when measurement invariance between countries was assumed. When country-specific models were assumed for the HRL index, some equating procedures between the country-specific distributions of the HRL index first had to be applied to make the transformation of these values meaningful. However, analyzing the influence of different equating procedures on the HRL index and thus on the GLMM results was beyond the scope of this paper.

Table 2 Descriptive statistics for mathematics achievement, early literacy/numeracy tasks, and home resources for learning index under different scaling models

Country	Variable															
	N		MAT		ET		HRL									
							RP		M1		M2		M3		M4	
	ST	SL	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Azerbaijan	4871	169	458.1	6.05	9.5	0.08	9.0	0.03	−0.7	0.02	−0.7	0.02	−0.6	0.02	−0.7	0.01
Australia	5943	280	514.2	2.94	9.4	0.03	11.2	0.03	0.6	0.02	0.6	0.02	0.6	0.02	0.5	0.02
Austria	4587	158	505.2	2.65	9.3	0.03	10.5	0.05	0.3	0.03	0.2	0.03	0.3	0.03	0.1	0.03
Chinese Taipei	4265	150	590.3	1.86	11.2	0.02	10.3	0.05	0.2	0.03	0.1	0.03	0.2	0.03	−0.0	0.03
Croatia	4545	152	486.1	1.89	10.5	0.03	10.0	0.04	−0.0	0.02	−0.2	0.02	−0.0	0.02	−0.2	0.03
Czech Republic	4433	177	508.3	2.48	9.9	0.03	10.6	0.04	0.3	0.02	0.2	0.02	0.3	0.02	0.1	0.03
Finland	4541	145	543.5	2.33	10.4	0.04	11.2	0.03	0.7	0.02	0.6	0.02	0.6	0.02	0.5	0.02
Georgia	4774	173	444.2	3.54	9.8	0.05	10.1	0.05	−0.0	0.03	−0.1	0.03	0.0	0.03	−0.2	0.03
Germany	3928	197	524.8	2.21	9.5	0.03	10.5	0.05	0.3	0.03	0.2	0.03	0.3	0.03	−0.0	0.03
Honduras	3830	147	388.7	5.71	10.7	0.05	8.1	0.07	−1.1	0.04	−0.7	0.02	−1.1	0.04	−0.7	0.02
Hungary	5149	149	512.4	3.40	9.3	0.03	10.1	0.07	0.1	0.04	−0.0	0.04	0.0	0.04	−0.2	0.04
Iran, Islamic Rep. of	5734	244	423.5	3.53	9.6	0.06	8.7	0.07	−0.7	0.04	−0.6	0.03	−0.7	0.04	−0.6	0.03
Ireland	4383	150	525.8	2.82	9.4	0.03	10.8	0.05	0.4	0.03	0.3	0.03	0.4	0.03	0.3	0.03
Italy	4125	202	503.8	2.71	9.2	0.02	9.9	0.04	−0.1	0.02	−0.2	0.02	−0.1	0.02	−0.3	0.02
Lithuania	4584	154	531.5	2.63	10.1	0.04	10.1	0.04	0.0	0.02	−0.1	0.02	0.0	0.02	−0.1	0.02
Malta	3492	96	491.8	1.27	10.2	0.03	10.3	0.02	0.2	0.01	0.1	0.01	0.1	0.01	−0.4	0.01
Morocco	7614	284	321.9	4.02	9.7	0.08	8.2	0.04	−1.0	0.02	−0.8	0.02	−1.0	0.02	−0.8	0.02

Table 2 continued

Country	Variable															
	N		MAT		ET		HRL									
							RP		M1		M2		M3		M4	
	ST	SL	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Oman	10,237	327	376.5	2.94	10.6	0.03	9.2	0.03	−0.5	0.02	−0.6	0.01	−0.5	0.02	−0.6	0.01
Poland	4962	150	477.0	2.26	9.9	0.04	10.1	0.05	0.1	0.03	−0.1	0.03	0.0	0.03	−0.2	0.03
Qatar	4104	166	406.4	3.46	10.8	0.03	10.4	0.04	0.2	0.02	0.1	0.02	0.2	0.02	0.3	0.02
Romania	4643	148	476.4	6.01	9.6	0.10	9.2	0.06	−0.5	0.03	−0.5	0.03	−0.5	0.03	−0.5	0.03
Saudi Arabia	4470	171	403.0	5.31	10.6	0.08	9.5	0.06	−0.4	0.03	−0.4	0.03	−0.3	0.03	−0.4	0.03
Singapore	6208	176	606.4	3.35	11.3	0.04	10.8	0.03	0.4	0.02	0.3	0.02	0.4	0.02	0.3	0.02
Slovak Republic	5561	197	503.1	3.93	9.0	0.04	10.1	0.05	0.1	0.03	−0.1	0.03	0.1	0.03	−0.1	0.03
Slovenia	4433	195	509.8	1.99	9.3	0.03	10.5	0.03	0.3	0.02	0.2	0.02	0.3	0.02	0.1	0.02
Spain	4105	151	478.8	2.81	10.6	0.04	10.3	0.05	0.2	0.03	0.1	0.03	0.2	0.03	−0.1	0.03
Sweden	4482	152	502.0	2.20	10.3	0.04	11.2	0.04	0.7	0.02	0.6	0.02	0.7	0.02	0.5	0.02
Abu Dhabi, UAE	4100	164	409.7	4.91	10.5	0.04	10.1	0.06	−0.0	0.03	−0.1	0.03	0.0	0.03	0.1	0.03

ST, student sample size; SL, school sample size; MAT, mathematics achievement of fourth grade; ET, early literacy/numeracy tasks index; RP, HRL index replicating the scaling model from Martin et al. (2013); M1, HRL index using scaling model 1; M2, HRL index using scaling model 2; M3, HRL index using scaling model 3; M4, HRL index using scaling model 4

Table 3 Correlations between the different HRL indices and mathematics achievement of fourth-grade students (average values across countries)

Variable	M1	M2	M3	M4	MAT
RP	0.99	0.99	1.00	0.97	0.41
M1		1.00	0.99	0.95	0.40
M2			0.99	0.95	0.40
M3				0.97	0.41
M4					0.40

RP, HRL index replicating the scaling model from Martin et al. (2013); M1, HRL index using scaling model 1; M2, HRL index using scaling model 2; M3, HRL index using scaling model 3; M4, HRL index using scaling model 4; MAT, mathematics achievement of fourth graders

According to these criteria, the model that best fitted that data was the least restrictive scaling Model—Model 4 (Table 4). We observed virtually no difference for Models 2 and 3. Model 1 (strict measurement invariance across the countries) had the worst fit. The analyses therefore support the assumption of country-specific scaling models for the HRL index and challenge the assumption of cross-national invariance of the HRL index.

With respect to the differential estimation of the fit of the four models, Table 5 shows the distribution of the varying discrimination parameters c_g, c_i and c_{gi}. When strict measurement invariance was assumed (Model 1), the estimated discrimination parameter was $c = 1.55$. When the discrimination parameter was allowed to vary across countries but was still constant between items (Model 2), cross country variance in this parameter (c_g) was observed ($s^2_{c_g} = 0.31$; $CI_l = 0.19, CI_u = 0.57$). In some countries (e.g., Australia, Ireland, Morocco, Romania), the HRL index measured the underlying construct with a higher degree of separation when a more country-specific scaling model was used. In other countries (e.g., Czech Republic, Georgia, Germany, Malta, Qatar, Slovenia), the differentiation became less distinct. Hence, in the first instance, the original HRL index underestimated the difference in HRL for Grade 4 students whereas in the second instance the original HRL index overestimated this difference.

With regard to the assumption that the contribution of the HRL items to the HRL index would vary while the influence of the items remained constant across countries (c_{gi}), we found that the indicator "number of home study supports" was least informative with respect to the measured construct. This result supports the findings from the descriptive statistics: having a connection to the internet and/or one's own room at home seem to have been standards and not exceptions for the fourth-grade students both within and across the countries participating in PIRLS/TIMSS 2011. The educational

Table 4 Model fit statistics for the partial credit model of the HRL index

Fit-statistics	Model			
	1	2	3	4
Log likelihood	−90,588.33	−89,914.61	−90,360.41	−88,523.81
AIC (smaller is better)	181,212.66	179,919.22	180,764.82	177,361.61
BIC (smaller is better)	181,389.71	180,361.83	180,981.21	178,905.83

Model 1, constant discrimination parameter across countries and items; Model 2, constant discrimination parameters across items but vary across countries; Model 3 discrimination parameters are constant across countries but vary over items; Model 4, discrimination parameters vary across countries and across items

Table 5 Distribution of slope parameters c_g, c_i and c_{gi} for the indicators of the HRL index

Item	Country																			
	1		2		3		4		5		6		7		8		9		10	
	c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}	
	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*
ASBG04	1.08	0.10	3.09	0.20	0.99	0.10	1.86	0.12	1.32	0.11	1.20	0.11	1.58	0.12	1.15	0.11	1.05	0.11	1.40	0.12
ASDG05	0.92	0.14	2.23	0.14	0.52	0.11	2.68	0.16	0.91	0.15	0.50	0.12	0.47	0.11	0.21	0.10	0.42	0.12	1.27	0.13
ASBH15	1.58	0.12	3.43	0.23	0.97	0.10	2.42	0.16	1.90	0.13	1.41	0.12	1.90	0.14	1.25	0.11	0.98	0.10	1.42	0.12
ASDHED	1.98	0.16	6.15	0.30	2.20	0.14	0.78	0.11	1.40	0.13	1.14	0.12	1.63	0.12	1.33	0.11	1.20	0.12	0.97	0.11
ASDHOC	1.86	0.17	2.55	0.16	1.83	0.15	2.73	0.20	1.27	0.13	1.17	0.13	1.78	0.16	1.77	0.15	1.79	0.17	2.38	0.21
c_g	1.47	0.07	3.05	0.12	1.14	0.06	1.83	0.08	1.43	0.07	1.07	0.06	1.40	0.07	1.02	0.06	1.03	0.06	1.35	0.06
Item	**Country**																			
	11		**12**		**13**		**14**		**15**		**16**		**17**		**18**		**19**		**20**	
	c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}	
	Est.	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***
ASBG04	1.06	0.10	1.38	0.11	2.38	0.14	1.21	0.11	1.03	0.10	0.50	0.10	3.00	0.17	1.13	0.11	0.89	0.09	0.13	0.11
ASDG05	0.40	0.11	0.66	0.11	2.16	0.13	0.56	0.11	0.82	0.12	-0.12	0.10	2.89	0.15	1.94	0.14	0.78	0.11	0.21	0.11
ASBH15	1.03	0.11	1.40	0.11	2.42	0.15	1.53	0.12	1.22	0.11	-0.02	0.11	2.92	0.17	1.95	0.15	0.89	0.09	0.08	0.12
ASDHED	2.84	0.19	4.27	0.24	4.22	0.21	2.12	0.16	1.46	0.12	3.15	0.17	5.01	0.23	3.44	0.19	3.09	0.17	3.83	0.28
ASDHOC	1.34	0.12	2.46	0.18	2.69	0.18	2.03	0.18	2.27	0.19	1.86	0.14	2.17	0.14	1.67	0.13	2.58	0.19	2.07	0.18
c_g	1.26	0.06	1.74	0.08	2.54	0.10	1.44	0.07	1.20	0.06	1.00	0.06	3.00	0.09	1.83	0.08	1.37	0.06	1.06	0.06

Table 5 continued

Item	Country																	
	21		22		23		24		25		26		27		28		c_i	
	c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}		c_{gi}			
	Est.	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***
ASBG04	2.02	0.13	1.31	0.12	0.95	0.10	1.24	0.11	0.86	0.10	0.98	0.10	1.28	0.10	0.48	0.12	1.48	0.03
ASDG05	1.54	0.12	1.22	0.12	0.27	0.10	0.88	0.12	0.30	0.11	0.34	0.10	1.35	0.17	0.15	0.11	1.02	0.03
ASBH15	2.13	0.14	1.64	0.16	1.25	0.11	1.14	0.10	1.03	0.10	1.03	0.10	1.89	0.13	0.59	0.14	1.65	0.04
ASDHED	2.82	0.15	3.44	0.21	2.29	0.18	1.76	0.12	1.38	0.12	3.32	0.20	1.87	0.15	3.25	0.24	1.89	0.04
ASDHOC	3.15	0.23	1.38	0.12	2.01	0.18	2.04	0.17	1.81	0.16	1.95	0.16	2.10	0.19	1.75	0.16	1.79	0.04
c_g	2.04	0.08	1.62	0.07	1.24	0.06	1.28	0.06	1.00	0.06	1.31	0.06	1.68	0.08	1.16	0.06	1.55	0.01

Values for Model M1 in the last two columns and the last row, values for Model M2 in the last row of each table, values for Model M3 in the last two columns and values for Model M4 in each of the 28 country columns. ASBG04, number of books in the home (student-reported); ASDG05, number of home study supports (student-reported); ASBH15, number of children's books in the home (parent-reported); ASDHED, highest level of education of either parent (parent-reported); ASDHOC, highest level of occupation of either parent (parent-reported); 1, Azerbaijan; 2, Australia; 3, Austria; 4, Chinese Taipei; 5, Croatia; 6, Czech Republic; 7, Finland; 8, Georgia; 9, Germany; 10, Honduras; 11, Hungary; 12, Iran, Islamic Rep. of; 13, Ireland; 14, Italy; 15, Lithuania; 16, Malta; 17, Morocco; 18, Oman; 19, Poland; 20, Qatar; 21, Romania; 22, Saudi Arabia; 23, Singapore; 24, Slovak Republic; 25, Slovenia; 26, Spain; 27, Sweden; 28, Abu Dhabi, UAE

status of the students' parents best explained the differences in the HRL index. The duality between parents' educational status and number of home study supports increased when the country-specific measurement models (c_{gi}) were assumed (Model 4). In this case, parents' highest educational level contributed to the HRL index in most countries approximately two to four times more than the number of home study supports did. This finding suggests that the original HRL index did overestimate the influence of all indicators, with the exception of "highest level of education of either parent" (the influence of which, in turn, was underestimated).

However, if we take a closer look at the distribution of the item-specific discrimination parameters across countries, that is, the variance of c_{gi} given item i, then it becomes obvious that the strong discriminating effect of parental highest educational level was not constant across countries (Table 6). The discrimination parameter was exceptionally high for Australia, Iran (Islamic Rep. of), Ireland, Malta, Morocco, Oman, Qatar, Saudi Arabia, Spain, and Abu Dhabi (United Arab Emirates; UAE) and lowest for Chinese Taipei and Honduras. Despite this indicator working very well for most (if not all) countries, it worked better in some of these countries than in others. The reverse was also observable for the low discriminating power of the number of home study supports: overall, this indicator differentiated poorly among Grade 4 students. Nonetheless, we could still observe a slight discrimation capacity in some countries (i.e., Australia, Chinese Taipei, Ireland, Morocco, Oman), although virtually no discriminating capacity in several other countries [i.e., Georgia, Germany, Hungary, Malta, Qatar, Singapore, Slovenia, Spain, Abu Dhabi (UAE)]. The psychometric property of the indicator "highest level of education of either parent" exhibited the strongest discriminating capacity across most countries. These findings can perhaps be attributed to challenges to the cross-national validity of these indicators.

Finer-grained detail about the country-specific discriminating power of the HRL indicator became evident when we inspected the variance of the discrimination parameter c_{gi} across items given country g (Table 7). We observed highly differential discrimination parameters for the items for Qatar, Australia, Iran (Islamic Rep. of), Malta, Abu Dhabi (UAE), Spain, Poland, and Morocco. In these countries, parental highest educational level had the strongest influence on the HRL index. However, in most of the remaining countries (around two-thirds), the variance across the estimated item discrimination

Table 6 Variance of the discrimination parameter c_{gi} across countries (given item i), χ^2 -value and asymmetric confidence interval (CI_l lower bound, CI_u upper bound; items ordered in descending order of $s^2_{gi|i}$)

Item	$s^2_{gi\|i}$	χ^2	CI_l	CI_u	
ASDHED	1.75	11,800.07	1.09	3.24	
ASDG05	0.63	4285.24	0.40	1.18	
ASBH15	0.58	3896.91	0.36	1.07	
ASBG04	0.44	2966.86	0.27	0.81	
ASDHOC	0.22	1498.92	0.14	0.41	▲

▲, item variance is statistically different from ASDHED; ASBG04, number of books in the home (student-reported); ASDG05, number of home study supports (student-reported); ASBH15, number of children's books in the home (parent-reported); ASDHED, highest level of education of either parent (parent-reported); ASDHOC, highest level of occupation of either parent (parent-reported)

Table 7 Variance of the discrimination parameter c_{gi} across items (given country g), χ^2-value and asymmetric confidence interval (CI_l lower bound, CI_u upper bound; countries ordered in descending order of $s^2_{gi|g}$)

Country	$s^2_{gi\|g}$	χ^2	CI_l	CI_u	
Qatar	2.75	2748.28	0.99	22.69	
Australia	2.42	2424.20	0.87	20.02	
Iran, Islamic Rep. of	1.98	1975.20	0.71	16.31	
Malta	1.97	1971.18	0.71	16.28	
Abu Dhabi, UAE	1.62	1617.25	0.58	13.35	
Spain	1.34	1337.09	0.48	11.04	
Poland	1.21	1208.22	0.43	9.98	
Morocco	1.13	1132.15	0.41	9.35	
Saudi Arabia	0.88	878.82	0.32	7.26	
Hungary	0.83	826.57	0.30	6.83	
Oman	0.73	734.58	0.26	6.07	
Ireland	0.69	690.84	0.25	5.70	
Singapore	0.66	663.00	0.24	5.47	
Chinese Taipei	0.66	660.04	0.24	5.45	
Austria	0.48	476.37	0.17	3.93	
Romania	0.42	415.65	0.15	3.43	▲
Italy	0.41	409.38	0.15	3.38	▲
Finland	0.33	330.66	0.12	2.73	▲△
Georgia	0.33	327.40	0.12	2.70	▲△
Slovenia	0.32	319.78	0.11	2.64	▲△
Lithuania	0.31	314.68	0.11	2.60	▲△
Honduras	0.28	278.98	0.10	2.30	▲△●
Germany	0.24	238.93	0.09	1.98	▲△●
Slovak Republic	0.22	223.86	0.08	1.85	▲△●
Azerbaijan	0.21	218.44	0.08	1.80	▲△●
Sweden	0.13	132.81	0.05	1.10	▲△●○
Croatia	0.13	126.27	0.05	1.04	▲△●○
Czech Republic	0.12	118.46	0.05	0.98	▲△●○

▲, country variance is statistically different from Qatar; △, country variance is statistically different from Australia; ●, country variance is statistically different from Iran, Islamic Rep. of, Malta and Abu Dhabi, UAE; ○, country variance is statistically different from Spain, Poland, Morocco, Saudi Arabia and Hungary

parameters was moderate or even low, indicating that the assumption of a one-dimensional construct for the HRL index was acceptable for these countries. Nevertheless, the observed significant difference in $s^2_{gi|g}$ across the countries participating in PIRLS/TIMSS 2011 again confirms the assumption of measurement non-invariance of the HRL index, with that non-invariance apparently mostly attributable to the indicators of the highest level of education of either parent and the number of home study supports.

Prediction model

Figure 1 shows the distributions of the estimated fixed effects across countries for the different scaling models. Noticeably, there were no differences in the distribution for the fixed effects $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. It seems that the different scaling procedures used for the HRL index left untouched all the fixed effects that were not associated with the HRL index. However, the effects of the scaling model on the distribution of the fixed effects across

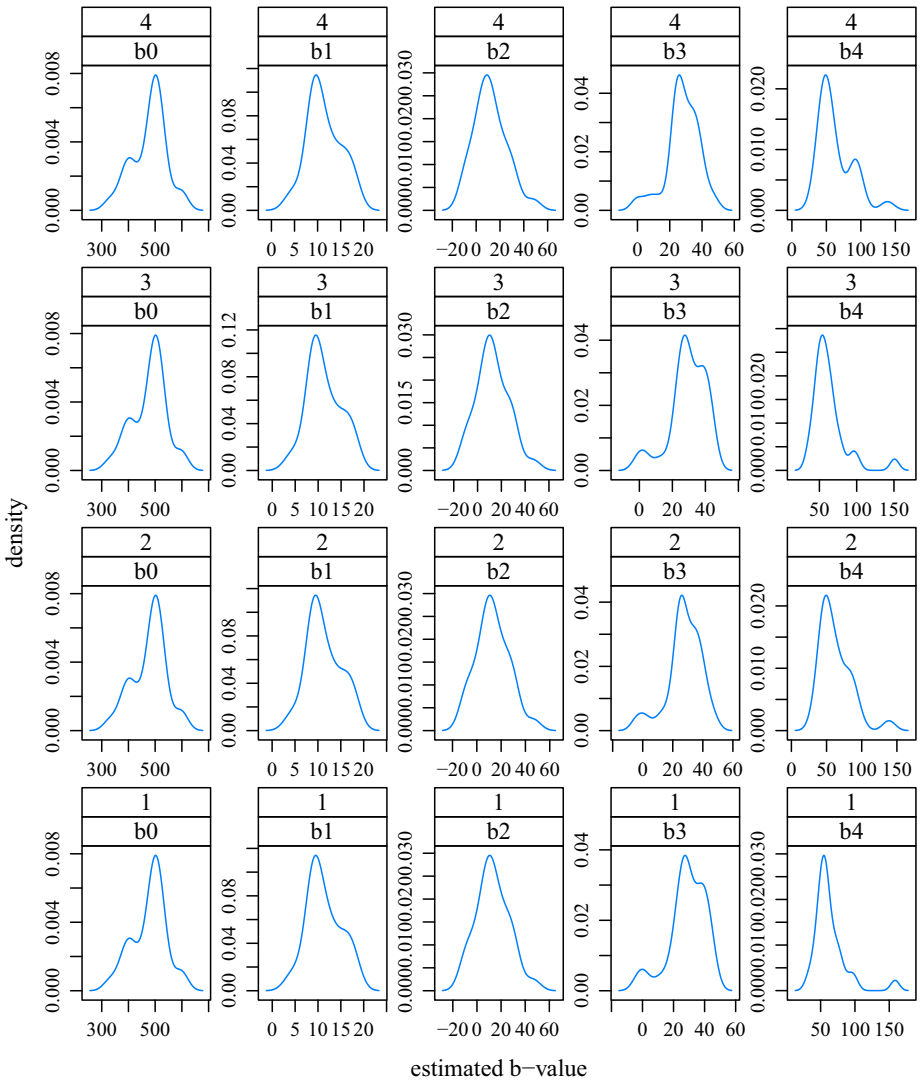

Fig. 1 Distribution of the fixed effects across countries given scaling model. *1* Scaling model 1—constant discrimination parameter across countries and items. *2* Scaling model 2—constant discrimination parameter across items but vary across countries. *3* Scaling model 3—discrimination parameter are constant across countries but vary over items. *4* Scaling model 4—discrimination parameter vary across countries and across items. *b0* Intercept. *b1* Individual effect of early literacy/numeracy tasks. *b2* School effect of early literacy/numeracy tasks. *b3* Individual effect of home resource for learning index. *b4* School effect of home resource for learning index

countries could be observed for those coefficients associated with the HRL index, either on an individual level ($\hat{\beta}_3$) or on the school level ($\hat{\beta}_4$). The scaling models thus affected both the mean and the variance of the distribution.

When conducting a statistical comparison of the distribution, we used a global F-type statistic in the first step. However, none of the $G \times z!/2!(z-2)! = 168$ derived F values were statistically significant. Thus, the overall hypotheses $\mathrm{H_0} : \boldsymbol{L_g}(\boldsymbol{\beta_{gw}} - \boldsymbol{\beta_{gq}}) = \mathbf{0}$ cannot be rejected in any of the cases. This finding corresponds with the invariance of the observed distribution of the fixed effects $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ across scaling models: when three out of five fixed effects are virtually unaffected by the scaling procedure, no overall

effects (as measured by the *F*-type statistic) can be expected. When we took a closer look at the results emerging from the use of the variance of the different estimated fixed effects across scaling models given the country, that is $s^2_{\hat{\beta}_{jgz.g\hat{}}}$, we found virtually no variation across the models for the estimated fixed effects $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. We can therefore assume that this lack of variation explains the results of the *F*-type statistic.

However, for those fixed effects that were associated with the HRL index ($\hat{\beta}_3$ for the individual effect of the HRL index on mathematics achievement and $\hat{\beta}_4$ for the school-level effect of HRL), we found that the scaling procedure had a strong influence. Table 8 shows the variance of the estimated fixed effect $\hat{\beta}_3$ across scaling models calculated for each country separately. As can be seen, for each country, the measurement model used to scale the HRL index did influence the size of the estimated fixed effect. The effect was remarkably high for Iran (Islamic Rep. of), Malta, Slovenia, Czech Republic, Abu Dhabi (UAE), Qatar, and Romania: the estimated fixed effects changed by up to 10 points when we used a country-specific measurement model to scale the HRL index. The direction of this change was not always the same, however, for some countries (Malta, Slovenia, Czech Republic), the estimated fixed effects decreased from measurement Model 1 to measurement Model 4; for others (Iran (Islamic Rep. of), Abu Dhabi (UAE), Qatar, Romania), the fixed effects increased.

Overall, the variance in the estimated fixed effect $\hat{\beta}_3$ across countries (with the scaling model held constant) decreased from $s^2_{\hat{\beta}_{3g1.1}} = 135.82$ to $s^2_{\hat{\beta}_{3g4.4}} = 102.71$ when we used the country-specific measurement models for the HRL index instead of the measurement invariance model. The differences across the countries in the observed association between the HRL index on the individual level and mathematics achievement reduced by approximately 30% when non-invariance models were used to scale the HRL index. However, for some countries (Chinese Taipei, Finland, Sweden), the influence of the scaling model on the estimated fixed effects $\hat{\beta}_3$ was very low. This finding was not surprising because the country-specific measurement model for these countries strongly agreed with the measurement invariance model (with the exception of the indicator "number of home study supports"). As such, no variation between the fixed effects should have been observed.

Table 9 displays the distribution of the school-level effects of the HRL index $\hat{\beta}_4$ across the scaling models. As observed for the individual effect of the HRL index, the scaling model influenced the size of the GLMM coefficients for all countries. The effect was largest for Morocco, Honduras, Iran (Islamic Rep. of), Qatar, Malta, Czech Republic, Romania, and Abu Dhabi (UAE). For these countries, the scaling model had an impact on $\hat{\beta}_3$ and $\hat{\beta}_4$. In addition, the effects followed the same pattern. For example, when the estimated coefficient of $\hat{\beta}_3$ decreased from scaling Model 1 to scaling Model 4, the coefficient from $\hat{\beta}_4$ also decreased from Model 1 to Model 4. However, the variance across countries in the estimated slope parameter $\hat{\beta}_4$ increased slightly from scaling Model 1 to scaling Model 4 ($s^2_{\hat{\beta}_{4g1.1}} = 577.60$ to $s^2_{\hat{\beta}_{4g4.4}} = 597.73$). Again, for those countries for which Model 4 strongly corresponded with Model 1 (i.e., Chinese Taipei, Sweden, Finland) virtually no variation between the fixed effects could be observed.

Our final step involved an analysis of the impact of the scaling procedure on the random effects of the GLMM. Tables 10 and 11 depicted the distribution of the $\boldsymbol{G}$ matrices across the countries and scaling models. Table 12 presents the fit-values of the applied

Table 8 Distribution of $\hat{\beta}_3$ across scaling models and countries, χ^2-value and asymmetric confidence interval (CI_l lower bound, CI_u upper bound; countries ordered in descending order of the conditional variance of $\hat{\beta}_3$ across scaling models given country *g*)

Country	$\hat{\beta}_3$				$s^2_{\hat{\beta}_{3gz.g}}$	χ^2	CI_l	CI_u	
	1	2	3	4					
Iran, Islamic Rep. of	20.77	26.94	22.35	30.60	19.96	14,972.88	6.41	277.54	
Malta	37.08	32.20	35.90	27.47	18.72	14,038.53	6.01	260.23	
Slovenia	42.95	36.89	43.05	35.31	16.29	12,218.83	5.23	226.49	
Czech Republic	38.89	33.79	38.75	31.75	12.90	9676.91	4.14	179.37	
Abu Dhabi, UAE	18.66	16.57	21.90	24.43	12.07	9050.66	3.87	167.76	
Qatar	29.36	25.81	30.85	33.90	11.29	8467.50	3.62	156.95	
Romania	35.52	41.32	36.55	42.18	11.18	8383.35	3.59	155.39	
Austria	39.15	34.52	39.11	33.51	8.89	6669.48	2.85	123.63	
Germany	32.59	30.02	31.43	25.93	8.42	6313.93	2.70	117.03	
Ireland	43.67	42.27	41.90	37.96	5.99	4492.46	1.92	83.27	
Morocco	1.60	2.08	3.83	7.00	5.98	4482.99	1.92	83.10	
Slovak Republic	40.91	37.52	41.50	36.85	5.52	4142.23	1.77	76.78	
Oman	33.44	36.40	34.36	38.34	4.78	3583.85	1.53	66.43	
Croatia	25.08	21.49	25.84	22.19	4.54	3406.78	1.46	63.15	
Italy	27.06	23.71	27.00	23.40	4.04	3029.52	1.30	56.16	
Georgia	25.31	23.78	26.93	22.26	4.03	3020.26	1.29	55.98	
Lithuania	26.17	23.48	27.10	24.04	2.95	2210.88	0.95	40.98	
Hungary	44.69	47.21	44.15	47.28	2.71	2033.31	0.87	37.69	
Honduras	−2.39	−3.81	−0.59	−0.36	2.65	1984.88	0.85	36.79	
Singapore	29.05	26.54	29.88	27.58	2.22	1664.81	0.71	30.86	
Azerbaijan	22.50	24.54	22.70	25.43	2.04	1530.08	0.65	28.36	▲
Poland	37.16	35.03	36.58	34.20	1.86	1397.34	0.60	25.90	▲△
Spain	26.46	24.59	26.00	23.60	1.71	1285.60	0.55	23.83	▲△●
Australia	40.50	39.20	40.92	38.52	1.25	935.29	0.40	17.34	▲△●○
Saudi Arabia	11.24	11.27	12.91	13.28	1.15	865.93	0.37	16.05	▲△●○◀
Chinese Taipei	28.56	27.29	29.20	28.07	0.65	487.64	0.21	9.04	▲△●○◀◁
Finland	25.44	24.89	25.37	23.78	0.59	440.63	0.19	8.17	▲△●○◀◁
Sweden	28.58	29.69	28.26	28.55	0.40	299.94	0.13	5.56	▲△●○◀◁
$s^2_{\hat{\beta}_{3gz.z}}$	135.82	129.56	119.69	102.71					
CL_l	84.90	80.99	74.82	64.20					
CL_u	251.64	240.04	221.75	190.28					

1, scaling model 1; 2, scaling model 2; 3, scaling model 3; 4, scaling model 4; ▲, country variance is statistically different from Iran, Islamic Rep. of; △, country variance is statistically different from Malta; ●, country variance is statistically different from Slovenia; ○, country variance is statistically different from Czech Republic and Abu Dhabi, UAE; ◀, country variance is statistically different from Qatar and Romania; ◁, country variance is statistically different from Austria, Germany, Ireland and Morocco

structural equation models. With the exception of Sweden, the applied scaling model affected the random coefficients of the GLMM in every country. The impacts were highest for Morocco, Malta, Honduras, and Iran (Islamic Rep. of), and lowest for Australia, Chinese Taipei, Finland, Ireland, Poland, and Sweden. Hence, there seems to be a weak relationship between the influence of the scaling model on the fixed effects and the random effects, in the sense that small impacts on the fixed effects (e.g., for Chinese Taipei, Finland, Poland, Sweden) correlated slightly with small impacts on the random components of the GLMM. Nevertheless, the impact of the scaling model on the random

Table 9 Distribution of $\hat{\beta}_4$ across scaling models and countries, χ^2-value and asymmetric confidence interval (CI_l lower bound, CI_u upper bound; countries ordered in descending order of the conditional variance of $\hat{\beta}_4$ across scaling models given country g)

Country	$\hat{\beta}_4$				$s^2_{\hat{\beta}_{4gz.g}}$	χ^2	CI_l	CI_u	
	1	2	3	4					
Morocco	58.63	81.35	63.61	95.99	292.81	219,608.17	93.97	4070.68	
Honduras	59.33	84.15	61.26	90.95	255.90	191,922.86	82.12	3557.50	
Iran, Islamic Rep. of	67.99	89.82	67.52	90.78	169.56	127,170.62	54.41	2357.25	
Qatar	158.42	138.73	150.80	138.54	94.71	71031.94	30.39	1316.65	
Malta	73.36	63.28	67.27	55.85	53.95	40,465.24	17.31	750.07	
Czech Republic	79.55	69.36	78.16	64.74	50.34	37,756.23	16.16	699.85	
Romania	59.73	69.87	60.93	72.69	41.54	31,153.02	13.33	577.46	
Abu Dhabi, UAE	92.34	82.41	93.06	94.07	29.40	22,047.72	9.43	408.68	▲
Croatia	55.14	47.33	52.95	44.27	25.01	18,753.96	8.02	347.62	▲
Austria	61.39	54.21	59.05	50.74	22.92	17,191.50	7.36	318.66	▲
Slovenia	57.10	49.43	56.50	47.98	22.23	16,669.13	7.13	308.98	▲
Hungary	79.65	84.35	77.74	87.68	20.32	15,242.15	6.52	282.53	▲
Germany	71.93	66.12	70.09	62.12	19.05	14,287.74	6.11	264.84	▲
Singapore	58.35	53.59	56.26	50.03	12.90	9674.75	4.14	179.33	▲△
Italy	51.74	45.44	50.29	44.61	12.43	9322.94	3.99	172.81	▲△
Azerbaijan	42.30	45.40	43.83	49.96	10.96	8218.86	3.52	152.35	▲△
Lithuania	48.05	43.30	46.25	40.65	10.65	7988.31	3.42	148.07	▲△
Slovak Republic	53.89	49.75	54.01	48.02	9.04	6781.11	2.90	125.70	▲△●
Oman	53.24	57.78	52.77	53.87	5.23	3925.58	1.68	72.76	▲△●○
Spain	47.14	43.82	44.89	41.87	4.83	3622.23	1.55	67.14	▲△●○
Ireland	66.86	64.70	65.21	61.93	4.20	3148.12	1.35	58.35	▲△●○◀
Georgia	47.75	44.71	49.56	48.01	4.12	3086.43	1.32	57.21	▲△●○◀
Saudi Arabia	32.06	32.66	35.73	34.77	3.00	2253.27	0.96	41.77	▲△●○◀◁
Australia	99.67	96.50	99.84	97.20	2.90	2176.08	0.93	40.34	▲△●○◀◁
Poland	49.70	46.89	47.99	46.38	2.16	1619.93	0.69	30.03	▲△●○◀◁
Chinese Taipei	50.96	48.85	49.95	48.21	1.47	1102.95	0.47	20.44	▲△●○◀◁
Sweden	56.36	58.75	56.76	57.87	1.17	880.25	0.38	16.32	▲△●○◀◁
Finland	38.91	38.08	38.31	36.58	0.98	736.06	0.31	13.64	▲△●○◀◁
$s^2_{\hat{\beta}_{4gz.z}}$	577.60	511.62	518.80	597.73					
CL_l	361.05	319.80	324.29	373.63					
CL_u	1070.11	947.88	961.18	1107.41					

1, scaling model 1; 2, scaling model 2; 3, scaling model 3; 4, scaling model 4; ▲, country variance is statistically different from Morocco and Honduras; △, country variance is statistically different from Iran, Islamic Rep. of; ●, country variance is statistically different from Qatar; ○, country variance is statistically different from Malta and Czech Republic; ◀, country variance is statistically different from Romania; ◁, country variance is statistically different from Abu Dhabi, UAE

effects, and thus on the institutional variation of the estimated relationship between the HRL index and mathematics achievement on the student-level, was remarkably high.

Discussion

This paper investigated the relationships between different procedures for scaling the "home resources for learning index" (HRL) and the prediction accuracy of this index in explaining the mathematics achievement of the fourth-grade students who participate in IEA's combined PIRLS/TIMSS survey of 2011. As work by Lüdtke et al. (2011) and

Table 10 Distribution of random effects *G* across scaling models and countries (Part I)

Country	Effect	1		2		3		4	
		$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$
Azerbaijan	$\hat{\alpha}_0$	5248.25	153.86	5244.57	162.60	5226.53	74.15	5248.85	291.87
	$\hat{\alpha}_1$	153.86	443.62	162.60	549.16	74.15	417.92	291.87	650.26
Australia	$\hat{\alpha}_0$	590.11	24.08	588.47	25.68	592.42	−4.31	587.72	10.40
	$\hat{\alpha}_1$	24.08	475.07	25.68	592.42	−4.31	428.68	10.40	415.55
Austria	$\hat{\alpha}_0$	407.73	−6.63	402.97	−7.05	423.97	−3.26	415.20	−7.77
	$\hat{\alpha}_1$	−6.63	51.65	−7.05	35.44	−3.26	47.70	−7.77	30.53
Chinese Taipei	$\hat{\alpha}_0$	176.44	−47.49	176.16	−46.11	171.42	−47.08	166.76	−42.88
	$\hat{\alpha}_1$	−47.49	74.59	−46.11	69.47	−47.08	77.28	−42.88	71.36
Croatia	$\hat{\alpha}_0$	212.37	−3.58	210.72	−3.33	211.57	−5.62	211.39	−6.56
	$\hat{\alpha}_1$	−3.58	55.29	−3.33	43.16	−5.62	55.03	−6.56	40.84
Czech Republic	$\hat{\alpha}_0$	244.49	−74.17	244.84	−63.68	244.47	−82.20	251.91	−86.89
	$\hat{\alpha}_1$	−74.17	344.00	−63.68	260.64	−82.20	361.47	−86.89	280.36
Finland	$\hat{\alpha}_0$	350.47	−49.36	350.52	−48.31	347.46	−41.26	344.18	−34.64
	$\hat{\alpha}_1$	−49.36	66.14	−48.31	63.19	−41.26	62.60	−34.64	64.96
Georgia	$\hat{\alpha}_0$	2429.86	−279.28	2432.33	−259.49	2428.40	−270.16	2420.54	−264.22
	$\hat{\alpha}_1$	−279.28	376.21	−259.49	329.27	−270.16	370.75	−264.22	312.49
Germany	$\hat{\alpha}_0$	416.96	47.33	414.62	43.03	423.58	36.03	475.08	15.70
	$\hat{\alpha}_1$	47.33	73.62	43.03	60.61	36.03	68.99	15.70	45.10
Honduras	$\hat{\alpha}_0$	2131.60	293.87	2181.93	413.57	2096.88	239.06	2164.52	261.29
	$\hat{\alpha}_1$	293.87	372.70	413.57	852.45	239.06	297.90	261.29	716.51
Hungary	$\hat{\alpha}_0$	578.70	−190.88	578.38	−198.93	584.08	−184.64	672.32	−168.77
	$\hat{\alpha}_1$	−190.88	183.69	−198.93	206.38	−184.64	174.82	−168.77	257.79
Iran, Islamic Rep. of	$\hat{\alpha}_0$	1536.11	73.79	1559.85	109.15	1564.47	47.99	1572.80	37.30
	$\hat{\alpha}_1$	73.79	275.83	109.15	514.43	47.99	242.12	37.30	446.03
Ireland	$\hat{\alpha}_0$	643.26	−56.05	643.50	−55.20	650.27	−55.71	664.92	−50.11
	$\hat{\alpha}_1$	−56.05	51.56	−55.20	48.60	−55.71	47.17	−50.11	45.44
Italy	$\hat{\alpha}_0$	1346.42	−112.52	1349.74	−98.70	1339.77	−114.57	1339.30	−108.21
	$\hat{\alpha}_1$	−112.52	267.16	−98.70	206.35	−114.57	245.99	−108.21	170.05
Lithuania	$\hat{\alpha}_0$	334.13	−63.92	333.83	−56.78	337.52	−69.82	346.11	−72.32
	$\hat{\alpha}_1$	−63.92	278.92	−56.78	226.04	−69.82	277.76	−72.32	222.45
Malta	$\hat{\alpha}_0$	468.51	41.21	472.58	30.92	470.84	34.23	518.49	2.89
	$\hat{\alpha}_1$	41.21	29.89	30.92	19.98	34.23	30.08	2.89	8.49

1, scaling model 1; 2, scaling model 2; 3, scaling model 3; 4, scaling model 4; $\hat{\alpha}_0$, random intercept; $\hat{\alpha}_1$, random slope of home resource for learning index

van den Heuvel-Panhuizen et al. (2009) has shown, scaling social background indicators into a latent variable enhances the validity of large-scale educational assessment studies. The content validity and the reliability of such an index are usually much higher than those of single indicators. Because both aspects are particularly important within the context of cross-national comparative studies of educational achievement, using a scaled index for PIRLS/TIMSS home environment (social background) variables provided a framework that enabled meaningful cross-national comparisons.

While the scaling of the social background indicators into a latent variable is without dispute, and probably without a reasonable alternative, the assumption of measurement invariance evident in scaling the HRL index needs to be challenged. As prior research on the scaling of social background indicators into latent indices in large-scale assessments

Table 11 Distribution of random effects *G* across scaling models and countries (part II)

Country	Effect	1		2		3		4	
		$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$
Morocco	$\hat{\alpha}_0$	4358.44	85.63	4433.22	104.05	4314.43	53.49	4311.60	−97.25
	$\hat{\alpha}_1$	85.63	432.09	104.05	959.00	53.49	392.12	−97.25	825.05
Oman	$\hat{\alpha}_0$	2187.73	17.59	2186.26	15.53	2176.12	−38.90	2217.85	−108.31
	$\hat{\alpha}_1$	17.59	319.17	15.53	377.71	−38.90	294.01	−108.31	349.03
Poland	$\hat{\alpha}_0$	322.04	−40.50	322.36	−37.82	321.52	−44.02	317.37	−36.29
	$\hat{\alpha}_1$	−40.50	42.59	−37.82	39.27	−44.02	40.80	−36.29	37.58
Qatar	$\hat{\alpha}_0$	1823.03	124.61	1824.27	107.88	1843.10	124.99	2290.76	154.27
	$\hat{\alpha}_1$	124.61	217.05	107.88	169.92	124.99	186.00	154.27	181.57
Romania	$\hat{\alpha}_0$	3132.69	−574.41	3133.08	−652.34	3119.28	−558.05	3138.69	−577.43
	$\hat{\alpha}_1$	−574.41	863.19	−652.34	1159.58	−558.05	818.38	−577.43	1098.26
Saudi Arabia	$\hat{\alpha}_0$	3929.55	−146.25	3930.06	−147.93	3891.32	−174.19	3900.55	−180.56
	$\hat{\alpha}_1$	−146.25	379.72	−147.93	392.95	−174.19	340.87	−180.56	290.76
Singapore	$\hat{\alpha}_0$	275.71	−9.82	273.85	−10.80	275.53	−16.42	281.66	−19.49
	$\hat{\alpha}_1$	−9.82	134.89	−10.80	112.66	−16.42	123.45	−19.49	86.69
Slovak Republic	$\hat{\alpha}_0$	1359.48	−314.15	1353.27	−287.64	1355.48	−318.11	1381.23	−298.18
	$\hat{\alpha}_1$	−314.15	308.61	−287.64	256.59	−318.11	304.41	−298.18	250.48
Slovenia	$\hat{\alpha}_0$	234.97	−33.21	234.17	−32.55	232.33	−43.19	234.77	−61.34
	$\hat{\alpha}_1$	−33.21	123.54	−32.55	91.74	−43.19	126.89	−61.34	98.47
Spain	$\hat{\alpha}_0$	454.26	−8.45	453.99	−7.85	459.55	−9.76	494.53	−8.72
	$\hat{\alpha}_1$	−8.45	94.17	−7.85	80.19	−9.76	81.59	−8.72	74.55
Sweden	$\hat{\alpha}_0$	134.78	−11.58	134.80	−11.66	135.76	−14.16	139.87	−15.65
	$\hat{\alpha}_1$	−11.58	159.41	−11.66	172.39	−14.16	167.74	−15.65	163.90
Abu Dhabi, UAE	$\hat{\alpha}_0$	2138.29	46.38	2136.88	43.58	2006.65	28.57	1795.27	−51.16
	$\hat{\alpha}_1$	46.38	210.85	43.58	170.91	28.57	219.04	−51.16	221.58

1, scaling model 1; 2, scaling model 2; 3, scaling model 3; 4, scaling model 4; $\hat{\alpha}_0$, random intercept; $\hat{\alpha}_1$, random slope of home resource for learning index

have shown, assuming a measurement invariance model across countries results in latent variables that are less reliable than those that occur when assuming measurement non-invariance (Caro and Sandoval-Hernandez 2012; Hansson and Gustafsson 2013; Lakin 2012). In our study, rescaling the HRL index with four different measurement models with different degrees of assumed measurement invariance also showed that the measurement non-invariance model fitted the data best. Thus, with respect to our first research question we can assume that measurement invariance across participating countries for the HRL index would not hold for the Grade 4 students assessed in PIRLS/TIMSS 2011.

From a methodical perspective, we were not surprised to find that our less restrictive model (the measurement non-invariance model) was superior to our more restrictive model (the measurement invariance model) in terms of fitting indices. Everything else being equal, a model where the parameters can take on any value will always fit at least as well as a model where some of the parameters are fixed to some value or where some of the parameters are set to constraints. It could be argued that the measurement invariance assumption is merely a practical matter because it makes cross-national comparative studies of educational achievement possible through use of model that most parsimoniously describes the data yet also describes the data sufficiently well to explain

Table 12 Fit-values for equality test of G_{gz} across scaling models *z* given country *g*

Country	Fit-value							
	χ^2	*df*	*p*	SRMR	GFI	RMSEA	CI_l	CI_u
Azerbaijan	323.41	9	<0.0001	0.11	0.98	0.08	0.08	0.09
Australia	42.21	9	<0.0001	0.03	1.00	0.02	0.02	0.03
Austria	426.86	9	<0.0001	0.13	0.98	0.10	0.09	0.11
Chinese Taipei	21.22	9	0.01	0.03	1.00	0.02	0.01	0.03
Croatia	177.11	9	<0.0001	0.08	0.99	0.06	0.06	0.07
Czech Republic	195.84	9	<0.0001	0.08	0.99	0.07	0.06	0.08
Finland	37.48	9	<0.0001	0.03	1.00	0.03	0.02	0.04
Georgia	64.84	9	<0.0001	0.05	1.00	0.04	0.03	0.04
Germany	355.66	9	<0.0001	0.13	0.98	0.10	0.10	0.11
Honduras	1561.82	9	<0.0001	0.32	0.90	0.21	0.20	0.22
Hungary	817.51	9	<0.0001	0.09	0.96	0.13	0.12	0.14
Iran, Islamic Rep. of	1157.06	9	<0.0001	0.21	0.95	0.15	0.14	0.16
Ireland	24.14	9	0.004	0.03	1.00	0.02	0.01	0.03
Italy	259.85	9	<0.0001	0.11	0.98	0.08	0.07	0.09
Lithuania	125.05	9	<0.0001	0.07	0.99	0.05	0.05	0.06
Malta	1701.07	9	<0.0001	0.49	0.87	0.23	0.22	0.24
Morocco	2346.54	9	<0.0001	0.27	0.92	0.18	0.18	0.19
Oman	324.51	9	<0.0001	0.06	0.99	0.06	0.05	0.06
Poland	31.48	9	0.0002	0.03	1.00	0.02	0.01	0.03
Qatar	153.41	9	<0.0001	0.08	0.99	0.06	0.05	0.07
Romania	237.43	9	<0.0001	0.09	0.99	0.07	0.06	0.08
Saudi Arabia	136.58	9	<0.0001	0.07	0.99	0.05	0.04	0.06
Singapore	357.01	9	<0.0001	0.11	0.99	0.08	0.07	0.09
Slovak Republic	128.30	9	<0.0001	0.06	0.99	0.05	0.04	0.06
Slovenia	354.37	9	<0.0001	0.09	0.98	0.09	0.08	0.10
Spain	70.70	9	<0.0001	0.05	1.00	0.04	0.03	0.05
Sweden	11.68	9	0.23	0.02	1.00	0.01	0.00	0.02
Abu Dhabi, UAE	192.68	9	<0.0001	0.09	0.99	0.07	0.06	0.08

χ^2, under the null hypothesis of equality of the covariance matrix G_{gz} across scaling models, this value should not be statistically different from zero; SRMR, standardized root mean square residual should be zero under the null hypothesis; GFI, goodness-of-fit index should be one under the null hypothesis; RMSEA, root mean square error of approximation should be zero under the null hypothesis; CI_l, CI_u, lower and upper bound of RMSEA

any observed achievement differences. However, viewing this matter from the perspective of predictive validity challenges this argument. Given the general inconsistency of measurement invariance and predictive invariance that Millsap (1995, 1997, 1998, 2007) found, we could expect that the most parsimonious model (the measurement invariance model) for latent variables would affect ability to compare the prediction coefficients of this latent variable across countries. Accordingly, with regard to the HRL index, we need to establish whether the hierarchical linear model applied by Martin et al. (2013) was sensitive to the assumption of measurement invariance.

To investigate that question, we rescaled the HRL index four times, with each scaling allowing a different degree of measurement invariance. We then introduced these indices as predictors in a generalized linear mixed model (GLMM) with mathematics achievement as the dependent variable. Overall, we observed a strong influence of the scaling model on the prediction outcomes of the GLMM. Assuming country-specific

measurement models for the HRL index decreased the cross-national variance of the individual effect of the HRL index on student mathematics achievement. The variance across countries of this effect was $s^2_{\hat{\beta}_{3gz.z}} = 135.82$ for the measurement invariance model. However, the strength of the effect dropped to $s^2_{\hat{\beta}_{3gz.z}} = 102.71$ for the measurement non-invariance model. Accordingly, the cross-national differences of this effect, expressed in terms of the cross-national variance of $\hat{\beta}_3$, can be reduced by approximately 25% when a measurement non-invariance model is assumed for the HRL index. This finding implies that those countries classified as unequal with respect to this effect when the measurement-invariance assumption applied, that is, Iran (Islamic Rep. of) and Slovenia, would be categorized as equal under the assumption of measurement non-invariance.

The results for the school-level effect of the HRL index were not as conclusive. Although we observed only a small difference in the cross-national variance of this effect when we compared the measurement invariance with the country-specific and item-specific measurement model (Model 1 vs. Model 4), we found the reduction in variance was substantial when a country-specific (but not an item-specific measurement model) was assumed (Model 2), or when an item-specific measurement model (but not a country-specific model) was assumed (Model 3). In both cases, the cross-national variance of the school-level effect of the HRL index reduced by about 11%. One explanation for these somewhat unpredictable results could be that the four HRL indices were scaled in the same way as in the study by Martin et al. (2013), that is, without taking the multilevel structure of the data into account. Loosely speaking, this possibility implies that the applied scaling procedure "ignored" the between-school part of the HRL index. Further research directed toward differentiating between a level one measurement invariance assumption and a level two measurement invariance assumption is needed. Nevertheless, application of the scaling procedure that Martin et al. used will result in school-level prediction effects of the HRL index that are obviously sensitive to the assumed degree of measurement invariance.

Although the effect of the measurement invariance assumption on cross-national comparisons of the fixed effects of the GLMM was the main focus of the present study, we also investigated country-specific differences in the effect of the measurement invariance assumption on the prediction coefficients. We were not surprised to find this effect was not constant across countries. For example, the influence of the measurement model on both the individual- and the school-level HRL coefficients was relatively strong in Iran (Islamic Rep. of), Malta, Czech Republic, Abu Dhabi (UAE), Qatar, and Romania, but was relatively weak in Australia, Saudi Arabia, Chinese Taipei, Finland, and Sweden. We can express this point in another way by stating that the regression coefficients for Finland, for example, were relatively robust with respect to the different assumptions about measurement invariance, while the coefficients for Iran (Islamic Rep. of) were very sensitive with respect to the assumed scaling model. The implication of this finding is that even when only the country-specific regression coefficients are of interest, we need to take the assumed degree of measurement invariance into account when interpreting the coefficients.

We were also able to observe the country-specific effects of the measurement invariance assumption on the prediction validity of the GLMM's random slope coefficients. In most countries, the random variance of this coefficient decreased when a non-invariance model was assumed. The fact that we can interpret the random coefficient as a measure

of the school-specific effect on the relationship between the individual HRL index and mathematics achievement, basically implies that, under the non-invariance model, differences between schools are a less suitable way of explaining the relationship between the HRL index and mathematics achievement. Accordingly, under the non-invariance assumption, we can expect that this relationship would be nearly the same in all schools of most of the participating countries, while under the measurement invariance model the relationship between the HRL index and mathematics achievement would vary across these schools. In short, researchers and others may draw completely different conclusions with respect to this effect because the nature of the effect will depend solely on the assumed measurement model.

The important point here is that the results of the hierarchical linear model that Martin et al. (2013) applied are very sensitive in terms of the assumed degree of measurement invariance. According to Millsap's (1995, 1997, 1998, 2007) findings this degree of sensitivity can be expected. However, if researchers agree that using latent variables in educational research is sound practice, and if assuming measurement invariance is a necessary requirement for cross-national comparisons of latent variables, it is vital to consider the question of how researchers engaged in large-scale assessment studies can control for these effects or take them into account.

While a comprehensive answer to this question will rely on further research and on more expertise, and although the research agenda of the IEA-ETS Research Institute calls for "a more scientific approach to the development, use and interpretability of background questionnaires" (http://ierinstitute.org/research-agenda.html, Accessed 04 May 2016), we can still offer some general ideas. For example, according to Brennan's (2001) generalizability theory, the variance in the GLMM coefficients that can be traced back to different assumptions about measurement invariance should be added to the standard errors of these coefficients. In regard to the results of the present study, this advice implies that, for example, the variance of $s^2_{\hat{\beta}_{3gz.g}} = 19.96$ for Iran (Islamic Rep. of) (see Table 8) should be added to the standard error of $\hat{\beta}_3$. Of course, more reliable estimates of this component are possible if we undertake a more exhaustive analysis where we implement a broader range of possible measurement models and also account for the random sample of students (by, for example, using bootstrapping methods).

Another approach that we could use to capture the dependency between measurement invariance and predictive invariance in large-scale assessment studies is the assumption of partial measurement invariance. This approach implies, for example, that measurement invariance across countries can be assumed for only some of the HRL index items and that the parameters of the other items will be left to vary freely across countries. This linking or equation procedure means that while the latent variable across countries may still be compared, it must be acknowledged that dependency between the measurement invariance and the predictive invariance will decrease (if not vanish). Again, taking the present study as an example, the parameters of the HRL indicators "highest level of education of either parent" and "number of home study supports" would need to vary freely across countries, because these indicators are the ones that exhibit the highest variance in the discrimination parameter across countries (see Table 6). However, as we stated above, more exhaustive analysis are necessary before decision as concrete as this one can be made. One requirement that would need to be in place before this degree of

analysis could be implemented for the HRL is surely that of defining the item sampling space for the HRL. Achieving this requirement, in turn, implies the need to develop a theoretical framework for the HRL index that is coherent and valid and reliable cross-nationally, but whether this aim can be credibly achieved is a moot point.

Limitations of the present study

Although our study is the first study to provide a deeper insight into the relationship between measurement invariance and predictive invariance in large-scale assessment studies and thus contributes, for example, to the research agenda of the IEA-ETS Research Institute, it has some limitations. The first is the index that we used. While it made sense for us to focus on the HRL index, it could be interpreted as a formative variable. As such, studying the relationship of measurement invariance and predictive invariance with the more reflective indices that are also part of, for example, TIMSS and PIRLS, seems advisable. In addition, the applied measurement model could be more exhaustive if it took into account the multilevel structure of the data and gave consideration to scaling models that have more parameters (or dimensions). In general, we did not know the true parameters of the models (both the scaling model and the prediction model) when we conducted our study. This lack of knowledge meant that we were unable to estimate the unbiased effect of the scaling model on the prediction coefficients. This consideration calls for implementation of another design, such as that used in simulation studies. Despite these limitations, we consider that the general inconsistency of measurement invariance and predictive invariance found in this study will remain valid even when these limitations have been satisfactorily resolved. We therefore think it safe to state that assuming measurement invariance of background indicators in cross-national studies of educational achievement is a challenge that needs to be addressed by anyone endeavoring to interpret cross-national differences in achievement.

Abbreviations

AIC: Akaike's information criterion; BIC: Bayesian information criterion; GLMM: generalized linear mixed model; ET: early literacy tasks/early numeracy tasks; HRL: home resources for learning; IEA: International Association for the Evaluation of Educational Achievement; MAP: maximum a posterior probability; PIRLS: Progress in International Reading Literacy Study; TIMSS: Trends in International Mathematics and Science Study; WLE: weighted likelihood estimate.

Authors' contributions

All authors made substantial contributions to the conception and the design of the study. In addition, HW provided the data sets for the analysis and DK conducted the analysis. DK drafted the manuscript. All authors made substantial contribution to the interpretation of the results. All authors read and approved the final manuscript.

Author details

[1] Institute for School Development Research, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany. [2] Federal Institute for Education Research, Innovation & Development of the Austrian Schooling System (BIFIE), Alpenstraße 121, 5020 Salzburg, Austria.

Acknowledgements

The authors acknowledge the PIRLS/TIMSS International Study Center and Boston College for providing the technical documentation that allowed the replication of the keyreference models published in Martin et al. (2013). The authors further acknowledge Wilfried Bos and the anonymous reviewers for the attention and expertise they generously shared to support the production of this paper. We finally thank Daniel Scott Smith and Paula Wagemaker for pre-submission English editing support.

Competing interests

The authors declare that they have no competing interests.

Received: 14 August 2015 Accepted: 9 February 2017
Published online: 16 March 2017

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Autom Control*, *19*, 716–723.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). New York: Greenwood.

Bos, W., Wendt, H., Köller, O., & Selter, C. (2012). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Gundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Caro, D., Sandoval-Hernandez, A., & Lütke, O. (2014). Cultural, social and economic capital constructs: An evaluation using exploratory structural equation modeling. *Sch Eff Sch Improv*, *25*, 433–450.

Caro, D., & Sandoval-Hernandez, A. (2012). A exploratory structural equation modeling approach to evaluate sociological theories in international large-scale assessment studies. In: Paper presented at the annual meeting of the American educational research association 2012

Çetin, B. (2010). Cross-cultural structural parameter invariance on PISA 2006 student questionnaire. *Eurasian J Educ Res*, *38*, 71–89.

Coleman, J. S. (1988). Social capital in the creation of human capital. *Am J Sociol*, *94*, 95–120.

Fischer, G. H., & Molenaar, I. W. (1995) *Rasch models. Foundations, recent developments, and applications*. New York: Springer

Foy, P. (2013). *TIMSS and PIRLS 2011 user guide for the fourth grade combined international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Hansson, Å., & Gustafsson, J.-E. (2013). Measurement invariance of socioeconomic status across migrational background. *Scand J Educ Res*, *57*, 148–166.

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects salamander mating revisited. *Biometrics*, *48*, 631–644.

Kasper, D. (2017). Multiple group comparisons of the fixed effects from the generalized linear mixed model. (**In preparation**)

Lakin, J. M. (2012). Multidimensional ability tests and culturally and linguistically diverse students: Evidence of measurement invariance. *Learn Individ Differ*, *22*, 397–403.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychol Methods*, *16*, 444–467.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. http://timss.bc.edu/methods/index.html. Accessed 20 Feb 2017.

Martin, M. O., & Mullis, I. V. S. (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M. O., Mullis, I. V. S., Foy, P., Olson, J. F., Erbeber, E., & Preuschoff, C. (2008). *TIMSS 2007 international science report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivar Behav Res*, *30*, 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychol Methods*, *2*, 248–260.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivar Behav Res*, *33*, 403–424.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erbeber, E., et al. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012a). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012b). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Appl Psychol Meas*, *16*, 159–176.

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis, Chap. 15* (pp. 318–344). Boca Raton: Chapman and Hall/CRC.

OECD. (2014a). *PISA 2012 results: What students know and can do—student performance in mathematics, reading and science (Vol. I, Revised edition, February 2014)*. Paris: PISA OECD Publishing.

OECD. (2014b). *PISA 2012: Technical report*. Paris: PISA, OECD Publishing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. London: Sage Publications.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Schulte, K., Nonte, S., & Schwippert, K. (2013). Die Überprüfung von Messinvarianz in international vergleichenden Schulleistungsstudien am Beispiel der Studie PIRLS [Testing measurement invariance in international large scale assessments using the example of PIRLS data]. *Zeitschrift für Bildungsforschung*, *3*, 99–118.
Schulz, W. (2005). Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory. Paper prepared for the Annual Meetings of the American Educational Research Association in San Francisco. http://files.eric.ed.gov/fulltext/ED493509.pdf. Accessed 20 Feb 2017.
Schwarz, G. (1978). Estimating the dimension of a model. *Ann Stat*, *6*(6), 461–464.
Segeritz, M., & Pant, H. A. (2013). Do they feel the same way about math? Testing measurement invariance of the PISA "students' approaches to learning" instrument across immigrant groups within Germany. *Educ Psychol Meas*, *73*, 601–630.
Smith, D. S., Wendt, H., & Kasper, D. (2016). Social reproduction and sex in German primary schools. *Compare J Comp Int Educ*,. doi:10.1080/03057925.2016.1158643.
van den Heuvel-Panhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika*, *74*, 351–365.
Wang, S., & Wang, T. (2001). Precision of warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Appl Psychol Meas*, *25*, 317–331.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; A Gibbs sampling approach. *J Am Stat Assoc*, *86*, 79–86.