Structural equation modelling trees for invariance assessment

W. Holmes Finch

Department of Educational Psychology,
Ball State University,
Muncie, IN, USA
Email: whfinch@bsu.edu

Abstract: Large-scale assessment data have become increasingly popular in educational research. Factor model invariance testing is also a key feature of educational research, as scholars seek to identify situations where scales work comparably for different subgroups in the population. There exist a variety of methods for assessing invariance; however, standard approaches for this purpose can be cumbersome with a large number of groups, and do not typically accommodate invariance assessment across multiple variables simultaneously. The purpose of this study is to demonstrate the use of Structural Equation Modelling Trees for invariance assessment with complex large scale assessment data. The results of the study, involving Programme for International Student Assessment reading interest inventory data, found a lack of invariance based on nation of residence, language spoken in the home and family socioeconomic status. Advantages, and disadvantages, of SEMtree when compared to other methods for invariance assessment are discussed in light of these findings.

Keywords: factor analysis; invariance testing; recursive partitioning; structural equation modelling.

Reference to this paper should be made as follows: Holmes Finch, W. (2017) 'Structural equation modelling trees for invariance assessment', *Int. J. Quantitative Research in Education*, Vol. 4, Nos. 1/2, pp.72–93.

Biographical note: Holmes Finch is the George and Frances Ball Distinguished Professor of Educational Psychology at Ball State University. He conducts research in latent variable modelling and multivariate statistics.

1 Introduction

A foundational principle underlying the use of educational and psychological measurements is that the scales used to assess latent constructs provide the same information in the same way for individuals in the population, regardless of the subgroup(s) to which they belong (Millsap, 2011). Put another way, a psychologist using a scale to ascertain whether a child may have Autism needs to be certain that the score obtained from that scale provides the same quality of information for males as it does for females. Indeed, such equivalence of measurement is a key component that must be made in any validity argument about the scale (Horn and McArdle, 1992). The area of invariance testing has been the subject of much research, with a number of approaches suggested for this purpose (see Millsap, 2011 for a discussion of these). Most frequently, these methods make use of comparisons of fit to the data for models that have different levels of constraints placed upon parameters of interest. For example, researchers interested in determining whether the loadings for a factor are equivalent between two groups (e.g. males and females) would fit one model in which the loadings are constrained to be equal for the groups, and another for which the loadings are allowed to vary. The fits of the two models are then compared using a difference of some fit statistic between the two models, such as the Chi-square goodness of fit statistic or the comparative fit index (CFI). If this difference is statistically meaningful (i.e. a statistically significant Chi-square or CFI change of 0.01 or more) then it is determined that at least some of the parameters being assessed are not equal for the two groups.

A limitation of the traditional invariance assessment approach as outlined above is that it can be quite cumbersome to use when there are many groups for which invariance testing is desired (Asparouhov and Muthén, 2014). For example, researchers working with large international educational assessments who are interested in the extent to which there is invariance of model parameters across several nations may find the standard technique for invariance assessment to be difficult to use. In addition, the traditional approach to the invariance problem does not easily accommodate continuous variables, such as family income, or age of the subject. Recent work by Merkle and Zeileis (2013) has demonstrated how the invariance testing paradigm can be extended to address this situation. However, this approach, which is based upon an extension of the Lagrange Multiplier, does not address the aforementioned issue of having a large number of groups for which invariance assessment is desired, though it does allow for invariance assessment with respect to a continuous variable.

Given the increasing popularity of large scale data collection efforts such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMMS), Progress in International Literacy (PIRLS), the Early Childhood Longitudinal Study (ECLS) and the National Assessment of Educational Progress (NAEP), among others, in educational research, and the importance of assuring that assessments perform equivalently across groups, as well as levels of continuous variables, it is crucial that researchers have access to tools for assessing invariance in such complex cases. The purpose of the current study is to describe the work by Brandmaier et al. (2013) in the area of recursive partitioning for structural equation models (SEM), and to demonstrate how it can be applied to invariance assessment in the context of large scale databases, with both continuous and categorical variables for which invariance is of interest. In addition, the results from recursive partitioning will be compared with those of the alignment procedure (Asparouhov and Muthén, 2014), which has been shown to be a viable option in the assessment of invariance for a large number of groups. The manuscript is organised as follows: first a very brief review of traditional invariance assessment is described, after which the alignment procedure is discussed, followed by structural equation modelling trees. The research problem that is the focus of the current study as well as the methods used to address them are then described, followed by a detailed review of the results of these analyses. Finally, the results are discussed in the context of both the current dataset, as well as more broadly in terms of application in a wide array of contexts.

1.1 Traditional invariance testing

The basic factor model that is assumed to link observed indicator variables, such as items on a scale, and the latent trait being measured, such as reading interest, takes the form

$$x = \tau + \Lambda\xi + \delta \tag{1}$$

where x = Vector of observed indicator variables; e.g. items on a scale or subscale scores in a battery,

ξ = Vector of latent traits being measured by x,

Λ = Matrix of factor loadings linking x and ξ,

τ = Vector of intercepts associated with x,

δ = Vector of unique errors associated with x.

The model in Eq. (1) implies the following covariance matrix for the observed indicators:

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \tag{2}$$

where Σ = Covariance matrix of the observed indicators, x,

Ψ = Covariance matrix of the latent traits ,

Θ = Covariance matrix of unique error terms, assumed to be diagonal .

As mentioned above, the most common methodology for assessing factorial invariance (FI) is based on multiple group confirmatory factor analysis (MGCFA). For the parameters in (1) and (2), invariance assumptions build upon one another, and can be assessed using increasingly constrained MGCFA models. The weakest type of invariance under the FI umbrella is configural invariance (CI), for which the number of latent variables is the same across groups in the population, as is the correspondence of observed indicators to factors. However, no assumptions are made regarding group equality of the individual parameters in (1) and (2). If CI holds, these parameters can then be constrained in a stepwise fashion, leading to increasingly strong levels of invariance (i.e. parameter equality) across groups. The parameters that research has traditionally focused on are factor loadings (Λ), known as metric invariance (MI) and factor intercepts (τ), called scalar invariance or SI (Horn and McArdle, 1992; Millsap, 2011; Meredith, 1993; Steenkamp and Baumgartner, 1998; Wicherts and Dolan, 2010; Widaman and Reise, 1997). In order to assess MI, the Λ estimates in (1) are constrained to be equal across groups. The χ^2 fit statistic is then obtained for this model and compared with the χ^2 value from the CI model using

$$\chi^2_\Delta = \chi^2_{\mathrm{MI}} - \chi^2_{\mathrm{CI}} \tag{3}.$$

χ^2_Δ is distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom for the two models. If χ^2_Δ is statistically significant, then we conclude that the fit of the two models differs and constraining the loadings to be equal across groups resulted in degraded model fit. If there is not degradation in fit, and thus MI holds, a similar approach is used to assess SI by comparing fit of the model in which τ is allowed to differ across groups, and the model for which it is constrained to be equal. It should be noted here that some authors have recommended using differences in other fit statistics, such as the CFI (Cheung, 2005) in order to ascertain whether more and less constrained models are statistically different.

1.2 Alignment procedure

The approach outlined above for ascertaining whether the model in (1) is invariant across groups has been shown to work well in a variety of conditions (French and Finch, 2006). However, as Asparouhov and Muthén (2014) point out, it can be very difficult to use when the number of groups is large, because finding a good fitting model can become exceedingly challenging, particularly in the context of SI. They demonstrated that an alternative approach, known as the alignment methodology, can be a particularly effective tool in the context of invariance assessment for a relatively large number of groups, such as nations in the PISA assessment (Asparouhov and Muthén, 2014).

The alignment procedure is based upon MGCFA, as described above and which can be expressed as:

$$y_{kig} = \tau_{kg} + \lambda_{kg}\eta_{ig} + \varepsilon_{kig} \tag{4}$$

where y_{kig} = Indicator variable *k* for individual *i* in group g ,

τ_{kg} = Threshold for item *k* in group g ,

λ_{kg} = Factor loading for indicator *k* in group g ,

η_{ig} = Latent trait for individual *i* in group g ,

ε_{kig} = Random error for individual *i* on item *k* in group g .

The alignment procedure is carried out using the following steps:

1 Estimate the configural factor model allowing all λ_{kg} and τ_{kg} to vary across groups, but constraining group factor structure to be the same. This is model M0.

2 Transform factor means (α_g) to 0 and factor variances (ψ_g) to 1 for each group.

3 Express item variances (*V*) and means ($E\left(y_{kg}\right)$) in terms of factor variances, means and item parameter values for M0:

$$V\left(y_{kg}\right) = \lambda_{kg}^2\psi_g = \lambda_{kg0}^2 \tag{5}$$

$$E\left(y_{kg}\right) = \tau_{ik} + \lambda_{kg}\alpha_g = \tau_{kg0} \tag{6}$$

where loadings and intercepts for M0 are

$$\lambda_{kg0} = \lambda_{kg}\sqrt{\psi_g} \tag{7}$$

$$\tau_{kg0} = \tau_{kg} + \frac{\lambda_{kg0}}{\sqrt{\psi_g}}\alpha_g \tag{8}$$

4 For every α_g and ψ_g there are an infinite number of threshold and loading sets that will yield the same fit to the data as does M0, without constraining any loadings or

thresholds to be equal between the groups. The goal of the alignment procedure is to find an alternative model (M1) that yields the same (or close) fit as M0, but does so using a set of factor loadings and thresholds that are as similar among the groups as is possible, unlike was the case for M0. The loadings and thresholds for M1 can be expressed in terms of M0 loadings and thresholds:

$$\lambda_{kg1} = \frac{\lambda_{kg0}}{\sqrt{\psi_g}} \tag{9}$$

$$\tau_{kg1} = \tau_{kg0} - \frac{\lambda_{kg0}}{\sqrt{\psi_g}}\alpha_g \tag{10}$$

5 Values of α_g and ψ_g for step 4 are selected to minimise the amount of group difference in the factor model parameters (i.e. λ_{kg1} and τ_{kg1}), which is done by minimising the loss function, *F*.

$$F = \sum_p \sum_{g1<g2} w_{g1,g2} f\left(\lambda_{pg1,g1} - \lambda_{pg1,g2}\right) + \sum_p \sum_{g1<g2} w_{g1,g2} f\left(\tau_{pg1,g1} - \tau_{pg1,g2}\right) \tag{11}$$

where

$$f(x) = \sqrt{\sqrt{x^2 + 0.0001}}$$

$w_{g1,g2} = \sqrt{N_{g1}N_{g2}}$; weight reflecting groups' sizes.

For every pair of groups (e.g. $g1$, $g2$) F is incremented by the difference in model parameter estimate values.

The alignment procedure is designed to yield a model in which a small number of parameters exhibit relatively large differences across groups, and the great majority of parameters are equal (or nearly so) across groups. In order to ensure model identification, $\prod \psi_g = 1$. The α_g values can be either freely estimated (free alignment) or fixed to 0 for all groups (fixed alignment). This distinction and its implications will be discussed in more detail below.

Once model alignment is completed, individual parameters can be tested for invariance across groups. This testing is done using the following algorithm:

1 For a given parameter (i.e. loading/discrimination, threshold/difficulty) on a target item, compare estimates for each pair of groups with a formal hypothesis test (discussed below). Groups are connected on that parameter if $p > 0.01$ for the test statistic.

2 The largest connected set of groups is identified for the target item parameter.

3 The mean of the target item parameter is calculated for the connected set identified in step 2.

4 The target item parameter for each group is compared with the mean for the connected set using a test statistic. If $p > 0.001$, the group is added to the connected set, and if not the group is removed from the connected set.

5 Repeat steps 3 and 4 until the connected set does not change.

6 Group pairs in which one is not in the connected set and the other is are said to exhibit DIF on the target item parameter.

Parameter estimation for the alignment procedure can be carried out using either maximum likelihood estimation, or the Bayesian framework based on the Markov Chain Monte Carlo (MCMC) methodology. Given prior research (Asparouhov and Muthén, 2014), the Bayesian approach appears to yield more accurate parameter estimates, and more accurate tests of group equality, in the form of the posterior distribution of parameter group differences. For this reason, it was used in the current study.

Although the alignment method has proven to be quite effective in the context of invariance assessment with a large number of groups, as is the case with large scale assessment programs such as PISA and PIRLS, it does have two limitations that may restrict its utility in many situations. First, the alignment procedure only allows for assessment of invariance for one variable at a time. This limitation is certainly also true for the standard MGCFA approach outlined earlier in this paper. However, it does mean that researchers who would like to ascertain whether invariance is present for multiple variables simultaneously (e.g. nation and gender) cannot do so with the alignment or traditional methods. Second, the alignment method is limited to testing invariance for categorical variables only. Therefore, researchers interested in determining whether invariance holds for a variable such as socioeconomic status (as measured on a continuous scale), or family income cannot use the alignment procedure for this purpose. In addition, the alignment method cannot be employed with a continuous variable, such as income, for example. For these reasons, although quite useful in some circumstances, the alignment procedure may not be appropriate for all scenarios involving large scale assessments in which invariance testing may be of interest.

1.3 Structural equation modelling trees

Bandmaier et al. (2013) described an approach for building decision trees, in the context of latent variable modelling, that are an extension of model-based recursive partitioning (MBRP; Zeileis et al., 2008). This approach is known as structural equation modelling tree (SEMtree). With MBRP a researcher is able to assess the extent to which parameters of some model, such as slopes in a linear regression, are stable across the levels of one or more covariates. For example, a researcher may wish to know whether the regression coefficients are different for some nations in the PISA assessment, and/or across values of family income. Thus, we can view MBRP conceptually as a process of repeatedly partitioning the observations within the data set by levels of the covariates in order to create very homogeneous groupings of individuals based upon parameters in the model of interest. This use of covariates to repeatedly divide the data into ever more homogeneous groups with regard to an outcome of interest is frequently referred to as recursive partitioning. MBRP allows for multiple covariates, which can be either continuous or categorical. The result of an MBRP analysis is a decision tree with branches leading through a series of decision points (known as nodes) to a final set of terminal nodes that are differentiated by their values of the model parameters.

In general (including with latent variable models) MBRP is carried out using the following algorithm:

1 Fit the model of choice (e.g. CFA) to all observations in the initial node (i.e. node 1).

2 Assess parameter instability (i.e. inequality) for each possible partition of each covariate (e.g. nation, income) for each independent variable in the model, select the partition with the highest instability, and proceed to step 3. If no instability is present, the algorithm stops.

3 Compute the split point of the model parameter value for the variable identified in step 2 that optimises partitioning group (e.g. nations) separation by yielding the most homogeneous (with respect to model parameters) child nodes possible.

4 Divide the node based upon this split point to create two child nodes in which members of the partitioning variable with the most similar parameter estimates from step 3 are placed together.

5 Repeat steps 1 through 4 until stability is achieved for all model parameters. In the context of SEMtrees, the model parameters of interest can be those in Eq. (1), or structural coefficients linking latent variables in a SEM.

By default, model parameter estimation in step 1 is carried out using maximum likelihood estimation. However, as Brandmaier et al. (2013) note, this is not a requirement of the SEMtree approach, and indeed use of any alternative method for estimation, such as robust weighted least squares, is possible with SEMtrees. In the context of invariance assessment, parameter stability with SEMtree is assessed using a likelihood ratio test comparing the fit of a model in which parameters in Eq. (1) are constrained to be equal across all potential subgroups in the sample, and a model in which a split based upon a covariate is made. To provide context, the initial model can be referred to as M_0 and the observations in the node treated as a single group. The parameters in (1) are estimated and a log-likelihood value is obtained for M_0. The node is then split into two parts based upon one of the covariates. For example if nation is a covariate, then the first split might put residents of Nation 1 in one group, and residents of the other nations in the sample in the other group. The factor model is then fit such that the parameters are allowed to differ between the two splits. The log-likelihood for this second model, M_1, is then obtained. This step of dividing the data into two parts is repeated for every possible split of each of the covariates, and the resulting log-likelihood values are obtained. It can be shown that model M_0, in which all members of the sample are treated as a single group, is nested within the family of all possible splits models described above (Brandmaier et al., 2013). This fact allows for the comparison of model fit through the log-likelihood ratio as calculated below:

$$LR = LL_1 - LL_0 \tag{12}$$

where $LL_1 = \text{Loglikelihood value for model } M_1$,

$LL_0 = \text{Loglikelihood value for model } M_0$,

$LR \text{ is distributed as a } \chi^2 \text{ statistic with } (k-1)m \text{ degrees of freedom}$,

$k =$ Number of splits ,

$m =$ Number of freely estimated parameters.

If no splits yields a statistically significant *LR* result, then growing of the tree stops. If there are multiple statistically significant *LR* results, the split that produces the greatest maximisation in the likelihood is chosen.

An important issue that must be dealt with in this algorithm is control of the Type I error rate due to the multiple testing that is inherently a part of the tree building process. Brandmaier et al. (2013) discuss two approaches for dealing with this problem. The first is the standard correction of the Type I error rate using the Bonferroni approach. However, they note that this is a conservative technique for dealing with the Type I error inflation problem, and may lead to trees that are not sufficiently complex so as to capture true group differences in the model parameters. Another approach that they describe involves the use of cross-validation samples. With this approach, the original dataset will be divided into *j* cross-validation subsets (S_j). Then, for each node, each potential split based upon each covariate will be tested using the full set of observations in that node, *S*, minus set S_j. The value of *LR* will then be calculated using the observations in subset S_j, and the model parameters obtained from the rest of the observations, as described in the previous step. The overall value for *LR* will then be taken as the average of the *JLR* estimates. Brandmaier et al. (2013) recommend using the cross-validation approach with SEMtree. They do note that the distribution of LR is no longer χ^2 in this case, so that the decision to accept a split should be based upon whether the split yields an *LR* value greater than 0, indicating that treating the subgroups differently yields superior fit to treating them as a unitary group. In this way, the splits are descriptive in nature, rather than inferential.

In addition to Type I error rate inflation, a second potential problem associated with tree-based models is their tendency to use variables with more categories for splitting more often than variables with fewer categories, even if the latter actually are more important in the population (Jensen and Cohen, 2000). A number of approaches have been suggested for dealing with this issue, including use of hypothesis testing results rather than reduction in node heterogeneity for selecting the variable to use in a split (Loh and Shih, 1997), separation of the variable selection and split point decisions (Shih, 2004), and examination of model residuals produced by using specific covariates for splitting (Loh, 2002). SEMtree uses an approach described by Kim and Loh (2001) involving the initial, independent identification of the optimal split point for each variable, and then selection of the optimal variable to use at a given split.

1.4 Study goals

The purpose of the current study is to demonstrate the use of SEMtree for investigating a lack of factor invariance for multiple variables simultaneously in the context of large scale assessments. Specifically, data from the 2009 administration of the PISA assessment (OECD, 2009) were used, with a particular focus on the 11 items of the reading interest scale. Invariance was investigated for the variables nation of residence, examinee sex, language spoken in the home (language of test or not) and a continuous indicator of socioeconomic status (SES). Of primary interest was the extent to which invariance held across these variables, and whether there were interactions of the variables that were associated with a lack of invariance. In other words, it was of interest to determine whether, and how, model parameters differed for combinations of nation, sex, language spoken in the home and SES. Standard invariance analyses would examine each of these variables individually, but SEMtree can ascertain whether there exists a lack of invariance for combinations of them, which is often expressed in the context of an interaction among the variables. In addition to SEMtree, the alignment method described by Asparouhov and Muthén (2014) were also used to address the invariance question.

2 Method

The current study was conducted using the individual level data from PISA 2009, which is an international assessment of academic achievement focused on 15-year-old examinees from 63 nations, and their performance in reading and mathematics. A total of 475,460 students (50.3% female) were included in the initial PISA data collection (OECD, 2009). For the purposes of the current study, 8,447 participants from the 20 wealthiest nations, as measured by national GDP, were included in the study. The list of these nations appears in Table 1.

Table 1 Percent female, percent language of the test spoken at home and mean (SD) SES index by nation

Nation	*Female (%)*	*Test language spoken at home (%)*	*Mean (SD) SES index*
Australia	46.9	90.6	0.29 (0.77)
Austria	51.1	85.9	0.06 (0.79)
Belgium	52.2	75.4	0.23 (0.92)
Canada	54.5	82.1	0.51 (0.83)
Denmark	52.4	82.5	0.15 (0.92)
Finland	51.7	91.4	0.30 (0.82)
France	52.6	87.2	−0.21 (0.90)
Germany	50.6	78.1	0.16 (0.83)
Greece	49.2	94.2	0.06 (0.99)
Hong Kong	45.5	90.6	−0.83 (1.01)
Ireland	48.2	89.8	0.08 (0.85)
Japan	51.2	97.9	−0.02 (0.74)
The Netherlands	49.1	90.6	0.35 (0.85)
Norway	49.4	93.9	0.44 (0.77)
Singapore	49.5	39.4	−0.50 (0.86)
Spain	47.8	83.7	−0.28 (1.06)
Sweden	51.0	89.5	0.31 (0.83)
Switzerland	50.4	79.2	0.07 (0.87)
UK	52.3	91.3	0.15 (1.08)
USA	52.6	83.4	0.07 (1.01)
Overall	50.5	84.4	0.09 (0.94)

In addition to the achievement testing variables, PISA also involves the collection of other information, including demographic data, and scales designed to measure a variety of constructs such as metacognitive strategy use, relationships with teachers and aspects of motivation. For the current study, a reading attitude scale will be assessed for model invariance. This scale is comprised of 11 ordinal items, measured on 4-point scale, where 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Agree* and 4 = *Strongly Agree*. The items appear in Table 2. When a total score is calculated, several items are reverse coded so that higher total scores reflect a more positive attitude towards reading. In this study, reading attitude was treated as a latent construct, and thus no reverse coding was necessary.

Table 2 Standardised factor loadings and intercepts for the attitudes towards reading items for full sample

Item	*Loading (SE)*	*Intercept (SE)*
I1: I read only if I have to	0.769 (0.005)	2.360 (0.021)
I2: Reading is one of my favourite hobbies	−0.817 (0.004)	2.179 (0.020)
I3: I like talking about books with other people	−0.754 (0.005)	2.277 (0.021)
I4: I find it hard to finish books	0.507 (0.009)	2.209 (0.020)
I5: I feel happy if I receive a book as a present	−0.753 (0.005)	2.393 (0.022)
I6: For me, reading is a waste of time	0.754 (0.005)	2.028 (0.019)
I7: I enjoy boing to a bookshop or a library	−0.739 (0.006)	2.305 (0.021)
I8: I read only to get information that I need	0.654 (0.007)	2.545 (0.023)
I9: I cannot sit still and read for more than a few minutes	0.594 (0.008)	2.026 (0.019)
I10: I like to express my opinions about books I have read	−0.584 (0.008)	2.662 (0.023)
I11: I like to exchange books with my friends	−0.693 (0.006)	2.174 (0.020)

Several variables were assessed for invariance, including nation of residence, examinee sex, language spoken in the examinee's home (language of the test or not) and the socioeconomic status of the examinee's family. The PISA SES index is derived from a factor analysis of variables that include parental education levels, parental occupations and the possessions in the home. This SES variable is scales with a mean of 0 and a standard deviation of 1, so that positive values indicate that an examinee's family SES is above average and a negative value means that it is below average. It is on a common scale across the world.

As mentioned previously, invariance assessment was conducted using the alignment procedure (Asparouhov and Muthén, 2014) and SEMtree (Brandmaier et al., 2013). Because the alignment procedure is limited to assessing invariance for one variable at a time, it was only used to determine whether invariance held cross-nationally. The MCMC algorithm was used to fit the alignment model, with 2 separate chains, and a total of 20,000 links in each chain. The first 5,000 links in the chain served as the burnin period, and the chains were thinned such that every 100th value was sampled, leading to a total of 1,500 observations in the posterior distribution of each parameter. The median of the posterior was taken as the point estimate for each of the parameters. Model convergence was assessed using convergence plots, as well as the proportional scale reduction (PSR), which had to be less than 1.05 for all parameters in order for convergence to be attained. The alignment procedure did in facto reach convergence for all model parameters. SEMtree was also be used in this context so that the results of the methods could be compared with one another. In addition, SEMtree was also be used with nation, as well as examinee sex, language spoken in the home, and SES as potential splitting variables. The settings used for SEMtree were the defaults. Therefore, the significance level for determining splits was 0.05, the Bonferroni correction was used to control the Type I error rate, the minimum number of individuals to be included in each node was 20, splits resulting in Heywood cases were excluded from the tree, and the method for determining splits as described above and by Kim and Loh (2001) were used. In order to ensure the generalisability of the final tree in terms of identification of optimal splits, fivefold cross-validation was used, per recommendations by Brandmaier et al. (2013), and as described above. For both the alignment procedure and SEMtree, a one factor CFA model was fit to the reading attitude items.

3 Results

3.1 Sample description

Table 1 includes the percentage of examinees who were female, and the percent for whom the test language was also the primary language spoken in the home, by nation. In addition, Table 1 also includes the mean and standard deviation of the SES index by nation. From these results, it is apparent that Singapore had the lowest percent of examinees who spoke the test language in their homes (39.4%) and Japan had the highest such percentage (97.9%). Overall, 84.4% of the sample reported the test language as the primary tongue spoken in the home. Across the nations included in the sample, 50.5% were female, with Hong Kong reporting the lowest percent of female examinees, and Canada reporting the highest percent. Finally, the mean and standard deviation of the SES index by nation appears in the final column of Table 1. The nation in this sample with the lowest mean SES was Singapore, whereas the nation with the wealthiest students in the sample was Canada. The overall mean SES index across nations was 0.09, with a standard deviation of 0.94.

The CFA model for reading was fit to the full sample, ignoring nation, sex, home language and SES. The purpose behind fitting this model was to ascertain whether it fit the full sample. The CFI and TLI were 0.915 and 0.894, respectively, whereas RMSEA was 0.104, with a confidence interval of (0.101, 0.107). The SRMR for this model was 0.048. Considered together, these fit indices suggest that the model did not provide a very good fit to the full sample, based on commonly used guidelines for what constitutes reasonable fit (Kline, 2016); i.e. RMSEA ≤ 0.08, CFI and TLI > 0.95 and SRMR < 0.10. The standardised factor loadings and intercepts, as well as their associated standard errors, appear in Table 2. These loadings were all statistically significant ($\alpha = 0.05$), indicating that the individual items were each related to the reading attitude factor. In addition, the absolute values of the standardised loadings were all above 0.58, and generally above 0.60, except for the item "I find it hard to finish books". The items that were associated with a positive attitude towards reading displayed negative loadings, whereas the measures of a negative reading attitude had positive loadings. This is because the first item, "I read only if I have to" serves as the reference indicator giving the factor its scale, and it is a measure of negative reading attitude. Thus, higher values on the latent trait would indicate more negative attitudes towards reading.

3.2 Alignment

The alignment procedure was used first to determine whether there was a lack of factor model invariance by nation. Recall that this method seeks to identify subsets of the grouping variable of interest (nation in this case) for which each of the factor model parameters are invariant. Table 3 contains the standardised factor loadings for the attitudes towards reading items by nation. Loadings found to be noninvariant are highlighted in bold. For item 1, France and Greece were found to have noninvariant factor loadings when compared to those of the other nations, with their loadings being somewhat lower than those of the other countries included in the sample. These results suggest that for Greece and France item 1 ("I read only if I have to") was more weakly related to the reading factor than it was for the other nations. A similar result for Greece was found for item 2 ("Reading is one of my favourite hobbies"), whereby the item was more weakly related to the factor than was the case for the other nations. In addition to Greece, noninvariance was also identified for item 2 with respect to Japan. However, in this case, the absolute value of the loading for Japan was larger than for the other nations (except Hong Kong), indicating that it was more strongly related to the reading factor. Note that although the loading estimate for Hong Kong was larger than for Japan, it was not found to be significantly different from the loadings for the other nations. This is due to the relatively large standard error for the Hong Kong item 2 loading (0.091) as compared to that of Japan (0.050), and indeed all of the other nations, which had values comparable to that of Japan. Similar interpretations of the results can be made for the loadings of each of the other items on the scale.

Table 3 Standardised factor loadings for the attitudes towards reading items by nation

Nation	*I1*	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*
Overall	0.769	−0.817	−0.754	0.507	−0.753	0.754	−0.739	0.654	0.594	−0.584	−0.693
Australia	0.805	−0.822	−0.690	0.549	−0.667	0.839	−0.809	0.682	**0.699**	−0.603	−0.609
Austria	0.899	−0.782	−0.720	0.561	−0.785	0.822	−0.703	0.649	0.466	−0.624	−0.728
Belgium	0.838	−0.758	−0.654	0.560	−0.784	0.896	−0.780	**0.497**	0.602	−0.642	−0.684
Canada	0.786	−0.821	−0.747	0.485	−0.759	0.758	−0.781	0.652	0.626	−0.619	−0.738
Denmark	0.738	−0.786	−0.805	0.466	−0.838	0.783	−0.888	0.683	0.531	−0.493	−0.720
Finland	0.822	−0.886	−0.782	0.505	−0.729	0.785	−0.835	0.712	0.478	−0.580	−0.615
France	**0.604**	−0.714	−0.765	0.551	−0.755	0.810	−0.622	0.579	0.698	−0.669	−0.849
Germany	0.869	−0.866	−0.741	0.554	−0.754	0.825	−0.710	0.763	0.497	−0.541	−0.671
Greece	**0.476**	**−0.599**	−0.780	0.439	−0.794	**0.480**	−0.900	0.610	0.501	−0.731	−0.784
Hong Kong	0.684	−1.047	−0.750	0.842	−0.731	0.618	−0.881	0.453	0.852	−0.400	−0.617
Ireland	0.838	−0.789	−0.726	0.553	−0.655	0.804	−0.710	0.554	**0.866**	−0.663	−0.768
Japan	0.928	**−0.974**	−0.808	0.660	−0.720	**0.637**	−0.631	**0.488**	**0.648**	−0.535	−0.619
The Netherlands	0.927	−0.760	−0.670	0.492	−0.744	0.902	−0.820	0.692	0.569	−0.475	−0.735
Norway	0.810	−0.771	−0.815	0.407	**−0.946**	0.740	−0.790	0.779	0.591	−0.522	−0.648
Singapore	0.857	−0.848	−0.699	0.534	−0.777	0.723	−0.765	0.765	0.565	−0.635	−0.547
Spain	0.776	−0.876	**−0.828**	**0.426**	**−0.848**	**0.648**	−0.703	0.617	0.527	−0.582	**−0.805**

Table 3 Standardised factor loadings for the attitudes towards reading items by nation (continued)

Nation	*I1*	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*
Sweden	0.848	−0.845	−0.727	0.432	−0.727	0.737	−0.817	0.718	0.604	−0.641	−0.663
Switzerland	0.890	−0.807	−0.678	0.608	−0.777	0.809	−0.765	0.664	0.558	−0.552	−0.660
UK	0.764	−0.809	−0.766	0.560	−0.696	0.742	−0.757	0.599	0.706	−0.665	−0.735
USA	0.800	−0.787	−0.689	0.607	−0.746	0.835	−0.745	0.647	0.685	−0.575	−0.750

Bold indicates a lack of loading invariance

Table 4 includes the factor intercept estimates by nation, with noninvariant results highlighted in bold. Perhaps most notable is that there were many more noninvariant intercepts than there were noninvariant loadings (57 versus 16). Intercepts reflect the location of the item, so that this relatively large number of noninvariant results indicate that respondents from different nations were likely to provide different mean responses to the items, after accounting for the factor means. A detailed review of these results is beyond the scope of the current manuscript. However, a review of item 1 can provide guidance for understanding the other results in the table. For item 1 ("I read only if I have to"), the intercepts for Belgium, Hong Kong, Japan, Norway, Sweden and Switzerland were found to be noninvariant when compared to the other nations in the sample. A further examination of the values of these estimates shows that those of Belgium, Norway, Sweden and Switzerland were lower than those of most other nations, whereas the estimates for Hong Kong and Japan were larger than most others. Therefore, it would appear that respondents in Hong Kong and Japan were more likely than respondents in the other nations to agree that they read only if they had to, and those in Belgium, Norway, Sweden and Switzerland were less likely to agree with this statement. Again, a detailed review of these results can reveal other, similar types of results.

Table 4 Factor intercepts for the attitudes towards reading items by nation

Nation	*I1*	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*
Overall	2.360	2.179	2.277	2.209	2.393	2.028	2.305	2.545	2.026	2.662	2.174
Australia	2.103	2.417	2.388	1.991	2.611	1.786	**2.608**	2.189	1.802	**2.530**	2.194
Austria	2.100	**2.202**	2.266	1.751	2.650	1.772	2.306	2.223	**1.498**	**2.953**	2.286
Belgium	**1.977**	2.267	2.339	1.957	**2.432**	1.748	2.511	2.188	1.784	2.781	2.346
Canada	2.142	2.385	2.410	1.958	**2.474**	**1.824**	**2.611**	2.186	1.845	2.619	**2.400**
Denmark	2.178	2.307	2.447	**1.716**	2.489	1.726	2.505	2.254	1.699	3.103	2.138
Finland	2.030	2.356	**2.232**	1.895	2.564	**1.853**	2.554	2.158	1.640	2.710	**2.050**
France	1.969	2.249	2.492	**2.159**	**2.374**	1.770	2.542	2.223	1.866	2.829	**2.615**
Germany	2.013	2.246	2.232	1.765	2.641	1.848	**2.205**	2.170	**1.554**	2.848	2.249
Greece	2.227	2.286	2.310	**2.181**	**2.415**	1.568	2.540	2.173	1.948	**2.952**	**2.507**
Hong Kong	**2.425**	**2.658**	2.529	2.073	2.525	1.767	2.680	**2.404**	1.948	2.571	**2.472**
Ireland	2.022	**2.416**	2.377	2.098	2.517	1.731	2.594	2.192	1.913	2.530	2.396
Japan	**2.273**	**2.384**	2.360	1.930	2.344	1.683	**2.910**	**1.881**	1.709	**1.966**	2.229
The Netherlands	2.150	2.136	**2.155**	1.691	2.663	1.747	2.471	2.083	1.678	2.458	2.309

Table 4 Factor intercepts for the attitudes towards reading items by nation (continued)

Nation	*I1*	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*	*I10*	*I11*
Norway	**1.977**	2.151	2.291	1.801	2.543	1.771	2.268	2.157	1.635	2.868	2.126
Singapore	2.232	2.623	2.456	**2.237**	2.591	1.679	**2.958**	**2.426**	1.860	2.550	2.376
Spain	2.174	2.348	2.401	**2.198**	**2.389**	1.660	**2.244**	2.196	1.758	**2.828**	**2.456**
Sweden	**1.944**	2.328	2.349	1.738	**2.352**	1.807	2.445	**1.931**	1.635	2.669	2.162
Switzerland	**1.993**	**2.240**	**2.190**	1.847	2.605	1.751	2.448	2.201	**1.639**	2.759	2.222
UK	2.069	2.340	2.449	**2.035**	2.625	1.704	2.450	2.231	1.844	2.604	2.311
USA	2.276	2.408	2.408	1.946	**2.431**	1.756	**2.710**	2.202	1.898	2.668	**2.438**

Bold indicates a lack of loading invariance

3.3 SEMtree

The SEMtree approach was also used to assess the invariance of model parameters for the reading data, with two models being fit. First, a tree that only included nation as a splitting variable was used in order to obtain results that were comparable to those from the alignment procedure, which can only examine one variable at a time. The second model included nation, family SES, student sex and home language (language of test or not). The SEMtree based on nation only appears in Figure 1, and the parameter estimates for the terminal nodes from this initial model appear in Table 5. Note that the original tree created by the SEMtree function in R is difficult to read. For this reason, it was recreated here using another software package. With regard to the organisation of the nations by the SEMtree, Japan (Node 1) and Hong Kong (Node 5) were each placed in their own unique terminal nodes, whereas Greece and Spain were grouped together in Node 4. Node 2 consisted of Austria, Denmark, Finland, Germany, the Netherlands, Norway, Sweden and Switzerland, and Node 3 included Australia, Belgium, Canada, France, Ireland, Singapore, the USA and the UK. These nodes reflect groupings of nations that had statistically equivalent factor loadings and intercepts. In other words, within the nodes, the factor model parameters were found to be invariant. Thus, we can conclude that Japan and Hong Kong each had unique patterns of loadings and intercepts that differentiated them from all of the other nations included in the sample. In addition, Greece and Spain also had different factor model parameters from all of the other sampled nations, but were not different from one another. Finally, within this sample there were two large groupings of nations that had statistically equivalent factor loadings and intercepts within the groups, but which differed from the other nations in the sample.

Figure 1 SEMtree for model including nation only (see online version for colours)

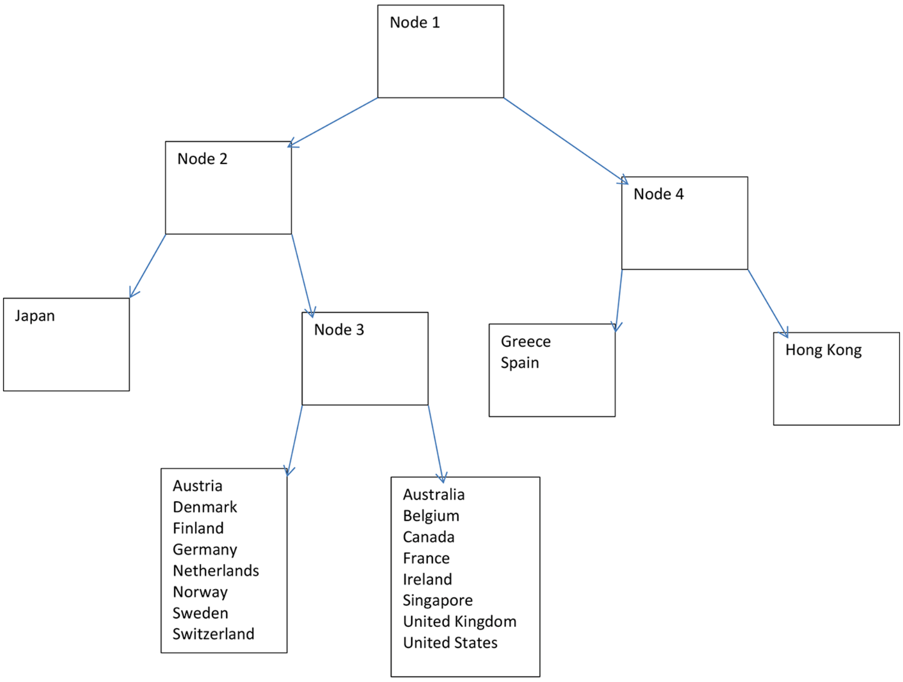


Table 5 Factor loadings for the attitudes towards reading items by SEMtree terminal node for nation only tree

Item	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*
I1	0.938	0.850	0.787	0.626	0.684
I2	−0.974	−0.813	−0.794	−0.738	−1.047
I3	−0.808	−0.742	−0.717	−0.804	−0.750
I4	0.660	0.503	0.550	0.433	0.842
I5	−0.720	−0.788	−0.730	−0.821	−0.731
I6	0.637	0.800	0.801	0.564	0.618
I7	−0.631	−0.791	−0.746	−0.802	−0.881
I8	0.488	0.708	0.622	0.614	0.435
I9	0.648	0.537	0.681	0.514	0.852
I10	−0.535	−0.554	−0.634	−0.657	−0.400
I11	−0.619	−0.680	−0.710	−0.795	−0.618

It is important to note here that although we can say with confidence that there are statistically significant differences among the nations with respect to the factor loadings and intercepts, it is not possible to identify specific loadings and intercepts that differ among the terminal nodes presented in Figure 1. In other words, we may see patterns of loadings and intercepts that appear to differ across the nodes, but we cannot say that any one of them is statistically significantly different among the nations. For this reason, we will discuss the trees in terms of patterns of loadings and intercepts that we see. From

Figure 1, we can see that there were a total of five terminal nodes in the tree including nation only. The nations included in each terminal node appear at the base of the tree. Table 5 includes the factor loadings for each terminal node in the nation only tree. From these results, several distinct patterns can be seen. First, the loading for Node 1, which contains Japan, were larger for item 1 than was the case for any of the other terminal nodes, indicating that this item was more strongly related to the latent trait of reading attitude for individuals in this node than for those in other nodes. In addition, loadings for Node 5, containing only Hong Kong, were larger for items 2, 4, 7 and 9 than was true for the other terminal nodes. Nodes 2 and 3 had the highest loadings for item 6, as did Node 2 for item 8. Node 4 had the largest loadings for items 5 and 11.

Table 6 contains the factor intercepts for each item by terminal node. As with the loadings, it is possible to examine these intercepts with an eye towards patterns that differentiate the nodes from one another. For example, the intercepts for item 7 was largest for Node 1 (Japan). This result indicates that holding the factor mean constant, the mean of this item was larger for Node 1 than for the other nodes. Node 2 had the lowest intercepts among the nodes for items 4 and 9. Node 5 had the largest intercept values for items 1, 2, 3, 7 and 8. Nodes 3 and 4 had generally similar intercept estimates, with the exception of items 7 and 10.

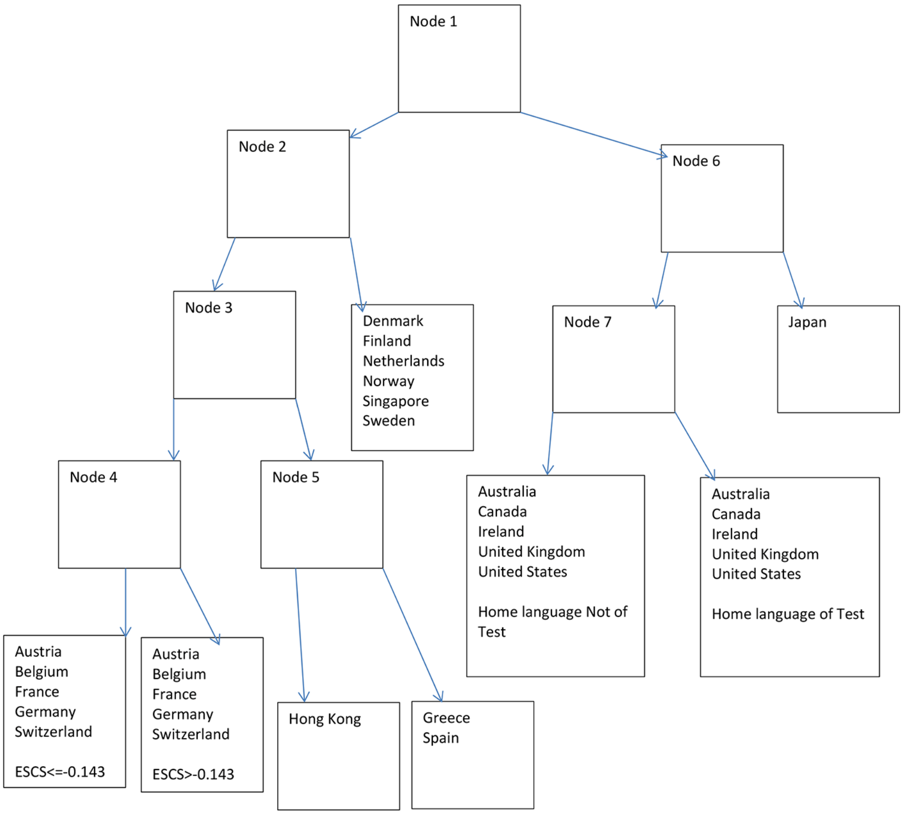
Table 6 Factor intercepts for the attitudes towards reading items by SEMtree terminal node for nation only tree

Item	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*
I1	2.273	2.048	2.100	2.201	2.425
I2	2.384	2.246	2.388	2.317	2.658
I3	2.360	2.270	2.415	2.356	2.529
I4	1.930	1.776	2.048	2.190	2.073
I5	2.344	2.563	2.507	2.402	2.525
I6	1.683	1.784	1.750	1.614	1.767
I7	2.910	2.400	2.623	2.392	2.680
I8	1.881	2.147	2.230	2.185	2.404
I9	1.709	1.622	1.852	1.853	1.948
I10	1.966	2.796	2.639	2.890	2.571
I11	2.229	2.193	2.385	2.482	2.472

As mentioned previously, one of the strengths of the SEMtree approach is the ability to include multiple covariates of any variable type (dichotomous, ordinal, nominal, count, continuous) for creating the tree. Therefore, a second tree was grown using the aforementioned PISA data, using nation, student sex, family SES (ESCS index), and language spoken in the home (language of test or not). The resulting tree appears in Figure 2. A total of 8 terminal nodes were needed for this second tree. Terminal nodes 1 and 2 each include Austria, Belgium, France, Germany and Switzerland. They are differentiated by the SES index value (ESCS), such that Node 1 includes those individuals from the aforementioned nations whose family ESCS value was less than or equal to −0.143. Node 2 includes those whose ESCS was greater than −0.143. Nodes 3 and 4 include Hong Kong, and Greece and Spain, respectively. Node 5 includes individuals from Denmark, Finland, the Netherlands, Norway, Singapore and Sweden.

Node 8 includes only Japan. Finally, Node 6 includes individuals from Australia, Canada, Ireland, the UK and the USA whose home language is not the language of the test, whereas Node 7 consists of examinees from those five nations whose home language is the language of the test.

Figure 2 SEMtree for model including nation, SES, sex and home language (see online version for colours)



When interpreting the results of the SEMtree, it again needs to be pointed out that these differences are for the sample and not necessarily true in the population, and that the SEMtree does not isolate specific differences among model parameters, but rather identifies differences in whole patterns of model parameters. Thus, while it is not incorrect to discuss apparent sample differences as just that, differences for the current sample, the different groups represented by the terminal nodes are representative of collective differences across the subgroups. Table 7 contains the factor loadings for each terminal node created by the SEMtree in Figure 1. When considering this tree, it might be of particular interest to consider how the model parameters differ between individuals from the same nations who differ on one of the demographic variables used in the splitting. For example, Nodes 1 and 2 differ based upon the level of family SES, with those in Node 1 having a lower value than those in Node 2. In terms of the factor loadings, we can see differential patterns across items such that individuals in Node 2 had somewhat larger loadings on items 1, 6 and 8. In contrast, the loadings for Node 1 were larger for items 2 and 9. It should be noted that none of these sample differences in loadings were particularly large in magnitude. Nodes 6 and 7 also include individuals from the same set of nations, but who have different home languages. Node 6 includes those individuals whose home language is not the language of the test, whereas Node 7 includes examinees who speak the language of the test in their homes. As was true for Nodes 1 and 2, the factor loadings for these nodes are generally fairly similar across items, with Node 6 having somewhat larger values for items 3, 4, 6, 8 and 9. Conversely, Node 7 had larger values for items 2 and 10. Given space limitations, we will not carefully dissect factor loading differences between each of the nodes, though this could certainly be done, given these results.

Table 7 Factor loadings for the attitudes towards reading items by SEMtree terminal node for tree including nation, sex, language spoken in the home and family SES

Item	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*	*Node 6*	*Node 7*	*Node 8*
I1	0.790	0.833	0.684	0.626	0.834	0.787	0.799	0.938
I2	−0.822	−0.794	−1.047	−0.738	−0.816	−0.781	−0.810	−0.974
I3	−0.710	−0.721	−0.750	−0.804	−0.750	−0.738	−0.715	−0.808
I4	0.545	0.570	0.842	0.433	0.473	0.577	0.548	0.660
I5	−0.783	−0.764	−0.731	−0.821	−0.794	−0.711	−0.696	−0.720
I6	0.821	0.855	0.618	0.564	0.778	0.810	0.786	0.637
I7	−0.714	−0.717	−0.881	−0.802	−0.819	−0.760	−0.773	−0.631
I8	0.600	0.659	0.435	0.614	0.725	0.674	0.558	0.488
I9	0.578	0.553	0.852	0.514	0.556	0.738	0.699	0.648
I10	−0.619	−0.603	−0.400	−0.657	−0.558	−0.611	−0.648	−0.535
I11	−0.724	−0.709	−0.618	−0.795	−0.655	−0.721	−0.720	−0.619

The factor intercept estimates for the terminal nodes in the second SEMtree appear in Table 8. Recall that these represent the means of the items for the nodes when we hold the factor means constant. From these values, we can see that the intercepts for Node 1 were generally larger than for Node 2 on items 1, 4, 6 and 9, whereas the intercepts for Node 2 were larger for items 2, 3, 7, 10 and 11. Considering the content of the items, it appears that the intercepts for items that reflected a more negative attitude towards reading were larger for individuals from Austria, Belgium, France, Germany and Switzerland who were relatively less well off financially, whereas for examinees from these nations who had a relatively higher family SES, intercepts were larger on the items reflecting a more positive attitude towards reading. In terms of Nodes 6 and 7, larger intercept values were found for Node 7 on items 2, 7, 10 and 11. Node 6 had larger intercept estimates for items 4, 6 and 9. Taken together, it appears that individuals from Australia, Canada, Ireland, the UK and the USA who did not speak the language of the test in their homes had larger intercept estimates for items that were generally less positive with respect to attitudes towards reading. However, examinees from these nations who did speak the test language in their homes had larger intercept values on several items associated with more positive attitudes towards reading. Similar reflections on all of the terminal nodes could be made. However, in the interest of keeping the manuscript size manageable, this will not be done here.

Table 8 Factor intercepts for the attitudes towards reading items by SEMtree terminal node for tree including nation, sex, language spoken in the home and family SES

Item	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*	*Node 6*	*Node 7*	*Node 8*
I1	2.305	1.909	2.425	2.201	2.085	2.104	2.291	2.273
I2	2.033	2.417	2.658	2.317	2.317	2.065	2.559	2.384
I3	2.108	2.556	2.529	2.356	2.322	2.301	2.489	2.360
I4	1.999	1.750	2.073	2.190	1.846	2.100	1.898	1.930
I5	2.472	2.671	2.525	2.402	2.534	2.527	2.544	2.344
I6	1.803	1.688	1.767	1.614	1.764	1.886	1.639	1.683
I7	2.296	2.632	2.680	2.392	2.534	2.334	2.791	2.910
I8	2.214	2.185	2.404	2.185	2.168	2.185	2.238	1.881
I9	1.893	1.522	1.948	1.853	1.691	1.947	1.676	1.709
I10	2.767	2.998	2.571	2.890	2.726	2.419	2.638	1.966
I11	2.117	2.548	2.472	2.482	2.194	2.205	2.888	2.229

4 Discussion

The goal of this study was to demonstrate the utility of SEMtree (Brandmaier et al., 2013), a new approach for latent variable modelling, for invariance testing with structural equation modelling with large scale assessments, and to compare this method with another relatively new technique that has been shown to be useful in similar circumstances, the alignment procedure (Asparouhov and Muthén, 2014). The results of this study show that SEMtree presents several advantages to the researcher working with large scale assessment data, and who is interested in invariance assessment. First, as was demonstrated here, SEMtree provides for the identification of subgroups within the data, based upon model parameter differences. For example, when considering Figure 1, we can conclude that there exist subsets of nations with statistically equivalent factor model parameters. In other words, for these subgroups the reading interest scale was found to be invariant. In this case, when considering the PISA 2009 data and focusing on nation only, SEMtree identified five groupings of countries, represented by the terminal nodes in Figure 1. Nations within the same node have a completely invariant set of factor model parameters, whereas nations in different nodes have at least one factor model parameter difference. Thus, we can see that Japan and Hong Kong have unique factor models when compared to all other nations in the sample. Greece and Spain have equivalent factor models to one another, that differ from those of the other sampled nations, and there exist two larger groups of nations that have invariant factor model parameter values.

These results would seem to have several implications for practice. First, the fact that subgroups were found within the data, based on nation, SES and language spoken in the home, suggest that researchers consider such differences when comparing scores on the attitudes towards reading scale across nations, and subgroups based upon language spoken in the home and SES, as identified by the trees. Essentially, the results presented here provide the same information as would a standard invariance study (Millsap, 2011). Thus, for example, when nations appear in different terminal nodes, we can conclude that a lack of CFA model parameter invariance has been found. Given this fact, we must then consider the impact of this lack of invariance on the scale scores, and on the meaning of the scale for individuals in different nodes. The impact of such partial invariance on scale interpretation is one of ongoing research, without a clearly defined answer as of this writing (Steinmetz, 2013; Millsap and Meredith, 2007). Millsap and Meredith describe three options for researchers faced with the issue of partial invariance

1. drop indicators that are shown to be noninvariant across groups and calculate composite scores using only those that are fully invariant
2. avoid using the scale altogether, or
3. ignore the presence of noninvariant parameters and argue that they are not sufficiently different or sufficiently numerous as to be problematic.

Each of these approaches to dealing with PI presents the researcher with problems, and no clear solution that works in every instance of noninvariance has been identified. Millsap and Kwok (2004) described how, in the context of classification using a latent variable, invariance can be dealt with. Likewise, Byrne (1989) described the modelling of partial invariance when estimating factors for members of different subgroups. However, as stated above, there does not exist a well-defined rule for determining when invariance might be problematic, though work on effect sizes for this purpose are ongoing (Millsap and Olivera-Aguilar, 2012).

With these issues in mind, let us consider just one example from Figure 2. Nodes 6 and 7 contain individuals from the same nations, but who differ based upon whether the language of the test is spoken in the home, or not. For those who do not speak the test language at home, the loading for item 8 (I read only to get information that I need) is larger (by 0.12) than for those who do speak the test language at home. This suggests that the relationship between the attitudes towards reading factor and the item is stronger for those who do not speak the test language at home. In contrast, the intercept of item 11 (I like to exchange books with my friends) is higher for those who do speak the test language at home than those who do not. In other words, when the attitude towards reading is the same for respondents in both groups, those who speak the test language are more likely to agree with this item. Such results may be interesting in themselves, as they might highlight differences in the way that children with different language backgrounds might feel about reading. In this example, it would appear that there is some evidence to support the notion that for respondents who do not speak the test language and who live in Australia, Canada, Ireland, the UK or the USA, reading may be seen as somewhat more utilitarian than for those who do speak the test language. In addition, as discussed previously it is not clear to what extent these noninvariant findings might impact the total reading attitude score. We might consider the impact in this case to be fairly minor given that there were not large differences in loadings and intercepts for many items. We must keep in mind, however, that it is not clear how much invariance is needed in order to fundamentally alter the meaning of scales. For this reason, employment of effect sizes (Millsap and Olivera-Aguilar, 2012) may prove to be useful. In order to fully describe the results of SEMtree, we would engage in a similar detailed exploration of model parameter differences across the various nodes, in order to characterise the individuals contained therein.

A second major advantage of SEMtree for invariance assessment, in addition to the identification of subgroups with invariant model parameters, is its ability to include multiple variables of various types when determining where invariance does (and does not) hold. This was represented by the SEMtree model appearing in Figure 2. In addition to nation, the student sex, family SES and language spoken in the home were also included in the analysis. These results showed that for some nations these variables were salient with respect to the invariance of factor model parameters, and for other nations they were not. As an example, for Austria, Belgium, France, Germany and Switzerland, family SES was associated with noninvariant loadings and/or intercepts, such that those with relatively lower SES index values had different model parameters than did those with relatively higher family SES. Similarly, noninvariance associated with the language spoken in the home was found to be present for Australia, Canada, Ireland, the UK and the USA. For the other nations in the sample, however, no such demographic sources of factor model parameter noninvariance were identified. Therefore, SEMtree allows for a more tailored identification of noninvariance by combinations of variables, rather than in a more broad brush fashion by looking at one variable (e.g. nation) at a time.

A third positive aspect of SEMtree is the identification of subsets of the variable of interest (e.g., nation) for which the groups had equivalent factor model parameter values. Other methods, such as the alignment procedure, identify specific model parameters that differ across the groups, but do not provide information about which of the individual groups might have statistically equivalent model parameters. SEMtree does provide such information, allowing the researcher to make some conclusions regarding for which nations, in the case of the current study, examinee scores could be compared. This creation of subsets for which model parameters are equivalent could be quite useful for researchers interested in working further with the scale(s) of interest.

5 Limitations

Despite these advantages that are presented by SEMtree, some weaknesses in the method should also be noted. First, SEMtree identifies patterns of factor model noninvariance, but does not yield results for specific comparisons on each of the model parameters. The alignment procedure was able to identify each difference on each parameter for each nation included in the sample. SEMtree, in contrast, identified groups of nations for which there were differences in at least one of the factor model parameters. However, unlike with the alignment method, it was not possible to identify specifically which of these parameters were different for which pair of nations. A second potential weakness of the SEMtree approach to doing invariance testing in the large scale assessment context is the amount of computing time that is necessary to actually conduct the analysis. For the nation only SEMtree model represented in Figure 1, the total computing time was 36 h and 23 min, using the full PISA dataset. For the more complex model including nation, sex, ESCS, and home language, the total computing time was 51 h and 19 min. Therefore, the researcher working with a large scale assessment including a big sample, such as with PISA, PIRLS or TIMMS, must allow for a sufficient amount of time to fit the model. It is also true that SEMtree converged much more quickly (less than 30 min) for smaller datasets.

6 Conclusion

SEMtree represents a powerful tool for researchers working with large scale assessment data and who need to investigate factor model invariance. It allows for invariance assessment of variables with many categories (e.g. nations), as well as multiple variables of various types (continuous, binary, multinomial). In addition, it provides subsets of the groups of interest so that when a large number of them are present, the researcher is able to identify those that are statistically equivalent, and which can thus be treated as the same. These qualities make SEMtree potentially quite useful for researchers working with large scale assessment data, where such variable types are common, and complex interactions among them might be anticipated.

References

Asparouhov, T. and Muthén, B. (2014) 'Multiple-group factor analysis alignment', *Structural Equation Modeling,* Vol. 21, p.1014.

Brandmaier, A.M., von Oertzen, T., McArdle, J.J. and Lindenberger, U. (2013) 'Structural equation model trees', *Psychological Methods*, Vol. 18, No. 1*,* pp.71–86.

French, B.F. and Finch, W. (2006) 'Confirmatory factor analytic procedures for the determination of measurement invariance', *Structural Equation Modeling,* Vol. 13, pp.378–402.

Horn, J.L. and McArdle, J.J. (1992) 'A practical and theoretical guide to measurement invariance in aging research', *Experimental Aging Research,* Vol. 18, Nos. 3–4, pp.117–144.

Jensen, D. and Cohen, P. (2000) 'Multiple comparisons in induction algorithms', *Machine Learning,* Vol. 38, pp.309–338.

Kim, H. and Loh, W. (2001) 'Classification trees with unbiased multiway splits', *Journal of the American Statistical Association,* Vol. 96, pp.589–604.

Loh, W. (2002) 'Regression trees with unbiased variable selection and interaction detection', *Statistica Sinica,* Vol. 12, pp.361–386.

Loh, W. and Shih, Y. (1997) 'Split selection methods for classification trees', *Statistica Sinica,* Vol. 7, pp.815–840.

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

Merkle, E.C. and Zeileis, A. (2013) 'Tests of measurement invariance without subgroups: a generalization of classical methods', *Psychometrika,* Vol. 78, No. 1, pp.59–82.

Millsap, R.E. (2011) *Statistical Approaches to Measurement Invariance,* A Taylor and Francis Group, New York: Routledge.

OECD. *PISA 2009 Technical Report*. OECD Publishing, Paris. 2012. Available at: http://dx.doi.org/10.1787/9789264167872-en.

Shih, Y. (2004) 'A note on split selection bias in classification trees', *Computational Statistics and Data Analysis,* Vol. 45, pp.457–466.

Steenkamp, J.E.M. and Baumgartner, H. (1998) 'Assessing measurement invariance in cross-national consumer research', *Journal of Consumer Research,* Vol. 25, pp.78–90.

Wicherts, J.M. and Dolan, C.V. (2010) 'Measurement invariance in confirmatory factor analysis: an illustration using IQ test performance of minorities', *Educational Measurement Issues and Practice*, Vol. 29, No. 3, pp.39–47.

Widaman, K.F. and Reise, S.P. (1997) 'Exploring the measurement invariance of psychological instruments: applications in the substance abuse domain', in Bryant, K.J.(Ed.): *Alcohol and Substance Use Research*, American Psychological Association, Washington, DC, pp.281–324.

Zeileis, A., Hothorn, T. and Hornik, K. (2008) 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics,* Vol. 17, No. 2, pp.492–514.