Large-scale Assessments in Education

RESEARCH Open Access

A comparison of three approaches to covariate effects on latent factors

Ze Wang[1,2*]

*Correspondence: zewang@gmail.com

[1] Department of Educational, School and Counseling Psychology, University of Missouri, MO 65211 Columbia, United States
[2] VA 20148 Ashburn, United States

Abstract

In educational and psychological research, it is common to use latent factors to represent constructs and then to examine covariate effects on these latent factors. Using empirical data, this study applied three approaches to covariate effects on latent factors: the multiple-indicator multiple-cause (MIMIC) approach, multiple group confirmatory factor analysis (MG-CFA) approach, and the structural equation model trees (SEM Trees) approach. The MIMIC approach directly models covariate effects on latent factors. The MG-CFA approach allows testing of measurement invariance before latent factor means could be compared. The more recently developed SEM Trees approach partitions the sample into homogenous subsets based on the covariate space; model parameters are estimated separately for each subgroup. We applied the three approaches using an empirical dataset extracted from the eighth-grade U.S. data from the Trends in International Mathematics and Science Study 2019 database. All approaches suggested differences among mathematics achievement categories for the latent factor of mathematics self-concept. In addition, language spoken at home did not seem to affect students' mathematics self-concept. Despite these general findings, the three approaches provided different pieces of information regarding covariate effects. For all models, we appropriately considered the complex data structure and sampling weights following recent recommendations for analyzing large-scale assessment data.

Keywords: Confirmatory factor analysis, MIMIC model, Multiple group analysis, SEM Trees, TIMSS

Background

In educational and psychological research, it is common to use latent factors to represent constructs. Latent factors are often established using the common factor model that includes both exploratory and confirmatory factor models. After factor models are run and tested against empirical data, there is usually a need for further analysis that involves effects of other covariates. For example, researchers may be interested in knowing whether the same factor structure would work for a normative sample vs. a referral sample (e.g., Parkin & Wang, 2021) or whether student sex and grade would be significant predictors of classroom engagement (e.g., Wang et al., 2014b). Within the framework of structural equation modeling (SEM), there are typically two methods for covariate

effects on latent factors. The first is the multiple-indicator multiple-cause (MIMIC) approach. With this approach, the covariates are included in the model as predictors of the latent factors; the direct effects of covariates on the latent factors are interpreted in the same way as regression coefficients. Statistical significance and effect sizes can also be obtained. The second approach, particularly when the covariates are categorical variables, invokes multiple group analysis where data are divided according to the values on the categorical variables and equality of model parameters (e.g., factor loading, indicator intercepts, latent factor means) across groups can be tested.

These two approaches have been widely used. With the MIMIC approach, it is easy to accommodate many covariates and both continuous and categorical covariates can be used. However, the MIMIC model assumes that latent factors are measured in the same way for different values of the covariates. Further, only linear (and variants of linear) relationships between the covariates and latent factors are allowed.

In contrast, for multiple group analysis, model parameters are allowed to vary and invariance between groups can be tested. It is also advised that measurement invariance testing precedes comparisons of the groups on the latent factors (Meredith, 1993; Millsap, 1997; Rensvold & Cheung, 1998). Compared to MIMIC, multiple group analysis is typically limited to a small number of groups, although Bayesian methods have been proposed for handling many groups (Muthen & Asparouhov, 2014).

Recently, structural equation model trees (SEM Trees; Brandmaier et al., 2013) have been proposed. SEM Trees are a generalization of decision trees that build a tree structure to separate data into subsets. The same SEM model is fit to data from each subset, but model parameters are separately estimated for each subset. The splitting of the data into subsets is based on covariates and done recursively with some criteria and stopping rules. SEM Trees have advantages in examining covariate effects because they can handle many different types of covariates and the relationships between the covariates and latent factors can be nonlinear. Further, it is not necessary to pre-specify the relationships, allowing data-driven explorations.

In this paper, we compare and contrast the three methods to examine the effects of covariates on the latent factor of mathematics self-concept using the U.S. eighth-grade data from the Trends in International Mathematics and Science Study (TIMSS) 2019 database (Fishbein et al., 2021).

Confirmatory factor analysis

Confirmatory Factor Analysis (CFA) is a popular measurement model used by researchers in educational, psychological, and social science fields. Under CFA, it is hypothesized that a latent factor is measured by multiple indicator variables. The latent factors would typically represent some type of unobserved constructs (e.g., motivation, engagement, attitudes) that are manifested by the observed indicator variables. One of the main advantages of CFA is that the latent factors are free from measurement errors. With CFA, all measurement error is assumed to be part of the observed indicator variables, and the latent factors represent the pure, shared variance among the indicators. Due to this, the effects of covariates on the latent factors are not affected by attenuated relationships.

CFA is a type of common factor model (Brown, 2006; Thurstone, 1947). The common factor postulates that each measured variable is a linear function of one or more common factors and a unique variable. Once the common factor(s) are removed, the observed variables are uncorrelated with each other. The unique variable is a combination of measurement error and specific error that is due to the selection of the measured variable. Suppose there are data of N participants on p observed variables and the score for the ith person on the jth variable is denoted as Y_{ij}. The linear factor model can be written as

$$Y_{ij} = v_j + \lambda_{j1}\eta_{1i} + \lambda_{j2}\eta_{2i} + ... + \lambda_{jk}\eta_{ki} + \varepsilon_{ij}$$

In matrix form, the response vector of participant i can be written as

$$\mathbf{y}_i = \mathbf{v} + \Lambda\eta_i + \varepsilon_i \quad (1)$$

where $\mathbf{y}_i$ is a $p \times 1$ vector of p observed variables, $\mathbf{v}$ is the $p \times 1$ vector of item intercepts, $\mathbf{\Lambda}$ is a $p \times$ m matrix of factor loadings, $\boldsymbol{\eta}_i \sim N(\kappa, \Phi)$ is an $m \times 1$ vector of common factors and $\mathbf{\Phi}$ is an $m \times m$ matrix of factor covariance matrix, $\boldsymbol{\varepsilon}_i \sim N(0, \Theta)$ is a $p \times 1$ vector of unique factors and $\mathbf{\Theta}$ is a $p \times p$ matrix of unique variances and covariances.

Further, it is assumed that

$E(\boldsymbol{\eta}_i) = 0$, $E(\boldsymbol{\varepsilon}_i) = 0$, and $Cov(\boldsymbol{\eta}_i, \boldsymbol{\varepsilon}_i) = 0$.

Under these assumptions, the population mean vector $\boldsymbol{\mu}$ and the population covariance matrix $\mathbf{\Sigma}$ of the p observed variables can be written, respectively, as

$$\mu = \mathbf{v} + \Lambda\kappa \quad (2)$$

$$\mathbf{\Sigma} = Cov\left(\mathbf{y}_i\mathbf{y}_i'\right) = \mathbf{\Lambda}E(\boldsymbol{\eta}_i\boldsymbol{\eta}_i')\mathbf{\Lambda}' + E(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i') = \Lambda\Phi\Lambda\prime + \Theta \quad (3)$$

where $\mathbf{\Sigma}$ is a $p \times p$ population covariance matrix of the observed variables, $\mathbf{\Phi}$ is an $m \times m$ matrix of factor covariance matrix, $\mathbf{\Theta}$ is a $p \times p$ matrix of unique variances and covariances.

Because latent factors are unobserved, it is necessary to set a location and metric for each latent factor. Two methods are commonly used: (a) putting the latent factors on a scale of zero mean and standard deviation of 1; and (b) choosing a marker indicator and set its loading to 1 and intercept to 0. Another method, called the effects coding method, imposes linear constraints on the unstandardized pattern coefficients to identify the model and can also be used (Little et al., 2006).

Covariate effects on latent factors

Whereas CFA, as a measurement technique, is often used for scale development and validation (typically together with exploratory factor analysis—another common factor model; e.g., Pratscher et al., 2019), it is also widely used in examining covariate effects. These covariates could represent demographical differences among individuals, or they could be attitudinal, psychological, situational, or trait variables. For example, the researcher may be interested in whether age is related to the latent factor means (e.g., Frisby & Wang, 2016). Covariates could be observed variables, or they themselves could

be unobserved and constructed from measurement models such as CFA. For this study, only observed covariates are considered.

MIMIC approach to covariate effects

For (observed) covariate effects on latent factors, based on Eq. (1), we further have

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \Gamma \mathbf{x}_i + \boldsymbol{\zeta}_i \tag{4}$$

where $\mathbf{x}_i$ is a $q \times 1$ vector of observed covariates, $\boldsymbol{\Gamma}$ is an $m \times q$ matrix of regression coefficients representing the covariate effects on latent factors, $\boldsymbol{\zeta}_i$ is an $m \times 1$ vector of disturbances, $\boldsymbol{\zeta}_i \sim N(0, \boldsymbol{\Psi})$, and $\boldsymbol{\alpha}$ is an $m \times 1$ vector of intercepts of the latent factors that are typically set to be zero.

The model parameter vector then is $\boldsymbol{\theta} = (\mathbf{v}, \boldsymbol{\Lambda}, \kappa, \Phi, \boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\alpha})$. For model identification, it is often the case that $\text{diag}(\boldsymbol{\Phi}) = \mathbf{I}$, $\boldsymbol{\kappa} = 0$, $\mathbf{v} = 0$, $\boldsymbol{\alpha} = 0$ (see Wu & Estabrook, 2016).

The MIMIC model is a single-group analysis and a special type of the full SEM model. In a MIMIC model, covariates directly affect the latent factor(s) and the path coefficients from the covariates to the latent factor(s) represent their effects. With a categorical covariate with more than two categories, some coding scheme (e.g., dummy coding) is used to create dummy variables. The effects of dummy variables on the latent factor(s) represent group differences, controlling for the other covariate(s).

The MIMIC approach is a direct extension of the linear regression model. The regular assumptions for regression models (independence of observations, linearity, and no correlations between covariates and the disturbance) also apply to the MIMIC model. A practical difference between the MIMIC model and the regression model is that the coefficients from dummy variables to the latent factor(s) in the MIMIC model should be standardized with respect to the latent factors because the scale of the latent factors is arbitrary, whereas in regression the unstandardized coefficients reflect group comparisons on the dependent variable.

The MIMIC approach is a standard method in SEM software packages such as Mplus (Muthén & Muthén, 1998–2017) and the "lavaan" R package (Rosseel, 2012).

Multiple group confirmatory factor analysis

When the covariates are categorical variables with a relatively small number of categories, their effects can be and often are examined using multiple group CFA (MG-CFA). The advantage of using MG-CFA is that cross-group equality of different types of parameters (e.g., factor means, factor variances, and covariates) can be tested. In addition to structural level parameters that involve latent factors and relationships between them, measurement level parameters—which represent relationships between latent factors and the observed indicators variables—are often investigated as well. There is a large body of literature on measurement invariance under the CFA framework, both methodologically (e.g., Liu et al., 2017; Meredith, 1993; Millsap, 2011), and empirical applications (e.g., Chan et al., 2019).

MG-CFA for covariate effects can be thought of as an extension of the analysis of variance (ANOVA) for group differences on observed means. Population parameters for

different groups are specified and tested, typically through null hypothesis significance testing (NHST). For ANOVA, the population parameters for NHST are the means, and the testing assumes that the groups have the same variance on the outcome variable in the population. When MG-CFA is used for covariate effects, the population parameters to be tested under NHST usually include mean differences on the latent factors (for identification purposes, the latent factor means for a reference group are usually constrained to zero), factor variances and covariances; however, other parameters can also be tested.

For MG-CFA, group sizes should be large enough to run CFA using data from individual groups. In addition, when there are many groups, even small differences between model parameters would be statistically significant, although Bayesian methods could be used for testing measurement invariance among many groups (Muthen & Asparouhov, 2014). When the covariate is continuous, some categorization is necessary before conducting MG-CFA.

When a covariate x represents group membership, instead of explicitly modeling the effect of x on latent factors as in Eq. (4), the covariate is used to subset data in MG-CFA. When there are multiple covariates, the researcher can either run multiple MG-CFA models, each time with a single covariate, or construct groups based on these covariates before conducting MG-CFA. The latter method may suffer from small sample sizes when the data are sliced in more ways. With G groups, Eqs. (5) and (6) show the population mean vector and the population covariance matrix of the p observed indicator variables, respectively, for a specific group g.

$$\boldsymbol{\mu}_g = \mathbf{v}_g + \Lambda_g \boldsymbol{\kappa}_g \tag{5}$$

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g{}' + \Theta_g \tag{6}$$

The parameter vector $\boldsymbol{\theta}$ is expanded to include parameters for multiple groups. For model identification, it is necessary to constrain parameters for each group (Millsap, 2011). When there are no equality constraints across groups, identification constraints for each group are similar to those for single group CFA (e.g., identifying the scale of latent factors). With equality constraints across groups (e.g., equal factor loadings, equal item intercepts), identification constraints are typically different for one group (e.g., the first group) compared to the other groups.

The biggest advantage of using MG-CFA is testing equality of different types of parameters across groups (i.e., invariance testing). In fact, invariance testing has been increasingly used in the development and validation and scales that involve CFA (e.g., Wang et al., 2014b). Like the MIMIC approach, MG-CFA is a standard method in SEM software packages such as Mplus (Muthén & Muthén, 1998–2017) and the "lavaan" R package (Rosseel, 2012).

Decision trees

Decision trees, also called trees, classification and regression trees (CART; Breiman et al., 2017; Loh, 2011), or recursive partitioning, are methods to split (i.e., partition) the space of covariates into subsets. Response values are similar within each subset but

different between subsets. The partitioning is repeated recursively until no splitting could be done based on some stopping criteria. When the outcome is a categorical variable, classification trees are built; when the outcome is numeric, regression trees are built.

Decision trees have been extended to incorporate parametric models (model-based recursive partitioning; MOB; Zeileis et al., 2008). With MOB, a stochastic model (e.g., a regression model) is assumed (called the template model); and the sample is split into groups with different values of model parameters. For example, if the template model is a regression model, the intercept and slopes may vary between subgroups according to some covariates. Therefore, in an example of regressing achievement on motivation, the intercept and slope may differ for students with different socioeconomic status (SES); therefore, a tree could use SES to divide participants based on differences in the regression model parameters.

MOB has been used to incorporate different stochastic models (e.g., Item Response Theory models) and decision trees (Brandmaier et al., 2013; Jeon & De Boeck, 2019; Merkle et al., 2014; Spratto et al., 2021; Wang et al., 2014a). SEM Trees (Brandmaier et al., 2013) combine recursive partitioning and SEM. SEM Trees use the likelihood ratio test or score-based tests to split observations based on covariates.

SEM Trees

For each covariate, data are split along all possible points of that covariate to create homogeneous groups according to some criteria (typically the likelihood but score-based tests are also available). These splits are binary splits, meaning that when a split happens, data are split into two groups (i.e., two nodes). For each candidate split, the log-likelihood values before and after the potential split are obtained. Because the model before the split is nested within the model after the split, a likelihood ratio test can be used to compare models. The partition of the covariate that leads to the greatest improvement in the model is retained. The process continues until a stopping criterion is reached. Stopping criteria could be a maximum tree depth, a minimum number of observations in a node, the p-value for the likelihood ratio test, etc.

There are a few different packages that can be used to implement SEM Trees. The "semtree" R package (Brandmaier et al., 2013) is a tree algorithm designed specifically for SEM. The package is based on the "OpenMx" package (Boker et al., 2011), which is a flexible R package that allows estimation of a wide variety of advanced multivariate statistical models including SEM. The "semtree" package can also be used together with the "lavaan" (Rosseel, 2012), a most popular R package for SEM. Another R package, "partykit" (Zeileis et al., 2008), is a general framework for MOB. To implement SEM Trees with the "partykit" package, some preliminary work is necessary to set up the SEM. It is possible to set up the SEM model with "lavaan".

Both "semtree" and "partykit" are solely based on the R language. Another package, "MplusTrees" is based on Mplus (Muthén & Muthén, 1998–2017) and the "MplusAutomation" R package (Hallquist & Wiley, 2018) that serves as an interface between Mplus and R (Serang et al., 2021). Mplus Trees taking advantage of the comprehensive Mplus software, allows users to specify complex SEM models using the regular Mplus syntax. The splitting procedure for the tree to grow is determined by the complexity parameter (*cp*) due to the package's reliance on the "rpart" package (Therneau & Atkinson, 1997).

cp reflects the relative improvement in the model fit for the split to be retained. If a candidate split improves the -2logL of the root node by a factor of at least *cp*, the split is made. The smaller the *cp*, the more complex the final tree is likely to be. Other stopping criteria such as the minimum number of observations within a node needed to attempt a split, the minimum observations within a terminal node, the maximum depth of the tree, the p-value for likelihood ratio tests can also be used/added.

Methods

Dataset description

To illustrate the three approaches, we used an empirical dataset from TIMSS 2019 (Martin et al., 2020). Specifically, we used the eighth grade U.S. data and only considered a subset of variables but data from all eighth graders who participated and were included in the public-use database were used. The sample size is 8,698. The variables for this study were seven indicators for the latent factor of mathematics self-concept, student's sex, home resources, language spoken at home, and mathematics achievement category. The seven indicator variables were: (a) I usually do well in mathematics; (b) Mathematics is not one of my strengths; (c) I learn things quickly in mathematics; (d) Mathematics makes me nervous; (e) I am good at working out difficult mathematics problems; (f) My teacher tells me I am good at mathematics; and (g) Mathematics makes me confused. They were rated on a 4-point Liker scale (1 = Agree a lot, 2 = Agree a little, 3 = Disagree a little, 4 = Disagree a lot). The four positively worded items (a, c, e, and f) were reverse coded so that a higher numeric rating would represent more mathematics self-concept. We coded the student sex variable as "0" for boys and "1" for girls. For the language spoken at home variable, always or almost always speaking English at home was coded as "1" and sometimes or never speaking English at home was coded as "0". Home resources was a categorical variable with three categories "Many resources", "Some resources", and "Few resources". It was dummy coded with "Some resources" as the reference group for MIMIC analysis. There were five mathematics achievement categories based on the first plausible value of mathematics achievement (Level 1 = Below 400, Level 2 = At or above 400 but below 475, Level 3 = At or above 475 but below 550, Level 4 = At or above 550 but below 625, Level 5 = At or above 625. These cutoffs and levels are from TIMSS. See Martin et al., 2020). The mathematics achievement category variable was dummy coded with Level 3 as the reference group for MIMIC analysis.

The study was reviewed by the Institutional Review Board at the author's university, which determined that this project does not constitute human subjects research according to the Department of Health and Human Services regulatory definitions. Informed consent is not applicable as this study analyzes publicly available data which do not include identifiable information.

Data analysis

For all models, we appropriately considered the complex data structure and sampling weights following recent recommendations (Stapleton, 2006a; Wang et al., 2019). Missing data for SEM models, including CFA, MIMIC, MG-CFA, and the template SEM model for SEM Trees were dealt with using the full information maximum likelihood

estimation for which both complete and partial data points were used. Specifically, observations with a missing value on *any* of the covariates involved in a particular analysis were deleted and observations with missing values on *all* indicator variables were removed. Missing data on the covariates for partitioning for the SEM Trees were dealt with using surrogate split. These are the default methods for these approaches.

All SEM models were estimated in Mplus using the "MplusAutomation" R package. For SEM Trees, we used the "MplusTrees" R package. The estimator for all SEM models was the robust maximum likelihood estimator (MLR) which adjusts standard errors of parameter estimates and rescales model chi-square values.

CFA without any covariates was applied to obtain an initial model that was later used as a template for MIMIC, MG-CFA, and SEM Trees. For the MIMIC model, all covariates, dummy coded if necessary, were tested simultaneously. For MG-CFA models, three measurement invariance tests (configural, metric, and scalar) were conducted for each of the covariates separately. We evaluated model fit in two ways. First, model fit for each model according to commonly used methods: comparative fit index (CFI) ≥ 0.95, Tucker–Lewis index (TLI) ≥ 0.95, root mean square error of approximation (RMSEA) < 0.06, and standardized root mean square residual (SRMR) < 0.08 (Hu & Bentler, 1999). Second, decreases in CFI and/or increases in RMSEA for testing factor loading and item intercept invariance. Chen (2007) and Cheung and Rensvold (2002) recommended that a decrease of at least 0.01 in CFI and an increase of at least 0.015 in RMSEA would suggest that the more constrained model fit the data significantly worse than the less constrained model. It should be noted that $\Delta\chi^2$ tests for nested models that are based on the Satorra-Bentler scaled χ^2 values could have been used as well. However, it is well known that the $\Delta\chi^2$ test, just like the χ^2 test, is sensitive to sample size. When the sample size is large, as is the case in this project, a small discrepancy would likely lead to a statistically significant $\Delta\chi^2$ test.

For SEM Trees in this study (i.e., Mplus Trees), we tested two *cp* values of 0.001 and 0.01, following recommendations by Serang et al. (2021). In addition, it is common to use other stopping criteria such as the minimum number of observations within a node needed to attempt a split, the minimum observations within a terminal node, the maximum depth of the tree, the *p* value for likelihood ratio tests, etc. For this study, we set the maximum depth of the tree to four, and the minimum number of observations in any terminal node to 100.

Results

Confirmatory factor analysis

The original single-factor CFA model with seven indicators for the construct of mathematics self-concept did not fit the data well: CFI = 0.879, TLI = 0.819, SRMR = 0.062, RMSEA = 0.090 with 90% CI [0.085, 0.095]. After examining the modification indices, covariances among the three negatively worded indicators were added, resulting in a well-fitting model: CFI = 0.989, TLI = 0.980, SRMR = 0.019, RMSEA = 0.030 with the 90% confidence interval [0.025, 0.036]. Table 1 has model fit information. Standardized factor loadings ranged from 0.398 to 0.802 with an average of 0.670.

Table 1 Model fit

Model	# Free parameters	Chi-square value	Chi-square scale factor	Chi-square DF	Chi-square p-value	CFI	TLI	RMSEA estimate	90% CI RMSEA		RMSEA p-value	SRMR
1. Single-factor CFA	21	931.68	2.85	14	< 0.001	0.879	0.819	0.090	0.085	0.095	< 0.001	0.062
2. Single-factor CFA with Correlated Residuals	24	91.97	2.50	11	< 0.001	0.989	0.980	0.030	0.025	0.036	1.000	0.019
3. MIMIC Model	32	553.43	1.73	59	< 0.001	0.963	0.951	0.032	0.030	0.035	1.000	0.028
4. Multiple Group Measurement Invariance—Student Sex												
4a. Configural Invariance Model	48	133.62	2.11	22	< 0.001	0.986	0.973	0.035	0.030	0.041	1.000	0.021
4b. Metric Model	42	135.36	2.37	28	< 0.001	0.987	0.980	0.031	0.026	0.036	1.000	0.031
4c. Scalar Model	36	228.70	2.22	34	< 0.001	0.976	0.970	0.037	0.033	0.042	1.000	0.043
5. Multiple Group Measurement Invariance—Language At Home												
5a. Configural Invariance Model	48	1.94	126.03	22	< 0.001	0.990	0.980	0.034	0.029	0.040	1.000	0.019
5b. Metric Model	42	1.83	138.77	28	< 0.001	0.989	0.983	0.031	0.026	0.037	1.000	0.023
5c. Scalar Model	36	1.74	163.93	34	< 0.001	0.987	0.984	0.031	0.026	0.036	1.000	0.024
6. Multiple Group Measurement Invariance—Home Resources												
6a. Configural Invariance Model	72	2.02	139.50	33	< 0.001	0.990	0.980	0.035	0.029	0.041	1.000	0.020
6b. Metric Model	60	1.85	190.56	45	< 0.001	0.986	0.980	0.035	0.030	0.040	1.000	0.042
6c. Scalar Model	48	1.75	240.19	57	< 0.001	0.982	0.980	0.035	0.030	0.039	1.000	0.042
7. Multiple Group Measurement Invariance—Achievement Category												
7a. Configural Invariance Model	120	1.66	213.79	55	< 0.001	0.984	0.969	0.042	0.036	0.048	0.986	0.021
7b. Metric Model	96	1.79	537.53	79	< 0.001	0.954	0.939	0.060	0.055	0.064	< 0.001	0.107
7c. Scalar Model	72	1.76	825.92	103	< 0.001	0.927	0.926	0.066	0.061	0.070	< 0.001	0.115

The estimator used was MLR. All models except Model 1 had correlated residuals among the three negatively worded indicators. Correlated residuals were not constrained to be equal across groups in multiple group analysis. In all models, the first factor loading was fixed to 1 to identify the latent factor in each group. For configural invariance and metric invariance models, the latent factor mean was fixed to 0 and the latent factor variance was free in all groups. For scalar invariance models, the latent factor mean for the reference group was fixed to 0 and was free for the other groups; the latent factor variance was free in all groups.
CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

MIMIC model

The MIMIC model was based on the one-factor CFA with correlated residuals among the three negatively worded items. The model fit the data well: CFI = 0.963, TLI = 0.951, SRMR = 0.028, RMSEA = 0.032 with 90% CI [0.030, 0.035]. Regarding covariate effects, girls had statistically significantly lower mathematics self-concept than boys (standardized coefficient = -0.107, $p < 0.001$). Language spoken at home and home resources did not have statistically significant effects on math self-concept. Students in lower mathematics achievement categories (Levels 1 and 2) had statistically significantly lower mathematics self-concept than those in the middle category (Level 3). Students in higher mathematics achievement categories (Levels 4 and 5) had statistically significantly higher mathematics self-concept than those in the middle category (Level 3).

Multiple group CFA

Table 1 has model fit information. Based on the model fit of individual models as well as model comparisons, scalar invariance existed between boys and girls, between those who mainly spoke English at home and those who mainly spoke another language at home, and between students with many, some, or few home resources. However, measurement invariance did not seem to exist among the achievement category groups. In other words, students' mathematics self-concept could be compared across sex, language, and home resources groups, but not among the achievement groups. Further examination suggested that girls had lower mathematics self-concept than boys. Students' mathematics self-concept did not differ statistically significantly between those who mainly spoke English at home and those who spoke another language at home. For students with many resources at home, their mathematics self-concept would be significantly higher than those with some or few resources at home.

SEM Trees

The template CFA model was the single-factor CFA with correlated residuals among the three negatively worded items (i.e., Model 2 in Table 1). The covariates were the same as in the MIMIC model and as the grouping variables for MG-CFA models. We grew two trees. The first had a *cp* of 0.001, and the second had a *cp* of 0.01. Figures 1 and 2 show the final trees for $cp = 0.001$ and $cp = 0.01$, respectively.

Figure 2 has the first three nodes and therefore a subtree of Fig. 1. In both trees, the initial split was based on mathematics achievement categories. Those in Levels 1, 2, and 3 were more homogenous than those in Levels 4 and 5. In Fig. 1, following the initial split, those in the lower mathematics achievement categories (Levels 1, 2, and 3) were further split by whether they were in the lowest mathematics achievement category (Level 1), and then whether they were boys or girls. For those in the higher mathematics achievement categories (Levels 4 and 5), they were also split by sex. Variables for the language spoken at home and home resources did not show up as split variables in the tree.

There were six terminal nodes in the first tree and two terminal nodes in the second tree. The estimates of parameters (factor loadings, item intercepts, variance of the latent factor, residual variances, and residual covariances for the three negatively worded items) for the terminal nodes are in Tables 2 and 3 for Tree 1 and Tree 2, respectively.

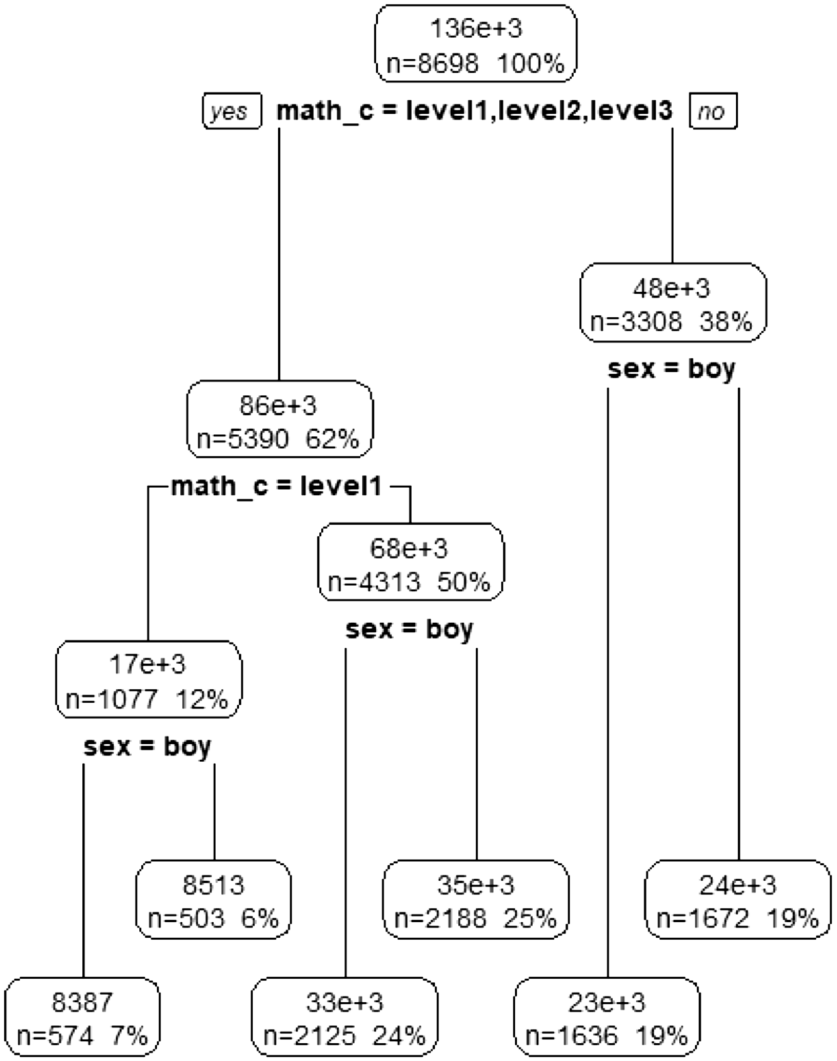

Fig. 1 Mplus Tree with Complexity Parameter (*cp*) = 0.001. The left branch at each partition point has observations that meet the condition (i.e. the result of testing the condition is "yes"). The right branch at each partition point has observations that do not meet the condition (i.e. the result of testing the condition is "no")

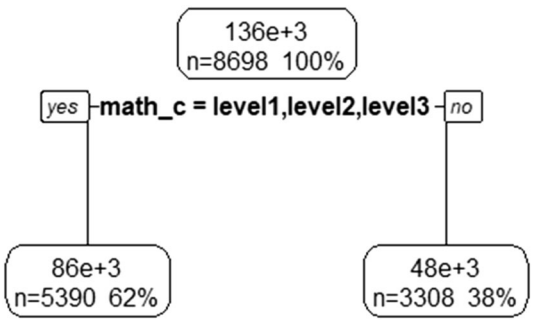

Fig. 2 Mplus Tree with Complexity Parameter (*cp*) = 0.01. The left branch at each partition point has observations that meet the condition (i.e. the result of testing the condition is "yes"). The right branch at each partition point has observations that do not meet the condition (i.e. the result of testing the condition is "no")

One advantage of SEM Trees is that we do not have to specify the interaction effects beforehand. Instead, the tree would identify the interaction effects. From Tree 1 (Fig. 1), the covariate effects can be thought of as interaction effects. There was an interaction effect between achievement category and student sex on the factor structure of the latent factor of mathematics self-concept. From Tree 2 (Fig. 2), there was only a main effect of mathematics achievement category. The terminal nodes of both trees include the splitting of observations. The parameter estimates for those nodes should be consistent with those obtained for the model if we subset the sample according to the splitting rules from the trees.

Discussion

The MIMIC model and MG-CFA have been widely used to examine covariate effects on latent factors. In this study, we included an additional method, SEM Trees, to examine covariate effects. SEM Trees allow examination of nonlinear interaction effects through recursive partitioning of the sample. SEM Trees separate the template SEM model from potential covariates. The sample is split in the covariate space based on data; and the SEM model is theory-driven. Therefore, SEM Trees combine theory-based and data-based approaches and allow theory-driven exploration (Brandmaier et al., 2016).

We used an empirical dataset to illustrate the three approaches to covariate effects on latent factors. The empirical data were about eighth-grade students' mathematics self-concept and two personal (sex and mathematics achievement) and two environmental (language spoken at home and home resources) covariates. There were some consistent findings across the three approaches. First, mathematics achievement was related to mathematics self-concept, both in terms of the factor structure (lack of measurement invariance from MG-CFA and data partitioning in SEM Trees) and in terms of the magnitude (statistically significant coefficient in the MIMIC model). This is not surprising because when students responded to items on mathematics self-concept, they would likely evaluate their mathematics ability. In this study, we used mathematics achievement as a covariate and mathematics self-concept as the outcome. This is consistent with the line of inquiry of the big-fish-little-pond effect (BFLPE; e.g., in Koivuhovi et al., 2020; Wang, 2020; Wang & Bergin, 2017) that looks into both personal and contextual effects in the formation of academic self-concept. In contrast, in the research area of achievement motivation, self-concept is typically modeled as a predictor of mathematics achievement and studies have found relationships between the two constructs as well (e.g., Wang et al., 2012; Wigfield & Eccles, 2000).

The other consistent finding across the three approaches was that the language spoken at home did not seem to have an effect on mathematics self-concept. This may be good news as the U.S. is a "melting pot" with many different languages and cultures. However, the nonsignificant finding may be due to the academic subject. It would be interesting to see if the finding could be generalized to other academic subjects such as English language and literacy.

Regarding differences between sex groups, the factor structure of mathematics self-concept was likely to be similar (scalar invariance from MG-CFA, Mplus Tree 2) but that factor structure may be unstable for different achievement groups. Mplus Tree 1 suggests an interaction between mathematics achievement and student sex. In addition,

Table 2 Parameter estimates for terminal nodes from Mplus tree with complexity parameter of 0.001

Parameter	Node 8	Node 9	Node 10	Node 11	Node 6	Node 7
SC.BY BSBM19A	1.000	1.000	1.000	1.000	1.000	1.000
SC.BY BSBM19C	0.367	0.252	0.915	0.956	1.354	1.681
SC.BY BSBM19D	1.085	0.985	1.101	1.119	1.322	1.363
SC.BY BSBM19E	− 0.360	0.095	0.485	0.622	0.655	1.012
SC.BY BSBM19F	0.663	1.023	1.164	1.073	1.328	1.310
SC.BY BSBM19G	0.904	0.906	0.864	0.917	0.926	0.944
SC.BY BSBM19I	0.097	0.322	0.810	0.812	1.118	1.272
BSBM19C.WITH BSBM19E	0.359	0.379	0.303	0.216	0.131	0.101
BSBM19C.WITH BSBM19I	0.498	0.451	0.343	0.261	0.105	0.090
BSBM19E.WITH BSBM19I	0.389	0.424	0.354	0.299	0.241	0.200
Intercepts BSBM19A	2.680	2.506	3.042	2.954	3.539	3.541
Intercepts BSBM19C	2.125	1.845	2.422	2.176	3.161	3.004
Intercepts BSBM19D	2.472	2.211	2.770	2.564	3.224	3.129
Intercepts BSBM19E	2.555	2.310	2.782	2.521	3.188	2.898
Intercepts BSBM19F	2.452	2.017	2.592	2.337	3.101	2.897
Intercepts BSBM19G	2.462	2.310	2.713	2.479	2.895	2.829
Intercepts BSBM19I	1.973	1.926	2.423	2.236	2.950	2.743
Variances SC	0.659	0.589	0.496	0.539	0.262	0.238
Residual.Variances BSBM19A	0.357	0.460	0.309	0.303	0.171	0.188
Residual.Variances BSBM19C	1.035	1.014	0.836	0.721	0.452	0.363
Residual.Variances BSBM19D	0.286	0.463	0.299	0.256	0.217	0.239
Residual.Variances BSBM19E	1.026	1.114	0.951	0.923	0.652	0.716
Residual.Variances BSBM19F	0.803	0.447	0.293	0.346	0.229	0.319
Residual.Variances BSBM19G	0.614	0.623	0.667	0.676	0.694	0.735
Residual.Variances BSBM19I	0.966	0.968	0.809	0.737	0.553	0.512

The stopping criteria for tree growth were a maximum tree depth of four, a minimum number of observations in any terminal node of 100, and complexity parameter (*cp*) = 0.001. The first factor loading was fixed to 1 for model identification. Node 8 was the group of boys in mathematics achievement category 1. Node 9 was the group of girls in mathematics achievement category 1. Node 10 was the group of boys in mathematics achievement categories 2 and 3. Node 11 was the group of girls in mathematics achievement categories 2 and 3. Node 6 was the group of boys in mathematics achievement categories 4 and 5. Node 7 was the group of girls in mathematics achievement categories 4 and 5

from both MIMIC and MG-CFA, girls tended to have lower mathematics self-concept than boys. Other studies have also found gender differences favoring boys (e.g., Koivuhovi et al., 2020) but there is also research that showed no gender differences (e.g., Ghasemi & Burley, 2019).

Having many home resources seemed to be beneficial from MG-CFA results, although home resources was not found to be a significant predictor of mathematics self-concept using the MIMIC or SEM Trees approaches. This could be due to the positive relationship between home resources and academic achievement. In MG-CFA, when home resources was examined, it was considered separately from the other covariates, whereas in MIMIC and SEM Trees, it was considered simultaneously with the other covariates.

One particular limitation of Mplus Trees is parameter constraints across nodes. After the sample is split, the SEM model is essentially estimated with each subset of the data. Although it is possible to fix and/or constrain parameters within the SEM model for each node, our understanding is that it is not possible to constrain and test parameter relationships between nodes. This could be an important limitation for analysis such

Table 3 Parameter estimates for terminal nodes from Mplus tree with complexity parameter of 0.01

Parameter	Node 2	Node 3
SC.BY BSBM19A	1.000	1.000
SC.BY BSBM19C	0.823	1.529
SC.BY BSBM19D	1.105	1.361
SC.BY BSBM19E	0.442	0.852
SC.BY BSBM19F	1.060	1.346
SC.BY BSBM19G	0.895	0.943
SC.BY BSBM19I	0.702	1.216
BSBM19C.WITH BSBM19E	0.315	0.126
BSBM19C.WITH BSBM19I	0.365	0.101
BSBM19E.WITH BSBM19I	0.379	0.230
Intercepts BSBM19A	2.920	3.540
Intercepts BSBM19C	2.237	3.083
Intercepts BSBM19D	2.604	3.177
Intercepts BSBM19E	2.608	3.044
Intercepts BSBM19F	2.419	3.000
Intercepts BSBM19G	2.553	2.862
Intercepts BSBM19I	2.255	2.847
Variances SC	0.555	0.245
Residual.Variances BSBM19A	0.335	0.184
Residual.Variances BSBM19C	0.858	0.416
Residual.Variances BSBM19D	0.290	0.226
Residual.Variances BSBM19E	1.017	0.704
Residual.Variances BSBM19F	0.390	0.275
Residual.Variances BSBM19G	0.667	0.716
Residual.Variances BSBM19I	0.846	0.536

The stopping criteria for tree growth were a maximum tree depth of four, a minimum number of observations in any terminal node of 100, and complexity parameter (cp) = 0.01. The first factor loading was fixed to 1 for model identification. Node 2 was the group of students in mathematics achievement categories 1, 2, and 3. Node 3 was the group of students in mathematics achievement categories 4 and 5

as measurement invariance testing, for which equality constraints on parameters are routinely tested. The "semtree" package allows invariance testing through either global invariance (fixing selected parameters to the sample estimation *before* the tree is grown) and local invariance (chosen parameters cannot differ while growing the tree). Such invariance options are supported with OpenMx, but not with lavaan to specify the SEM model.

As one reviewer pointed out, the mathematics achievement category variable is categorized based on the first plausible value. Technically speaking, that plausible value, along with the other four plausible values for the same measure, is not an observed variable; and results might change if a different plausible value were to be used. Plausible values in large-scale assessments are generated based on item response modeling, latent regression, and multiple imputation (von Davier, 2020). Multiple plausible values are typically used in analysis to allow for the calculation of total variances of estimates which consist of within- and between-imputation variances. However, for the purpose of this article, we focus on comparing the three approaches instead of using plausible values. Readers interested in best practices of using plausible values from large-scale

assessments can refer to many resources on this topic. Among them are Rutkowski et al. (2014), von Davier et al. (2009), Wang (2020) and Wu (2005).

In this article, we only used observed covariates. Although conceptually the methods can be extended to include latent covariates, we think there are still some technical challenges. The MIMIC approach is relatively easy to include latent covariates. The MG-CFA approach inherently relies on observed grouping variables as covariates although it can be extended to latent classes as covariates under the mixture modeling framework. It is challenging, in our opinion, to using latent, instead of observed, covariates with the SEM Trees approach due to the recursive partitioning of the sample. However, we are optimistic and look forward to new advancements in SEM (as an example, Merkle & Zeileis, 2013).

The data used were nationally representative and had a complex structure. Such large-scale assessment data are available for researchers to conduct substantive and methodological research (Rutkowski et al., 2014; Wang, 2017). There have been methodological advancements on how to analyze such data (e.g., Asparouhov, 2006; Asparouhov & Muthen, 2006; Hahs-Vaughn et al., 2011; Muthen & Satorra, 1995; Stapleton, 2006b; Trendtel & Robitzsch, 2020; Wu & Kwok, 2012), as well as software development (Bailey et al., 2020; Caro & Biecek, 2017; Oberski, 2014). In this study, we considered both the complex data structure and sampling weights, taking advantage of Mplus. The "lavaan" and "OpenMx" packages can both handle some aspects of complex data structures but additional research is needed for more user-friendly tools.

Conclusions

Using empirical data from TIMSS 2019, this study applied MIMIC, MG-CFA, and SEM Trees approaches to covariate effects on latent factors. The MIMIC and MG-CFA have been widely used in educational and psychological research. The SEM Trees approach is a more recent development. Applied researchers can take advantage of the new method with the availability of several packages. This study is one application that demonstrates the use of SEM Trees and hopefully will generate more interest in using it with large-scale assessment data.

Abbreviations

ANOVA	Analysis of variance
CFA	Confirmatory factor analysis
CFI	Comparative fit index
cp	Complexity parameter
NHST	Null hypothesis significance testing
MG-CFA	Multiple group confirmatory factor analysis
MIMIC	Multiple-indicator multiple-cause
MLR	Robust maximum likelihood estimator
MOB	Model-based recursive partitioning
RMSEA	Root mean square error of approximation
SEM	Structural equation model(ing)
SES	Socioeconomic status
SRMR	Standardized root mean square residual
TIMSS	Trends in International Mathematics and Science Study
TLI	Tucker–Lewis index

Acknowledgements
Not applicable.

Author contributions
ZW conceived the idea, conducted the analysis, and wrote the manuscript. The author read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Data used are publicly available at https://timss2019.org/international-database/.

Declarations

Competing interests
The author has no conflicts of interest/competing interests to disclose.

Received: 26 August 2021 Accepted: 9 December 2022
Published online: 21 December 2022

References

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods, 35*(3), 439–460.
Asparouhov, T., & Muthen, B. (2006). Multilevel modeling of complex survey data. the American Statistical Association, Seattle, WA: American Statistical Association.
Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., Zhang, T., Bundsgaard, J., & C'deBaca, R. (2020). *EdSurvey: Analysis of NCES education survey and assessment data. R package version 2.5.0.* In https://CRAN.R-project.org/package=EdSurvey
Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., & Bates, T. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317.
Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*(4), 566–582. https://doi.org/10.1037/met0000090
Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86. https://doi.org/10.1037/a0030001
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.
Brown, R. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.
Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software, 81*(7), 1–44. https://doi.org/10.18637/jss.v081.i07
Chan, M.H.-M., Gerhardt, M., & Feng, X. (2019). Measurement invariance across age groups and over 20 years' time of the Negative and Positive Affect Scale (NAPAS). *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000529
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
Frisby, C. L., & Wang, Z. (2016). The g factor and cognitive test session behavior: Using a latent variable approach in examining measurement invariance across age groups on the WJ III. *Journal of Psychoeducational Assessment, 34*(6), 524–535. https://doi.org/10.1177/0734282915621440
Ghasemi, E., & Burley, H. (2019). Gender, affect, and math: A cross-national meta-analysis of Trends in International Mathematics and Science Study 2015 outcomes. *Large-Scale Assessments in Education, 7*(1), 10. https://doi.org/10.1186/s40536-019-0078-1
Hahs-Vaughn, D. L., McWayne, C. M., Bulotsky-Shearer, R. J., Wen, X., & Faria, A.-M. (2011). Complex sample data recommendations and troubleshooting. *Evaluation Review, 35*(3), 304–313. https://doi.org/10.1177/0193841x11412070
Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2017.1402334
Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology, 72*(3), 517–537. https://doi.org/10.1111/bmsp.12182
Koivuhovi, S., Marsh, H. W., Dicke, T., Sahdra, B., Guo, J., Parker, P. D., & Vainikainen, M.-P. (2020). Academic self-concept formation and peer-group contagion: Development of the big-fish-little-pond effect in primary-school classrooms and peer groups. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000554
Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3), 486–506. https://doi.org/10.1037/met0000075

Loh, W.-Y. (2011). Classification and regression trees. *Wires Data Mining and Knowledge Discovery, 1*(1), 14–23. https://doi.org/10.1002/widm.8

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

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

Merkle, E., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79*(4), 569–584. https://doi.org/10.1007/s11336-013-9376-7

Merkle, E., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika, 78*(1), 59–82. https://doi.org/10.1007/s11336-012-9302-4

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248–260. https://doi.org/10.1037/1082-989x.2.3.248

Millsap, R. E. (2011). *Statistical approach to measurement invariance*. Routledge.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Muthen, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978. https://doi.org/10.3389/fpsyg.2014.00978

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). American Sociological Assocaition.

Oberski, D. (2014). lavaan.survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software, 57*(1), 27. https://doi.org/10.18637/jss.v057.i01

Parkin, J. R., & Wang, Z. (2021). Confirmatory factor analysis of the WIAT-III in a referral sample. *Psychology in the Schools, 58*(5), 837–852. https://doi.org/10.1002/pits.22474

Pratscher, S. D., Wood, P. K., King, L. A., & Bettencourt, B. A. (2019). Interpersonal mindfulness: Scale development and initial construct validation. *Mindfulness, 10*(6), 1044–1061. https://doi.org/10.1007/s12671-018-1057-2

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement model for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58*(6), 1017–1034. https://doi.org/10.1177/0013164498058006010

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *2012*, *48*(2), 36. https://doi.org/10.18637/jss.v048.i02

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Chapman Hall/CRC Press.

Serang, S., Jacobucci, R., Stegmann, G., Brandmaier, A. M., Culianos, D., & Grimm, K. J. (2021). Mplus trees: Structural equation model trees using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 127–137. https://doi.org/10.1080/10705511.2020.1726179

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement, 81*(1), 39–60. https://doi.org/10.1177/0013164420918655

Stapleton, L. M. (2006a). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13*(1), 28–58. https://doi.org/10.1207/s15328007sem1301_2

Stapleton, L. M. (2006b). *Using multilevel structural equation modeling techniques with complex sample data*. Information Age Publishing.

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines*.

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.

Trendtel, M., & Robitzsch, A. (2020). A Bayesian item response model for examining item position effects in complex survey data. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998620931016

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item response theory, population models, and linking across modes. In *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 11.11–11.25). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series, 2*, 9–36.

Wang, Z. (2017). Editorial: Large-Scale Educational Assessments. *International Journal of Quantitative Research in Education*, *4*(1/2), 1–2. https://www.inderscience.com/info/inarticletoc.php?jcode=ijqre&year=2017&vol=4&issue=1/2

Wang, Z. (2020). When large-scale assessments meet data science: The big-fish-little-pond effect in fourth- and eighth-grade mathematics across nations. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.579545

Wang, T., Merkle, E. C., & Zeileis, A. (2014a). Score-based tests of measurement invariance: use in practice [Original Research]. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.00438

Wang, W., Liao, M., & Stapleton, L. M. (2019). Incidental second-level dependence in educational survey data with a nested data structure. *Educational Psychology Review, 31*, 571.

Wang, Z., Bergin, C., & Bergin, D. A. (2014b). Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly, 29*(4), 517–535. https://doi.org/10.1037/spq0000050

Wang, Z., & Bergin, D. A. (2017). Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: Analysis of TIMSS 2011 data. *Learning and Individual Differences, 57*, 141–156. https://doi.org/10.1016/j.lindif.2017.04.003

Wang, Z., Osterlind, S. J., & Bergin, D. A. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *International Journal of Science and Mathematics Education, 10*(5), 1215–1242. https://doi.org/10.1007/s10763-011-9328-6

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Wu, J.-Y., & Kwok, O.-M. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(1), 16–35. https://doi.org/10.1080/10705511.2012.634703

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of EDucational and Behavioral Statistics*. https://doi.org/10.3102/10769986017002155

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of ComputAtional and Graphical Statistics, 17*(2), 492–514.

Publisher's Note