# Chapter 8 Analysing PIAAC Data with Structural Equation Modelling in Mplus



**Ronny Scherer** 

**Abstract** Structural equation modelling (SEM) has become one of the most prominent approaches to testing substantive theories about the relations among observed and/or unobserved variables. Applying this multivariate procedure, researchers are faced with several methodological decisions, including the treatment of indicator variables (e.g. categorical vs. continuous treatment), the handling of missing data, and the selection of an appropriate level of analysis. The PIAAC data pose additional issues, such as the clustering of individual-level data, the large number of participating countries, the representation of performance scores by a set of plausible values, and the differences in the selection probabilities. Therefore, a flexible software package is required to handle them. This chapter introduces readers to analysing PIAAC data with SEM in the software Mplus by (a) presenting the key concepts behind SEM, (b) discussing the complexities of the PIAAC data and their possible handling, (c) illustrating the specification and evaluation of measurement and structural models, and (d) pointing to current developments in the areas of measurement invariance testing and multilevel SEM. Sample input and output files are provided.

Structural equation modelling (SEM) represents a broad range of multivariate approaches that allow researchers to test hypotheses related to the means, variances, and covariances of manifest and latent variables (Kaplan 2009). It includes approaches such as path analysis, confirmatory factor analysis, and structural models that are based on researchers' hypotheses and theories about the relations among variables. In his seminal book, Kline (2016) emphasised that SEM requires

**Electronic Supplementary Material** The online version of this chapter (https://doi.org/10.1007/978-3-030-47515-4\_8) contains supplementary material.

R. Scherer (⊠)

University of Oslo, Oslo, Norway e-mail: ronny.scherer@cemo.uio.no

three inputs: first, a set of hypotheses about the relations among variables (based on theory or informed by the results of empirical studies); second, a set of specific questions about these relations (e.g. To what extent does an indirect effect of a variable X on a variable Y via a variable M exist?); and third, appropriate datasets to test these hypotheses and answer these questions. Ultimately, the process of SEM generates three outputs (Kline 2016): numeric estimates of model parameters, a set of logical implications of the model, and information about the extent to which the data support the model. Given the richness of outputs, SEM has become a prominent tool for researchers to test substantive theories and assumptions about the relations among variables. Moreover, SEM is considered a flexible modelling approach that allows for the inclusion of both manifest (observable) and latent (unobservable) variables in the measurement and structural models (Raykov and Marcoulides 2006). Due to this flexibility, researchers are faced with several methodological decisions, including the treatment of indicator variables (e.g. categorical vs. continuous treatment), the handling of missing data, and the selection of an appropriate level of analysis (e.g. individual vs. country level). Besides these decisions, international large-scale assessment data add further complexities, such as the weighting of samples and the use of plausible values as performance scores (Rutkowski and Zhou 2014).

In the light of these considerations, this chapter seeks to (1) draw attention to the data issues associated with the SEM of PIAAC data; (2) illustrate ways to address these issues in the software package Mplus (Version 8.2); and (3) exemplify the application of typical classes of models within SEM using PIAAC data. All examples are supplemented by explanations of the Mplus syntax and the interpretation of the outputs. Although this chapter provides a brief introduction to the classes of SEM approaches, it does not deliver a complete introduction to SEM. Readers are encouraged to refer to the seminal SEM literature to learn more about model specification, identification, estimation, and interpretation (e.g. Hancock and Mueller 2013; Hoyle 2012; Kaplan 2009; Kline 2016; Loehlin and Beaujean 2017; Raykov and Marcoulides 2006). Moreover, this chapter does not deliver an introduction to the software Mplus. Readers are encouraged to review the material provided by Muthén et al. (2017) and Muthén and Muthén (1998–2017).

The first section of this chapter highlights the complexities associated with the PIAAC data and reviews the options Mplus offers to handle them. The second section briefly reviews the application of SEM using PIAAC data and draws attention to the extent to which these data complexities have been addressed in the extant literature. The third section focuses on specifying and estimating measurement models by means of confirmatory factor analysis (CFA). This section also showcases approaches to the testing of measurement invariance across few or many groups. The fourth and final section introduces classes of structural models, including path models, structural equation models, and multi-group versions thereof. However, the examples and structural equation modelling approaches will focus mainly on measurement models. A short summary concludes this chapter. All syntax files are contained in the Supplementary Material.

#### 8.1 Issues with the SEM of PIAAC Data

As noted in the previous chapters in this volume, the complexity of the PIAAC data is driven by several elements. These elements include, but are not limited to, the use of survey weights; the nested data structure with study participants nested in, for instance, countries; the use of a set of plausible values to represent participants' performance on the literacy, numeracy, and problem-solving assessments; and the occurrence of missing data in the background questionnaire data. The PIAAC Technical Report notes that 'inferences will not be valid unless the corresponding variance estimators appropriately reflect all of the complex features of the PIAAC sample design' (OECD 2013, p. 26). These issues are by no means unique to the PIAAC data—several international large-scale assessments, including the Programme for International Student Assessment (PISA), the Teaching and Learning International Survey (TALIS), the Trends in International Mathematics and Science Study (TIMSS), the International Computer and Information Literacy Study (ICILS), and the Progress in International Reading Literacy Study (PIRLS), follow similar study designs (Rutkowski et al. 2010; Rutkowski and Zhou 2014). In the following, I will briefly review these issues and describe ways to deal with them in Mplus. Table 8.1 provides an overview of the relevant Mplus commands.

• Weighting. The PIAAC data accommodate two types of weights, a final participants' weight (SPFWT0) and a set of replicate weights (SPFWT1-SPFWT80). The former were created by a base weight that included the selection probabilities of households and several adjustment factors (OECD 2013). The latter represent a set of weights that can be used for improving the variance estimation through jackknifing or other approaches. Several authors have suggested examining how informative sampling weights are before including them in the analysis of international large-scale assessment data—for instance, by evaluating the effective sample sizes and design effects (Laukaityte and Wiberg 2018; Rutkowski and Svetina 2014). The inclusion of the final weight and the replicate weights in Mplus is straightforward: In the VARIABLE section, researchers can

Data issue	Mplus sample options
Weighting	WEIGHT = SPFWT0;
	REPWEIGHTS = SPFWT1-SPFWT80;
	<pre>REPSE = JACKKNIFE;</pre>
	<pre>REPSE = JACKKNIFE2;</pre>
	<pre>REPSE = BOOTSTRAP;</pre>
Nested data structure	CLUSTER = CNTRYID;
	TYPE = COMPLEX;
	TYPE = TWOLEVEL;
	TYPE = TWOLEVEL RANDOM;
Plausible values	TYPE = IMPUTATION;
Missing data	MISSING ARE ALL;

Table 8.1 Overview of Mplus options to address PIAAC data issues

specify the final weight using the WEIGHT option and the REPWEIGHTS option for the replicate weights. Replicate weights can be accompanied by several additional specifications, such as the type of standard error adjustment (REPSE). Furthermore, weights can be scaled using the WTSCALE option.

- Nested data structure. To account for the clustering of the individual data in, for instance, regions or countries, researchers have at least two options: First, they may account for the nested data structure by adjusting the standard errors of the SEM parameters using the TYPE = COMPLEX option. This option does not call for a multilevel model that models the level of nesting explicitly. It is accompanied by the robust maximum likelihood estimator (MLR) and the specification of the clustering variable (e.g. CLUSTER = CNTRYID). Second, researchers may want to model the nested data structure through multilevel modelling in order to quantify and explain between-country variation in PIAAC variables, or relations among them. The corresponding commands for the two-level models with random intercepts and/or slopes are TYPE = TWOLEVEL and/or TYPE = TWOLEVEL RANDOM.
- *Plausible values.* PIAAC uses plausible values to represent literacy, numeracy, and problem solving in technology-rich environments. In the extant literature, several procedures have been applied to include these sets of performance scores. Among these procedures, the following deals best with the variation within and between the sets of plausible values (Laukaityte and Wiberg 2017; Rutkowski et al. 2010): The SEM analyses are conducted for each of the ten datasets containing the ten plausible values. The resultant model parameters are subsequently pooled as the means across all ten sets of model parameters, and their variances are quantified according to Rubin's combination rules. These rules incorporate the variances within and between plausible values and the number of plausible values (e.g. Laukaityte and Wiberg 2017). Mplus offers a convenience option (TYPE = IMPUTATION) that performs SEM for each set of plausible values and combines the resultant model parameters. Although combining means, variances, covariances, and path coefficients may be straightforward with this procedure (Enders 2010), the combined fit statistics require further adjustments (Enders and Mansolf 2018; Meng and Rubin 1992). As far as the Mplus documentation goes, the adjustments of the chi-square statistic and the model deviance are performed by default in the software with (robust) maximumlikelihood estimation (Asparouhov and Muthén 2010).
- *Missing data*. Missing data may occur in the background variables for several reasons. Without reviewing the details behind the mechanisms of missingness, I note that *Mplus* has several options to deal with missing data. They include multiple imputation and model-based approaches with or without auxiliary variables (Enders 2010). Researchers also have the opportunity to perform multiple imputation in alternative software packages (e.g. the R package 'mice') and submit the resulting complete datasets to *Mplus* for SEM (e.g. Enders et al. 2016; Grund et al. 2018).

# 8.2 A Brief Review of PIAAC Secondary Data Analyses Using SEM

To review the current status of how SEM is utilised for the analysis of PIAAC data, I performed a search in the databases PsycINFO and ERIC using the search terms PIAAC AND (structural equation model\* OR path model\* OR factor analysis OR CFI OR RMSEA OR indirect effect OR mediation) and retrieved seven publications (as of 25 February 2019). An additional search for the term 'PIAAC' in the Elsevier Scopus database yielded 17 further publications and cross-references to two further publications. Of these 26 publications, 12 presented the application of SEM to PIAAC data. Table 8.2 shows a description of these publications, including the models the authors specified and the extent to which the complexity of the data was addressed.

Most secondary analyses were based on multiple PIAAC samples (75%), comprising 18–29 participating countries. The types of structural equation models covered single-level path models (33.3%); single-level structural equation models (50.0%), including exploratory (8.3%) and confirmatory factor analyses (25.0%); and multilevel SEM (16.7%). The software package Mplus dominated the list of analytic tools (80%), next to LISREL (10%) and the R package lavaan (10%). Only 1 of the 12 publications did not make explicit whether and how the complexity of the PIAAC data was considered during SEM. With some exceptions (25%), the authors included sampling weights in their analyses. In the analyses involving plausible values, Rubin's combination rules were mainly applied; however, one study averaged the ten plausible values provided by the PIAAC database, and one study used these values as manifest indicators of a latent variable to represent participants' skills. Finally, the procedures for handling missing data varied considerably and included multiple imputation, full information maximum-likelihood estimation, and listwise deletion procedures.

Overall, this brief, and by no means complete, review suggested that SEM is making its way into the secondary analysis of PIAAC data. At the same time, the procedures for handling the data complexities varied between studies and call for a framework that may guide researchers in their SEM analysis.

# **8.3** PIAAC Data and Measures Used in the Illustrative Examples

The following illustrative examples of classes of structural equation models are based on two datasets: the first contains the Norwegian (N=5128) and German (N=5465) PIAAC data; the second contains the data from 27 countries participating in PIAAC (N=181,236), excluding the data from Austria, Cyprus, Russia, and Turkey (the reasoning for this exclusion can be found in Borgonovi and Pokropek 2017b). These data have been made publicly available by the OECD

data
PIAAC
alyse P
M to an
using SEN
papers usi
nple pa
8.2 San
Table 8

			,		
References	Description	Data	Model	Complexity	Software
Borgonovi and Pokropek (2017a)	Relations among education, generalised trust, mediated by literacy skills, income, and occupational prestige; country differences due to birthplace diversity, and income inequality	PIAAC background questionnaire and skills data from 29 participating countries	Multilevel SEM (within-level, individuals; between-level, countries)	C, M, P, W	Mplus
Borgonovi and Pokropek (2017b)	Country differences and disparities in external political efficacy, mediated by parental socio-economic status and cognitive abilities	PIAAC background questionnaire and skills data from 28 participating countries	Multilevel SEM (within-level, individuals; between-level, countries)	C, M, P, W	Mplus
Cincinnato et al. (2016)	Relations among parents' education, educational attainment, readiness to learn, and participation in adult education	PIAAC background questionnaire data from 23 participating countries	Path model	C, M, W	Mplus
Duchhardt et al. (2017)	Relations among the use of mathematics and numeracy skills, controlled for background characteristics, mathematical requirements on the job, and education	PIAAC background questionnaire and skills data from Germany	Path model	P, W	Mplus
Ganzach and Patel (2018)	Role of general mental ability (g) and specific abilities in predicting wages, controlled for gender and age	PIAAC background questionnaire and skills data from 19 participating countries	Path model	M	NA
Gorges et al. (2016)	Invariance testing of the motivation-to-learn scale across countries and relations to learning engagement, controlled for literacy	PIAAC background questionnaire data from 18 participating countries	Multi-group CFA	C, M, W	Mplus

imputation, listwise deletion, maximum-likelihood-based procedures), P plausible values handling through Rubin's combination rules (other procedures excluded), W weighting (e.g. replicate weights, case weights) (2016) and contain the full study samples in these countries covering a broad age range (16–65 years). For details of the descriptive sample statistics, readers are referred to the public use files provided by the OECD. The illustrative examples in this chapter explore the relations among several cognitive skills measures in PIAAC and measures from the PIAAC background questionnaire (Table 8.3).

**Table 8.3** Overview of the variables included in the illustrative examples

Variable	Description
CNTRYID	Country ID (ISO 3166, numeric code)
GERMAN	Dummy-coded variable ( $1 = Germany, 0 = Norway$ )
FEMALE	Participant's gender ( $I = Female, 0 = Male$ )
HOMLANG	Test language same as language spoken most often at home (derived; $l = Test$ language same as home language, $0 = Test$ language not the same as home language)
Curiosity	PIAAC scale 'Learning strategies' ( $0 = Not \ at \ all, \ 1 = Very \ little, \ 2 = To \ some \ extent, \ 3 = To \ a \ high \ extent, \ 4 = To \ a \ very \ high \ extent)$
I_Q04b	Relate new ideas into real life
I_Q04d	Like learning new things
I_Q04h	Attribute something new
I_Q04j	Get to the bottom of difficult things
I_Q041	Figure out how different ideas fit together
I_Q04m	Looking for additional info
Skills use at work	PIAAC scale 'Skill use work – ICT – Internet' ( $0 = Never$ , $1 = Once\ a$ month, $2 = Less\ than\ once\ a\ week\ but\ at\ least\ once\ a\ month, 3 = At\ least once a week but not every day, 4 = Every\ day)$
G_Q05a	How often—for mail
G_Q05c	How often—work-related info
G_Q05e	How often—spreadsheets
G_Q05f	How often—word
Skills use everyday	PIAAC scale 'Skill use everyday life – ICT – Internet' ( $0 = Never$ , $1 = Once$ a month, $2 = Less$ than once a week but at least once a month, $3 = At$ least once a week but not every day, $4 = Every$ day)
H_Q05a	How often—for mail
H_Q05c	How often—in order to better understand various issues
H_Q05e	How often—spreadsheets
H_Q05f	How often—word
PVLIT1-	Literacy (LIT) scale scores—plausible values 1–10
PVLIT10	
PVPSL1-	Problem solving in technology-rich environments (PS-TRE) scale
PVPSL10	scores—plausible values 1–10
SPFWT0	Final full sample weight
SPFWT1- SPFWT80	Final replicate weights (1–80)

#### **8.4** Measurement Models

# 8.4.1 Confirmatory Factor Analysis with Categorical or Continuous Indicators

To establish a measurement model of a construct, researchers may choose among several procedures. These include, but are not limited to, confirmatory factor analysis (CFA), exploratory factor analysis (EFA), and exploratory structural equation modelling (ESEM)—the latter bringing together the features of CFA and EFA (Brown 2015; Marsh et al. 2014). In this section, I will focus on CFA as a means to develop a suitable measurement model that represents the latent (unobserved) variable of interest.

A CFA model comprises one or more latent variables that are measured by a set of categorical or continuous indicators, such as item responses, subscale scores, or item parcels. For the *i*th indicator and the *j*th person in the dataset,  $y_{ij}$ , a one-factor model with only one latent variable  $\eta$  is specified as  $y_{ij} = v_i + \lambda_i \eta_j + \varepsilon_{ij}$ , where  $\lambda_i$  denotes the factor loading of this indicator,  $v_i$  the item intercept, and  $\varepsilon_{ij}$  the residual. Using matrix notation, the resultant model can be described as  $\mathbf{y} = \mathbf{v} + \lambda \eta + \varepsilon$  with  $\mathbf{y} \sim N(\mathbf{0}, \mathbf{\Sigma})$ ,  $\eta \sim N(\mathbf{0}, \mathbf{\Psi})$ , and  $\varepsilon \sim N(\mathbf{0}, \mathbf{\Theta})$ . The underlying covariance structure of this model is  $\mathbf{\Sigma} = \Lambda \Psi \Lambda' + \mathbf{\Theta}$ , where  $\Lambda$  represents the matrix of factor loadings,  $\Psi$  the matrix of factor variances and covariances, and  $\mathbf{\Theta}$  the matrix of residual variances and covariances (for more details, please refer to Brown 2015).

When performing CFA, researchers are faced with several decisions, such as the treatment of the indicators as categorical or continuous variables. Despite the option of robust maximum likelihood estimation (Mplus: ESTIMATOR = MLR) to account for possible deviations from a normal distribution of the continuous indicators, the choice of an appropriate estimator is especially relevant for indicators that are semi-continuous. More specifically, when researchers intend to use item responses that are based on frequency, agreement, or rating scales with a categorical set of response options (e.g. ranging from 0 = not at all to 5 = always), they have to decide whether to treat these responses categorically or continuously. Without reviewing the extant literature on this topic to the full extent, the existing body of research suggests that five or more response options and tolerable deviations of the item response distributions from normality may justify the treatment of item responses as continuous variables (e.g. Finney and DiStefano 2013). In such a situation, the CFA model becomes more parsimonious because only one item intercept is estimated in the model instead of several thresholds between the response categories (Kline 2016). In Mplus, maximum-likelihood-based estimators (e.g. Robust ML [MLR]) and the weighted least square mean and variance adjusted (WLSMV) estimator are available to treat item responses categorically (Brown 2015). However, these two estimators may not perform equally well in CFA with ordinal data. Li (2016), for instance, found that the WLSMV estimator was less biased in estimating factor loadings yet overestimated the correlations between factors (see also Beauducel and Herzberg 2006). The following example illustrates the specification and performance of these two estimation approaches.

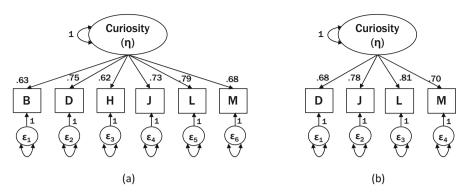


Fig. 8.1 Measurement models of curiosity with categorically treated item indicators (WLSMV estimator) based on (a) six and (b) four items

The PIAAC participants were asked to indicate on a five-point rating scale the degree to which they agreed with six statements concerning their curiosity (see Table 8.3). A CFA model assuming a single latent variable measured by the six item responses and treating these responses continuously resulted in a marginal fit to the data,  $\chi^2(9) = 516.4$ , p < 0.001, RMSEA = 0.074, CFI = 0.926, SRMR = 0.043. Gorges et al. (2017) reviewed the theoretical anchoring of the six items in existing curiosity frameworks and argued that four items (I\_Q04D, J, L, M) represented the construct. Moreover, evaluating the factor loadings revealed that items I\_Q04B and I\_Q04H showed lower values in comparison to the remaining items—this observation indicates that they represent what is shared among all items to a smaller extent. As a consequence, the initial measurement model was refined by deleting these two items,  $\chi^2(2) = 90.9$ , p < 0.001, RMSEA = 0.066, CFI = 0.979, SRMR = 0.022 (Syntax 8.1).

The same model exhibited an acceptable fit to the data when treating the four-item responses categorically through WLSMV estimation,  $\chi^2(2) = 151.7$ , p < 0.001, RMSEA = 0.085, CFI = 0.988, SRMR = 0.0170. To specify this model in Mplus, the item responses have to be defined as categorical using the CATEGORICAL ARE I\_Q04d-I\_Q04m command. Moreover, the WLSMV is called by ESTIMATOR = WLSMV, and the theta parameterisation is selected by the PARAMETERIZATION = THETA command (for more details about this parameterisation, please refer to Kline 2016). The factor loadings of the models based on six and four items are shown in Fig. 8.1.

```
Syntax 8.1: CFA Model Describing the Factor Structure of Curiosity VARIABLE: [...]
USEVARIABLES ARE I_Q04d I_Q04j I_Q04l I_Q04m;
```

```
! Missing data coding
MISSING ARE ALL(-99);
! Final participant weight
WEIGHT = SPFWT0;

ANALYSIS:

ESTIMATOR = MLR;
H1ITERATIONS = 10000;
PROCESSORS = 4;

MODEL:
! Measurement model
CURIOUS BY I_Q04d I_Q04j I_Q04l I_Q04m;

OUTPUT:

SAMPSTAT; ! Sample statistics
STDYX; ! Fully standardized parameters
MOD(ALL); ! Modification indices
```

The secondexample illustrating the specification and estimation of CFA models in Mplus concerns PIAAC participants' exposure to certain skills (see Table 8.2). Differentiating between skills needed at work and in everyday life, researchers may specify a factor model with two correlated factors (Fig. 8.2). Given that items are formulated similarly for work and everyday life situations, a covariance structure is

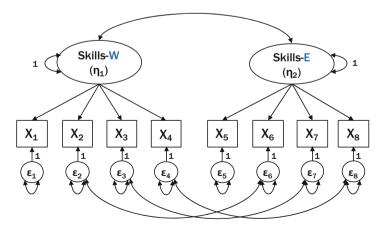


Fig. 8.2 Measurement model of participants' skills use at work (Skills-W) and in everyday life (Skills-E)

added to the residuals to account for any relations among items after controlling for the two latent variables. These residual covariances are specified in the two-factor model as shown in Syntax 8.2.

```
Syntax 8.2: Model Syntax of the Skills Use Measurement Model

MODEL:

! Measurement model

! Two-factor model with correlated residuals

WORK BY G_Q05a G_Q05c G_Q05e G_Q05f;

LIFE BY H_Q05a H_Q05c H_Q05e H_Q05f;

! Modifications
! Residual covariances

G_Q05C WITH H_Q05C;

G_Q05E WITH H_Q05F;

G_Q05F WITH H_Q05F;
```

Treating item responses continuously, this model exhibited a substantial fit to the data,  $\chi^2(16) = 429.6$ , p < 0.001, RMSEA = 0.052, CFI = 0.936, SRMR = 0.038. Treating item responses categorically, and using the WLSMV estimator, the model showed an acceptable fit,  $\chi^2(16) = 422.9$ , p < 0.001, RMSEA = 0.052, CFI = 0.975, SRMR = 0.034. In the first model, the correlation between the two factors was positive and significant ( $\rho = 0.501$ ); in the second model, this correlation was only marginally smaller ( $\rho = 0.479$ ). Given the existence of residual covariances in the model, the latent variables  $\eta_1$  and  $\eta_2$  can no longer be considered unidimensional constructs—alternative model specifications with nested factors can facilitate a clearer interpretation of these constructs (Koch et al. 2018).

Overall, the specification of measurement models in Mplus allows researchers to treat indicators categorically or continuously. Deviations from the multivariate normality assumption can be compensated (at least partly) by the robust ML estimation. Deviations from a simple structure—that is, a factor structure without any residual covariances and possible cross-loadings—can also be implemented in the software package.

# 8.4.2 Measurement Invariance Testing with Few Groups

In many scenarios, group comparisons are of major interest to researchers. Such comparisons may refer to the differences in means of variables or differences in the relations among constructs. In both cases, researchers have to establish that the variables used in group comparisons are comparable to a sufficient degree. More specifically, mean differences or differences in structural relations across

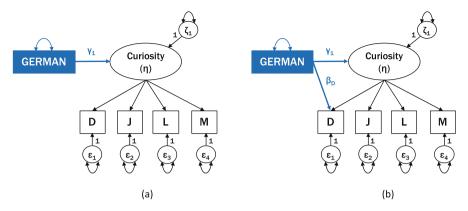


Fig. 8.3 (a) MIMIC and (b) MIMIC-DIF model of curiosity with GERMAN as covariate

groups can have several causes, including actual mean differences and differences in structural relations, but also possible differences in the functioning of items, scales, or entire tests (Meredith 1993). To examine the extent to which possible differential item functioning may affect the group differences found by researchers, several procedures have been developed under the umbrellas of 'measurement invariance' and 'differential item functioning' (Millsap 2011). In the following, I will illustrate how to implement these procedures in Mplus, focusing on scenarios with few groups. The full input and output files can be accessed through the Supplementary Material.

#### **MIMIC Models and Differential Item Functioning**

Multiple causes multiple indicator (MIMIC) models are usually comprised of a latent variable (measured by multiple indicators) and one or more covariates (multiple causes; Brown 2015). These models represent probably the simplest of structural equation models and allow researchers to examine the effects of covariates on the latent trait—for example, to identify possible group differences. The latter, however, is based on the assumption that the measurement model holds for the different groups—in other words, the model is based on measurement invariance assumptions that can facilitate the meaningful interpretation of factor means (Kim et al. 2012b).

In the following data example, differences in participants' curiosity between the Norwegian and the German PIAAC samples are examined. The binary variable GERMAN (I = Germany, O = Norway) serves as the covariate of the latent variable 'Curiosity' (Fig. 8.3 (a)). Using the regression command in Mplus, curiosity is simply regressed on the covariate (Syntax 8.3). Given that GERMAN is binary, the regression coefficient  $\gamma_1$  is partially standardised (STDY standardisation; Muthén et al. 2017).

Syntax 8.3: Model Syntax of the MIMIC Model of Curiosity with GER-MAN as Covariate

```
MODEL:

! Measurement model

CURIOUS BY I_Q04d I_Q04j I_Q04l I_Q04m;

! Structural part

CURIOUS ON GERMAN;

OUTPUT:

TECH1;
TECH8;
STDY; ! Due to the binary predictor GERMAN
```

The MIMIC model exhibited an acceptable fit to the data,  $\chi^2(5) = 386.6$ , p < 0.001, RMSEA = 0.086, CFI = 0.942, SRMR = 0.032. The relation between the binary country variable GERMAN and curiosity was statistically significant ( $\gamma_1 = -0.315$ , SE = 0.024, 95% CI [-0.361, -0.268], p < 0.001) and suggested that the Norwegian sample showed higher curiosity than the German sample. However, this interpretation does not consider the possibility of differential item functioning across the two countries.

To identify whether specific manifest indicator variables (i.e. items) exhibit differential item functioning between the German and the Norwegian samples that is, a situation in which the probability of responding to an item differs between groups although they have the same level on the latent variable (Millsap 2011)—the MIMIC model can be extended to a MIMIC-DIF model by adding structural paths to the specific variables (see Fig. 8.3 (b)). This model provides information not only about possible factor mean differences but also about differences in the item intercepts (Kim et al. 2012b). Hence, the MIMIC-DIF model allows researchers to test for the DIF of specific items and extract factor mean differences controlling for item DIF. In Mplus, the item intercepts are regressed on the covariate, in this case by adding the command line I Q04d ON GERMAN to obtain the parameter  $\beta_D$ . (Note: for categorical indicators, the item thresholds are regressed on the covariate.) In this example, the DIF parameter of item D was  $\beta_D = -0.140$ , SE = 0.018, 95% CI [-0.176, -0.105], p < 0.001. The corresponding differences in the factor mean of curiosity were  $\gamma_1 = -0.276$ , SE = 0.025, 95% CI [-0.325, -0.227], p < 0.001. The MIMIC-DIF model outperformed the MIMIC model in terms of model fit,  $\Delta \chi^2(1) = 54.4$ , p < 0.001. Hence, there is evidence for cross-country differences in curiosity favouring the Norwegian sample and the differential functioning of item

The MIMIC-DIF approach has been further developed to test not only so-called uniform DIF effects but also non-uniform DIF effects by including an interaction

term between the latent variable and the covariate (Woods and Grimm 2011). Bauer (2017) proposed the more general framework of moderated factor analysis to examine uniform and non-uniform DIF effects of categorical, continuous, or both types of covariates.

#### Multi-group Confirmatory Factor Analysis

Besides the testing of differential item functioning with the help of MIMIC- and MIMIC-DIF models, there is multi-group CFA, a procedure that allows researchers to specify and estimate a set of models for which the parameters in the CFA model can be constrained to equality across groups. Multi-group CFA has become the standard approach to measurement invariance testing in education and psychology (Putnick and Bornstein 2016; Scherer and Greiff 2018) and forms the basis for several extensions, such as multi-group ESEM, the alignment optimisation method, and Bayesian measurement invariance testing (Marsh et al. 2013, 2018; Muthén and Asparouhov 2012).

Typically, three multi-group CFA models are specified to test for measurement invariance based on continuously treated item indicators of a latent variable  $\eta$  (Van de Schoot et al. 2012): (1) The configural invariance model assumes the same factor structure (i.e. number of factors and the pattern of the links between the latent variable and the manifest indicators) across groups. This model is often used as the baseline model against which all other models with additional parameter constraints are compared. All model parameters are freely estimated across groups. Specifically, for the ith item indicator and the ith person in the kth group, a configural one-factor model is specified for the manifest indicator variable  $y_{ijk}$ , the latent variable  $\eta_{ik}$ , the group-specific intercept  $\nu_{ii}$ , and the residual term  $\varepsilon_{iik}$ as  $y_{ijk} = v_{ik} + \lambda_{ik}\eta_{jk} + \varepsilon_{ijk}$ , where  $\lambda_{ik}$  denotes the factor loading of the *i*th item for the kth group. The factor means are fixed to zero, and the factor variance are fixed to 1 for all groups. (2) The *metric invariance model* constrains the factor loadings  $\lambda_{ik}$ to equality across groups based on the configural model,  $y_{ijk} = v_{ik} + \lambda_i \eta_{ik} + \varepsilon_{ijk}$ . Again, the factor means are constrained to zero, yet the factor variance is freely estimated to identify the model. If metric invariance holds, factor variances and covariances can be compared across groups. (3) The scalar invariance model further constrains the item intercepts  $v_i$  to equality across groups,  $y_{iik} = v_i + \lambda_i \eta_{ik} + \varepsilon_{iik}$ . To identify the mean structure in the model, factor means are freely estimated; factor variances are also freely estimated. If scalar invariance holds, factor means can be compared across groups. In all models, residuals are assumed to be uncorrelated to the latent variable and to have a mean zero (Muthén and Asparouhov 2018). Marsh et al. (2009) proposed extending this measurement invariance framework by systematically testing additional parameter constraints—these constraints involve the factor means, variances, covariances, and item residuals. Apart from these extensions, the measurement literature often includes the invariance of item residual (co-)variances  $\theta$  in addition to the scalar invariance constraints to test whether the measurement models indicate the same reliabilities (Raykov and Marcoulides 2006). The resulting model is referred to as the strict invariance model. If strict invariance holds, (manifest) scale means can be compared across groups.

To determine which level of measurement invariance holds for a given dataset, several indicators are available, including the results of chi-square difference testing and the differences in fit indices between models with different parameter constraints (Brown 2015). For instance, if comparing the configural and metric invariance models results in an insignificant chi-square difference test, this can be interpreted as evidence that the constraints on the factor loadings do not deteriorate the overall model fit—hence, metric invariance can be retained. However, in large samples, and for complex datasets, the chi-square difference test may result in a significant test statistic although the constraints on model parameters do not substantially deteriorate the model fit (Yuan and Chan 2016). As a consequence, differences in fit indices provide additional sources of information. For these differences, several cut-off criteria were suggested: (a)  $\Delta$ CFI less than -0.010 (Cheung and Rensvold 2002); (b)  $\Delta$ CFI less than -0.010,  $\Delta$ RMSEA less than 0.015, and  $\Delta$ SRMR less than 0.030 (Chen 2007); and (c)  $\Delta$ CFI less than -0.008 (Meade et al. 2008). However, these criteria should not be considered to be 'golden rules', as they depend on several factors, such as the type of the factor model (Khojasteh and Lo 2015), the types of invariance models that are compared (Rutkowski and Svetina 2014), or whether the invariance of mean or covariance structures is examined (Fan and Sivo 2009). Moreover, the application of these guidelines varies in that some researchers compare all models against the configural model, whereas others compare adjacent models to identify the effects of additional parameter constraints given the constraints in the previous model. Note that similar criteria apply to the measurement invariance testing in situations where item indicators are treated categorically. Please review Liu et al. (2017) for more details.

Example 1: Gender as the Grouping Variable The following example uses the variable FEMALE as the grouping variable and focuses on the three standard measurement invariance models (i.e. configural, metric, and scalar invariance) using the Mplus convenience option MODEL = CONFIGURAL METRIC SCALAR. This option specifies all three models with either continuously treated (MLR estimator) or categorically treated item responses (WLSMV estimator) and compares them with the help of chi-square difference testing. Syntax 8.4 shows the corresponding commands. These models can also be specified by imposing the parameter constraints directly. The corresponding syntax files are part of the Supplementary Material (from file 'MM6c-Curiosity-MG-Gender.inp').

```
ANALYSIS:
    ESTIMATOR = MLR;
    H1ITERATIONS = 10000;
    PROCESSORS = 4;
    MODEL = CONFIGURAL METRIC SCALAR;
    ! Invariance models

MODEL:
    ! Measurement model
    CURIOUS BY I_Q04d I_Q04j I_Q04l I_Q04m;

OUTPUT:
    TECH1;
    TECH8;
    SAMPSTAT;
```

To further test whether strict invariance holds, this syntax can be modified so that the equality of residual variances across groups is imposed (Syntax 8.5).

```
Syntax 8.5: Multi-group CFA Model Assuming Strict Invariance Across
Gender
VARIABLE: [...]
    ! Grouping specification
   GROUPING IS FEMALE (0 = Men 1 = Women);
ANALYSIS:
   ESTIMATOR = MLR;
   H1ITERATIONS = 10000;
   PROCESSORS = 4;
MODEL:
   ! Measurement model
   ! Factor loadings constrained to equality across
    ! Factor loadings labelled as L2-L4
   CURIOUS BY
         I Q04d
         I Q04j(L2)
         I Q041(L3)
         I Q04m(L4);
    ! Item intercepts constrained to equality across
  groups
```

```
! labelled as I1-I4
   [I O04d-I O04m] (I1-I4);
   ! To identify the mean structure, the factor
  mean is now
   ! freely estimated.
   [CURIOUS*];
   ! Item residual variances constrained to
  equality
   ! across groups
   ! labelled as R1-R4
   I Q04d-I Q04m(R1-R4);
MODEL WOMEN:
   ! Nothing needs to be specified here
OUTPUT:
   TECH1;
   TECH8;
   SAMPSTAT;
   STDYX;
```

The resulting model fit indices and their comparisons between models are shown in Tables 8.4 and 8.5. All models exhibited an acceptable fit to the data. Considering the results of the chi-square difference testing, metric invariance could be assumed. Considering the differences in the CFI, RMSEA, and the SRMR between the configural model and all other models, strict invariance could be assumed. Considering the changes in the CFI, RMSEA, and the SRMR after imposing more constraints on the model parameters (i.e. between adjacent models), strict invariance could be assumed. Overall, the invariance testing suggested that strict measurement invariance holds across gender. In this situation, researchers can interpret possible gender differences in the means of curiosity as actual mean differences.

Example 2: Country as the Grouping Variable Following the same procedure, the testing of measurement invariance across the two PIAAC participating countries Germany and Norway resulted in a good model fit for the configural and metric models, but not for the scalar and strict invariance models (Table 8.4). Moreover, the model comparisons suggest that metric invariance can be retained (Table 8.5). In this case, mean differences in curiosity between the two countries are camouflaged by the differential functioning of the scale or, more precisely, the non-invariance of the measurement model.

### 8.4.3 Measurement Invariance Testing with Many Groups

With the increasing number of countries, language groups, and educational systems participating in international large-scale assessments comes the challenge of establishing that the measures used for comparisons are sufficiently invariant (Rutkowski et al. 2018). However, the commonly used approach of multi-group CFA to establish measurement invariance across many groups may increase the chances of falsely detecting non-invariance due to the large number of pairwise comparisons of model parameters (Rutkowski and Svetina 2013). Addressing this issue, several alternative approaches to invariance testing with many groups have been developed. These include, but are not limited to, (a) the alignment method (Asparouhov and Muthén 2014), (b) the alignment-within-CFA method (Marsh et al. 2018), (c) multilevel CFA (Kim et al. 2012a), (d) multilevel factor mixture modelling (Kim et al. 2016b), and (e) Bayesian approximate invariance testing (Van de Schoot et al. 2013). These approaches have strengths and weaknesses, a thorough review of which is beyond the scope of this chapter. Readers are referred to the extant literature comparing the performance of measurement invariance testing procedures (e.g. Desa 2014; Kim et al. 2017; Muthén and Asparouhov 2018).

In the following, I will illustrate the application of the alignment optimisation method, the alignment-within-CFA method (AwC), and multilevel CFA to the PIAAC data, focusing on the invariance of the curiosity scale across 27 of the participating countries. (Note: Due to quality issues, the data from Austria, Cyprus, Russia, and Turkey were excluded; Borgonovi and Pokropek 2017b). The Mplus syntax files can be found in the Supplementary Material.

**Alignment Optimisation Method** The alignment optimisation method represents an approach to multi-group CFA or item response theory that estimates the factor means and variances for each group, based on the assumption of the configural measurement invariance model (Asparouhov and Muthén 2014). This method is aimed at minimising the departures from the invariance of the model parameters.

				=		
Model	$\chi^2(df)$	CFI	RMSEA	SRMR	AIC	BIC
Grouping is I	FEMALE					·
Configural	95.2 (4)*	0.978	0.066	0.023	96,883	97,057
Metric	103.6 (7)*	0.977	0.052	0.027	96,891	97,043
Scalar	115.5 (10)*	0.975	0.045	0.030	96,904	97,034
Strict	118.1 (14)*	0.975	0.038	0.029	96,914	97,016
Grouping is (	GERMAN					'
Configural	117.1 (4)*	0.983	0.074	0.019	93,906	94,080
Metric	133.8 (7)*	0.981	0.059	0.026	93,918	94,070
Scalar	954.6 (10)*	0.861	0.135	0.086	94,983	95,113
Strict	978.6 (14)*	0.858	0.116	0.102	95,059	95,160

Table 8.4 Fit indices of the multi-group CFA models for curiosity

*Note.* The scaling correction factors and information criteria can be found in the Supplementary Material

p < 0.001

Table 8.5	Comparisons	of the multi-group	CFA models fo	r curiosity

Model	$\Delta \chi^2(df)$	$\Delta CFI$	$\Delta RMSEA$	$\Delta SRMR$				
Grouping is FEMALE								
Configural vs. metric	6.3(3), p = 0.10	-0.001	-0.006	+0.004				
Configural vs. scalar	14.8(6), p = 0.02	-0.003	-0.021	+0.007				
Configural vs. strict	21.3 (10), p = 0.02	-0.003	-0.028	+0.006				
Metric vs. scalar	8.6(3), p = 0.04	-0.002	-0.007	+0.003				
Metric vs. strict	15.0(7), p = 0.04	-0.002	-0.017	+0.002				
Scalar vs. strict	2.7(4), p = 0.14	0.000	-0.007	-0.001				
Grouping is GERMAN								
Configural vs. metric	13.4 (3), <i>p</i> < 0.01	-0.002	-0.015	+0.007				
Configural vs. scalar	876.5 (6)*	-0.122	+0.061	+0.067				
Configural vs. strict	871.1 (10)*	-0.125	+0.042	+0.083				
Metric vs. scalar	916.0 (3)*	-0.120	+0.076	+0.060				
Metric vs. strict	849.8 (7)*	-0.125	+0.057	+0.076				
Scalar vs. strict	55.7 (4)*	-0.003	-0.019	+0.016				

p < 0.001

It begins with specifying a null model—that is, the configural model with freely estimated factor loadings and item intercepts, the factor means constrained to zero, and the factor variances constrained to 1. Without deteriorating the fit of this model, the alignment method performs a transformation of the factor means so that the non-invariance of the factor loadings and item intercepts is minimised with the help of a simplicity function (Muthén and Asparouhov 2014). Asparouhov and Muthén (2014) explain the details of this transformation and how it reduces non-invariance. Overall, the extant literature on the performance of the alignment optimisation suggests that this method is suitable for estimating group-specific factor means and variances without relying on the often unrealistic assumptions of scalar invariance across countries. It also estimates the model parameters efficiently and is less computationally demanding than alternative methods, such as multilevel CFA. Among alternative methods, it can detect non-invariance reasonably well (Kim et al. 2017).

For the example of measuring curiosity across the 27 PIAAC countries, the Mplus syntax to specify the alignment method without setting a reference country (i.e. free alignment) is shown in Syntax 8.6. The grouping must be specified in the context of a mixture model using the KNOWNCLASS option. Once this has been set, the alignment method is called (ALIGNMENT = FREE) and the corresponding output requested (ALIGN). In this example, the free alignment method estimates the factor mean of the first group freely; Mplus also offers a fixed alignment option (ALIGNMENT = FIXED()) that constrains the factor mean of the first group or that of another group to zero.

# Syntax 8.6: Alignment Optimisation Method Applied to the Curiosity Measurement Model Across Countries

```
VARIABLE: [...]
   ! Grouping specification
   ! Define the 27 countries by the ISO code
   CLASSES = c(27);
   KNOWNCLASS = c(CNTRYID=56 124 152 203 208 233
  246 250 276 300 372 376
   380 392 410 440 528 554 578 616 702 703 705 724
  752 826 840);
  ANALYSIS:
   TYPE = MIXTURE;
   ESTIMATOR = MLR;
   H1ITERATIONS = 10000;
   PROCESSORS = 4;
   ALIGNMENT = FREE;
   ! Call the free alignment method
   ! Alternative specification
   ! ALIGNMENT = FIXED(56);
   ! Call the fixed alignment method with Belgium
  (ISO code 56)
   ! as the reference group
   ! similar to the scalar invariance model
MODEL:
   %OVERALL%
   ! Measurement model
   CURIOUS BY I Q04d I Q04j I Q04l I Q04m;
OUTPUT:
   TECH1;
   TECH8;
   ALIGN;
   SVALUES;
```

The output file contains information about the degree of non-invariance (Output 8.1) and the estimated factor means (Output 8.2) for the free alignment method. Output 8.1 indicates the (non-)invariance of the factor loadings and item intercepts among the 27 countries. (Note: non-invariance is indicated by the country's ISO code shown in brackets.) Output 8.2 shows the ranking of the countries based on their factor means.

Output 8.1: (Non-)Invariance of Factor Loadings and Item Intercepts

APPROXIN	MATE MEASUREMENT INVARIANCE (NONINVARIANCE) FOR GROUPS
Item interce	epts
I_Q04B	(56) 124 (152) 203 (208) (233) (246) 250 276 (300) (372) 376 (380) (392) (410) (440) (528) (554) (578) 616 (702) (703) (705) 724 (752) (826) 840
I_Q04D	(56) 124 (152) (203) (208) (233) 246 250 276 (300) (372) (376) (380) (392) (410) (440) 528 (554) (578) (616) (702) (703) (705) (724) (752) (826) 840
I_Q04H	(56) (124) 152 (203) (208) (233) (246) (250) 276 300 372 (376) (380) 392 (410) (440) (528) 554 (578) 616 702 (703) (705) 724 (752) 826 840
I_Q04J	(56) 124 152 (203) 208 (233) (246) (250) (276) (300) (372) (376) (380) (392) (410) (440) (528) 554 (578) 616 (702) (703) 705 (724) (752) (826) 840
I_Q04L	(56) 124 152 203 208 233 (246) 250 276 (300) (372) 376 380 (392) (410) (440) (528) 554 (578) (616) 702 (703) 705 (724) (752) (826) 840
I_Q04M	56 124 (152) (203) (208) 233 (246) 250 (276) (300) 372 376 (380) (392) 410 (440) 528 (554) (578) (616) (702) (703) (705) (724) (752) 826 840
Factor load	lings
I_Q04B	56 124 152 203 208 233 246 250 (276) 300 372 376 380 392 410 (440) 528 554 578 (616) (702) (703) 705 724 752 826 840
I_Q04D	56 124 152 203 208 (233) 246 250 (276) 300 372 376 380 (392) 410 (440) 528 554 578 616 (702) 703 (705) 724 (752) 826 840
I_Q04H	56 124 152 203 208 233 (246) (250) (276) 300 372 376 380 392 410 (440) 528 554 578 616 702 703 705 724 (752) 826 840
I_Q04J	(56) 124 152 203 208 (233) 246 (250) 276 300 372 (376) (380) 392 410 440 528 554 578 616 (702) 703 705 724 (752) 826 840
I_Q04L	(56) 124 (152) 203 208 233 246 250 276 (300) 372 376 (380) (392) (410) (440) 528 554 578 (616) (702) (703) (705) (724) (752) 826 840
I_Q04M	56 124 152 (203) 208 (233) 246 (250) 276 300 372 376 380 392 (410) 440 528 554 578 (616) 702 (703) 705 724 752 826 840

Alignment-Within-CFA Method Similar to the conceptualisation of exploratory SEM, Marsh et al. (2018) extended the alignment method in order to make accessible analyses that could not be conducted with the original alignment optimisation approach. These analyses include, for instance, testing the invariance of residual or factor variances and covariances, the estimation of covariate effects in MIMIC models, the direct testing of factor mean differences, and the relations to other variables and constructs. Essentially, the extended alignment-within-CFA (AwC) method comprises two analytic steps: In the first step, the alignment optimisation method is performed, and the resulting parameters of the measurement models across countries are saved. These parameters form the starting values for a standard multi-group CFA model in the second step; in this model, some parameters are fixed to identify the model and mimic the exploratory alignment estimates. The starting values from the fixed alignment method with Belgium (i.e. the first group, ISO code 56) are requested using the SVALUES option in the output section of the Mplus

Output 8.2: Ranking of Countries Based on the Curiosity Factor Means

# FACTOR MEAN COMPARISON AT THE 5% SIGNIFICANCE LEVEL IN DESCENDING ORDER

Results fo	r factor C	URIOUS		
Ranking	Latent class	Group value	Factor mean	Groups with significantly smaller factor mean
1	7	246	0.635	840 752 152 208 124 554 724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
2	27	840	0.580	152 208 124 554 724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
3	25	752	0.521	124 554 724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
4	3	152	0.488	724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
5	5	208	0.469	724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
6	2	124	0.466	724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
7	18	554	0.448	724 578 376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
8	24	724	0.379	376 250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
9	19	578	0.365	250 380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
10	12	376	0.322	380 703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
11	8	250	0.284	703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
12	13	380	0.261	703 300 826 372 705 616 203 276 233 528 440 56 702 392 410
13	22	703	0.207	705 616 203 276 233 528 440 56 702 392 410
14	10	300	0.200	616 203 276 233 528 440 56 702 392 410
15	26	826	0.176	616 203 276 233 528 440 56 702 392 410
16	11	372	0.174	616 203 276 233 528 440 56 702 392 410
17	23	705	0.149	203 276 233 528 440 56 702 392 410
18	20	616	0.124	203 276 233 528 440 56 702 392 410
19	4	203	0.058	233 528 440 56 702 392 410
20	9	276	0.056	233 528 440 56 702 392 410
21	6	233	-0.013	56 702 392 410
22	17	528	-0.052	702 392 410
23	16	440	-0.054	702 392 410
24	1	56	-0.092	702 392 410
25	21	702	-0.156	392 410
26	14	392	-0.776	410
27	15	410	-0.931	

syntax. These values are pasted into the syntax of the second AwC step, and some factor loadings and intercepts are fixed to these values for identification (Syntax 8.7).

```
Syntax 8.7: AwC Method Applied to the Curiosity Measurement Model
Across Countries
VARIABLE: [...]
   ! Grouping specification
   ! Define the 27 countries by the ISO code
   GROUPING IS
   CNTRYID(56 124 152 203 208 233 246 250 276 300
  372 376 380 392
   410 440 528 554 578 616 702 703 705 724 752 826
  840);
ANALYSIS:
   ESTIMATOR = MLR;
   H1ITERATIONS = 10000;
   PROCESSORS = 4;
MODEL:
   ! Overall measurement model
   CURIOUS BY I Q04d I Q04j I Q04l I Q04m;
   ! What follows are the measurement models for
   ! country with the starting values from the free
   ! alignment method and the first factor loading
   ! its starting value for identification
  purposes.
   ! The same goes for the intercept of the first
  item.
MODEL 56:
   ! Belgium
   curious BY i q04d@0.58087; !fixed to identify
  the model
   curious BY i q04j*0.87936;
   curious BY i q041*0.77616;
   curious BY i q04m*0.64888;
```

```
[ i q04d@2.74282 ]; !fixed to identify the mean
  structure
    [ i q04j*2.10687 ];
    [ i q041*2.08939 ];
    [i q04m*2.75272];
    [ curious*0 ];
   i q04d*0.47764;
   i q04j*0.39830;
   i q041*0.33275;
   i q04m*0.38442;
   curious*1;
MODEL 124:
    ! Canada
   curious BY i q04d@0.60476;
   curious BY i q04j*0.79433;
   curious BY i q041*0.84478;
   curious BY i q04m*0.61792;
    [ i q04d@2.74774 ];
    [i q04j*2.35664];
    [ i q041*2.14924 ];
    [i q04m*2.69570];
    [ curious * 0.64712 ];
   i q04d*0.40232;
   i q04j*0.36653;
   i q041*0.30741;
   i q04m*0.37838;
   curious * 0.79731;
    [...]
```

The overall fit of this model was acceptable,  $\chi^2(54) = 1715.1$ , p < 0.001, RMSEA = 0.068, CFI = 0.985, SRMR = 0.019. In fact, the AwC model fit was identical to the fit of the configural multi-group CFA model; however, the AwC model improved the comparability of factor means across countries based on the initial alignment optimisation (Step 1). This model forms the basis for further country comparisons of, for instance, factor means. Marsh et al. (2018) noted that the factor means differences between countries obtained using the AwC method are similar to those of the scalar invariance model, although the latter may not be accepted due to marginal model fit. For the curiosity scale, the scalar invariance model did indeed show a marginal fit to the data,  $\chi^2(210) = 11682.1$ , p < 0.001,

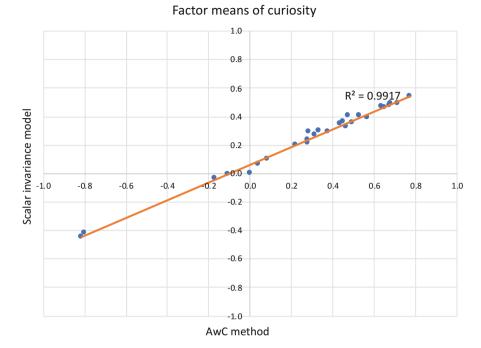


Fig. 8.4 Factor means of the PIAAC countries obtained from the scalar invariance model and the AwC method

Note. each data point represents a country

RMSEA = 0.091, CFI = 0.896, SRMR = 0.088. Figure 8.4 shows that the factor means obtained from this model and the AwC method correlated highly, r = 0.996.

As noted earlier, the AwC method can be extended to multi-group structural equation models in order, for example, to examine the effects of covariates (e.g. socio-economic status), additional group differences (e.g. across gender), or relations to outcome variables (e.g. numeracy). This flexibility represents a major strength of the method and allows researchers to approximate scalar invariance without imposing parameter constraints that are often not met in large-scale assessment datasets.

Multilevel Confirmatory Factor Analysis Both multi-group CFA and the alignment methods represent approaches to the invariance across many groups assuming groups as fixed and thereby drawing inferences to the groups in the sample (Muthén and Asparouhov 2018). In contrast to this assumption, multilevel CFA models allow researchers to draw inferences to the population level, assuming that the groups were drawn from a population of groups. Independent of the level of inferences, multilevel models can include random effects of the parameters in the measurement models (i.e. factor loadings and item intercepts or thresholds) to quantify possible between-group variation (Fox and Verhagen 2010). For instance, for a given

factor model, loadings may vary between the groups and thus indicate that metric invariance may not hold (Kim et al. 2017). In the psychometric literature, this scenario is often referred to as 'cluster bias' (Jak et al. 2014). Besides the random effects in factor loadings, random effects in the item intercepts can be specified at the between level. In a situation where a random-intercept model fits the data better than a model with fixed intercepts, researchers have gained some evidence for violations of the scalar invariance assumption (Muthén and Asparouhov 2018). Besides these invariance tests of model parameters using random effects, testing whether the measurement model is similar across the levels of analysis represents another critical step. Muthén and Asparouhov (2018) argued that different factor structures may well occur for the within level and the between level in a multilevel model. Kaplan (2009), for instance, observed a two-factor model for students' self-efficacy in mathematics assessed in the PISA 2003 study at the within (i.e. student) level, while a one-factor model held at the between (i.e. school) level. Establishing that the same factor structure holds between levels represents an invariance test that is similar to the configural invariance model. Jak (2018) emphasised the importance of establishing cross-level measurement invariance by constraining the factor loadings of the measurement model to equality across levels. She showed that the lack of this constraint may result in estimation issues, biased parameter estimates, and biased standard errors. Ultimately, the meaning of the latent variables differs between the levels of analysis. To establish whether cross-level invariance holds, researchers can compare a model with the equality constraints to a model without these constraints.

The general form of a multilevel CFA model with random intercepts and fixed factor loadings is as follows (see Fig. 8.5; e.g. Davidov et al. 2012): The *i*th item indicator  $y_{ij}$  of the *j*th person in the *k*th group can be decomposed into

Level 1 (Within level): 
$$y_{ijk} = v_{ik} + \lambda_{W_{ik}} \eta_{W_{jk}} + \varepsilon_{W_{ijk}}$$
 (8.1)

Level 2 (Between level): 
$$v_{ik} = v_i + \lambda_{B_i} \eta_{B_k} + \varepsilon_{B_{ik}}$$
 (8.2)

where  $v_{ik}$  is the *i*th item intercept of the *k*th group,  $v_i$  the *i*th item intercept across all groups,  $\lambda_W$  and  $\lambda_B$  the within- and between-level factor loadings,  $\eta_{Wjk}$  and  $\eta_{Bk}$  the within- and between-level scores of the latent variable, and  $\varepsilon_{W_{ijk}}$  and  $\varepsilon_{B_{ik}}$  the level-specific residuals. Assuming that the within- and between-level item scores are not correlated, the covariance structure of  $y_{ijk}$  ( $\Sigma_T$ ) can be decomposed into the sum of the within- ( $\Sigma_W$ ) and between-level ( $\Sigma_B$ ) covariances (e.g. Hox 2013; Muthén and Asparouhov 2018):

$$\Sigma_T = \Sigma_W + \Sigma_B = \Lambda_W \Psi_W \Lambda_W' + \Theta_W + \Lambda_B \Psi_B \Lambda_B' + \Theta_B$$
 (8.3)

This model forms the basis for testing the invariance of factor loadings, item intercepts, and residual variances across groups, as well as cross-level invariance (Kim et al. 2016a).

Evaluating the fit of multilevel CFA models—with or without random effects in the model parameters—is based on examining fit indices. However, validating

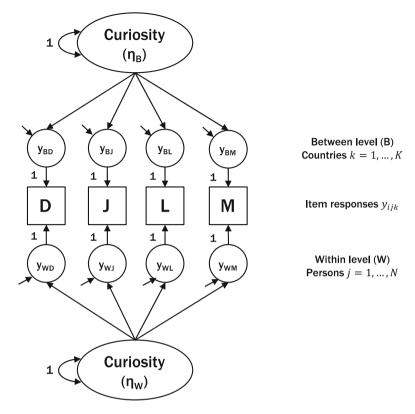


Fig. 8.5 Multilevel CFA with factor structures at both levels

the suggested guidelines for these indices still requires some research. Hsu et al. (2015) found that the common fit indices applied in SEM (i.e. CFI, TLI, RMSEA, SRMR) are not equally sensitive to model misspecifications in multilevel situations. For instance, whereas the CFI, TLI, and RMSEA were especially sensitive to misspecifications in the factor loadings and structural parameters, the SRMR-within flagged misspecifications of the factor covariance at the within level; at the between level, only the SRMR-between indicated model misspecifications. Ryu (2014b) suggested estimating level-specific fit indices and taking several steps to detect possible misfit in multilevel CFA (see also Ryu and West 2009). Her approach involves specifying and comparing at least three multilevel CFA models: (a) a model with the factor structure at the within level and the saturated between level (i.e. only variances and covariance among variables are estimated); (b) a model with the saturated within level and a factor structure at the between level; and (c) a model with the factor structure at both the within and the between levels. Comparing these models, and exploring the factor structure at one level while saturating the other, allows researchers to identify possible sources of misfit. Stapleton (2013) suggested specifying two independent baseline models against which the three models can

Model	Description	$\chi^2(df)$	CFI	RMSEA	$SRMR_w$	$SRMR_b$
1	W: Factor structure B: Saturated	252.3 (2)*	0.991	0.026	0.021	0.000
2	W: Saturated B: Factor structure	5.2 (2), p = 0.07	1.000	0.000	0.000	0.025
3	W: Factor structure B: Factor structure	469.0 (4)*	0.983	0.025	0.021	0.025
4	M3 with cross-level invariance	753.2 (8)*	0.973	0.023	0.021	2.835

**Table 8.6** Fit indices of the multilevel CFA models

Notes. The scaling correction factors and information criteria can be found in the Supplementary Material. B between level, W within level

be compared (see also Finch and Bolin 2017). The multilevel CFA approach can readily be extended to multi-group or MIMIC models by, for instance, introducing a grouping variable at the within or between level (Ryu 2014a) or by adding between-level covariates that may explain possible variation (i.e. random effects) in the model parameters and, ultimately, possible differential item functioning (Davidov et al. 2016).

In the following, I will illustrate the specification of the multilevel CFA models in Mplus. The PIAAC participating countries are considered to be the groups or, in Mplus terms, the clusters (CLUSTER = CNTRYID). Table 8.6 shows the resulting fit indices. The first model specifies the factor structure at the within level and saturates the between level (Syntax 8.8). Similarly, the second model saturates the within level and specifies the factor structure at the between level (Model 2). Model 3 assumes the same factor structure at both levels.

```
Syntax 8.8: Multilevel CFA: Factor Structure at the Within Level,
Saturated Between Level

VARIABLE: [...]

! Countries as clusters
CLUSTER = CNTRYID;

ANALYSIS:

TYPE = TWOLEVEL;
ESTIMATOR = MLR;
H1ITERATIONS = 10000;
PROCESSORS = 4;

MODEL:
```

p < 0.001

```
%WITHIN%
! Individual participant level
! Measurement model
CURIOUSw BY I_Q04d I_Q04j I_Q04l I_Q04m;
%BETWEEN%
! Country level
! Saturated measurement model
I_Q04d-I_Q04m WITH I_Q04d-I_Q04m;
```

Models 1–3 indicate that the factor structure of the curiosity construct is well-described at the within and the between level—neither the model parameters (see Supplementary Material) nor the fit indices indicate severe misfit. At the same time, the number of groups (i.e. countries) in these models is relatively small (i.e. k = 27), thus leading to large standard errors in some parameters. The resulting parameters may be biased and must therefore be interpreted with caution (Kim et al. 2016a). To further reduce the number of model parameters, researchers may consider saturating one level of analysis, especially when only country-level inferences are of major interest (Stapleton et al. 2016).

Regarding the cross-level invariance, Model 4 did not fit the data well (Table 8.6, Syntax 8.9), and the comparison between Models 3 and 4 suggested a significant deterioration in model fit after imposing the equality constraints on the factor loadings,  $\Delta\chi^2(4)=26.1,\ p<0.001,\ \Delta\text{CFI}=-0.010,\ \Delta\text{RMSEA}=-0.002,\ \Delta\text{SRMR}_w=0.000,\ \Delta\text{SRMR}_b=+2.810.$  Hence, cross-level invariance does not hold.

To test further levels of between-country invariance, additional constraints on the between-level intercepts (Mplus syntax:  $[I_Q04d-I_Q04m]$  (I1-I4);) and item residuals (Mplus syntax:  $I_Q04d-I_Q04m$  (R1-R4);) can be imposed on the multilevel CFA model with factor structures at both levels (Model 3). Finch and Bolin (2017) present and discuss the implementation of these constraints in Mplus.

```
Syntax 8.9: Multilevel CFA: Factor Structure at Both Levels with Cross-Level Invariance

VARIABLE: [...]

! Countries as clusters
    CLUSTER = CNTRYID;

ANALYSIS:
```

```
TYPE = TWOLEVEL;
   ESTIMATOR = MLR:
   H1ITERATIONS = 10000;
   PROCESSORS = 4;
MODEL:
   %WITHIN%
   ! Individual participant level
   ! Measurement model
   CURIOUSW BY
   I Q04d*(L1)
   I Q04j(L2)
   I Q041(L3)
   I Q04m(L4);
    ! Fixed factor variance
   CURIOUSw@1;
   %BETWEEN%
    ! Country level
   ! Measurement model
   CURIOUSD BY
   I Q04d*(L1)
   I Q04j(L2)
   I Q041(L3)
   I Q04m(L4);
    ! Fixed factor variance
   CURIOUSb@1;
```

#### 8.5 Structural Models

After specifying measurement models, specifying structural models that describe the relations among manifest and/or latent variables represents the second step in SEM (Kline 2016). In the following, several examples of structural equation models and their implementation in Mplus are presented, including path models, structural models with latent variables, and multi-group structural equation models. As these examples contain the PIAAC cognitive skills measures of problem solving and literacy, all analyses were conducted using the Mplus option TYPE = IMPUTATION. Hence, the model parameters presented here were combined using Rubin's rules.

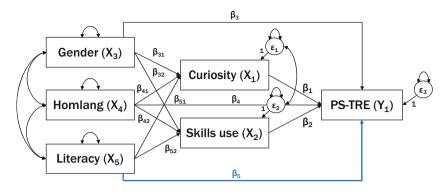


Fig. 8.6 Example of a path model

#### 8.5.1 Path Models

The following model describes the relationships among participants' background variables (i.e. gender, home language, literacy) and their performance on the PIAAC problem-solving test (PS-TRE). Researchers may hypothesise that these relationships follow a certain mechanism with indirect effects via curiosity and the use of skills. Figure 8.6 depicts the corresponding path model containing only manifest variables. Notably, this model serves only illustrative purposes and does not fully represent the theoretical assumptions that researchers may have about the relations among the variables.

Researchers may be interested in the existence of the direct path between literacy and problem solving, testing the null hypothesis that  $\beta_5 = 0$ . In a sequence of steps, several models may be specified to test this hypothesis:

- 1. The first model contains all regression paths and freely estimates the parameter  $\beta_5$  (see Syntax 8.10). This model is just-identified (df = 0) and fits the data perfectly. The corresponding, average information criteria were  $\overline{AIC} = 70,353$  and  $\overline{BIC} = 70,550$ . Across the ten datasets containing one plausible value of the cognitive skills measure each, the direct path between literacy and problem solving was  $\overline{\beta}_5 = 0.775$ , SE = 0.010, 95% CI [0.756, 0.794], p < 0.001.
- 2. The second model restricts the direct path to zero ( $\beta_5 = 0$ ). The resulting average information criteria were considerably higher than those in the first model,  $\overline{AIC} = 78,149$  and  $\overline{BIC} = 78,338$ . Moreover, this model exhibited a poor fit to the data,  $\overline{CFI} = 0.201$ ,  $\overline{RMSEA} = 0.622$ ,  $\overline{SRMR} = 0.130$ .
- 3. The third model uses the Wald test of parameter constraints to test whether  $\beta_5 = 0$ . This is implemented in Mplus by adding the following syntax after the model commands: MODEL TEST: b5 = 0; (Note: the path coefficient  $\beta_5$  is named b5 in this example; see Syntax 8.10). The resulting test was statistically significant (Wald- $\chi^2[1] = 2493.2$ , p < 0.001), indicating that the path between literacy and problem solving significantly deviates from zero. Together with the

superior fit of the model that freely estimates the path coefficient  $\beta_5$ , this provides evidence for the existence of the direct path.

Besides the direct path, researchers may also be interested in the indirect effects of the participants' background variables on problem solving via curiosity and skills use. To estimate these effects, Mplus offers several approaches, one of which is the MODEL INDIRECT option (Syntax 8.10). This option estimates all indirect, direct, and total effects in the model and can be used in conjunction with the standardisation commands and the call for confidence intervals. At the time of writing, the bootstrapped confidence intervals were not available for multiply imputed datasets. The total indirect effect of literacy on problem solving was  $\overline{\beta}_{TInd} = 0.044$ , SE = 0.006, 95% Wald CI [0.032, 0.056], p < 0.001. The specific indirect effect of literacy on problem solving via curiosity was  $\overline{\beta}_{SInd} = -0.001$ , SE = 0.004, 95% Wald CI [-0.010, 0.007], p = 0.72. The specific indirect effect of literacy on problem solving via skills use was  $\overline{\beta}_{SInd} = 0.045$ , SE = 0.005, 95% Wald CI [0.035, 0.055], p < 0.001.

```
Syntax 8.10: Model Commands for the Path Model Example
MODEL:
   ! Structural model
   ! Note: LIFE represents skills use in everyday
  life (X2).
   PSTRE ON LIFE CURIOUS FEMALE HOMLANG LIT(b5);
   LIFE ON FEMALE HOMLANG LIT;
   CURIOUS ON FEMALE HOMLANG LIT;
   ! Remaining covariances among residuals
   CURIOUS WITH LIFE;
   ! Covariances among predictors
   FEMALE WITH HOMLANG LIT;
   HOMLANG WITH LIT;
MODEL INDIRECT:
   ! Indirect, direct, and total effects
   PSTRE IND FEMALE;
   PSTRE IND HOMLANG;
   PSTRE IND LIT;
MODEL TEST:
   ! Calls the Wald test of parameter constraints
   b5 = 0;
```

```
OUTPUT:
STDY;
STDYX;
CINTERVAL; ! Wald confidence intervals
```

### 8.5.2 Structural Equation Models

As path models contain manifest variables that are prone to measurement error, substituting them with latent variables can improve the approximation of the true variances and covariances in a structural equation model (Kline 2016; Raykov and Marcoulides 2006). To accomplish this, researchers have several options, such as using the participants' item responses or item parcels as indicators of latent variables or single-indicator variables that are corrected for unreliability (Kline 2016; Little 2013). In the following example, item responses form the indicators of the latent variables (Fig. 8.7). The Supplementary Material contains additional structural equation models using item parcels to represent curiosity.

Except for the inclusion of latent variables, this model's structural part is identical to that of the path model with manifest variables. As a result, the model syntax only has to be modified by adding the measurement models of the latent variables 'Curiosity' and 'Skills use'. The specification of indirect effects and the Wald test remains (Syntax 8.11).

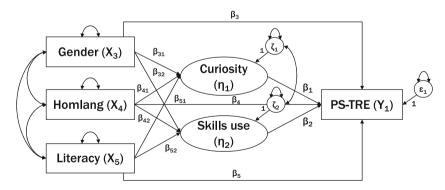


Fig. 8.7 Example of a structural equation model

### Syntax 8.11: Model Commands for the Structural Equation Model **Example**

```
MODEL:
   ! Measurement models
   ! Note: LIFE represents skills use in everyday
  life.
   LIFE BY H Q05a H Q05c H Q05e H Q05f;
   ! CURIOSITY
   CURIOUS BY I Q04d I Q04j I Q04l I Q04m;
   ! Structural model
   PSTRE ON LIFE CURIOUS FEMALE HOMLANG LIT;
   LIFE ON FEMALE HOMLANG LIT;
   CURIOUS ON FEMALE HOMLANG LIT;
   ! Remaining covariances among residuals
   CURIOUS WITH LIFE;
   ! Covariances among predictors
   FEMALE WITH HOMLANG LIT;
   HOMLANG WITH LIT;
```

This model exhibited a good fit to the data, as indicated by the following fit indices:  $\overline{\chi}^2(43) = 869.5$ , p < 0.001,  $\overline{CFI} = 0.936$ ,  $\overline{RMSEA} = 0.043$ ,  $\overline{SRMR} = 0.043$ 0.040,  $\overline{AIC} = 2,313,912, \overline{BIC} = 231,732$ . The corresponding indices for each of the ten plausible value datasets varied only marginally between datasets. Across the ten imputed datasets, the factor loadings of the curiosity variable ranged between 0.652 and 0.753 and between 0.571 and 0.678 for skills use. The average direct effect of literacy on problem solving was slightly smaller than that obtained from the path model,  $\overline{\beta}_5 = 0.755$ , SE = 0.011, 95% CI [0.732, 0.777], p < 0.001. The total indirect effect of literacy on problem solving via curiosity and skills use was  $\overline{\beta}_{TInd}$ = 0.064, SE = 0.008, 95% CI [0.048, 0.079], p < 0.001. Similar to the path model, only the specific indirect effect via skills use was statistically significant,  $\overline{\beta}_{SInd}$ = 0.070, SE = 0.008, 95% CI [0.054, 0.086], p < 0.001. This structural equation model may well be compared to alternative models through likelihood-ratio testing (Asparouhov and Muthén 2010; Enders and Mansolf 2018).

# 8.5.3 Multi-group Structural Equation Models and Structural Invariance Testing

Both path models and structural equation models can be extended to multi-group models in order to test possible differences in structural coefficients across groups. However, if latent variables are part of the model, the comparisons of structural coefficients are meaningful only if at least metric invariance holds (Guenole and Brown 2014; Raykov and Marcoulides 2006; Sass and Schmitt 2013). In the following example, the structural equation model displayed in Fig. 8.7 is specified for the German and the Norwegian sample assuming metric invariance across the two countries (Syntax 8.12). This specification uses the Mplus option GROUPING IS GERMAN (0 = Norway 1 = Germany).

```
Syntax 8.12: Model Commands for the Multi-group Structural Equation Model Example
```

```
MODEL:
   ! Measurement models
   ! Note: LIFE represents skills use in everyday
  life.
   LIFE BY H Q05a* H Q05c H Q05e H Q05f;
   ! CURIOSITY
   CURIOUS BY I Q04d* I Q04j I Q04l I Q04m;
   ! Factor variances fixed to 1 to identify the
  model
   LIFE@1:
   CURIOUS@1;
   ! Structural model
   PSTRE ON LIFE CURIOUS FEMALE HOMLANG LIT;
   LIFE ON FEMALE HOMLANG LIT;
   CURIOUS ON FEMALE HOMLANG LIT;
   ! Remaining covariances among residuals
   CURIOUS WITH LIFE;
   ! Covariances among predictors
   FEMALE WITH HOMLANG LIT;
   HOMLANG WITH LIT;
   ! Freely estimated item intercepts
   [H Q05a-H Q05f*];
   [I Q04d-I Q04m*];
```

```
! Constrained factor means to identify the mean
structure
  [LIFE@0];
  [CURIOUS@0];

! Freely estimated residual variances
  H_Q05a-H_Q05f*;
  I_Q04d-I_Q04m*;

MODEL NORWAY:
! Nothing is specified here.
```

This multi-group model resulted in an acceptable fit,  $\bar{\chi}^2(94) = 1715.9$ , p < 0.001,  $\overline{CFI} = 0.929$ ,  $\overline{RMSEA} = 0.058$ ,  $\overline{SRMR} = 0.053$ ,  $\overline{AIC} = 198,047$ ,  $\overline{BIC}$ = 198,539. The direct effect of literacy on problem solving was  $\overline{\beta}_5 = 0.763$ (SE = 0.014, 95% CI [0.736, 0.790], p < 0.001) for the German sample and  $\overline{\beta}_5 = 0.753$  (SE = 0.012, 95% CI [0.729, 0.777], p < 0.001) for the Norwegian sample. Although these two path coefficients are quite similar, researchers may want to test for their invariance across countries. Modifying the structural part under the MODEL command by labelling the structural coefficient between literacy and problem solving (e.g., PSTRE BY ... LIT (P1)) restricts this coefficient to equality across the two countries. The resulting model did not deteriorate the model fit substantially,  $\bar{\chi}^2(95) = 1718.6$ , p < 0.001,  $\overline{CFI} = 0.929$ ,  $\overline{RMSEA} = 0.058$ ,  $\overline{SRMR} = 0.053$ ,  $\overline{AIC} = 198,049$ ,  $\overline{BIC} = 198533.7$ . Reviewing the goodness-of-fit indices and the information criteria, there is no convincing evidence suggesting that the effects of literacy on problem solving differ between the German and Norwegian samples. As a consequence, researchers may conclude that the structural parameter  $\beta_5$  is invariant. To further back up this conclusion, researchers should conduct the model comparison for each of the ten datasets separately to rule out that the variation in the plausible values between these datasets may have affected these tests.

# 8.6 Some Recommendations for SEM Analyses

Given the aims associated with performing SEM, and considering the complexity of the PIAAC data, researchers are faced with several methodological decisions. The following, by no means exhaustive, recommendations may aid this decision-making process.

• *Treatment of item indicators*. Item indicators that are based on rating scales can be treated categorically or continuously using a range of estimation procedures. If researchers are primarily interested in the psychometric quality of a scale and

its items, the categorical treatment of item indicators provides more information than the continuous treatment (e.g. multiple item thresholds are estimated as opposed to one item intercept per indicator). Moreover, in many situations, the categorical treatment of indicators is preferred over the continuous treatment, especially when the item response distributions are skewed (Li 2016; Suh 2015). Nevertheless, if five or more response categories exist, the response distributions deviate from normality only to some extent, and researchers are primarily interested in controlling for measurement error in a structural equation model with latent variables, treating item indicators continuously is a viable alternative (Rhemtulla et al. 2012).

- Inclusion of performance scores. Whenever performance scores are included
  in any structural equation model, all plausible values should be used, model
  estimations should be performed for each set of plausible values, and the
  resulting model parameters should be combined following Rubin's combination
  rules (Laukaityte and Wiberg 2017; Rutkowski and Rutkowski 2016). Including
  only one plausible value in a structural equation model may be feasible when
  researchers want to perform initial checks of their hypotheses. However, it does
  not provide correct standard errors of the model parameters.
- Hierarchical data structure. In situations where multiple countries are included in the analyses, accounting for the hierarchical structural of the data (i.e. participants nested in countries) prevents ecological fallacies. The SEM framework offers several options to achieve this: (a) countries can be treated as groups, and country differences are modelled as fixed effects, for instance, through multiple-group SEM, MIMIC models, or alignment procedures; (b) countries can be treated as groups, and country differences are modelled as random effects through multilevel SEM. If researchers are aiming at comparing only a few countries, the former approaches are suitable. If researchers are aiming at quantifying and explaining the variance of certain SEM parameters (e.g. factor loadings, intercepts) between a large number of countries, the latter approach is suitable (Kim et al. 2017; Muthén and Asparouhov 2018).
- Effects of using weights. While the inclusion of weights is recommended
  for obtaining accurate model parameters and standard errors, the effects of
  not including any weights in specific analytic situations have not yet been
  fully understood (Laukaityte and Wiberg 2018; Stapleton 2002). Researchers
  are therefore encouraged to examine the effects of including weights on the
  parameters of their structural equation models.

### 8.7 Summary

Overall, SEM provides a flexible framework in order for researchers to address questions surrounding the psychometric quality of a scale, the structural relations among constructs, and ultimately the testing of specific theories and hypotheses. This framework is well applicable to international large-scale assessment data

(Rutkowski and Zhou 2014)—data sources that contain not only manifest scores and indices representing participants' background characteristics and test performance but also multiple-items scales representing constructs. Besides the great potential of SEM for analysing PIAAC data, several methodological issues must be considered. They include, but are not limited to, (a) handling missing data, (b) weighting individual-level and country-level data, (c) establishing the comparability of scores and/or relations among variables whenever comparisons between countries and subgroups within the PIAAC samples are conducted, (d) modelling the complex structure of the PIAAC data, and (e) handling multiple datasets of plausible values. As far as the range of structural equation models presented in this chapter goes, the software package Mplus can handle these complexities through a set of ready-made specification options, including the use of weights, imputed data, and convenience functions for investigating measurement invariance across multiple groups. Moreover, the software allows researchers to perform several types of estimation procedures, including maximum-likelihood, weighted least squares, and Bayesian estimation.

As the primary goal of this chapter was to present worked examples, I hope to encourage researchers to consider conducting SEM for testing their substantive theories and hypotheses using the PIAAC data. At the same time, researchers should be aware of the above-mentioned data issues and the current methodological developments to address them. In fact, I believe that using the PIAAC data for SEM requires a continuous review and updating of these developments.

### **Supplementary Material**

The Supplementary Material contains the M*plus* (version 8.2) input and output files for the models presented in this chapter. Readers can access these files through the Open Science Framework at https://doi.org/10.17605/OSF.IO/HGBFK. (ZIP 83340 kb)

### References<sup>1</sup>

Asparouhov, T., & Muthén, B. O. (2010). Chi-square statistics with multiple imputation. Los Angeles: Muthén & Muthén. Retrieved June 20, 2019, from https://www.statmodel.com/download/MI7.pdf.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. Structural Equation Modeling: A Multidisciplinary Journal, 21(4), 495–508. https://doi.org/10.1080/10705511.2014.919210.

<sup>&</sup>lt;sup>1</sup>References marked with an asterisk indicate examples of PIAAC secondary analyses using SEM.

- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22, 507–526. https://doi.org/10.1037/met0000077.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling:* A Multidisciplinary Journal, 13(2), 186–203. https://doi.org/10.1207/s15328007sem1302\_2.
- \*Borgonovi, F., & Pokropek, A. (2017a). Birthplace diversity, income inequality and education gradients in generalised trust. OECD Education Working Papers, 164, 1–40. https://doi.org/10.1787/f16a8bae-en.
- \*Borgonovi, F., & Pokropek, A. (2017b). Mind that gap: The mediating role of intelligence and individuals' socio-economic status in explaining disparities in external political efficacy in 28 countries. Intelligence, 62, 125–137. https://doi.org/10.1016/j.intell.2017.03.006.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 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: A Multidisciplinary Journal, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902\_5.
- \*Cincinnato, S De Wever, B., Van Keer, H., Valcke, M. (2016). The influence of social background on participation in adult education: Applying the cultural capital framework. Adult Education Quarterly, 66(2), 143–168. https://doi.org/10.1177/0741713615626714
- Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (2016). Explaining measurement nonequivalence using multilevel structural equation modeling: The case of attitudes toward citizenship rights. *Sociological Methods & Research*, 47(4), 729–760. https://doi.org/10.1177/0049124116672678.
- Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, 43(4), 558–575. https://doi.org/10.1177/0022022112438397.
- Desa, D. (2014). Evaluating measurement invariance of TALIS 2013 complex scales. *OECD Education Working Papers*, 103. https://doi.org/10.1787/5jz2kbbvlb7k-en.
- \*Duchhardt, C., Jordan, A.-K., Ehmke, T. J. I. J. o. S Education, M. (2017). Adults' use of mathematics and its influence on mathematical competence. International Journal of Science and Mathematics Education 15(1), 155–174. https://doi.org/10.1007/s10763-015-9670-1.
- Enders, C. K. (2010). Applied missing data analysis. New York: The Guilford Press.
- Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, 23(1), 76–93. https://doi.org/10.1037/met0000102.
- Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. https://doi.org/10.1037/met0000063.
- Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, 16(1), 54–69. https://doi.org/10.1080/10705510802561311.
- Finch, W. H., & Bolin, J. E. (2017). Multilevel modeling using Mplus. Boca Rayon: CRC Press.
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte: Information Age Publishing.
- Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), Cross-cultural analysis: Methods and applications (pp. 467–488). London: Routledge.
- \*Ganzach, Y., & Patel, P. C. (2018). Wages, mental abilities and assessments in large scale international surveys: Still not much more than g. Intelligence, 69, 1–7. https://doi.org/10.1016/j.intell.2018.03.014.

- \*Gorges, J., Koch, T., Maehler, D. B., & Offerhaus, J. J. (2017). Same but different? Measurement invariance of the PIAAC motivation-to-learn scale across key socio-demographic groups. *Large-scale Assessment in Education*, 5(1). https://doi.org/10.1186/s40536-017-0047-5.
- \*Gorges, J., Maehler, D. B., Koch, T., & Offerhaus, J. J. (2016). Who likes to learn new things? Measuring adult motivation to learn with PIAAC data from 21 countries. *Large-scale Assessment in Education*, 4(1). https://doi.org/10.1186/s40536-016-0024-4.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, 21(1), 111– 149. https://doi.org/10.1177/1094428117703686.
- Guenole, N., & Brown, A. (2014). The Consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, 5(980). https://doi.org/10.3389/fpsyg.2014.00980.
- \*Hahs-Vaughn, D. L. (2017). *Applied multivariate statistical concepts* [chapter 10]. New York: Routledge by Taylor & Francis.
- Hancock, G. R., & Mueller, R. O. (2013). *Structural equation modeling: A second course* (2nd ed.). Charlotte, NC: Information Age Publishing.
- \*Heisig, J. P. J. (2018). Measuring the signaling value of educational degrees: Secondary education systems and the internal homogeneity of educational groups. *Large-scale Assessment in Education*, 6(1). https://doi.org/10.1186/s40536-018-0062-1.
- Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology (Vol. 2: Statistical analysis)* (pp. 281–294). Oxford: Oxford University Press.
- Hoyle, R. H. (2012). Handbook of structural equation modeling. New York: The Guilford Press.
- Hsu, H.-Y., Kwok, O.-M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, 50(2), 197–215. https://doi.org/10.1080/00273171.2014.977429.
- Jak, S. (2018). Cross-level invariance in multilevel factor models. Structural Equation Modeling: A Multidisciplinary Journal, 1–16. https://doi.org/10.1080/10705511.2018.1534205.
- Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 31–39. https://doi.org/10.1080/10705511.2014.856694.
- Kaplan, D. (2009). Structural equation modeling: Foundations and extensions (2nd ed.). Thousand Oaks: SAGE Publications.
- Khojasteh, J., & Lo, W.-J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 531–541. https://doi.org/10.1080/10705511.2014.937791.
- Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822.
- Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016a). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, 51(6), 881– 898. https://doi.org/10.1080/00273171.2016.1228042.
- Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016b). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. Structural Equation Modeling: A Multidisciplinary Journal, 23(6), 870–887. https://doi.org/10.1080/10705511.2016.1196108.
- Kim, E. S., Kwok, O.-M., & Yoon, M. (2012a). Testing factorial invariance in multilevel data: A Monte Carlo study. Structural Equation Modeling: A Multidisciplinary Journal, 19(2), 250–267. https://doi.org/10.1080/10705511.2012.659623.
- Kim, E. S., Yoon, M., & Lee, T. (2012b). Testing measurement invariance using MIMIC. Educational and Psychological Measurement, 72(3), 469–492. https://doi.org/10.1177/0013164411427395.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

- Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods*, 23(3), 505–523. https://doi.org/10.1037/met0000146.
- Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. Communications in Statistics Theory and Methods, 46(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764.
- Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics – Theory and Methods*, 47(20), 4991–5012. https://doi.org/10.1080/03610926.2017.1383429.
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7.
- Little, T. D. (2013). Longitudinal structural equation modeling. New York: The Guilford Press.
- 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.
- Loehlin, J. C., & Beaujean, A. A. (2017). Latent variable models: An introduction to factor, path, and structural equation models (5th ed.). New York: Routledge.
- Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multigroup factor analysis comparison of latent means across many groups. *Psychological Methods*, 23(3), 524–545. https://doi.org/10.1037/met0000113.
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling. An integration of the best features of exploratory and confirmatory factor analysis., 10(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700.
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439–476. https://doi.org/10.1080/10705510903008220.
- Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, 49(6), 1194–1218. https://doi.org/10.1037/a0026913.
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568.
- Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1), 103–111. https://doi.org/10.1093/biomet/79.1.103.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. https://doi.org/10.1007/BF02294825.
- Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York: Routledge.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. https://doi.org/10.1037/a0026802.
- Muthén, 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.
- Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. Sociological Methods & Research, 47(4), 637–664. https://doi.org/10.1177/0049124117701488.
- Muthén, B., Muthén, L. K., & Asparouhov, T. (2017). Regression and mediation analysis using Mplus. Los Angeles: Muthén & Muthén.
- Muthén, L., & Muthén, B. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles: Muthén & Muthén.

- OECD. (2013). Technical report of the survey of adult skills (PIAAC). Paris: OECD Publishing. Retrieved June 20, 2019, from https://www.oecd.org/skills/piaac/\_Technical%20Report\_17OCT13.pdf.
- OECD. (2016). Programme for the international assessment of adult competencies (PIAAC): Public use files. Paris: OECD Publishing. Retrieved from: http://www.oecd.org/skills/piaac/publicdataandanalysis/.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. https://doi.org/10.1016/j.dr.2016.06.004.
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah: Lawrence Erlbaum Associates Publishers.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. https://doi.org/10.1037/a0029315.
- Rutkowski, D., Rutkowski, L., & Liaw, Y.-L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice*, 37(4), 40–48. https://doi.org/10.1111/emip.12225.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. https://doi.org/10.3102/0013189x10363170.
- Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, 45(4), 252–257. https://doi.org/10.3102/0013189x16649961.
- Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1), 31–57. https://doi.org/10.1177/0013164413498257.
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1), 31–57. https://doi.org/10.1177/0013164413498257.
- Rutkowski, L., & Zhou, Y. (2014). Using structural equation models to analyze ILSA data. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 425–450). Boca Raton: CRC Press.
- Ryu, E. (2014a). Factorial invariance in multilevel confirmatory factor analysis. British Journal of Mathematical and Statistical Psychology, 67(1), 172–194. https://doi.org/10.1111/bmsp.12014.
- Ryu, E. (2014b). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, 5(81). https://doi.org/10.3389/fpsyg.2014.00081.
- Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. Structural Equation Modeling, 16(4), 583–601. https://doi.org/10.1080/10705510903203466.
- Sass, D. A., & Schmitt, T. A. (2013). Testing measurement and structural invariance. In T. Teo (Ed.), Handbook of quantitative methods for educational research (pp. 315–345). Rotterdam: SensePublishers.
- Scherer, R., & Greiff, S. (2018). Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing. *European Journal of Psychological Assessment*, 34(3), 141–144. https://doi.org/10.1027/1015-5759/a000487.
- \*Scandurra, R., & Calero, J. (2017). Modelling adult skills in OECD countries. 43(4), 781–804. https://doi.org/10.1002/berj.3290.
- \*Sikora, J., Evans, M. D. R., Kelley, J. (2019). Scholarly culture: How books in adolescence enhance adult literacy, numeracy and technology skills in 31 societies. Social Science Research, 77, 1–15. https://doi.org/10.1016/j.ssresearch.2018.10.003.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 475–502. https://doi.org/10.1207/S15328007SEM0904\_2.

- Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 521–562). Charlotte: Information Age Publishing.
- Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist*, 51(3–4), 317– 330. https://doi.org/10.1080/00461520.2016.1207178.
- Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 568–580. https://doi.org/10.1080/10705511.2014.937669.
- \*Trapp, S., Blömeke, S., Ziegler, M. (2019). The openness-fluid-crystallized-intelligence (OFCI) model and the environmental enrichment hypothesis. Intelligence, 73, 30–40. https://doi.org/10.1016/j.intell.2019.01.009.
- Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4(770). https://doi.org/10.3389/fpsyg.2013.00770.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. European Journal of Developmental Psychology, 9(4), 486–492. https://doi.org/10.1080/17405629.2012.686740.
- Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple mndicator multiple cause models. *Applied Psychological Measurement*, 35(5), 339–361. https://doi.org/10.1177/0146621611405984.
- Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, 21(3), 405–426. https://doi.org/10.1037/met0000080.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

