Approximate measurement invariance

Kimberley Lek, Utrecht University

Daniel Oberski, Utrecht University

Eldad Davidov, University of Cologne and University of Zurich

Jan Cieciuch, University of Zurich and Cardinal Wyszynski University in Warsaw

Daniel Seddig, University of Zurich

Peter Schmidt, University of Giessen

Introduction

When comparing data from different countries, time points or groups, we run into two problems. First, we want to avoid large measurement artefacts that lead to erroneous substantive conclusions (Davidov et al., 2010, 2014). For example, when comparing Finnish to Columbian survey answers, we may want to account for any differences in exuberance. Second, we want to ignore the – likely plentiful – small measurement artefacts whose effect on substantive conclusions is negligible (Meuleman, 2012; Oberski, 2014). For example, when comparing Fins in 2002 with Fins in 2004 on an income question, most of the found differences are likely to be substantive; we would not want to spend an inordinate amount of time and modeling power on identifying all the little measurement differences between these already highly comparable groups. Tests for the presence or absence of measurement differences are typically called "measurement invariance tests", sometimes also known as tests of "differential item functioning" (Holland & Wainer, 2012) or "item bias" (Mellenbergh, 1989; Shealy & Stout, 1993).Techniques to test for measurement invariance are numerous (Van De Schoot et al., 2015), but, for the purposes of this chapter, can be described as broadly falling into two categories: *exact* and *approximate*.

In the *exact* methods (see Vandenberg & Lance, 2000; Vandenberg, 2002; Brown, 2015 and elsewhere in this volume), the researcher looks for a measurement model in which any "small" measurement differences are assumed to be exactly zero, while "large" differences are left completely free to be estimated from the data (termed 'partial' measurement invariance; Byrne, Shavelson & Muthén, 1989). Methods to establish the fit of such models include chi-square difference testing (Steenkamp & Baumgartner, 1998); CFI, RMSEA, and other fit measure comparisons (Cheung & Rensvold, 2002; Chen, 2007); and examination of local fit measures such as modification indices (MI) and the expected parameter changes (EPC; Byrne, Shavelson & Muthén, 1989), or the EPC of interest (Oberski, Vermunt & Moors, 2015). One way or another, all of these methods ultimately aim to find a model that balances two goals: accounting for large measurement differences while ignoring the small ones.

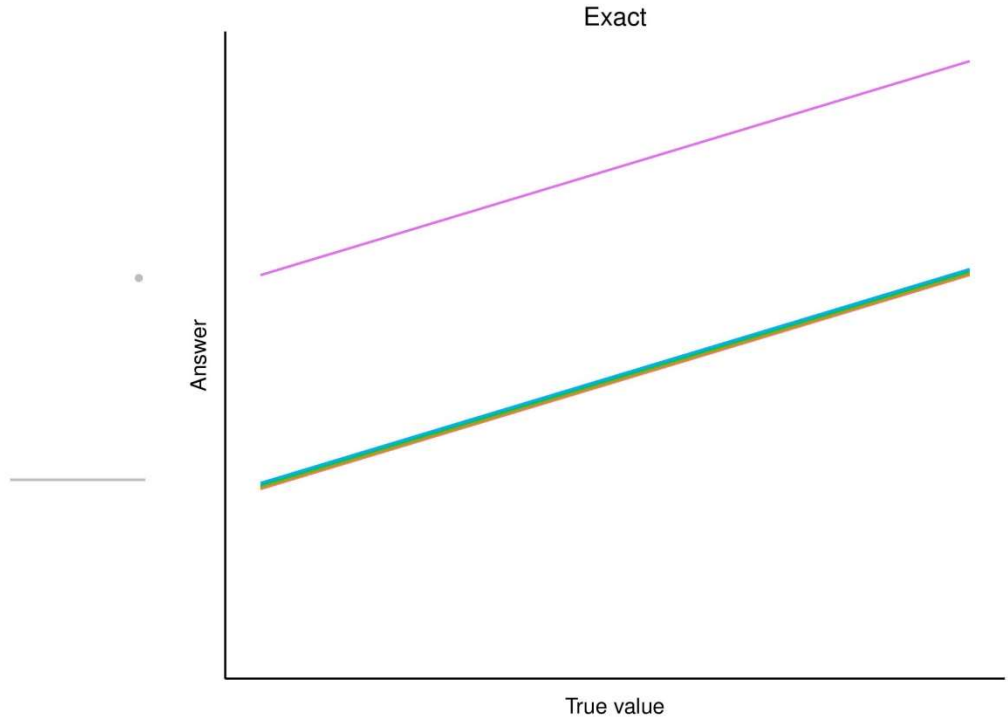
An alternative to the 'family' of exact methods, and the focus of this chapter, is the *approximate* approach. In this "approximate measurement invariance" model, "large" and "small" differences alike are assumed to follow a known distribution of nonzero values. Random effects distributions

(Fox & Verhagen, 2010), multilevel models (Davidov et al., 2012, 2016), and strong Bayesian priors (Muthén & Asparouhov, 2013; Van de Schoot et al., 2013) have all been used for this purpose. The idea in all of these techniques is that any smaller differences are automatically accounted for in the model; thus, approximate measurement invariance is primarily designed to deal with the second goal—that of ignoring small differences automatically. Achieving the first goal – dealing with large measurement artefacts – is currently an unresolved problem with this technique, although several existing proposals are discussed at the end of this chapter.

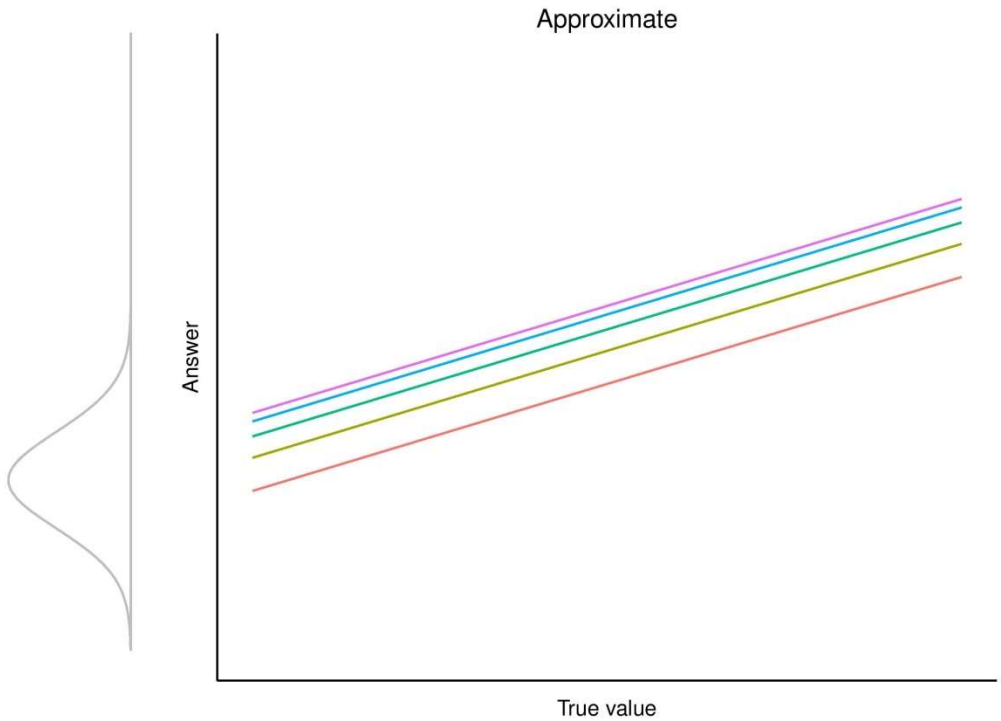
According to the advocates of approximate measurement invariance, exact zero constraints are overly strict, especially when there are many groups or time points involved (e.g., Davidov et al., 2015). One consequence is a frequent rejection of the exact invariance model, even when the parameter differences are ignorable (goal two). Another consequence is often a large series of model modifications that may capitalize on chance (MacCallum et al., 1992).In approximate measurement invariance, small differences in parameters are allowed. Moreover, the mindboggling search through all possible combinations of measurement restrictions is replaced by a relatively simple estimation procedure. With many groups and measurement parameters this practical advantage is considerable. For example, even in the simplest testing setup, a ten-factor analysis of 21 items over 19 countries (e.g. Davidov et al., 2008; Davidov 2008, 2010) yields 380 possible univariate violations of intercept equalities alone. The number of models resulting from all possible combinations of equality restrictions on intercepts and loadings is in the tens of millions. The corresponding approximate measurement invariance model aims to allow for measurement differences in these models by parameterizing them and imposing zero-mean, small-variance distributions, in a more manageable procedure.

Figure 1 illustrates the difference between the exact (A) and approximate models (B). Each graph shows the theoretical, unobserved, true value to be measured on the horizontal axis, and the obtained survey answer on the vertical axis. The lines thus correspond to the answer given by respondents with a particular true value: the response functions. These response functions may differ in intercept over groups to be compared (colors); if so, the same answer (point on the vertical axis) given by respondents from different groups (colored lines), could correspond to very different true values (corresponding points on the horizontal axis). Thus, comparing answers from these groups will compare not only true value differences but also differences in the intercepts of their response functions. If the true value differences are of the same order of magnitude as these measurement artefacts, the differences should be accounted for to prevent bias in the comparisons. Likewise, the slope of the response function may differ across groups (i.e., the loading of the survey item onto the latent factor; Brown, 2015). To keep matters simple, this chapter focusses on differences in intercepts only.

Figure 1A demonstrates the exact model: some (most) of the lines are held equal, while others (one in the example) is allowed to differ by any amount. How much it will differ is estimated from the data without restriction. The distribution of lines, shown in gray on the vertical axis, consists of a spike and a dot, since all intercepts are assumed equal except for some, which can differ by any amount. Figure 1B illustrates the corresponding approximate model. All lines are allowed to differ here, turning the spike into a normal distribution. This means that the lines that differed somewhat from the average are now allowed to differ by some amount. How much they will differ is determined in part by the data and in part by the restriction that the difference follows a normal distribution, shown on the vertical axis. This also implies that the pink group, which was estimated to differ considerably from the others in the exact model, is now pulled strongly



(A) Exact



(B) Approximate

Figure 1. Response functions (lines) for different groups (colors) under exact (A) vs. approximate (B) measurement invariance models.

towards the average by this prior. In other words, the goal of allowing for small measurement differences is accomplished, but traded off against reduced detection of large measurement differences.

The multigroup confirmatory factor analysis (MGCFA)

This chapter discusses the use measurement invariance testing as illustrated in Figure 1 in latent variable measurement models. In such models, the response functions are estimated through presumed conditional independence assumptions, and investigation of measurement invariance proceeds through restrictions on the parameters of these estimated functions. The most common model for this test is the confirmatory factor model, but this framework also includes IRT models, latent class models, and generalized multitrait-multimethod models (see Davidov et al. 2014). To simplify the discussion, we will limit ourselves to a multigroup CFA (MGCFA) here.

Given a survey response y_{igj} for respondent *i*, group *g*, and item *j*, a MGCFA measurement model is

$$y_{igj} = \tau_{gj} + \lambda_{gj}\eta_{igj} + \epsilon_{igj},$$

where

- η_{igj} is the unobserved true value (latent variable) for respondent *i;*
- ϵ_{igj} is the unobserved measurement error value (latent variable) for respondent *i;*
- τ_{gj} is the group-specific intercept for item *j;*
- λ_{gj} is the group-specific loading (slope) for item *j.*

Measurement invariance then imposes cross-group restrictions on respectively the item structure ("configural invariance"), the factor loadings ("metric invariance") and the intercepts ("scalar invariance"; Billiet, 2003; Millsap, 2011). Exact scalar invariance as in Figure 1a (for all groups except for the pink group), for example, may imply $\tau_{blue,j} = \tau_{green,j} = \tau_{yellow,j} = \tau_{red,j} \neq \tau_{pink,j}$. Since the intercept of the pink group is allowed to differ from the other groups, we speak of 'partial' rather than 'full' measurement invariance (Byrne, Shavelson & Muthén, 1989). We can test a similar assumption for the slopes, though we will simplify matters here by limiting ourselves to intercept differences, as in Figure 1, and assuming that all slopes are equal in the data.

Approximate measurement invariance suggests that the intercept differences follow a certain probability distribution, often normal (Gaussian):

$$\tau_{gj} - \tau_{gj'} \sim N(0, \sigma_j)$$

for all differing pairs of groups $g \neq g'$. This distribution corresponds to the distribution of differences shown on the vertical axis of Figure 1b. Like the exact procedure, *on average* intercept differences are expected to be zero. Differences may vary, however, and the standard deviation of these differences for item *j* is denoted here as σ_j. When σ_j is estimated from the data, a random effect (Verhagen & Fox, 2010) or a multilevel model (Davidov et al. 2012; Jak et al. 2014a, 2014b) results. When it is fixed in advance by the researcher, a "Bayesian

approximate measurement invariance" model results (Muthén & Asparouhov, 2013). An important question is how large the "typical difference" σ_j should be to appropriately balance the two goals of measurement invariance analysis: accounting for large measurement differences while ignoring the small ones.

In the remainder of this chapter, we will focus on a practical analysis of the Bayesian approximate measurement invariance model using standard software. The following section contains a worked example. We then discuss some of the outstanding pitfalls and issues with this technique in the discussion and conclusion section.

Illustration

For this illustration, we have simulated a simple dataset (called "dataset 1") consisting of continuous variables y1-y4, each believed to measure a certain continuous latent construct f1. Two groups are created, consisting of 500 respondents each. Mplus (Muthén & Muthén, 1998-2016) is used to apply the approximate measurement invariance testing procedure to this data. Together with the R package Blavaan (Merkle & Rosseel, 2016), Mplus is currently the only software package that allows you to test for approximate measurement invariance. The Mplus (version 7.4) input file which is used to simulate the data can be found in Fig. 2. Notice that the intercept differences are relatively small (0.1 versus -0.1) and cancel each other out between as well as within groups. The latent mean difference between group 1 and 2 is 0.5 (i.e., 0 in group 1 and 0.5 in group 2).

```
Montecarlo:
    names = y1-y4;
  ngroups = 2;
     nobs = 500 500;
    nreps = 1;
     save = dataset 1.dat;

Model montecarlo:
    f1 by y1@0.7  y2@0.6  y3@0.5  y4@0.4;
          y1@0.51 y2@0.64 y3@0.75 y4@0.84;     ! 1 - factor loading^2

  [y1@-0.1 y2@0.1 y3@-0.1 y4@0.1];
  [f1@0];
   f1@1;

Model montecarlo-g2:             !group 2
  [y1@0.1 y2@-0.1 y3@0.1 y4@-0.1];
  [f1@0.5];
```

Figure 2. Mplus input file containing the population parameter values for the intercepts, factor loadings, latent means and latent variances.

Using the MGCFA chi-square difference test procedure to test for measurement invariance (Vandenberg & Lance, 2000) – which is default in Mplus - one would conclude that exact measurement invariance does not hold in dataset 1. This can be seen in Figure 3, which shows that the chi-square difference test of scalar versus metric equivalence is statistically significant (at α = 0.05). Since chi-square tests are known to be sensitive to sample size and violations of the normality assumption (Brannick, 1995), some authors (e.g., Brown, 2015; Chen, 2007) have suggested to take into account commonly used fit indices such as the comparative fit index (CFI; Bentler & Bonett, 1980) and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) in the judgement of measurement invariance. Following the guidelines of

Chen (2007; p. 501), also based on the CFI and RMSEA differences we would conclude that scalar invariance does not hold ($\Delta\text{CFI} \geq -0.01$; $\Delta\text{RMSEA} \geq 0.015$; Table 1). Ignoring the absence of scalar invariance leads to an underestimation of the f1 mean difference between group 1 and 2 (i.e., 0.399 instead of 0.500).

```
                                          Degrees of
     Models Compared              Chi-square   Freedom     P-value

     Metric against Configural         0.797         3       0.8502
     Scalar against Configural        63.928         6       0.0000
     Scalar against Metric            63.131         3       0.0000
```

Figure 3. Mplus output of the MGCFA chi-square comparisons. The scalar equivalence model fits significantly worse than the metric equivalence model, hence exact measurement equivalence does not hold.

Table 1

RMSEA and CFI differences between the configural, metric and scalar model

	Configural	Metric	Scalar
CFI	0.991	0.995	0.873
RMSEA	0.047 (0.000 – 0.092)	0.025 (0.000 – 0.064)	0.112 (0.088 – 0.137)

Instead of forcing the differences in intercepts to be exactly zero, we could opt for approximate measurement invariance, by using the Mplus input file depicted in Figure 4 (based on Muthén & Asparouhov, 2013 and Van De Schoot et al., 2013b). This input file is a special application of Bayesian structural equation modeling (BSEM) in which strict zero constraints are replaced by probability distributions with zero-mean and small-variance (see Muthén & Asparouhov, 2012; Van Erp, Mulder & Oberski, in press).These probability distributions are called “priors” in the Bayesian terminology. The prior distributions are confronted with the data, reflected in the “likelihood”, to come to a “posterior” distribution which is essentially a compromise of the prior and the likelihood (for a more thorough discussion of Bayesian statistics see, amongst others, Gelman et al., 2003; Kaplan & Depaoli, 2013; Kruschke et al., 2012; Lee, 2007). Thus, when we place a small-variance prior with zero mean on the intercepts, the posterior balances model fit on the one hand (i.e., the likelihood) and measurement invariance restrictions (i.e., the prior) on the other. The smaller the prior variance, the more the posterior will be influenced by the prior measurement invariance restrictions. The key part of the input file in Figure 4 is

```
MODEL: [y1 - y4] (nu#_1 - nu#_4);

MODEL PRIOR: DO(1,4) DIFF (nu1_#-nu2_#) ~ N(0,.01);
```

In the MODEL statement, labels nu1_1 – nu1_4 are placed on the intercepts for group 1 and labels nu2_1 – nu2_4 are placed on the intercepts in group 2, using automatic labeling where # is replaced by the group number. In the MODEL PRIOR statement[1], a normal prior with mean 0 and relatively small variance .01 is placed on the *difference* in intercepts of group 1 (nu1_# refers to nu1_1 – nu1_4) and 2 (nu2_# refers to nu2_1 – nu2_4). This prior determines the

[1] For all other parameters in the model, we rely on the default prior settings of Mplus (see Asparouhov & Muthén, 2010).

“wiggle room” we allow in the intercept estimates of group 1 and 2 (Van De Schoot et al., 2013b).

When we run the input file of Fig. 4, posterior draws of the parameters are generated over and over again in each iteration of the Bayesian algorithm. As an illustration, Figure 5 shows the posterior draws of iteration 1-20 for the intercept of y1 in group 1 (left) and group 2 (right). Posterior draws for groups 1 and 2 in a specific iteration are connected by a line. The steepness of this line – i.e., the difference between y1 in group 1 and 2 - is restricted by the prior we have specified. If the parameters were equal in each draw, the lines would be horizontal; the steeper the lines, the larger the intercept differences between groups in each posterior draw. As can be seen in the Figure 5, these differences, in each posterior draw are present but modest – exactly as the Gaussian prior on these differences stipulates.

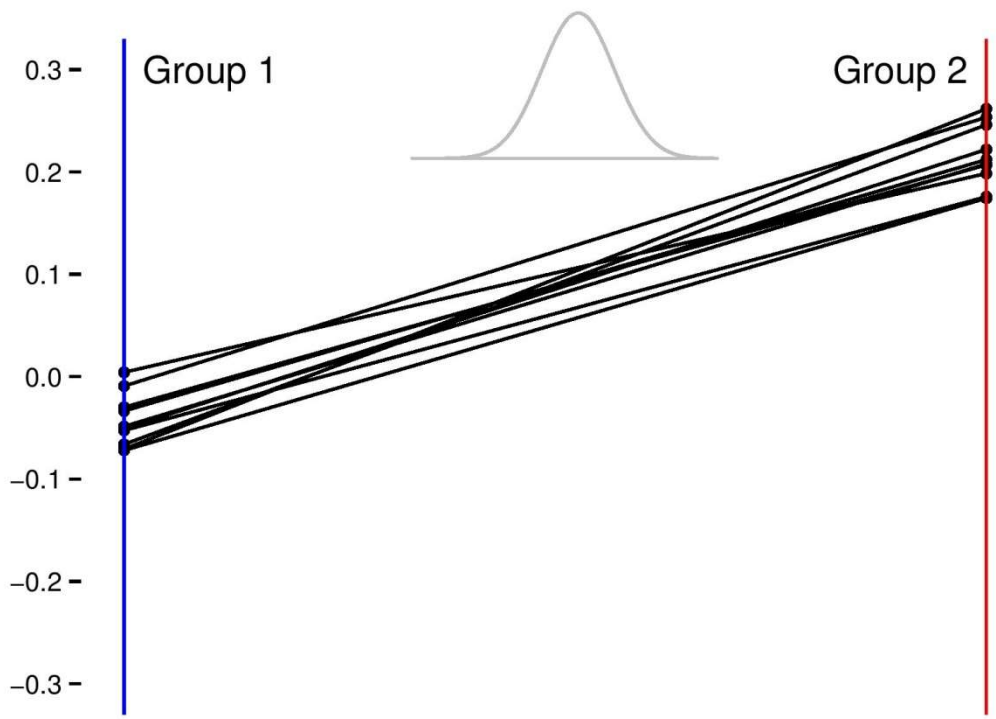


Figure 5. Visualization of the estimation of the intercept y1 in groups 1 and group 2

When the Bayesian algorithm is finished, we first need to check whether this algorithm has converged to the appropriate posterior (see Depaoli & Van De Schoot, 2015). In Mplus, convergence can be assessed visually - by looking at the traceplot for every parameter in the model – and statistically, by checking the potential scale reduction factor which should be close to 1 (PSR, Gelman, 1996). Mplus stops the Bayesian algorithm when the PSR drops below 1 + ϵ with a default ϵ between 0.05 and 1 for most of the models[2] (Asparouhov & Muthén, 2010). We choose a more stringent stopping rule by specifying `BCONVERGENCE = .01;` (Figure 4). We additionally force Mplus to run at least 100.000 iterations by specifying `BITERATIONS =`

[2] $\epsilon = fc$ where *c* is 0.05 by default and f is a multiplicity factor that takes into account the number of parameters in the model. `Bconvergence = .01;` replaces *c* by 0.01, hence yielding a more stringent convergence criterion.

`(100000);`. Mplus informs us that convergence has been reached according to the adjusted PSR criterion (`THE MODEL ESTIMATION TERMINATED NORMALLY`). Based on the traceplots of the intercepts we would also conclude that the algorithm has converged (Figure 6), allowing us to turn to the Mplus output.

```
DATA:  FILE = "dataset 1.dat";

VARIABLE:  NAMES ARE y1-y4 group;
           KNOWNCLASS IS g(group=1 group=2);
           CLASSES ARE g(2);

ANALYSIS: TYPE = MIXTURE;
          ESTIMATOR = BAYES;
          MODEL = ALLFREE;

          BCONVERGENCE = .01;
          BITERATIONS = 500000(100000);
          bseed = 123;

MODEL:

%OVERALL%
f1 by y1 y2 y3 y4 (lam#_1-lam#_4);
 [y1-y4](nu#_1-nu#_4);

%G#1%
[f1@0];
 f1@1;

%G#2%
[f1];
 f1@1;

MODEL PRIOR:
DO(1,4) DIFF (lam1_#-lam2_#) ~ N(0,.01);
DO(1,4) DIFF (nu1_#-nu2_#) ~ N(0,.01);
```

Knownclass is used to describe the grouping variable; needed when “type is mixture” is specified in the analysis command

MODEL = ALLFREE is needed for DIFF and automatic labeling with # (see MODEL statement)

Stricter convergence guidelines than default to reduce any bias due to precision

Labeling; the # makes sure labels are automatically specified for group 1 and 2

DO(1,4) loop applies the DIFF statement to all 4 variables. DIFF statement is used to place a prior on the differences in intercepts and factor loadings.

Figure 4. Input file in Mplus for the Bayesian approximate measurement equivalence test.

```
MODEL FIT INFORMATION

Bayesian Posterior Predictive Checking using Chi-Square

          95% Confidence Interval for the Difference Between
          the Observed and the Replicated Chi-Square Values

                        -14.418            27.098

          Posterior Predictive P-Value       0.269

MODEL RESULTS

                                Posterior  One-Tailed         95% C.I.
               Estimate       S.D.      P-Value   Lower 2.5%  Upper 2.5%
Significance

Latent Class 1 (1)

Means
    F1                 0.000       0.000      1.000       0.000       0.000

 Intercepts
    Y1                -0.069       0.044      0.056      -0.155       0.017
    Y2                 0.127       0.045      0.002       0.039       0.215      *
    Y3                -0.058       0.044      0.095      -0.145       0.029
    Y4                 0.112       0.044      0.006       0.025       0.197      *

Latent Class 2 (2)

Means
    F1                 0.477       0.122      0.000       0.242       0.721      *

 Intercepts
    Y1                 0.121       0.080      0.069      -0.041       0.272
    Y2                -0.084       0.072      0.121      -0.227       0.056
    Y3                 0.053       0.066      0.212      -0.079       0.177
    Y4                -0.011       0.056      0.421      -0.125       0.096

DIFFERENCE OUTPUT

                                    NU1_1       NU2_1
      5       0.026      0.053     -0.095*      0.095*

                                    NU1_2       NU2_2
      6       0.022      0.049      0.105*     -0.105*

                                    NU1_3       NU2_3
      7      -0.003      0.046     -0.055       0.055

                                    NU1_4       NU2_4
      8       0.050      0.041      0.062*     -0.062*
```

Figure 5. Part of the Mplus output resulting from the input file in Figure 4.

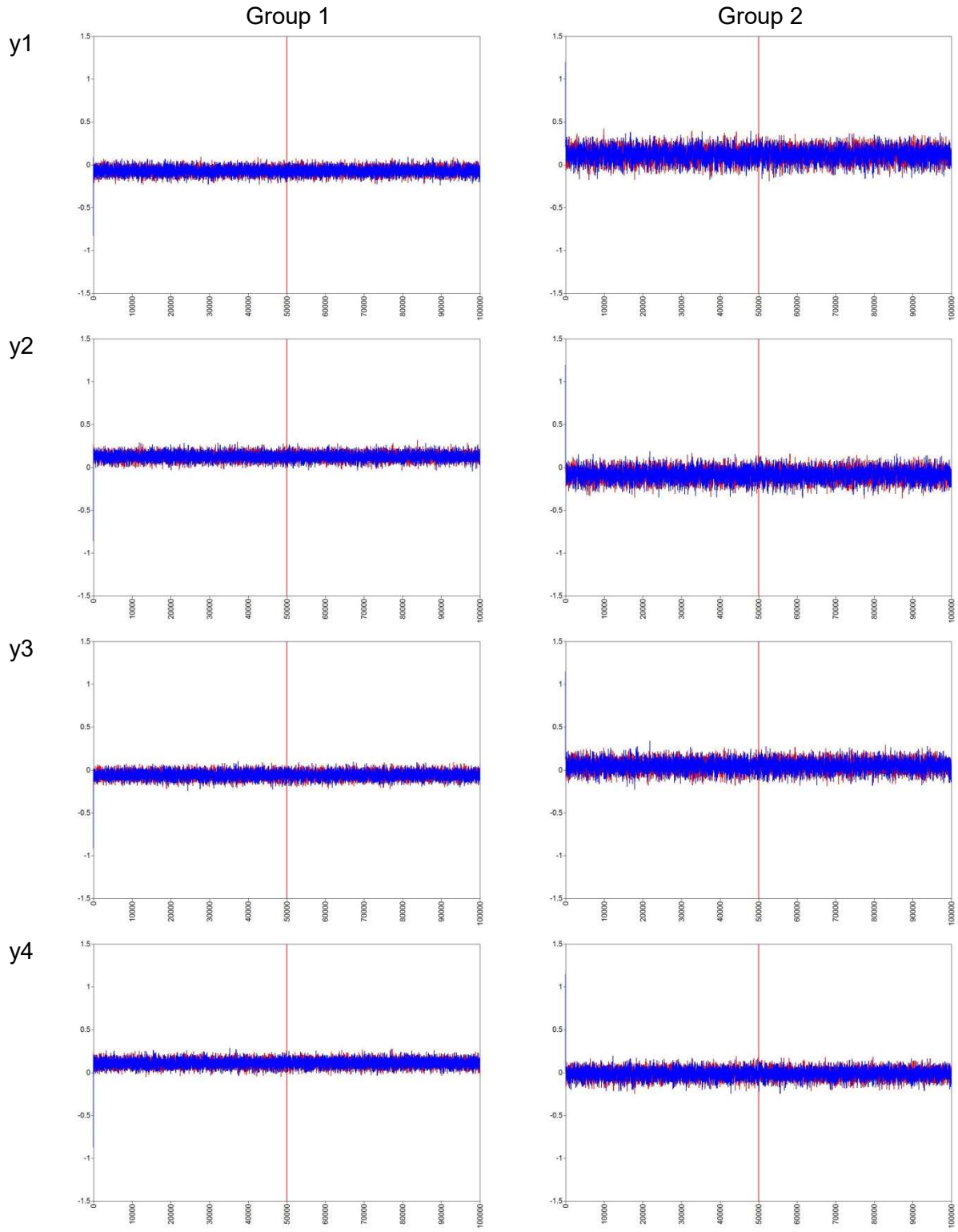


Figure 6. Traceplots to judge the convergence of intercept y1-y4 in group 1 and 2. Note that only the last 50.000 (after the red vertical line) are used for the parameter estimates.

A part of the Mplus output resulting from the input in Fig. 4 is shown in Fig. 7. Notice first that most of the fit indices usually provided by Mplus (RMSEA, CFI, et cetera) are not available anymore. To judge whether our Bayesian approximate measurement invariance model fits our data, we rely on a likelihood ratio test (LRT) between the approximate measurement invariance model and an unrestricted mean and (co)variance model (Asparouhov & Muthén, 2010). Specifically, in every iteration Mplus conducts two LRT's, using the current parameter estimates. The first of these LRT's (1) evaluates the fit between the current model and the original data. The second one (2) confronts the current model with a newly generated dataset, simulated on the basis of the estimated model. This latter one provides us with the LRT chi-square values we could reasonably obtain, would our approximate measurement invariance model be true. Chi-square values of (1) which are systematically higher than those of (2) are an indication of model misfit. To determine whether this is the case, we can either look at the PPP-value (Gelman et al., 1996) or the 95% credibility interval provided in the Mplus output (Fig. 7). The PPP expresses the proportion of chi-square values obtained with (2) that exceed (1). PPP-values around 0.5 are indicative of good model fit, low PPP-values close to zero should be avoided. In this case, we would be fairly satisfied with a PPP value of 0.269 (Fig. 7), although a PPP closer to 0.5 would be favorable. The 95% credibility interval is determined for the distribution of differences between (1) and (2). When (1) is not systematically higher than (2), zero is included in this 95% credibility interval, which is fortunately the case in the present example. Turning to the estimates (Fig. 7), we see that the intercepts of the two groups are estimated in line with their true values (Fig. 2) but are generally pulled closer to zero. The DIFFERENCE section of the output shows the mean intercept across groups and the amount by which every group-specific intercept deviates from this value. The latent mean difference is estimated to be 0.477, reasonably close to the true difference of 0.5. In this example, we initially allowed a prior variance of 0.01, taking into account the scale of the y1-y5 variables. Since the choice for a suitable prior variance is crucial to the Bayesian approximate measurement procedure, it is good practice to perform a sensitivity analysis with multiple plausible prior variances, as displayed in Figure 8 (Van De Schoot & Depaoli, 2014). In this way, it is possible to carefully balance model fit (i.e., PPP, 95% confidence interval) and the possibility to compare groups (i.e., keeping the prior variance as small as possible). When we increase the prior variance to 0.05 in this example, the PPP-value moves closer to 0.5 and the 95% credibility interval becomes more symmetric around zero. However, increasing the prior variance also enlarges the standard errors of the intercepts and the latent mean difference estimate. The resulting latent mean difference estimate (0.457) is slightly worse than the one we obtained with prior variance 0.01 (0.477). Increasing the prior variance to 0.1 does not yield a further improvement of the PPP / 95% credibility interval and only changes the parameter estimates slightly. Therefore, a prior variance of 0.01 or 0.05 seems the best choice here.

Altogether, Bayesian approximate measurement invariance seems to largely solve the problem of exact scalar noninvariance (Fig. 3) in dataset 1. Indeed, Bayesian approximate measurement invariance is suggested to be useful in situations in which there are many small parameter differences that cancel each other out both within and between groups (De Boeck, 2008; Muthén & Asparouhov, 2013; Van de Schoot et al., 2013b; Wolvers & Lugtig, 2016). What if the differences between intercepts become larger? Or if the differences between the groups are systematic (i.e., do not cancel each other out within groups)? To check the performance of Bayesian approximate measurement invariance in these situations, we altered the intercept values of "dataset 1" in the way described in Table 2.

	σ_j = 0.1		σ_j = 0.05		σ_j = 0.01	
	Est(se) G1	Est(se) G2	Est(se) G1	Est(se) G2	Est(se) G1	Est(se) G2
Intercepts						
y1	-.09(.04)	.17(.19)	-.08(.04)	.16(.14)	-.07(.04)	.12(.08)
y2	.15(.05)	-.10(.19)	.15(.05)	-.10(.14)	.13(.05)	-.08(.07)
y3	-.07(.05)	.09(.15)	-.07(.05)	.09(.11)	-.06(.04)	.05(.07)
y4	.13(.05)	-.02(.12)	.13(.05)	-.02(.09)	.11(.04)	-.01(.06)
Δf1	.447(.296)		.457(.216)		.477(.122)	
Model fit						
95% CI Observed - Replicated Chi-Square Values	-15.78	25.19	-15.86	24.80	-14.42	27.10
PPP-value	.322		.326		.269	

Figure 8. Sensitivity analysis on the influence of prior variance on parameter differences.

Table 2

Alteration of the intercept values of dataset 1

	Group 1				Group 2			
difference is	y1	y2	y3	y4	y1	y2	y3	y4
large	-0.5	0.5	-0.5	0.5	0.5	-0.5	0.5	-0.5
systematic	-0.1	-0.1	-0.1	-0.1	0.1	0.1	0.1	0.1

Regardless of prior variance choice, when intercept differences are systematic, the intercept estimates are no longer in line with their true values. With a prior variance of 0.01, the latent mean difference estimate is too high: 0.789. Interestingly, the PPP fails to detect the misfit (PPP = 0.368). As stated by Muthén and Asparouhov (2013), recovery of parameters is not expected when the non-invariance is not in line with BSEM. Enlarging the intercept differences as in the first row of Table 2 leads to a PPP-value of 0.000 with prior variance 0.01. Increasing the prior variance to 0.05 yields a PPP-value of 0.186 and a latent mean difference estimate of 0.642. Increasing the prior variance even further to 0.1 changes the PPP to 0.278 and a more acceptable latent mean difference estimate of 0.566. In sum, when differences are systematic or relatively large, one should be cautious in applying the approximate measurement testing procedure.

Discussion & Conclusion

The increasing availability of large cross-cultural and cross-country surveys in the last decades has significantly increased the possibilities for researchers to conduct comparative studies. However, they have also considerably increased the risk researcher may run into of drawing wrong conclusions. Therefore, the methodological literature on cross-cultural and cross-country analysis has recommended testing for measurement equivalence to guarantee that differences across groups are due to substantive true differences and not methodological artefacts. This recommendation has been increasingly applied by researchers, who tested for the measurement equivalence properties of various scales (e.g. Cieciuch et al. 2014; for an overview, see Davidov et al. 2014). Unfortunately, a new problem has come up: Many scales failed to display high levels of equivalence.

In this chapter we have discussed approximate measurement invariance as a possible solution to this problem. Instead of restricting the differences between all measurement parameters (i.e., factor loadings, intercepts) to be exactly zero, approximate measurement invariance assumes that these differences follow a (normal) distribution with mean zero and small variance σ_j. This variance σ_j can either be estimated from the data (Verhagen & Fox, 2010; Davidov, 2012) or be fixed in advance by the researcher (Asparouhov & Muthén, 2013). The latter is known as 'Bayesian' approximate measurement invariance and is illustrated in this chapter with standard software. Approximate measurement invariance seems especially advantageous when (1) the number of groups or repeated measurements is large, (2) there are many small differences in intercepts and factor loadings and (3) differences cancel each other out both within and between groups (Muthén & Asparouhov, 2013, Van de Schoot et al., 2013; Wolvers & Lugtig, 2016). Exact measurement invariance almost never holds in this scenario and is cumbersome to test for.

When there are additionally some large differences in intercepts or factor loadings, approximate measurement invariance may not establish equivalence. The small variance prior tends to pull strongly deviating parameter estimates towards the average across groups/time points. The result is that the deviating parameter will be smaller while the invariant parameters will be larger than their true values (Muthén & Asparouhov, 2012). This leads to a considerable bias in the latent mean estimates (Van de Schoot et al., 2013). As illustrated in this chapter, bias may also result from systematic differences between groups. A promising solution to reduce bias is to combine approximate measurement invariance testing with the newly developed alignment procedure of Asparouhov and Muthén (2014). This alignment procedure rotates the solution in such a way that there are many invariant parameters and a few (large) noninvariant parameters, using the same principles as used in CFA (see Jennrich, 2006 for technical details; for an application see Cieciuch et al. 2018, Munck et al. 2017). Another solution is to free non-invariant parameters and only apply approximate measurement invariance to the remaining parameters (see Muthén & Asparouhov, 2013).

Several studies have already applied the approximate measurement invariance test (e.g. Davidov et al. 2015, 2017, Zercher et al. 2015). These studies have demonstrated that approximate equivalence may be given also when exact equivalence is rejected by the data. However, as Davidov et al. (2015) mentioned, it "does not do magic": there is a point at which one must conclude that measurement invariance simply does not hold (Lommen, Van de Schoot & Engelhard, 2015). The key question is when *exactly* that point is reached. More

research into this key question, the role of large deviating parameters and the size of σ_j is necessary[3].

Acknowledgments

The authors thank Rens van de Schoot for comments on an earlier version of this chapter. The work of Eldad Davidov and Jan Cieciuch was supported by the University Research Priority Program Social Networks, University of Zurich. The work of Peter Schmidt was supported by the Alexander von Humboldt Polish Honorary Research Fellowship granted by the Foundation for Polish Science. Kimberley Lek is funded with a talent grant from the Netherlands Organization for Scientific Research (NWO): NWO Talent 406-15-062. Daniel Oberski is supported by NWO Veni grant 451-14-017.

References

Asparouhov, T., & Muthén, B. O. (2010). Bayesian analysis using Mplus: Technical implementation. Technical appendix. Los Angeles: Muthén & Muthén. Retrieved from www.statmodel.com

Asparouhov, T., & Muthén, B. O. (2014). Multi-group factor analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 1-14, doi: 10.1080/10705511.2014.919210

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. In J. A. Harkness, F. J. R. Van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 247–264). New York: John Wiley.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16, 201-213.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Publications.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi:10.1037/0033-2909.105.3.456

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. doi:10.1080/10705510701301834

[3] This chapter has introduced the concept of approximate measurement invariance and illustrated the use of its most basic variant. More complex variants, such as multilevel/hierarchical models and other types of Bayesian priors on differences, have fallen out of the scope of this chapter. For applications of multilevel/hierarchical models to measurement invariance, see Cheung and Au 2005; Davidov et al. 2012, in press; Jak et al. 2014a,b; Jak et al. 2013; Meuleman 2016. Furthermore, we have avoided issues external to measurement equivalence, such as overall model fit and concept equivalence (see, e.g., Meitinger 2014).

Cheung, M. W.-L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research, *12*(4), 598–619.

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. doi:10.1207/S15328007SEM0902_5

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (in press). Testing for approximate measurement invariance of human values in the European Social Survey. Unpublished manuscript submitted for publication.

Cieciuch, J., Davidov, E., & Schmidt, P. (in press). Alignment optimization: Estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In E. Davidov, P. Schmidt, & J. Billiet J. (Eds), *Cross Cultural Analysis: Methods and Applications*. NY: Routledge

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: A cross-country illustration with a new scale to measure 19 human values, *Frontiers in Psychology*, 5:982.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European social survey. *Survey Research Methods*, *2*(1), 33–46. doi:10.18148/srm/2008.v2i1.365

Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European social survey. *International Journal of Comparative Sociology*, *51*(3), 171–191. doi:10.1177/0020715210363534

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European social survey exact versus approximate measurement equivalence. *Public Opinion Quarterly*, *79*(S1), 244–266. doi:10.1093/poq/nfv008

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (in press). Explaining measurement non-equivalence using multilevel structural equation modeling.

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.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. doi:10.1146/annurev-soc-071913-043137.

Davidov, E., Schmidt, P., & Billiet, J. (2010). *Cross-cultural Analysis: Methods and Applications*. New York: Taylor and Francis.

Davidov, E., Schmidt, P., & Schwartz S. H. (2008). Bringing values back in: the adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly 72*(3), 420–445

Depaoli, S., & Van De Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, Advance online publication.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice*. London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition*. CRC Press.

Gelman, A., Meng, X. L., Stern, H. S., & Rubin, D. B. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733-807.

Holland, P. W., & Wainer, H. (2012). *Differential Item Functioning*. New York: Routledge.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 265–282. doi:10.1080/10705511.2013.769392

Jak, S., Oort, F. J., & Dolan, C. V. (2014a). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 31–39. doi:10.1080/10705511.2014.856694

Jak, S., Oort, F. J., & Dolan, C. V. (2014b). Using two-level factor analysis to test for cluster bias in ordinal data. *Multivariate Behavioral Research*, *49*(6), 544–553. doi:10.1080/00273171.2014.947353

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*(1), 173–191. doi:10.1007/s11336-003-1136-B

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 407-437). Oxford, UK: Oxford University Press.

Kruschke, J. K., Arguinis, H., & Joo, H. (2012). The time has come! Bayesian methods for data analysis in the organizational sciences. *Organisational Research Methods, 15,* 722-752.

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian approach*. New York: John Wiley & Sons.

Lommen, M. J. J., Van De Schoot, R., & Engelhard, I. M. (2014). The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. *Frontiers in Psychology*, *5,* 1-7. doi:10.3389/fpsyg.2014.01304

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalizing on chance. *Psychological Bulletin, 111,* 490-504.

Meitinger, K. (2014). Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. Presented at the XVIII ISA World congress of Sociology conference, July 13-19, Yokohama, Japan.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. doi:10.1016/0883-0355(89)90002-5

Meuleman, B. (2012). When are item intercept differences substantively relevant in measurement invariance testing? In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, Theories and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt* (pp. 97–104). Heidelberg, Germany: Springer VS.

Meuleman, B. (2016). Explaining cross-national inequivalence in factor loadings and intercepts: A Bayesian multilevel SEM approach. Paper presented at the 2nd 3MC conference, 25-29 July, Chicago.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. New York: Routledge.Munck, I., Barber, C. & Torney-Purta, J. (2017). *Measurement invariance in comparing attitudes towards immigrants among youth across Europe in 1999 and 2009: The Alignment methods applied to IEA CIVED and ICCS*. Unpublished manuscript submitted for publication.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi:10.1037/a0026802

Muthén, B. O., & Asparouhov, T. (2013). BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17. Available online at: http://www.statmodel.com

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Oberski, D. L., Vermunt, J. K., & Moors, G. B. D. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis*, *23*(4), 550–563.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. doi:10.1007/BF02294572

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. doi:10.1086/209528Van De Schoot, R., Kaplan, D., Denissen, J., Assendorpf, J. B., Neyer, F. J., & Van Aken, M. A. G. (2013a). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842-860.

Van De Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist, 16*(2), 75-84.

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013b). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00770

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.01064

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. doi:10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. doi:10.1177/109442810031002

Wolvers, R. J., & Lugtig, P. (2016). A comparison of four methods for testing measurement invariance across many groups. Unpublished master's thesis, Utrecht University.

Zercher, F.& Schmidt, P.& Cieciuch& Davidov E.(2015) The Comparability of the universalism value over time and across countries in the European Social Survey: exact versus approximate measurement invariance, Frontiers in Psychology,6, 207-217.