41

Approximate Measurement Invariance

Kimberley Lek[1], Daniel Oberski[1], Eldad Davidov[2,3], Jan Cieciuch[3,4], Daniel Seddig[2,3], and Peter Schmidt[5]

[1] *Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*
[2] *Institute of Sociology and Social Psychology, University of Cologne, Cologne, Germany*
[3] *Department of Sociology and University Research Priority Program "Social Networks", University of Zurich, Zurich, Switzerland*
[4] *Institute of Psychology, Cardinal Wyszynski University in Warsaw, Warsaw, Poland*
[5] *Department of Political Science, University of Giessen, Giessen, Germany*

41.1 Introduction

When comparing data from different countries, time points, or groups, we run into at least two problems. First, we want to avoid large measurement artifacts that lead to erroneous substantive conclusions [1, 2]. 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 artifacts whose effect on substantive conclusions is negligible [3, 4]. For example, when comparing Finns in 2002 with Finns in 2004 on an income question, most of the differences found are likely to be substantive; we would not want to spend an inordinate amount of time and modeling power on identifying all the small 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 [5] or item bias [6, 7]. Techniques to test for measurement invariance are numerous [8] but, for the purposes of this chapter, can be described as broadly falling into one of two categories: exact and approximate.

In the exact methods (see Refs. [9–11] and Chapter 40, this volume), the researcher looks for a measurement model in which any small measurement

Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC), First Edition. Edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A.L. Stoop, and Brita Dorer.

differences are assumed to be exactly zero, while large differences are left completely free to be estimated from the data (termed partial measurement invariance [12]; also see Chapter 40 in this volume). Methods to establish the fit of such models include chi-square difference testing [13], comparative fit index (CFI), root mean squared error of approximation (RMSEA), and other fit measure comparisons [14, 15] and examination of local fit measures such as modification indices (MI) and the expected parameter changes (EPC) [12], or the EPC of interest [16]. One way or another, all of these methods ultimately aim to find a model that balances two strategies, namely, 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 [17]. In this approximate measurement invariance model, large and small differences alike are assumed to follow a known distribution of nonzero values. Random effects distributions [18], multilevel models [19, 20], and strong Bayesian priors [21, 22] 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 strategy – that of ignoring small differences automatically. The first strategy – dealing with large measurement artifacts – is problematic, 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. Ref. [23]). One consequence is a frequent rejection of the exact invariance model, even when the parameter differences are ignorable (i.e. the second strategy). Another consequence is often a large series of model modifications that may capitalize on chance [24]. In approximate measurement invariance, small differences in parameters are allowed. Moreover, the mind-boggling 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 10-factor analysis of 21 items over 19 countries (e.g. Refs. [25–27]) 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 and small variance distributions in a more manageable procedure.

Figure 41.1 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

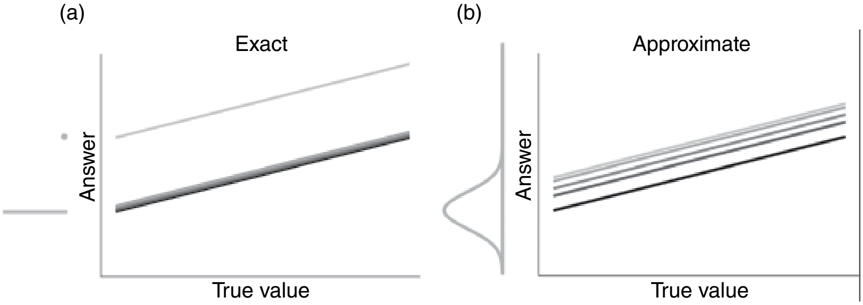


Figure 41.1 Response functions (lines) for different groups (grayscale) under exact (a) vs. approximate (b) measurement invariance models.

vertical axis. The lines thus correspond to the answers given by respondents with a particular true value: the response functions. These response functions may differ in intercept over groups to be compared (grayscale); if so, the same answer (point on the vertical axis) given by respondents from different groups (gray 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 artifacts, 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 [11]). To keep matters simple, this chapter focuses on differences in intercepts only.

Figure 41.1a demonstrates the exact model: most of the lines are held equal, while others (one in the example) are 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 41.1b 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 lightest gray group, which was estimated to differ considerably from the others in the exact model, is now pulled strongly toward the average by this prior. In other words, the strategy for allowing for small measurement differences is accomplished but traded off against reduced detection of large measurement differences.

41.2 The Multigroup Confirmatory Factor Analysis

This chapter discusses the use of measurement invariance testing as illustrated in Figure 41.1 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 item response theory (IRT) models, latent class models, and generalized multitrait–multimethod models (see Ref. [2]). To simplify the discussion, we will limit ourselves to a multigroup confirmatory factor analysis (MGCFA) here. Given a survey response y_{igj} for respondent *i*, group *g*, and item *j*, a MGCFA measurement model is

$$y_{igj} = \tau_{gj} + \lambda_{gj}\eta_{igj} + \epsilon_{igj} \tag{41.1}$$

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 the item structure (configural invariance), the factor loadings (metric invariance), and the intercepts (scalar invariance [28, 29]). Exact scalar invariance as in Figure 41.1a (for all groups except for the lightest gray group), for example, may imply $\tau_{1,j} = \tau_{2,j} = \tau_{3,j} = \tau_{4,j} \neq \tau_{5,j}$. Since the intercept of the lightest gray group ($\tau_{5,j}$) is allowed to differ from the other groups, we speak of "partial" rather than "full" measurement invariance (Ref. [12]; see also Chapter 40 in this volume). We can test a similar assumption for the slopes, though we will simplify matters here by limiting ourselves to intercept differences, as in Figure 41.1, 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):

$$\tau_{gj} - \tau_{g'j} \sim N(0, \sigma_j) \tag{41.2}$$

for all differing pairs of groups $g \neq g'$. This distribution corresponds to the distribution of differences shown on the vertical axis of Figure 41.1b. As in 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 [18] or a multilevel model [29–31] results. When it is fixed in advance by the researcher, a Bayesian approximate measurement invariance model results [21]. An important question is how large the typical difference *σj* should be to appropriately balance the two strategies of measurement invariance analysis: accounting for the 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.

41.3 Illustration

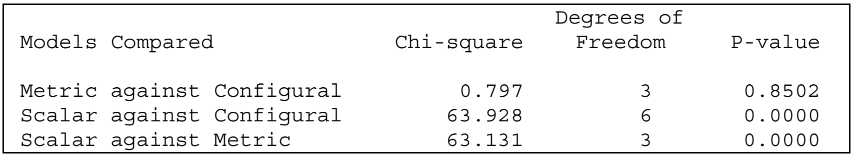
For this illustration, we have simulated a simple dataset (dataset 1) consisting of continuous variables y1–y4, each believed to measure a certain continuous latent construct f1. Two groups are created, consisting of 500 respondents each. Mplus [32] is used to apply the approximate measurement invariance testing procedure to this data. Together with the R package Blavaan [33], Mplus is currently the only software package that allows you to test for approximate measurement invariance. The Mplus (Version 7.4) input file that is used to simulate the data can be found in Figure 41.2. Notice that the intercept differences are relatively small (0.1 vs. –0.1) and cancel each other out between as well as within groups. The latent mean difference between groups 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 41.2 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 [9] – which is the default in Mplus – one would conclude that exact measurement invariance does not hold in dataset 1. This can be seen in Figure 41.3, which shows that the chi-square difference test of scalar versus metric equivalence is statistically significant ($\alpha = .05$). Since chi-square tests are known to be sensitive to sample size and violations of the normality assumption [34], some authors (e.g. Refs. [11, 15]) have suggested to take into account commonly used fit indices such as the CFI [35] and the RMSEA [36] in the judgment of measurement invariance. Following the guidelines of Chen [15, p. 501], also based on the CFI and RMSEA differences, we would conclude that scalar invariance does not hold (ΔCFI ≥ – 0.01; ΔRMSEA ≥ 0.015; Table 41.1). Ignoring the absence of scalar invariance leads to an underestimation of the f1 mean difference between groups 1 and 2 (i.e. 0.399 instead of 0.500).

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 41.4 (based on Refs. [21, 22]). 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 Refs. [37, 38]). 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 that is essentially a compromise of the prior and the likelihood (for a more thorough discussion of Bayesian statistics, see, e.g. Refs. [39–41] and [42]). Thus, when we place a small variance prior with zero mean on the intercepts, the posterior



```
                                            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 41.3 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 41.1 RMSEA and CFI differences between the configural, metric, and scalar models.

	Configural	Metric	Scalar
CFI	0.991	0.995	0.873
RMSEA	0.047	0.025	0.112

```
 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 groups 1 and 2

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

Figure 41.4 Input file in Mplus for the Bayesian approximate measurement equivalence test.

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 in the input file in Figure 41.1 is MODEL PRIOR: DO(1,4) DIFF (nu1_#-nu2_#) ~ *N*(0,0.01); the part where the small variance prior is specified. Let us disentangle this part of the input file step by step, beginning with the last comment "*N*(0,0.01)". This comment shows that in this input file, the prior follows a normal distribution with mean zero and variance 0.01. Remember that the choice of variance is important, since it is the variance that determines the wiggle room we allow in the intercept estimates of groups 1 and 2 [22]. In front of "*N*(0,0.01)" we find the statement "DIFF (nu1_#-nu2_#)." Because of this "DIFF" statement, we place the small variance prior on the

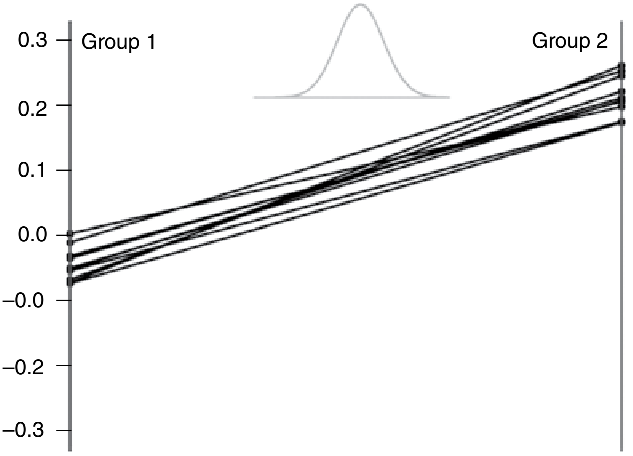


Figure 41.5 Visualization of the estimation of the intercept y1 in group 1 and group 2.

difference in intercepts between group 1 and group 2, instead of on the intercepts themselves. "nu1_#-nu2_#" between brackets are the labels referring to the intercepts for y1–y4 in group 1 and group 2. These labels were attached previously to the intercepts of y1–y4, under the "MODEL" statement. Since labeling can be cumbersome, in this input file automatic labeling is applied. Specifically, the # in the labels is automatically replaced by the number of the item (1, 2, 3, or 4). As such, Mplus automatically places the small variance prior on the difference between nu1_1 (the intercept of y1 in group 1) and nu2_1 (the intercept of y1 in group 2), the difference between nu1_2 and nu2_2, the difference between nu1_3 and nu2_3, and the difference between nu1_4 and nu2_4. Finally, the "DO(1,4)" comment makes sure Mplus correctly replaces the # of the automatic labeling by 1, 2, 3, and 4.

When we run the input file of Figure 41.4, posterior draws of the parameters are generated over and over again in each iteration of the Bayesian algorithm. As an illustration, Figure 41.5 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 groups 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 Figure 41.5, these differences in each posterior draw are present but modest – exactly as the Gaussian prior on these differences stipulates.

When the Bayesian algorithm is completed, we first need to check whether this algorithm has converged to the appropriate posterior (see Ref. [43]). 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 that should be close to 1 (PSR [44]). Mplus stops the Bayesian algorithm when the PSR drops below $1 + \epsilon$ with a default ϵ between 0.05 and 1 for most of the models[1] [45]. We choose a more stringent stopping rule by specifying BCONVERGENCE = 0.01 (Figure 41.4). We additionally force Mplus to run at least 100 000 iterations by specifying BITERATIONS = (100 000). Mplus informs us that convergence has been accomplished 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 41.6), allowing us to turn to the Mplus output.

A part of the Mplus output resulting from the input in Figure 41.4 is shown in Figure 41.7. Notice first that most of the fit indices usually provided by Mplus (RMSEA, CFI) 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 [45]. Specifically, in every iteration Mplus conducts two LRTs using the current parameter estimates. The first of these LRTs, (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 shows LRT chi-square values can reasonably be expected when approximate measurement invariance holds. Chi-square values of (1) that 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 [44] or the 95% credibility interval provided in the Mplus output (Figure 41.7). 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, and low PPP-values close to zero should be avoided. In this case, we would be fairly satisfied with a PPP-value of 0.269 (Figure 41.7), although a PPP closer to 0.5 would be preferable. 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 (Figure 41.7), we see that the intercepts of the two groups are estimated in line with their true values (Figure 41.2) but are generally pulled closer to zero. The DIFFERENCE OUTPUT shows the mean intercept across groups and the amount by which every group-specific intercept deviates from this value. The latent

[1] $\epsilon = fc$ 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 = 0.01 replaces c by 0.01, hence yielding a more stringent convergence criterion.

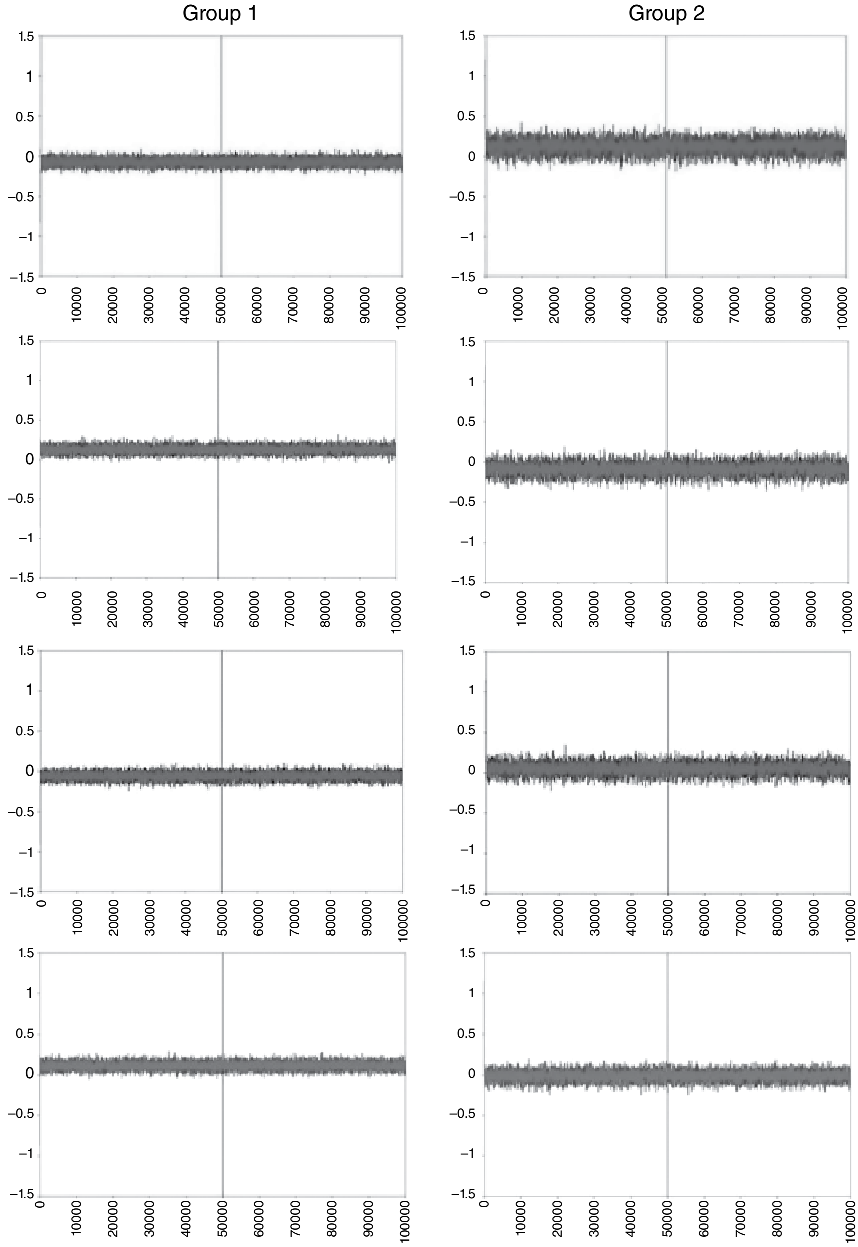


Figure 41.6 Traceplots to judge the convergence of intercept Y1–Y4 in groups 1 and 2. Note that only the last 50 000 (after the gray vertical line) are used for the parameter estimates.

```
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 41.7 Part of the Mplus output resulting from the input file in Figure 41.4.

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–y4 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 Table 41.2 [46]. In this way, it is

Table 41.2 The influence of prior variance on parameter differences.

	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	−0.09 (0.04)	0.17 (0.19)	−0.08 (0.04)	0.16 (0.14)	−0.07 (0.04)	0.12 (0.08)
y2	0.15 (0.05)	−0.10 (0.19)	0.15 (0.05)	−0.10 (0.14)	0.13 (0.05)	−0.08 (0.07)
y3	−0.07 (0.05)	0.09 (0.15)	−0.07 (0.05)	0.09 (0.11)	−0.06 (0.04)	0.05 (0.07)
y4	0.13 (0.05)	−0.02 (0.12)	0.13 (0.05)	−0.02 (0.09)	0.11 (0.04)	−0.01 (0.06)
Δf1	0.447 (0.296)		0.457 (0.216)		0.477 (0.122)	
Model fit						
95% CI[a]	−15.78	25.19	−15.86	24.80	−14.42	27.10
PPP-value	0.322		0.326		0.269	

[a] In each iteration, a chi-square value is obtained for (1) the difference between the current model and the original data and (2) the current model and a newly generated dataset, simulated on the basis of the current model. Subtracting (2) from (1) in every iteration leads to a distribution of observed (1) minus replicated (2) chi-square values. 95% of this distribution lies between the lower bound and the upper bound of the 95% credibility interval.

possible to carefully balance model fit (i.e. PPP, 95% credibility 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 (Figure 41.3) 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 [21, 22, 47, 48]. What if

Table 41.3 Alteration of the intercept values of dataset 1.

	Group 1				Group 2			
Differences	**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[a]	−0.1	−0.1	−0.1	−0.1	0.1	0.1	0.1	0.1

[a] Systematic in the sense that the direction of the intercepts (i.e. negative or positive) is systematically different in the two groups.

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 41.3.

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, 0.789, is too high. Interestingly, the PPP fails to detect the misfit (PPP = 0.368). As stated by Muthén and Asparouhov [21], recovery of parameters is not expected when the noninvariance is not in line with BSEM. Enlarging the intercept differences as in the first row of Table 41.2 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.

41.4 Discussion and Conclusion

The increasing availability of large cross-cultural and cross-country surveys in the last several decades has significantly increased the possibilities for researchers to conduct comparative studies. However, they have also considerably increased the risk of drawing wrong conclusions that researchers may run into. 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 artifacts. This recommendation has been increasingly applied by researchers, who tested for the measurement equivalence properties of various scales (e.g. Ref. [49]; for an overview, see Ref. [2]). Unfortunately, a new problem has come up, namely, that 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 [18, 19] or be fixed in advance by the researcher [21, 22]. 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 the number of groups or repeated measurements is large, there are many small differences in intercepts and factor loadings, and differences cancel each other out both within and between groups [21, 22, 48]. Exact measurement invariance almost never holds in this scenario and is cumbersome to test.

When additionally there are 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 toward the average across groups and/or time points. The result is that the deviating parameter will be smaller, while the invariant parameters will be larger than their true values [37]. This leads to a considerable bias in the latent mean estimates [22]. 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 in [50]. 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 Ref. [51] for technical details; for an application see Ref. [52]). Another solution is to free noninvariant parameters and only apply approximate measurement invariance to the remaining parameters (see Ref. [21]).

Several studies have already applied the approximate measurement invariance test (e.g. Refs. [23, 53]). These studies have demonstrated that approximate equivalence may be given also when exact equivalence is rejected by the data. However, as [23] mentioned, it “does not do magic”; there is a point at which one must conclude that measurement invariance simply does not exist [54]. 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.

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 Refs. [19, 20, 30, 31, 55–59]. Additionally, it is not yet clear how exactly to compare models with different priors in the context of approximate measurement invariance. Some preliminary results show that the PPP and DIC are not so well suited and alternatives have been proposed [60]. Furthermore, we have avoided issues external to measurement equivalence, such as overall model fit and concept equivalence (see, e.g. Ref. [61]).

Acknowledgment

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.

References

1 Davidov, E., Schmidt, P., Billiet, J., and Meuleman, B. (2018). *Cross-Cultural Analysis: Methods and Applications*, 2. New York: Taylor & Francis.

2 Davidov, E., Meuleman, B., Cieciuch, J. et al. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology* 40 (1): 55–75.

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

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

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

6 Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research* 13 (2): 127–143.

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

8 Van De Schoot, R., Schmidt, P., De Beuckelaer, A. et al. (2015). Editorial: measurement invariance. *Frontiers in Psychology* 6: doi: 10.3389/fpsyg.2015.01064.

9 Vandenberg, R.J. and 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.

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

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

12 Byrne, B.M., Shavelson, R.J., and 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.

13 Steenkamp, J.-B.E.M. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25 (1): 78–90.

14 Cheung, G.W. and Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling* 9 (2): 233–255.

15 Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling* 14 (3): 464–504.

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

17 Muthén, B. and Asparouhov, T. (2017). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research* doi:10.1177-0049124117701488.

18 Fox, J.-P. and Verhagen, A.J. (2010). Random item effects modeling for cross-national survey data. In: *Cross-Cultural Analysis: Methods and Applications* (ed. E. Davidov, P. Schmidt and J. Billiet), 467–488. London, UK: Routledge Academic.

19 Davidov, E., Dülmer, H., Schlüter, E. et al. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology* 43 (4): 558–575.

20 Davidov, E., Dülmer, H., Cieciuch, J. et al. (2016). Explaining measurement non-equivalence using multilevel structural equation modeling. *Sociological Methods & Research* doi: 10.1177/0049124116672678.

21 Muthén, B.O. and Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Note 17*. http://www.statmodel.com/examples/webnotes/webnote17.pdf (accessed 8 March 2018).

22 Van De Schoot, R., Kluytmans, A., Tummers, L. et al. (2013). 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.

23 Davidov, E., Cieciuch, J., Meuleman, B. et al. (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.

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

25 Davidov, E., Schmidt, P., and 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.

26 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.

27 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.

28 Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. In: *Cross-Cultural Survey Methods* (ed. J.A. Harkness, F.J.R. van de Vijver and P.P. Mohler), 247–264. New York: Wiley.

29 Millsap, R.E. (2011). *Statistical Approaches to Measurement Invariance.* New York: Routledge.

30 Jak, S., Oort, F.J., and Dolan, C.V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling* 21 (1): 31–39.

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

32 Muthén, L.K. and Muthén, B.O. (1998–2015). *Mplus User's Guide*, 7. Los Angeles, CA: Muthén and Muthén.

33 Merkle, E. and Rosseel, Y. (2016). *Blavaan (R Package Version 0.1–3)*, 1–16. Vienna, Austria: The Comprehensive R Archive Network.

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

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

36 Steiger, J.H. and Lind, J. M. (1980). Statistically based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA (28 May).

37 Muthén, B. and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods* 17 (3): 313–335.

38 Van Erp, S., Mulder, J., and Oberski, D.L. (2017). Prior sensitivity in default Bayesian structural equation modeling. *Psychological Methods* doi: doi 10.1037/met0000131.

39 Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*, 2. Boca Raton, FL: Chapman and Hall/CRC Press.

40 Kaplan, D. and Depaoli, S. (2013). Bayesian statistical methods. In: *Oxford Handbook of Quantitative Methods* (ed. T.D. Little), 407–437. Oxford, UK: Oxford University Press.

41 Kruschke, J.K., Arguinis, H., and Joo, H. (2012). The time has come! Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods* 15: 722–752.

42 Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach.* New York: Wiley.

43 Depaoli, S. and Van De Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: the WAMBS-checklist. *Psychological Methods* 22: 240–261.

44 Gelman, A. (1996). Inference and monitoring convergence. In: *Markov Chain Monte Carlo in Practice* (ed. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 131–144. London: Chapman and Hall.

45 Asparouhov, T. and Muthén, B.O. (2010). *Bayesian Analysis Using Mplus: Technical Implementation. Technical Appendix*. Los Angeles, CA: Muthén and Muthén.

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

47 De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73: 533–559.

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

49 Cieciuch, J., Davidov, E., Schmidt, P. et al. (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.

50 Asparouhov, T. and Muthén, B.O. (2014). Multi-group factor analysis alignment. *Structural Equation Modeling* 21: 1–14.

51 Jennrich, R.I. (2006). Rotation to simple loadings using component loss functions: the oblique case. *Psychometrika* 71 (1): 173–191.

52 Cieciuch, J., Davidov, E., and Schmidt, P. (2018). Alignment optimization: estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In: *Cross Cultural Analysis: Methods and Applications* (ed. E. Davidov, P. Schmidt, J. Billiet and B. Meuleman), 571–593. New York: Routledge.

53 Zercher, F., Schmidt, P., Cieciuch, J., and 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.

54 Lommen, M.J.J., Van De Schoot, R., and Engelhard, I.M. (2014). The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. *Frontiers in Psychology* 5: 1–7.

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

56 Jak, S., Oort, F.J., and Dolan, C.V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling* 20 (2): 265–282.

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

58 Meuleman, B. and Schlüter, E. (2018). Explaining cross-national measurement inequivalence: a Bayesian multilevel CFA with random loadings. In: *Cross Cultural Analysis: Methods and Applications* (ed. E. Davidov, P. Schmidt, J. Billiet and B. Meuleman), 363–390. New York: Routledge.

59 Davidov, E., Dülmer, H., Cieciuch, J. et al. (2016). Explaining measurement nonequivalence using multilevel structural equation modeling: the case of attitudes toward citizenship rights. *Sociological Methods & Research.* doi: 10.1177/0049124116672678.

60 Hoijtink, H. and Van De Schoot, R. (2017). Testing small variance priors using prior-posterior predictive p-values. *Psychological Methods* doi: 10.1037/met0000131.

61 Meitinger, K. (2014). Necessary but insufficient: why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly* 81 (2): 447–472.