# New Methods for the Study of Measurement Invariance with Many Groups

*Bengt Muthén & Tihomir Asparouhov*

Mplus

www.statmodel.com

October 1, 2013

## Abstract

This papers considers new factor analytic and item response theory (IRT) approaches to the study of invariance across groups. Two methods are described and contrasted. The alignment method considers the groups as a fixed mode of variation, while the random-intercept, random-loading two-level method considers the groups as a random mode of variation. Both maximum-likelihood and Bayesian analysis is applied. A survey of close to 50,000 subjects in 26 countries is used as an illustration. In addition, the two methods are studied by Monte Carlo simulations. A list of considerations for choosing between the two methods is presented.

# 1 Introduction

This papers considers new factor analytic and item response theory (IRT) approaches to the study of invariance across many groups. The analysis of many groups presents special difficulties in that it is often realistic to assume that there is a large degree of measurement non-invariance. This is typically the case with studies comparing countries in that quite different subject background and country characteristics cause potentially wide differences in response processes. Recent methodological developments attempt to take this into account, providing modeling that assumes only approximate measurement invariance while still making it possible to make group comparisons on latent variables.

To structure the presentation, it is useful to distinguish between two traditional strands of research viewing the groups as fixed or random modes of variation. With fixed mode, inference is to the groups in the sample (e.g. all U.S. states, all European countries) and usually there is a relatively small number of groups, leading to multiple-group factor analysis or multiple-group IRT. With random mode, inference is to a population from which the groups/clusters have been sampled (e.g. U.S public schools) and usually there is a relatively large number of groups/clusters, leading to two-level factor analysis or two-level IRT. Using either of the two views, two new techniques have recently been proposed that have in common the notion of approximate measurement invariance:

1. Fixed mode: Alignment (Asparouhov & Muthén, 2013)
2. Random mode: Two-level modeling with random item parameters (de Jong et al., 2007; Fox, 2010; Jak, 2013ab)

This paper gives an overview of the two approaches, describes how they relate to each other, and gives some recommendations for choosing between them.

The following example will be used throughout to illustrate the different analysis approaches. The data are from the European Social Survey as discussed in Beierlein et al. (2012). The survey intended to cover the 28 European Union countries and if possible all other European states including Russia and Israel. Due to cost issues, however, not all countries participated, resulting in 26 countries and 49,894 subjects with an average country sample size of 1,919. The latent variable constructs of tradition and conformity are measured by four items presented in portrait format, where the scale of the items is such that a high value represents a low level of tradition-conformity. The item wording is shown in Table 1. The two constructs have been found to correlate highly and are here viewed as forming a single factor.

[Table 1 about here.]

The structure of the paper is as follows. Section 2 applies conventional, fixed-mode multiple-group factor analysis to the 26-country data, presents the fixed-mode alignment method, and applies the alignment method to the 26-country data. Section 3 presents different two-level models, contrasts them, and applies them to the 26-country data. Section 4 presents Monte Carlo simulation studies of the two methods. Section 5 concludes with a comparison of the two methods on several practical criteria.

# 2 Fixed Mode Analysis

## 2.1 Conventional Multiple-Group Factor Analysis

With fixed mode analysis, it is well known that factor analysis of multiple groups commonly considers three different degrees of measurement invariance (see, e.g. Millsap, 2011): configural, metric (also referred to as weak factorial invariance), and scalar (strong factorial invariance). Configural invariance specifies the same location of the zero factor loadings of confirmatory factor analysis (CFA) commonly used with multiple-group analysis; see, however, "ESEM" (Exploratory Structural Equation Modeling) analysis of multiple groups (Asparouhov & Muthén, 2009). No equality restrictions across groups are present for any of the parameters. Metric invariance holds the values of the factor loadings equal across groups. This makes it possible to make group comparisons of factor variances and structural relationships in SEM. Scalar invariance specifies that both the factor loadings and the measurement intercepts (thresholds with categorical items) are invariant. This makes it possible to compare factor means and factor intercepts across groups.

The following introduces notation and gives a quick refresher of the corresponding three sets of factor analysis formulas for a particular item in the one-factor case for individual $i$ in group $j$.

Configural:

$$y_{ij} = \nu_j + \lambda_j \; f_{ij} + \epsilon_{ij}, \tag{1}$$

$$E(f_j) = \alpha_j = 0, V(f_j) = \psi_j = 1.$$

Metric:

$$
\begin{gathered}
y_{ij} = \nu_j + \lambda\, f_{ij} + \epsilon_{ij}, \qquad (2)\\
E(f_j) = \alpha_j = 0, V(f_j) = \psi_j.
\end{gathered}
$$

Scalar:

$$
\begin{gathered}
y_{ij} = \nu + \lambda\, f_{ij} + \epsilon_{ij}, \qquad (3)\\
E(f_j) = \alpha_j, V(f_j) = \psi_j,
\end{gathered}
$$

where $\nu$ is a measurement intercept, $\lambda$ is a factor loading, $f$ is a factor with mean $\alpha$ and variance $\psi$, and $\epsilon$ is a residual with mean zero and variance $\theta$, uncorrelated with $f$. The configural model has subscript $j$ for both intercepts and loadings, the metric model drops the subscript $j$ for the loadings, and the scalar model drops the subscript $j$ for both intercepts and loadings. Given the non-invariant intercepts and loadings, the configural model cannot identify the factor mean and variance, but sets the metric of the factor by fixing the factor mean to zero and the factor variance to one, while the metric model identifies group differences in the factor variances, and the scalar model identifies group differences in both factor means and variances.

For historical reasons, metric invariance has dominated multiple-group analysis given that mean structure modeling was introduced relatively late in SEM, initially having a covariance structure emphasis. In other fields such as IRT, the opposite is the case with a stronger emphasis on the categorical counterpart to measurement intercepts (referred to as difficulties in IRT). The emphasis on metric invariance is unfortunate because it is hard to imagine how an item can be perceived the same way by subjects if in the regression of an item on a factor only the regression slope (the factor loading) and not the regression intercept (the measurement intercept) is invariant. Scalar invariance, however, has been found to rarely fit the data well, especially in the analysis of many groups. This has hampered the comparison of factor means across groups. The new fixed-mode method referred to as alignment solves this problem. Interestingly, the method is not limited to the traditional domain of multiple-group CFA or IRT where only a few groups are typically studies, but the alignment method is suitable for the study of many groups, say up to 100.

Measurement invariance (referred to as "item bias" and "DIF" in IRT) has traditionally been concerned with comparing a small number of groups such as with gender or ethnicity using techniques such as likelihood-ratio chi-square testing of one item at a time (see, e.g., Thissen et al, 1993). Two common approaches have been discussed (Kim & Yoon, 2011; Lee et al., 2010; Stark et al., 2006):

- Bottom-up: Start with no invariance (configural case), imposing invariance one item at a time
- Top-down: Start with full invariance (scalar case), freeing invariance one item at a time, e.g. using modification indices (Sörbom, 1089)

Neither approach is scalable - both are very cumbersome when there are many groups, such as 50 countries ($50 \times 49/2 = 1225$ pairwise comparisons for each item). The correct model may well be far from either of the two starting points, which may lead to the wrong model.

## 2.2 Conventional Multiple-Group CFA of the 26-Country Example

Table 2 shows the model fit results for the configural, metric, and scalar models. The large sample size of $49,894$ produces zero p-values for all three models. The configural model, however, may be deemed to have reasonable RMSEA and CFI fit values. It is clear that the addition of invariant intercepts of the scalar model in particular adds greatly to the misfit.

[Table 2 about here.]

The scalar model shows many large modification indices: 83 in the range of 10-100, 15 in the range of 100-200, and 16 in the range of 200-457 (the largest value). The presence of so many large modification indices implies that a long sequence of model modifications is needed to reach a model with acceptable fit and the search for a good model may easily lead to the wrong model. We conclude that traditional multiple-group CFA fails due to too many necessary model modifications. This is a typical outcome when a scalar invariance model is applied to many groups. It is then impossible to compare factor means across the groups. A new method is needed. In this paper we review the radically different method of alignment as proposed in Asparouhov and Muthén (2013).

## 2.3 The Alignment Method

An advantage of the alignment method is that it has the same fit as the configural model. The alignment method minimizes the amount of measurement non-invariance by estimating the factor means $\alpha$ and factor variances $\psi$. This is possible despite the fact that these parameters are not identified without imposing scalar invariance because a different set of restrictions is imposed that optimizes a simplicity function. The simplicity function $F$ is optimized at a few large non-invariant parameters and many approximately invariant parameters rather than many medium-sized non-invariant parameters (compare with EFA rotations using functions that aim for either large or small loadings, not mid-sized loadings)

In the alignment optimization of the simplicity function, the factor means $\alpha_j$ and variances $\psi_j$ are free parameters, noting that for every set of factor means and variances the same fit as the configural model is obtained with loadings $\lambda_j$ and intercepts $\nu_j$ changed as:

$$\lambda_j = \lambda_{j,configural}/\sqrt{\psi_j}, \tag{4}$$

$$\nu_j = \nu_{j,configural} - \alpha_j \; \lambda_{j,configural}/\sqrt{\psi_j}. \tag{5}$$

The alignment method has two steps:

1. Estimate the configural model:
   - Loadings and intercepts free across groups, factor means fixed at zero, factor variances fixed at one
2. Alignment optimization:
   - Free the factor means and variances and choose their values to minimize the total amount of non-invariance using a simplicity function

$$F = \sum_p \sum_{j_1<j_2} w_{j_1,j_2} \; f(\lambda_{pj_1} - \lambda_{pj_2}) + \sum_p \sum_{j_1<j_2} w_{j_1,j_2} \; f(\nu_{pj_1} - \nu_{pj_2}), \tag{6}$$

for every pair of groups and every intercept and loading using a component loss function (CLF) $f$ from EFA rotations (Jennrich, 2006)

In this way, a non-identified model where factor means and factor variances are added to the configural model is made identified by adding a simplicity requirement. Simulation studies show that the alignment method works very well unless there is a majority of significant non-invariant parameters or small group sizes. For well-known examples with few groups and few non-invariances, the results agree with the alignment method.

In addition to the estimated aligned model, the alignment procedure as implemented in Mplus Version 7.11 gives measurement invariance test results produced by an algorithm that determines the largest set of parameters that has no significant difference between the parameters. Factor mean ordering among groups and significant differences produced by z-tests are also given. Information is further provided on each item's intercept and loading contribution to the optimized simplicity function. An $R^2$ measure is a useful descriptive statistic for the degree of invariance for a parameter, showing how much of the configural parameter variation across groups can be explained by variation in the factor means and factor variances. A high $R^2$ value indicates a high degree of measurement invariance. Further details of the alignment method are given in Asparouhov and Muthén (2013).

## 2.4 Critique of the Alignment Method

The assumption of the alignment method is that a majority of the parameters are invariant and a minority of the parameters are non-invariant. In some applications there may not be a clear invariance pattern of this kind to be found. For example, in achievement studies of civic education in different countries, country-specific curricula and history may cause non-invariance among most or all items and countries. A difficulty of the method is how to be aware that such a situation is at hand.

## 2.5 Alignment Analysis of the 26-Country Example

This section continues the analysis of the tradition-conformity items for $49,894$ subjects in 26 European countries that was introduced in Section 2.2. It is shown how the alignment method resolves the problem of comparing factor means found with the traditional multiple-group factor analysis under scalar invariance. Maximum-likelihood estimation was used for the initial configural model as discussed in Asparouhov and Muthén (2013).

Table 3 shows the (non-) invariance results for the measurement intercepts and factor loadings. The countries that are deemed to have a significantly non-invariant measurement parameter are shown as bolded within parentheses. As seen in Table 3, most of the items show a large degree of measurement non-invariance for the measurement intercepts and, to a lesser extent, the loadings. The large degree of non-invariance is in line with the findings of the traditional approach using the scalar model. However, Table 3 also shows that item IPBHPRP has no significant measurement non-invariance and this item is therefore particularly useful for comparing these countries on the factor.

Table 4 shows each item's intercept and loading contribution to the optimized simplicity function. These values add up to the total optimized simplicity function value. In line with Table 3, it is seen that the item IPBHPRP contributes by far the least, while the items IPMODST, IMPTRAD, and IPFRULE contribute roughly the same. This implies that IPMODST, IMPTRAD, and IPFRULE have a similar degree of measurement non-invariance. The $R^2$ column of Table 4 also indicates that the IPBHPRP item is the most invariant in that essentially all the variation across groups in the configural model intercepts and loadings for this item is explained by variation in the factor mean and factor variance across groups. The variance column of Table 4 again shows the variation in the alignment parameters across groups and again indicates invariance for item IPBHPRP. Taken together, these three columns give an indication of the plausibility of the assumption underlying the alignment method mentioned in Section 2.4, namely that an invariance pattern can be found. In this example, the inclusion of the IPBHPRP item makes this assumption plausible and ensures good performance of the alignment method. This is also supported by Monte Carlo simulation studies discussed in Section 4.2. Note, however, that simulation studies show that to obtain good alignment performance, it is not necessary that any item has invariant measurement parameters across all groups.

Table 5 shows the factor means as estimated by the alignment method. For convenience in the presentation, the factor means are ordered from high to low and groups that have factor means significantly different on the 5% level are shown. Figure 1 compares the estimated factor means using the alignment method with the factor means of the scalar invariance model (without relaxing any invariance restrictions). The correlation between the two sets is 0.943, but despite this seemingly high correlation there are several discrepancies. Recalling the reversed scale, the two methods agree that Sweden (country 23) has the lowest level of tradition-conformity and Cyprus (country 4) the highest level. The alignment method, however, finds that Portugal (country 21) has a significantly different mean from the Netherlands (country 18), whereas the scalar method finds essentially no difference between these countries. Other discrepancies between the two methods are found for France compared to Switzerland and for Norway compared to Russia.

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Figure 1 about here.]

## 3 Random Mode Analysis

Turning to random mode analysis, the question is what two-level factor analysis and two-level IRT can tell us about measurement invariance and how it can be used to compare groups with respect to group-specific factor values. As a refresher on two-level factor analysis and IRT it is useful to distinguish between three major types of models:

1. Random intercepts, non-random (invariant) loadings: Different within- and between-level factor loadings

2. Measurement invariance (non-random intercepts and loadings): Same within- and between-level factor loadings and zero between-level residual variances

3. Random intercepts and random loadings

## 3.1 Model Type 1: Random Intercepts, Non-Random Loadings, Different Within- and Between-Level Factor Loadings

As a background for model type 1, recall random effect ANOVA for individual $i$ in cluster $j$,

$$y_{ij} = \nu + y_{B_j} + y_{W_{ij}}, \tag{7}$$

where $y_{B_j}$ and $y_{W_{ij}}$ are uncorrelated. For a given item, two-level factor analysis generalizes this to

$$y_{ij} = \nu + \lambda_B \; f_{B_j} + \epsilon_{B_j} \; + \; \lambda_W \; f_{W_{ij}} + \epsilon_{W_{ij}} \tag{8}$$

with covariance structure $V(y_{ij}) = \Sigma_B + \Sigma_W$, where

$$\Sigma_B = \Lambda_B \; \Psi_B \; \Lambda'_B + \Theta_B,$$

$$\Sigma_W = \Lambda_W \; \Psi_W \; \Lambda'_W + \Theta_W.$$

It is clear that (8) can be equivalently expressed as a random intercept model:

$$\mathit{Level}\ 1: \; y_{ij} = \nu_j + \lambda_W \; f_{W_{ij}} + \epsilon_{W_{ij}}, \tag{9}$$

$$\mathit{Level}\ 2: \; \nu_j = \nu + \lambda_B \; f_{B_j} + \epsilon_{B_j}. \tag{10}$$

The variation in the random intercept $\nu_j$ is expressed in terms of variation in a between-level factor $f_{B_j}$ and a between-level residual $\epsilon_{B_j}$.

Figure 2 shows the model in diagram form. On the within level there are two factors (f1w and f2w), shown as circles, whereas on the between level there is one factor (fb). In an educational testing context with students clustered within schools, the within factors may correspond to verbal and mathematics achievement, while the between factor may correspond to school excellence. This illustrates that the factor loadings can be different on the two levels. The filled circles on the within level indicate that the intercepts of the factor indicators $y_1 - y_6$ are random effects. These random effects are latent continuous variables on the between level, where the figure shows a standard linear one-factor model albeit with latent instead of observed factor indicators. The short arrows show the residuals, labelled $\epsilon_{B_j}$ on the between level in (10). The idea of possibly different factor structures on the two levels is in line with the two-level factor analysis tradition starting with Cronbach (1976) and Härnqvist (1978) and carried further in Goldstein and McDonald (1988), McDonald and Goldstein (1989), Longford and Muthén (1992), Härnqvist et al. (1994), and Muthén (1994).

[Figure 2 about here.]

## 3.2 Model Type 2: Measurement Invariance, Same Within- and Between-Level Factor Loadings

Moving to model type 2, it is instructive to see the connections between random intercept two-level factor analysis, conventional two-level IRT, and measurement invariance. Conventional two-level IRT (see, e.g., Fox & Glas, 2001; Fox, 2005,

2010) considers the special case of $\lambda_W = \lambda_B = \lambda$ and $V(\epsilon_{B_j}) = 0$, so that (9) and (10) become

$$Level\ 1: \ y_{ij} = \nu_j + \lambda\ f_{W_{ij}} + \epsilon_{ij}, \tag{11}$$

$$Level\ 2: \ \nu_j = \nu + \lambda\ f_{B_j} + 0, \tag{12}$$

so that $\nu_j$ varies only as a function of $f_B$, that is, the intercept of the outcome is determined by the cluster factor value. In conventional two-level IRT contexts this is typically re-written as

$$y_{ij} = \nu + \lambda\ f_{ij} + \epsilon_{ij}, \tag{13}$$

$$f_{ij} = f_{B_j} + f_{W_{ij}}, \tag{14}$$

which shows that the model assumes invariance of the intercept $\nu$ and the loading $\lambda$ across clusters and that the same $\lambda$ multiples both $f_B$ and $f_W$. This conventional two-level IRT model has the covariance structure

$$\Sigma_B = \Lambda\ \Psi_B\ \Lambda', \tag{15}$$

$$\Sigma_W = \Lambda\ \Psi_W\ \Lambda' + \Theta_W. \tag{16}$$

so that $\Theta_B = 0$.

Testing of measurement invariance with random intercept two-level factor analysis is considered in Jak et al. (2013a,b). This involves testing the general model of (9) and (10) against the model with $\lambda_W = \lambda_B = \lambda$ and $V(\epsilon_{B_j}) = 0$ using likelihood-ratio $\chi^2$. Modification indices (Lagrange multipliers; Sörbom, 1989) are used to reveal model misfit due to non-zero $V(\epsilon_{B_j})$, pointing to factor indicators that have significant between-level residual variance and therefore non-invariant intercepts. This approach is illustrated in Section 3.5.1.

## 3.3 Model Type 3: Random-Intercepts, Random-Loadings

Model type 3 lets both intercepts and factor loadings vary across between-level units. This has been discussed in De Jong, Steenkamp and Fox (2007), De Boeck (2008), De Jong and Steenkamp (2010), Frederickx et al. (2010), Fox (2010), Fox and Verhagen (2011), Verhagen and Fox (2013), Verhagen (2013), and Asparouhov and Muthén (2012). Bayesian estimation is needed because random loadings with maximum-likelihood estimation gives rise to numerical integration with many dimensions which is computationally intractable. The proposed analysis implies a new conceptualization of measurement invariance where each measurement parameter varies across groups/clusters, but groups/clusters have a common mean and variance. As with the alignment method, only approximate measurement invariance is presumed. Different groups/clusters have different random deviations from the common mean. E.g., for a factor loading,

$$\lambda_j \sim N(\mu_\lambda, \sigma^2_\lambda). \tag{17}$$

This is illustrated in Figure 3, where the overall factor loading $\mu_\lambda = 1$, but there is a small variance $\sigma^2_\lambda = 0.01$ across groups/clusters. Nevertheless, 95% of the groups/clusters have a factor loading between 0.8 and 1.2.

[Figure 3 about here.]

Fox (2010) considered this approach in the context of IRT with binary indicators, where the random-intercepts, random-loadings model can be expressed for an outcome $y_{ij}$ for individual $i$ in group/cluster $j$ as

$$P(y_{ij} = 1) = \Phi(a_j\, \theta_{ij} + b_j), \tag{18}$$

$$a_j = a + \epsilon_{a_j}, \tag{19}$$

$$b_j = b + \epsilon_{b_j}. \tag{20}$$

where $\Phi$ is the standard normal distribution function, $\theta_{ij}$ is an ability factor, $\epsilon_{a_j} \sim N(0, \sigma_a)$, and $\epsilon_{b_j} \sim N(0, \sigma_b)$. This is a 2-parameter probit IRT model where both discrimination ($a$) and difficulty ($b$) vary across groups/clusters. The $\theta$ ability factor is decomposed into between- and within-group/cluster components as

$$\theta_{ij} = \theta_{B_j} + \theta_{W_{ij}}. \tag{21}$$

The mean and variance of the ability vary across the groups/clusters. The model preserves a common measurement scale while accommodating measurement non-invariance. The ability for each group/cluster can be obtained by factor score estimation.

As discussed by Fox (2010), special modeling considerations are needed to separately identify cluster/varying factor means and variances in the presence of random intercepts and loadings. Asparouhov and Muthén (2012) proposed a convenient way to accomplish this. This is described here for continuous factor indicators but carries over directly to binary indicators. For a certain continuous factor indicator $y_{ij}$, the model is specified as

$$y_{ij} = \nu_j + \lambda_j \, f_{W_{ij}} + \epsilon_{ij}, \tag{22}$$

$$\nu_j = \nu + \lambda \, f_{B_j} + \epsilon_{\nu_j}, \tag{23}$$

$$\lambda_j = \lambda + \lambda \, f_{\psi_j} + \epsilon_{\lambda_j}, \tag{24}$$

where $f_{W_{ij}} \sim N(0,1)$, $\epsilon_{ij} \sim N(0,\theta)$, $\epsilon_{\nu_j} \sim N(0,\sigma^2_\nu)$, $\epsilon_{\lambda,j} \sim N(0,\sigma^2_\lambda)$, $f_B \sim N(0,\psi)$, and $f_{\psi_j} \sim N(0,\sigma^2)$. The variation in intercepts is captured by $\sigma^2_\nu$, the variation in the loadings is captured by $\sigma^2_\lambda$, the variation in factor means is captured by $\psi$, and the variation in the factor variance is captured by $\sigma^2$. Cluster-specific factor values corresponding to factor means, can be obtained as factor score means for the between-level factor $f_{B_j}$ using draws of Bayesian plausible values.

The previous two types of two-level factor analysis and IRT models are easily related to the model in (22) - (24). Model type 2 of (13), (14) is obtained when setting $\sigma^2 = 0$, $\sigma^2_\lambda = 0$, and $\sigma^2_\nu = 0$, that is, requiring no factor loading variation so that $\lambda_j = \lambda$ and requiring no intercept variation that is not explained by $f_B$, so that $\nu_j = \nu + \lambda \, f_{B_j}$. Model type 1 of (9), (10) is obtained when setting $\sigma^2 = 0$, $\sigma^2_\lambda = 0$ and in addition letting $\lambda_j = \lambda_W$, that is, requiring no factor loading variation but allowing different factor loadings on the two levels. It may be noted that only model type 3 allows for cluster variation in the factor variances by letting $\sigma^2$ be freely estimated.

## 3.4 Critique of the Assumptions Behind Two-level Analysis

The random mode approach of two-level analysis builds on the assumptions of randomly sampled groups/clusters and normally distributed random measurement parameters. In some cases, these random mode assumptions are not well supported. The group of countries studied may not represent a random sample of a specific population and may in fact be a heterogeneous collection of different country types. Bou and Satorra (2010) criticize the random mode approach in favor of a fixed-mode, multiple-group approach. They argue from a substantive point of view in terms of comparing countries that it is not likely that the set of countries can be considered as random draws from a population. Non-normality of the distribution of a measurement parameter may be violated due to a set of outlying countries for which the survey question has quite different meaning. From this point of view, deviations from a common mean are not likely to follow a simple distribution such as the normal. For example, consider a situation such as shown in Figure 4. The figure can be seen as showing a set of measurement intercepts for a factor indicator, where a majority of the groups/clusters have a small intercept with some variation around it and a minority of the groups/clusters have a much larger intercept with some variation around it. In this way, there is a mixture of two unobserved subpopulations and treating this as a single population random intercept situation gives distorted results with an estimated mean that is incorrect for both subpopulations and a variance estimate that is inflated. The mixture case is considered in De Jong and Steenkamp (2010), but results in a very complex analysis.

[Figure 4 about here.]

## 3.5 Two-Level Analyses of the 26-Country Example

In this section the three types of two-level models discussed above are applied to the 26-country data. One factor is specified for both levels.

### 3.5.1 Random Intercept Analysis

Three random intercept models are fitted, following the suggestions of Jak et al. (2013a). Model 1 lets factor loadings be different on the two levels and lets the residual variances on the between level be free ($\lambda_B \neq \lambda_W$, $\theta_B$ free). Model 2 holds the factor loadings equal across levels, while still letting the between-level residual variances be free ($\lambda_B = \lambda_W$, $\theta_B$ free). Model 3 holds the factor loadings equal across levels and fixes the residual variances on the between level to zero ($\lambda_B = \lambda_W$, $\theta_B = 0$). The models are estimated by maximum-likelihood. The resulting fit statistics are shown in Table 6.

[Table 6 about here.]

Model 1 fits rather well given the large sample size of 49,894 subjects. The $\chi^2$ p-value is 0.000, but good fit is indicated by $RMSEA = 0.011$ and $CFI = 0.999$. A test of Model 2 against Model 1 leads to a $\chi^2$ test of 3.9 with 3 degrees of freedom so that equality of factor loadings across levels cannot be rejected. Testing Model 3 against Model 1, however, rejects zero between-level residual variances with a $\chi^2$ of 6,703 with 7 degrees of freedom.

The influence on model misfit for Model 3 due to non-zero residual variances on the between level is shown in Table 7. In addition to modification indices, the actual $\chi^2$ improvements (the drop in $\chi^2$)) when freeing the residual variances one at a time are shown. For these parameters the modification index values do not seem to give a good approximation of the actual model fit improvement, although the conclusions about which indicators are most in need of free residual variances are the same as for the actual $\chi^2$ improvement. The two factor indicators IPMODST and IPFRULE show a much stronger need for free residual variances than the other two indicators and are therefore exhibiting much stronger non-invariance of the measurement intercepts.

[Table 7 about here.]

### 3.5.2 Random Intercept and Random Loading Analysis

Bayesian analysis was applied to the random-intercept, random-loading model of (22) - (24). The intercept and loading variance estimates are shown in Table 8. The two factor indicators IPMODST and IPFRULE show larger intercept variances than the other two indicators. This is in line with the random intercept model, that is, allowing loadings to be random as well does not change the picture. For the loadings, the IPMODST item has the largest variance. The variance estimates are in line with those of the alignment method shown in Table 4.

Significant variation in factor means as well as factor variances is also found (not shown). The ordering of the countries based on factor means can be compared between the factor means of the alignment method and the factor score means of the Bayesian plausible values for the between-level factor $f_{B_j}$. For this example the correlation between the two sets is 0.987. Figure 5 shows the relationship. Some of the differences between the two approaches in the ordering of the countries are similar to those of Figure 1, with the two-level approach taking the role of the scalar model approach. The relationship between the scalar model approach and two-level approach is, however, not perfect, but the correlation is 0.980.

[Table 8 about here.]

[Figure 5 about here.]

# 4 Simulation Studies Comparing Fixed versus Random Mode Analysis

This section compares the alignment method and the random-intercept, random-loading method using simulated data. In the Monte Carlo studies it is useful to have a simple gauge of the quality of the estimation. An important goal is to correctly estimate the ordering of the groups with respect to the factor means/factor scores. In Monte Carlo simulations, an important statistic is therefore the correlation between the true factor means and the estimated factor means. As a first step, the relationship between this correlation and the error in the estimation of the factor mean is derived. This is followed by several simulation studies using the alignment approach and using the two-level approach. The results for the model with full measurement invariance are also shown as a comparison.

## 4.1 Correlation and Standard Error for Group-Specific Factor Means

Consider the alignment method, that is, a fixed-mode, multiple-group analysis and the goal of correctly estimating the ordering of the groups with respect to the factor means. The correlation between the true factor means and the estimated factor means can be computed for each replication and averaged over the replications. It can also be computed from the correlation between the true factor means and the average estimated factor means, where the average is over the replications. The latter value is largely independent of the sample size and therefore shows the potential of the alignment method to do a good job for the extent of non-invariance studied, whereas the former value shows the performance of the alignment method for the extent of non-invariance studied as well as the sample size studied.

Although the size of a correlation is easy to understand, it is also useful to consider the standard error of the factor mean estimate. Appendix 1 derives the relationship between the standard error and the correlation. Table 9 shows examples of correlation values and the corresponding limit of the estimation error for 95% of the groups, where the error is given in a standardized metric. It is seen that a rather high correlation is required to keep the absolute error small. For example, to achieve a relatively small absolute error limit of 0.277 for 95% of the groups, a correlation of 0.99 is required. A correlation of 0.95 gives a large error of 0.620. Figure 1 and Figure 5 exemplify the difference in ordering of the countries for a correlation of 0.943 and 0.987, respectively. A correlation of at least 0.99 has also shown to be a good requirement for low bias in estimating each group's factor mean.

[Table 9 about here.]

Using the factor mean correlation as a gauge of quality is also applicable to the two-level, random-intercept, random-slope method. In this case the factor means are replaced by factor score means from Bayesian plausible value draws for each group/cluster. Because of the random-mode approach, the true values vary across replications.

## 4.2 Simulations Based on the 26-Country Data

As discussed in Asparouhov and Muthén (2013), the quality of estimation can be studied based on the features of a particular real data set. The estimated parameter values for the data set are used to generate data for the simulation study. To study the alignment method, the real data are analyzed by the alignment method, data are generated in many replications from those estimates, and analyzed using the alignment method. The two-level method is studied analogously by analyzing the real data by the random-intercept, random-loading two-level method, generating data from those estimates over many replications, and analyzing using the random-intercept, random-loading two-level method. The real data used here is the 26-country data.

For the alignment method the correlation between the true factor means and the estimated factor means computed for each replication and averaged over the replications is 0.990 for the factor means. According to Table 9, the high factor mean correlation corresponds to a relatively small absolute error of 0.277. The correlation between the true factor means and the average estimated factor means, where the average is over the replications, is 0.999 for the factor means. The latter value approximates the quality of estimation for a very large sample, whereas the former value is sample-size specific. These values indicate very good performance of the alignment method. Analysis using the scalar model performs considerably less well with correlations of 0.940 and 0.943, respectively for the replication-specific and average computations.

Using the analogous approach when applying the random-intercept, random-loading two-level method, the correlation between the true factor scores and the estimated factor scores computed for each replication and averaged over the replications is only 0.950 corresponding to an absolute error of 0.620. The correlation using averages is not applicable in this case given that average scores are zero. The poor performance of the two-level method is most likely due to using only four factor indicators. The corresponding correlations when adding similar indicators to use 8, 12, 16, and 20 indicators are 0.977, 0.982, 0.985, and 0.988, respectively. This suggests that for indicators of the quality seen for the 26-country data, about 20 indicators are needed for good recovery of the factor scores.

Still generating the data according to the random-intercept, random-loading two-level method, but analyzing using the two-level model type 2, where both the intercepts and loadings are invariant (not random), a correlation of only 0.874 is obtained. This is akin to using the scalar model in the fixed-mode case. Applying model type 1, a correlation of 0.872 is obtained. These two results show the importance of using random measurement parameters. Note, however, that in this case using a model with random intercepts and non-random loadings that are equal across the two levels obtains a correlation of 0.951, that is, the same as when also letting the loadings be random. This means that this simpler model can be estimated by maximum likelihood in line with what was used for model type 1, leading to quicker computations.

In the above studies, data were analyzed by the same model that generated the data. It is useful to also study the methods when applied to data generated by a different model. Appendix 2 shows simulations where the data generation is based on multiple-group data suitable for the alignment method and a comparison is made between the results of analyzing by the alignment method versus analyzing by the two-level method. The analogous case of data generation based on a random-intercept, random-slope two-level model is also studied. In these comparisons between methods, the same data are used and it is therefore possible to compare the methods with respect to both correlation and a mean squared error (MSE) that describes in one statistic both the bias and variability of the estimates. The reader is referred to Appendix 2 for the results.

# 5 Conclusions

This paper discusses two new methodologies for studying invariance across many groups. Both are based on the idea of approximate measurement invariance and perform well under a large set of conditions. The availability of the two new methods should be a welcome contribution to the study of invariance across many groups. They represent a big step forward in the methodology and they are not difficult to use. The Mplus scripts for all analyses in this paper are available at www.statmodel.com.

The differences between the two methods discussed in this paper is in how the group-specific factor mean and variance parameter are obtained and what assumptions are added to the information in the data. The assumption of the alignment method is that a majority of the parameters are invariant and a minority of the parameters are non-invariant. The assumption of the random intercept and loading method is that all parameters are approximately the same, i.e., no parameters are exactly the same across the groups, but rather each parameter has random variation that makes it slightly different from the corresponding parameter in the rest of the groups. Thus when deciding which model to use for a practical application one should focus on deciding which of the above two assumptions is more appropriate for the particular application. The alignment method focuses on identifying the reason for non-invariance and produces a model that has clear interpretation in terms of invariance and non-invariance. The random intercept and slope method is not as detailed or focused on the actual parameter variations across the groups but instead looks at the entire population as a whole. In addition to these general considerations, there are several practical issues in deciding between the two methods as described below.

## 5.1 Practical Issues in Choosing Between Fixed and Random Approaches

There are several practical reasons for preferring either the alignment or the random-intercept, random-loading two-level approach. The pros and cons of the two methods are listed in Table 10. A plus sign denotes that the method has an advantage over the other method, and a minus sign denotes that it has a disadvantage.

[Table 10 about here.]

### 5.1.1 Number of Factor Indicators

As seen in the simulations, the two-level method needs a sufficiently large number of factor indicators to perform well. This is due to the need to estimate factor scores and is in this way analogous to scoring issues in IRT. Many survey instruments represent factors with only a few indicators in order to cover many factors without making the survey instrument too long. For achievement studies, however, the number of indicators is much larger and the two-level method would work well. The alignment method can work very well with a small number of indicators as seen in the simulations. For one factor, three indicators is sufficient in principle.

### 5.1.2 Number of Groups

If the number of groups is small the random intercept, random-loading model may not perform well and perhaps not even converge. Typically, at least 30 groups is recommended in the multilevel literature. If the number of groups is large the alignment method may have slow convergence and with more than say 100 groups computations are prohibitive due to the many parameters of the configural model. In many cases, however, both methods are possible and for any particular example it may be useful to compare the results to better understand the data.

### 5.1.3 Group Size

With small group sizes, the two-level method has an advantage over the alignment method. In contrast to the alignment method, the two-level method does not estimate parameters specific to each group. The two-level method borrows information from all groups in estimating the parameters which are common to all the groups, while allowing for random variation across groups. The group size requirement for the alignment method varies depending on how clear the invariance pattern is. For both alignment and two-level analysis, a notion of the actual group size needed in a specific example can be obtained by Monte Carlo simulation. Asparouhov and Muthén (2013) did Monte Carlo studies of the 26-country data and found good alignment results for group sizes as low as 100, but in other situations group sizes of several thousand observations may be needed.

### 5.1.4 Invariance Pattern

The type of measurement non-invariance pattern is an important factor in choosing between the two methods. As mentioned in Section 2.4, the assumption of the alignment method that a majority of the parameters should be invariant and a minority of the parameters should be non-invariant may not be at hand in all applications. In such situations, the two-level method is preferable.

### 5.1.5 Information About Groups Contributing to Non-Invariance

Measurement invariance studies benefit from information on which groups contribute to non-invariance. This information is readily obtained by the alignment method. The two-level method, however, has currently no such counterpart given that it only estimates the degree of measurement variance across groups.

### 5.1.6 Normality of Measurement Parameter Distributions

Normality of the distribution of measurement parameters across groups is assumed by the two-level method. In contrast, the alignment method allows any kind of measurement parameter distribution and is in this sense non-parametric.

### 5.1.7 Explanatory Variables for Non-Invariance

Group-level variables are sometimes hypothesized to influence measurement parameters and therefore explain part of the measurement non-invariance. Such variables can be incorporated in the two-level analysis, but currently this option is not available with the alignment method.

### 5.1.8 Complex Survey Data

Comparisons of many groups often arise in surveys of many countries where a complex survey design is used. For instance, with PISA, TIMSS and other surveys of school children, sampling of schools is carried out using probability proportional to size (PPS), giving rise to the need to use sampling weights. Complex survey features of weights, stratification, and clustering can be taken into account in the maximum-likelihood estimation of the alignment method. To date, however, Bayesian analysis can not accommodate complex survey features.

### 5.1.9 Computational Speed

Computational speed is a final important practical consideration. In most cases, the maximum-likelihood estimation with the alignment method gives much quicker computations than the Bayesian analysis with the two-level method. This is due to the simple, two-step procedure of alignment where a configural model is estimated first, followed by a computationally simple optimization of the alignment fit function. In contrast, the Bayesian analysis needed for the random-intercept, random-slope two-level model involves a complex model with many random effects.

# 6 Appendix 1: Correlation and Error of Estimation

This Appendix derives the relationship between the correlation between true and estimated factor means on the one hand and the standard error of the estimated factor mean on the other hand. Suppose that $f_j$ are the group-specific factor means standardized to mean zero and variance one. Suppose that $\hat{f}_j$ are the corresponding estimates and suppose that $\rho = Cor(f_j, \hat{f}_j)$. Then

$$f_j = \rho \hat{f}_j + \varepsilon$$

where $\varepsilon$ is a residual with mean zero and variance $1 - \rho^2$, uncorrelated with $\hat{f}_j$. Therefore

$$f_j - \hat{f}_j = (1 - \rho)\hat{f}_j + \varepsilon$$

so that the residual $f_j - \hat{f}_j$ has a variance of $(1 - \rho)^2 + 1 - \rho^2$ and a standard deviation

$$\sqrt{(1 - \rho)^2 + 1 - \rho^2}.$$

Thus the error in the estimate is limited by absolute value to

$$1.96\sqrt{(1 - \rho)^2 + 1 - \rho^2}$$

in 95% of the groups.

# 7 Appendix 2: Further Simulations

This appendix describes the means squared error for factor means and presents two simulation studies. In the first study, data are generated by the alignment model and in the second study, data are generated by the random-intercept, random-loading two-level model.

## 7.1 Mean Squared Error

From a practical perspective one of the most important results of the multiple-group factor analysis is the group specific factor mean parameter $\alpha_j$. Therefore the two methods are compared based on the following mean squared error (MSE) measure:

$$MSE = \sqrt{\sum_{i=1}^{G}(\hat{\alpha}_j - \alpha_j)^2/G}$$

where $\alpha_j$ is the true factor mean $\hat{\alpha}_j$ is the estimated factor mean and $G$ is the number of groups. The identification condition for the alignment method is that the factor mean in the first group is zero and the factor variance in the first group is 1. The data are generated using these conditions. The identification conditions for the factor intercept and loading method are quite different so these estimates need to be standardized before they can be compared to the alignment estimates using the MSE measure. Suppose the $\hat{\alpha}_{i,0}$ are the group specific factor mean estimates obtained with the random intercepts and loadings model. To get these estimates standardized in the same global metric as the alignment and the generation metric one can reparameterize the factor mean estimates as follows

$$\hat{\alpha}_j = \frac{\hat{\alpha}_{j,0} - \hat{\alpha}_{1,0}}{\sqrt{\hat{\psi}_{1,0}}}$$

where $\hat{\psi}_{1,0}$ is the estimated factor variance in the first group obtained with the random intercepts and loadings method. Now $\hat{\alpha}_j$ estimates are in the same metric as the alignment estimates because mean and the variance of the first group will be 0 and 1 respectively.

## 7.2 Two-Level Random using Multiple-Group Model Data

Data are generated by a one-factor model for 30 groups using three indicator variables. Each group contains 1000 observations. For a certain item, the model is given by

$$y_{ij} = \nu_j + \lambda_j \ f_{ij} + \varepsilon_{ij}$$

for individual $i$ in group $j$. The factor $f_{ij}$ has mean $\alpha_j$ and variance $\psi_j$ and the residual $\varepsilon_{ij}$ has mean 0 and variance $\theta_j$.

The generation parameters are set as follows, reflecting the situation in Figure 4 with some outlying groups with respect to intercept values. The residual variances $\theta_j$ are set to 1 for every indicator for every group. The indicator intercepts $\nu_j$ are 0 and the loadings $\lambda_j$ are 1 except in the non-invariant cases listed below. The parameters are set in groups of 10, that is, the parameters in the groups 1,...,10 are the same as the parameters in groups 11,...,20 and also in groups 21,...,30. The non-invariant parameters in the first 10 groups are as follows. The first intercept is non-invariant in groups 3, 6 and 9. The second intercept is non-invariant in groups 4, 7 and 10. The third intercept is non-invariant in groups 2, 5 and 8. All non invariant intercepts are set to 1. The first loading is non-invariant in groups 4, 7 and 10. The second loading is non-invariant in groups 2, 5 and 8. The third loading is non-invariant in groups 3, 6 and 9. All non invariant loadings are set to 1.5. All factor variances $\psi_j$ are set to 1 and the factor means $\alpha_j$ are set to 0 except for $\alpha_4 = -1$, $\alpha_5 = 1$ and $\alpha_{10} = 2$.

Using the above simulated example the MSE for the alignment method is 0.055 and for the random-intercept, random-loading two-level method it is 0.229. The correlation between the true and estimated means is 0.998 for the alignment method and 0.985 for the random-intercept, random-loading method. The correlation between the estimates of the two methods is 0.989. This illustrates the fact that the alignment method performs better in terms of recovering the model parameters more accurately when the data are generated by a multiple-group model. The random intercepts and loadings method is essentially driven by minimizing the variability of the group-specific parameter estimates across the groups, which is a very different goal than finding the simplest and most interpretable non-invariance patterns, which is the principle that the alignment methods is based on.

## 7.3 Alignment using Two-Level Model Data

In the above simulation, the data generation is favoring the alignment method. Data were generated according to a model where most parameters are invariant and few are non-invariant. This kind of non-invariance pattern is what the alignment method is searching for. In this section the two methods are compared based on data generation that instead favors the two-level, random-intercept, random-loading model.

Data are generated according to a random-intercept, random-loading single-factor model with three indicators using the same parameter estimates that were obtained in the random-intercept, random-loading model in the simulation of the previous section. In this data generation the factor analysis parameters vary randomly across groups, i.e., all the parameters are slightly different and no parameter is invariant. These data are analyzed with the alignment method and the random-intercept, random-loading method. The factor means are now standardized so that they add up to 0 and have a standard deviation of 1. The true factor means are standardized the same way and the MSE measure computed for both methods. The MSE for the alignment method is now 0.514 and the MSE for the random-intercept, random-loading method is 0.329. The correlation between the true and estimated means is 0.863 for the alignment method and 0.944 for the random-intercept, random-slope method. The correlation between the estimates of the two methods is 0.885. Thus with this alternative simulation one could conclude that the random-intercept, random-loading method recovers the parameters estimates better. Thus depending on which way the data are generated a different method can perform better. Note, however, that in this case where the two-level method is applied to the multiple-group data of Section 7.2 and data are generated from the two-level model estimates, any simple invariance pattern that is originally present is distorted and causes poor alignment performance. This shortchanging of the alignment method in this section does not have a counterpart for the two-level method in the previous section because the two-level method can be well fitted to multiple-group data.

# 8 Acknowledgement

We thank Peter Schmidt for providing the 26-country data.

# References

[1] Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. Structural Equation Modeling, 16, 397-438.

[2] Asparouhov, T. and Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Technical Report.
http://statmodel.com/download/Bayes3.pdf

[3] Asparouhov, T. & Muthén, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. http://www.statmodel.com/download/NCME_revision2.pdf

[4] Asparouhov, T. & Muthén (2013). Multiple-group factor analysis alignment. Accepted for publication in Structural Equation Modeling.
http://www.statmodel.com/examples/webnotes/webnote18.pdf

[5] Beierlein, C., Davidov, E., Schmidt, P. & Schwartz, S.H. (2012). Testing the discriminant validity of Schwartz' portrait value questionnaire items - A replication and extension of Knoppen and Saris (2009). Survey Research Methods, 6, 25-36.

[6] Bou, J.C. & Satorra, A. (2010). A multigroup structural equation approach: A demonstration by testing variation of firm profitability across EU samples. Organizational Research Methods, October 2010; vol. 13, 4: pp. 738-766., first published on January 26, 2010.

[7] Cronbach, L.J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Unpublished manuscript, Stanford University, Stanford Evaluation Consortium, School of Education.

[8] De Boeck, P. (2008). Random item IRT models. Psychometrika, 73, 533-559.

[9] De Jong, M.G., Steenkamp, J.-B.E.M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. Journal of Consumer Research, 34, 260-278.

[10] De Jong, M.G., & Steenkamp, J.-B.E.M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large-scale cross-cultural research. Psychometrika, 75, 3-32.

[11] Fox, J.P. (2005). Multilevel IRT using dichotomous and polytomous response data. British Journal of Mathematical and Statistical Psychology, 58, 145-172.

[12] Fox, J.P. (2010). Bayesian item response theory. New York: Springer.

[13] Fox, J.P. & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs. Psychometrika, 66, 269-286.

[14] Fox & Verhagen (2011). Random item effects modeling for cross-national survey data. In E. Davidov & P. Schmidt, and J. Billiet (Eds.), Cross-cultural Analysis: Methods and Applications.

[15] Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: a random item mixture model to detect differential item functioning. Journal of Educational Measurement, 47, 432-457.

[16] Goldstein, H. & McDonald, R.P. (1988). A general model for the analysis of multilevel data. Psychometrika, 53, 455-467.

[17] Härnqvist, K. (1978). Primary mental abilities of collective and individual levels. Journal of Educational Psychology, 70, 706-716.

[18] Härnqvist, K., Gustafsson, J. E., Muthén, B., & Nelson, G. (1994). Hierarchical models of ability at class and individual levels. Intelligence, 18, 165-187.

[19] Jak, S., Oort, F.J., & Dolan, C.V. (2013a). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. Structural Equation Modeling, 20, 265-282.

[20] Jak, S., Oort, F.J., & Dolan, C.V. (2013b). Measurement bias in multilevel data. To appear in Structural Equation Modeling.

[21] Jennrich R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. Psychometrika, 71, 173-191.

[22] Kim, E.S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18, 212-228.

[23] Lee, J., Little, T.D. & Preacher, K.J. (2010). Methodological issues in using structural equation models for testing differential item functioning. In Davidov, E., Schmidt, P., & Billiet, J. (eds.), pp. 55-84. Cross-cultural analysis. Methods and applications. New York: Routledge.

[24] Longford, N. T., & Muthén, B. (1992). Factor analysis for clustered observations. Psychometrika, 57, 581-597.

[25] McDonald, R.P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. British Journal of Mathematical and Statistical Psychology, 42, 215-232.

[26] Millsap, R.E. (2011). Statistical approaches to measurement invariance. New York: Taylor and Francis Group.

[27] Muthén, B. (1994). Multilevel covariance structure analysis. In J. Hox & I. Kreft (eds.), Multilevel Modeling, a special issue of Sociological Methods & Research, 22, 376-398.

[28] Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. Journal of Applied Psychology, 91(6), 12921306.

[29] Sörbom, D. (1989). Model modification. Psychometrika, 54, 371-384.

[30] Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

[31] Verhagen, A.J. (2013). Bayesian item response theory models for measurement invariance. Doctoral dissertation, University of Twente, the Netherlands.

[32] Verhagen, A. J. , & Fox, J.-P (2013). Bayesian tests of measurement invariance. Accepted for publication in The British Journal of Mathematical and Statistical Psychology.

# List of Figures

Figure 1: 26-Country Example: Factor Means for Alignment Method vs Scalar Model

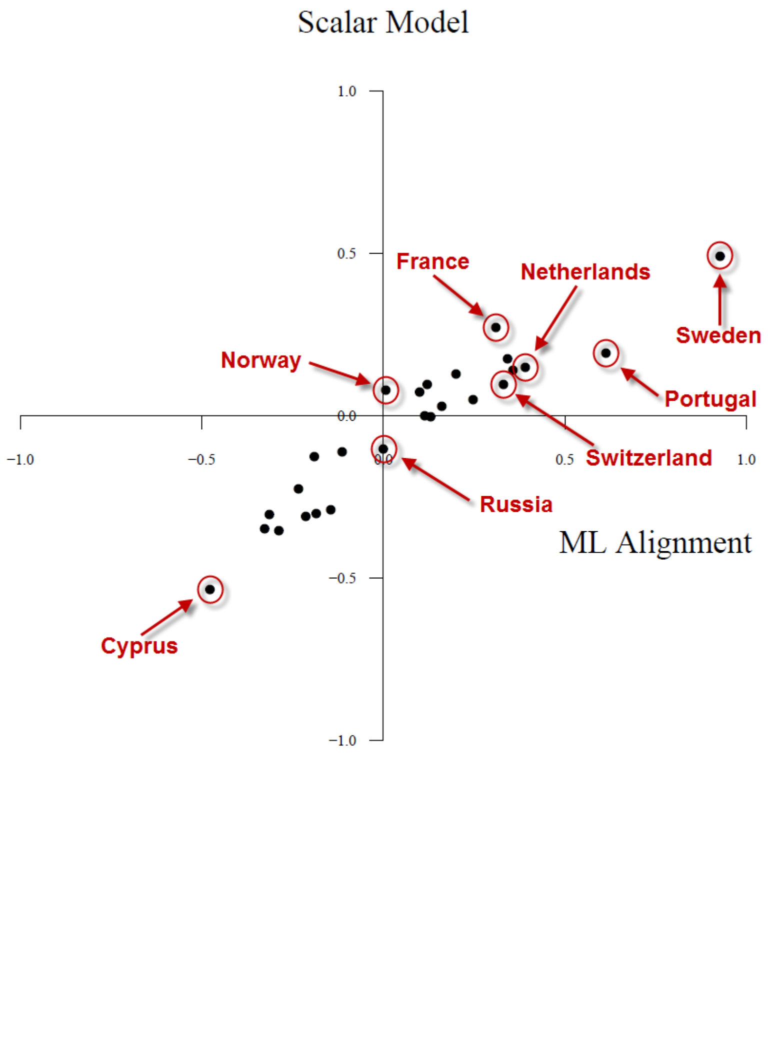


Figure 2: Random Intercept Two-Level Factor Analysis in Figure Form

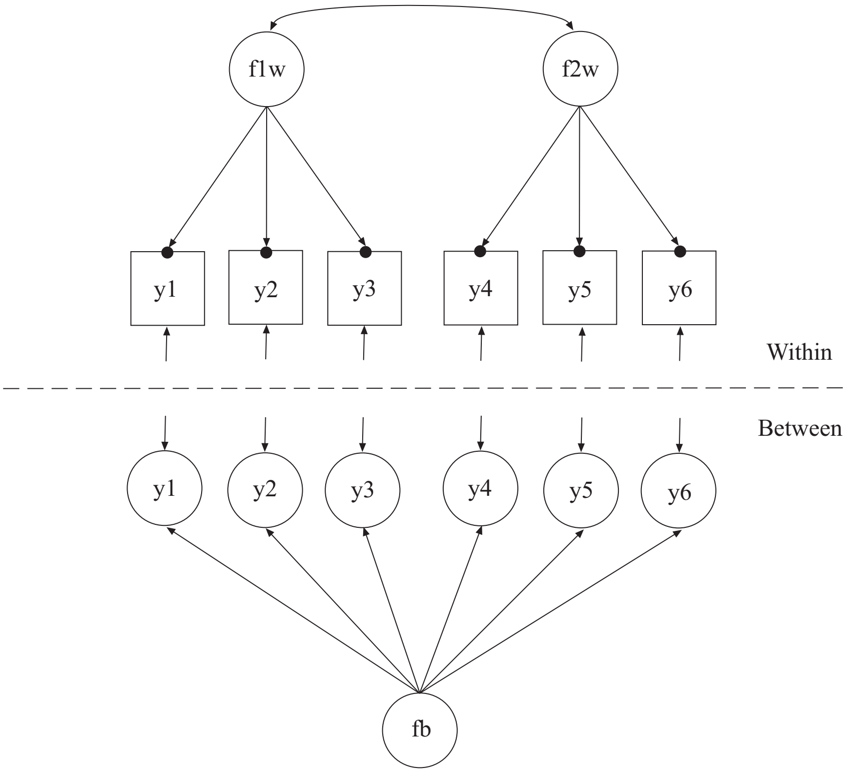


Figure 3: Random Measurement Parameter

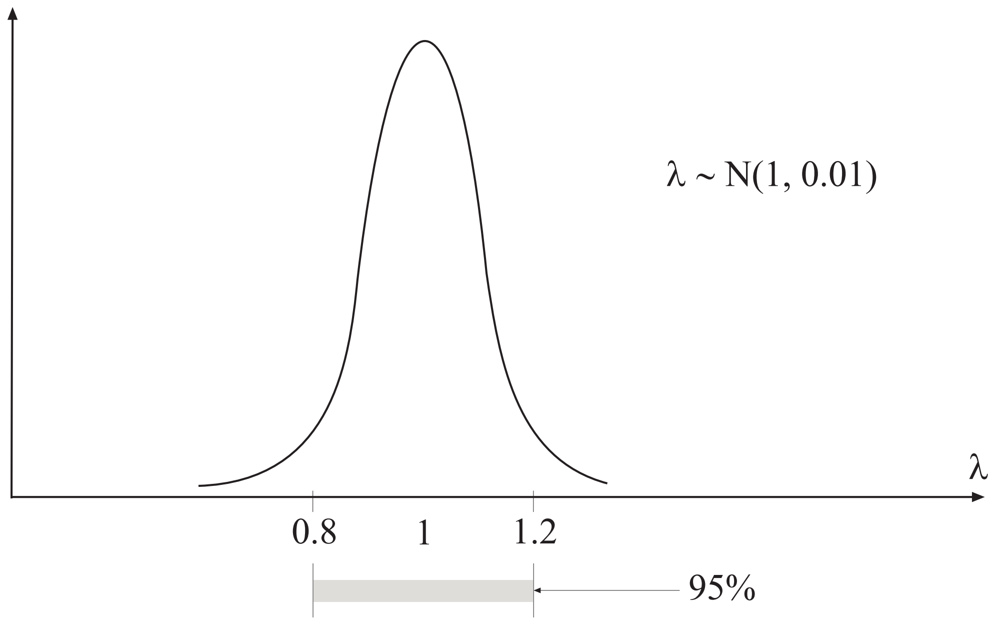


Figure 4: Group-Varying Intercepts

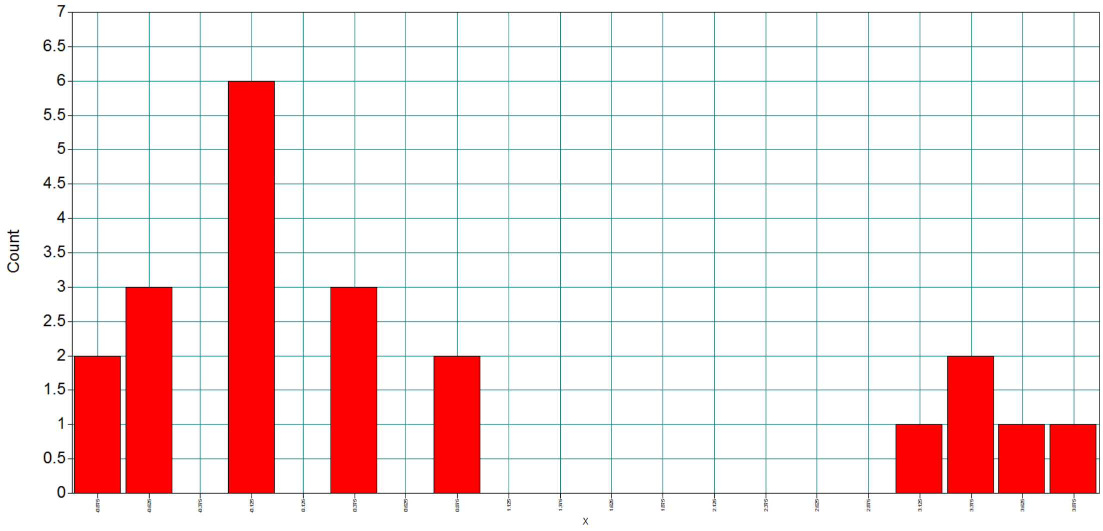
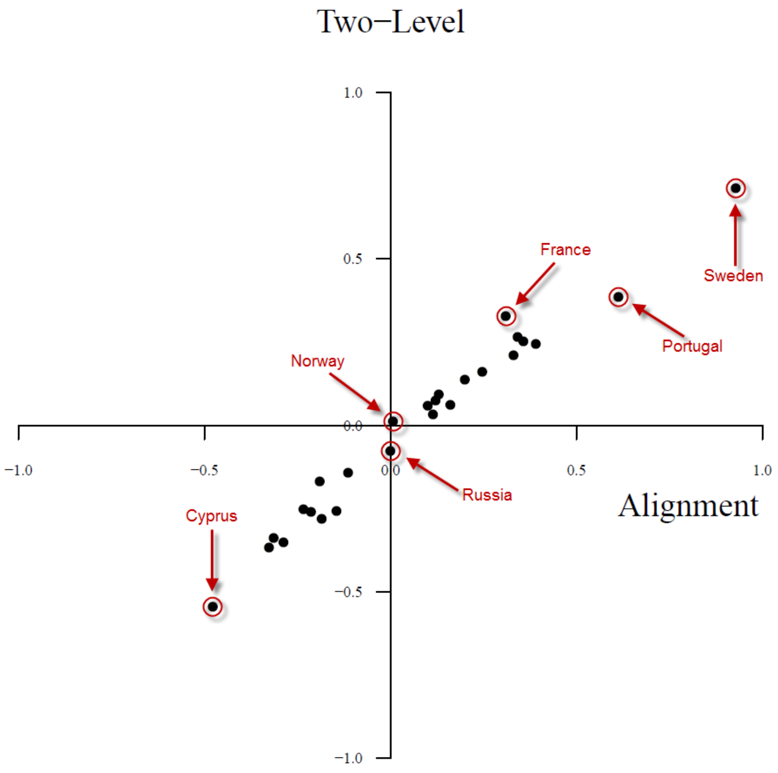


Figure 5: 26-Country Example: Factor Means for Two-Level vs Alignment Analysis



# List of Tables

Table 1: Tradition-Conformity Items from the 26-Country European Social Survey

| | |
|---|---|
| Tradition (TR): | 9. It is important for him to be humble and modest. He tries not to draw attention to himself (ipmodst). |
| | 20. Tradition is important to him. He tries to follow the customs handed down by his religion or family (imptrad). |
| Conformity (CO): | 7. He believes that people should do what they're told. He thinks people should follow rules at all times, even when no one is watching (ipfrule). |
| | 16. It is important for him to always behave properly. He wants to avoid doing anything people would say is wrong (ipbhprp). |

Table 2: 26-Country Example: Model Fit for Multiple-Group Model ($n = 49,894$)

| Model | $\chi^2$ | Df | P-value | RMSEA (Prob $\leq$ .05) | CFI |
|---|---|---|---|---|---|
| Configural | 317 | 52 | 0.000 | 0.052 (.311) | 0.990 |
| Metric | 1002 | 127 | 0.000 | 0.060 (.000) | 0.967 |
| Scalar | 8654 | 202 | 0.000 | 0.148 (.000) | 0.677 |
| Metric vs Config | 685 | 75 | 0.000 | | |
| Scalar vs Config | 8337 | 150 | 0.000 | | |
| Scalar vs Metric | 7652 | 75 | 0.000 | | |

Table 3: 26-Country Example: Approximate Measurement (Non-) Invariance for Intercepts and Loadings over Countries

| | |
|---|---|
| | Intercepts: |
| IPMODST | **(2)** **(3)** **(4)** 5 **(6)** **(7)** **(8)** 9 **(10)** **(11)** **(12)** 13 14 **(15)** 16 17 **(18)** **(21)** **(22)** **(23)** **(24)** 25 26 **(27)** 28 **(30)** |
| IMPTRAD | **(2)** **(3)** **(4)** **(5)** 6 **(7)** 8 9 **(10)** 11 **(12)** 13 **(14)** **(15)** **(16)** **(17)** 18 **(21)** **(22)** **(23)** **(24)** **(25)** 26 27 **(28)** **(30)** |
| IPFRULE | **(2)** 3 **(4)** **(5)** 6 **(7)** **(8)** **(9)** **(10)** 11 **(12)** **(13)** **(14)** **(15)** **(16)** **(17)** 18 **(21)** **(22)** **(23)** 24 **(25)** 26 **(27)** 28 30 |
| IPBHPRP | 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 21 22 23 24 25 26 27 28 30 |
| | Loadings: |
| IPMODST | **(2)** 3 **(4)** 5 6 **(7)** **(8)** 9 **(10)** **(11)** **(12)** **(13)** 14 15 16 17 18 21 22 23 24 25 **(26)** **(27)** 28 30 |
| IMPTRAD | 2 3 4 5 6 7 **(8)** 9 10 11 12 13 14 15 16 17 18 21 22 23 **(24)** 25 **(26)** 27 **(28)** 30 |
| IPFRULE | 2 3 4 5 6 **(7)** 8 9 10 **(11)** **(12)** 13 14 15 16 17 18 21 22 23 24 25 26 27 28 30 |
| IPBHPRP | 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 21 22 23 24 25 26 27 28 30 |

Table 4: 26-Country Example: Alignment Fit Statistics

| | Intercepts | | | Loadings | | |
|---|---|---|---|---|---|---|
| Item | Fit Function Contribution | R-Square | Variance | Fit Function Contribution | R-Square | Variance |
| IPMODST | -229.849 | 0.203 | 0.105 | -158.121 | 0.000 | 0.020 |
| IMPTRAD | -199.831 | 0.566 | 0.058 | -134.042 | 0.000 | 0.014 |
| IPFRULE | -213.806 | 0.198 | 0.103 | -113.305 | 0.263 | 0.008 |
| IPBHPRP | -32.836 | 1.000 | 0.000 | -33.941 | 0.999 | 0.000 |

Table 5: 26-Country Example: Factor Mean Comparisons of Countries

| Ranking | Group | Value | Groups with significantly smaller factor mean |
|---|---|---|---|
| 1 | 26 | 0.928 | 24 21 7 11 4 12 30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 2 | 24 | 0.613 | 21 7 11 4 12 30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 3 | 21 | 0.391 | 30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 4 | 7 | 0.357 | 30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 5 | 11 | 0.342 | 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 6 | 4 | 0.331 | 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 7 | 12 | 0.310 | 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 8 | 30 | 0.247 | 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 9 | 8 | 0.200 | 13 22 25 15 23 28 16 18 10 3 14 27 5 |
| 10 | 6 | 0.161 | 22 25 15 23 28 16 18 10 3 14 27 5 |
| 11 | 17 | 0.130 | 22 25 15 23 28 16 18 10 3 14 27 5 |
| 12 | 9 | 0.121 | 22 25 15 23 28 16 18 10 3 14 27 5 |
| 13 | 2 | 0.114 | 22 25 15 23 28 16 18 10 3 14 27 5 |
| 14 | 13 | 0.100 | 25 15 23 28 16 18 10 3 14 27 5 |
| 15 | 22 | 0.007 | 15 23 28 16 18 10 3 14 27 5 |
| 16 | 25 | 0.000 | 15 23 28 16 18 10 3 14 27 5 |
| 17 | 15 | -0.114 | 18 10 3 14 27 5 |
| 18 | 23 | -0.145 | 10 3 14 27 5 |
| 19 | 28 | -0.185 | 3 14 27 5 |
| 20 | 16 | -0.190 | 3 14 27 5 |
| 21 | 18 | -0.214 | 14 27 5 |
| 22 | 10 | -0.234 | 14 27 5 |
| 23 | 3 | -0.288 | 5 |
| 24 | 14 | -0.314 | 5 |
| 25 | 27 | -0.327 | 5 |
| 26 | 5 | -0.478 | |

Table 6: 26-Country Example: Two-Level Random Intercept Analysis

| Model | Chi-Square | Df | RMSEA | CFI |
|---|---|---|---|---|
| 1. | Different loadings across levels,<br>residual variances free on between | | | |
| | 28.010 | 4 | 0.011 | 0.999 |
| 2. | Equal loadings across levels,<br>residual variances free on between | | | |
| | 31.868 | 7 | 0.008 | 0.999 |
| 3. | Equal loadings across levels,<br>residual variances fixed at zero | | | |
| | 6731.072 | 11 | 0.111 | 0.723 |

Table 7: 26-Country Example: Two-Level Random Intercept Predicted and Actual Chi-Square Improvement for Model 3 Between-Level Residual Variances

| Item | Chi-Square Improvement | |
|---|---|---|
| | Predicted by Modification Index | Actual |
| IPMODST | 201,293 | 3,549 |
| IMPTRAD | 29,726 | 1,201 |
| IPFRULE | 161,347 | 2,924 |
| IPBHPRP | 14,347 | 852 |

Table 8: 26-Country Example: Two-Level Random Intercept and Random Loading Variance Estimates and 95% Credibility Intervals

| Item | Intercept | | Loading | |
|---|---|---|---|---|
| | Estimate | CI | Estimate | CI |
| IPMODST | 0.122 | [0.070, 0.240] | 0.022 | [0.012, 0.042] |
| IMPTRAD | 0.056 | [0.031, 0.116] | 0.010 | [0.005, 0.021] |
| IPFRULE | 0.100 | [0.055, 0.196] | 0.008 | [0.003, 0.019] |
| IPBHPRP | 0.008 | [0.000, 0.038] | 0.006 | [0.002, 0.016] |

Table 9: Relationship Between Factor Mean Correlation and Absolute Error Size

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Correlation | 0.95 | 0.96 | 0.97 | 0.98 | 0.99 | 0.995 | 0.999 |
| Error | 0.620 | 0.554 | 0.480 | 0.392 | 0.277 | 0.196 | 0.088 |

Table 10: Advantages and Disadvantages of Fixed versus Random Approaches in Terms of Estimating Factor Means/Scores

| Criterion | Alignment | Random-Intercepts, Random-Slopes |
|---|---|---|
| Small number of factor indicators | + | - |
| Number of groups | | |
| 2-30 | + | - |
| 30-100 | + | + |
| >100 | - | + |
| Small group size | - | + |
| Weak invariance pattern | - | + |
| Information about which groups contribute to non-invariance | + | - |
| Not requiring normality of measurement parameters | + | - |
| Ability to relate non-invariance to other variables | - | + |
| Complex survey data | + | - |
| Computational speed | + | - |