Special Issue Article

Sociological Methods & Research
2018, Vol. 47(4) 637-664

DOI: 10.1177/0049124117701488
journals.sagepub.com/home/smr

Recent Methods for the Study of Measurement Invariance With Many Groups: Alignment and Random Effects

Bengt Muthén[1] and Tihomir Asparouhov[2]

Abstract

This article reviews and compares recently proposed factor analytic and item response theory 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 analyses are 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.

Keywords

factor analysis, item response theory, random intercepts, random slopes, configural invariance, scalar invariance

[1] UCLA, Los Angeles, CA, USA
[2] Muthen & Muthen, Los Angeles, California

Corresponding Author:
Bengt Muthén, UCLA, 3463 Stoner Ave., Los Angeles, CA 90066, USA.
Email: bmuthen@statmodel.com

Introduction

This article 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 noninvariance. 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 (Asparouhov and Muthén 2014; Fox 2010), 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 been recently proposed that have in common the notion of approximate measurement invariance:

1. Fixed mode: Alignment (Asparouhov and Muthén 2014).
2. Random mode: Two-level modeling with random item parameters (De Jong, Steenkamp, and Fox 2007; Fox 2010; Jak et al. 2013, 2014).

This article 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.

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).

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.

The structure of this article is as follows. The second section 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. The third section presents different two-level models, contrasts them, and applies them to the 26-country data. The fourth section presents Monte Carlo simulation studies of the two methods. The fifth section concludes with a comparison of the two methods on several practical criteria.

Fixed-mode Analysis

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. A recent alternative to CFA is multiple-group "exploratory structural equation modeling" analysis (Asparouhov and Muthén 2009). With configural invariance, 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 Structural Equation Modeling (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. Strict measurement invariance also holds residual variances invariant across groups, but this case is not considered here.

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:

$$\begin{aligned} y_{ij} &= v_j + \lambda_j f_{ij} + \epsilon_{ij}, \\ E(f_j) &= \alpha_j = 0, V(f_j) = \psi_j = 1. \end{aligned} \tag{1}$$

Metric:

$$\begin{aligned} y_{ij} &= v_j + \lambda f_{ij} + \epsilon_{ij}, \\ E(f_j) &= \alpha_j = 0, V(f_j) = \psi_j. \end{aligned} \tag{2}$$

Scalar:

$$\begin{aligned} y_{ij} &= v + \lambda f_{ij} + \epsilon_{ij}, \\ E(f_j) &= \alpha_j, V(f_j) = \psi_j, \end{aligned} \tag{3}$$

where v is a measurement intercept, λ is a factor loading, f is a factor with mean α and variance ψ, and ϵ is a residual with mean zero and variance θ 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 noninvariant 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 0 and the factor variance to 1, 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

Table 2. Twenty-six-country Example: Model Fit for Multiple-group Model.

Models	χ^2	*df*	*p* Value	RMSEA (Probability $\leq$ .05)	CFI
Configural	317	52	.000	.052 (.311)	.990
Metric	1,002	127	.000	.060 (.000)	.967
Scalar	8,654	202	.000	.148 (.000)	.677
Metric vs. configural	685	75	.000		
Scalar vs. configural	8,337	150	.000		
Scalar vs. metric	7,652	75	.000		

Note. $n = 49{,}894$. RMSEA = root mean square error of approximation; CFI = comparative fit index.

problem. Interestingly, the method is not limited to the traditional domain of multiple-group CFA or IRT, where only a few groups are typically studied, but the alignment method is suitable for the study of many groups, say up to 100.

The study of measurement noninvariance (referred to as "item bias" and "Differential Item Functioning (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 χ^2 testing of one item at a time (see, e.g., Thissen, Steinberg, and Wainer 1993). Two common approaches have been discussed (Kim and Yoon 2011; Lee, Little, and Preacher 2010; Stark, Chernyshenko, and Drasgow 2006):

- Bottom-up approach: Start with no invariance (configural case), imposing invariance one item at a time.
- Top-down approach: Start with full invariance (scalar case), freeing invariance one item at a time, for example, using modification indices (Sörbom 1989).

Neither approach is scalable—both are very cumbersome when there are many groups, such as 50 countries ($50 \times 49/2 = 1{,}225$ 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.

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 root mean square error of approximation (RMSEA) and comparative fit

index (CFI) fit values. It is clear that the addition of invariant intercepts of the scalar model in particular adds greatly to the misfit.

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 makes it very difficult to properly identify the sources of noninvariance 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 article, we review the radically different method of alignment as proposed in Asparouhov and Muthén (2014).

The Alignment Method

To save space, only a brief description of the alignment method is given here; for a full account, the reader is referred to Asparouhov and Muthén (2014). 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 noninvariance by estimating group-varying factor means α and factor variances ψ. 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 noninvariant parameters and many approximately invariant parameters rather than many medium-sized noninvariant parameters (compare with Exploratory factor Analysis (EFA) rotations using functions that aim for either large or small loadings, not midsized loadings).

In the alignment optimization of the simplicity function, the factor means α_j and variances ψ_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 λ_j and intercepts v_j changed as:

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

$$v_j = v_{j,\text{configural}} - \alpha_j \, \lambda_{j,\text{configural}} / \sqrt{\psi_j}. \tag{5}$$

The alignment method has two steps:

1. Estimate the configural model:
    - Loadings and intercepts free across groups, factor means fixed at 0 in all groups, factor variances fixed at 1 in all groups.
2. Alignment optimization:
    - Free the factor means and variances and choose their values to minimize the total amount of noninvariance using a simplicity function

$$F = \sum_{p} \sum_{j_1 < j_2} w_{j_1,j_2} f(\lambda_{pj_1} - \lambda_{pj_2}) + \sum_{p} \sum_{j_1 < j_2} w_{j_1,j_2} f(v_{pj_1} - v_{pj_2}), \qquad (6)$$

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

In this way, a nonidentified model where factor means and factor variances are added to the configural model is made identified by adding a simplicity requirement. Our simulation studies show that the alignment method works very well unless there is a majority of significant noninvariant parameters or small group sizes. For well-known multiple-group examples with few groups and few noninvariances, such as with the classic Holzinger–Swineford data for two different schools, the results agree with the alignment method.

In addition to the estimated aligned model, the alignment procedure as implemented in Mplus Version 7.1 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 (2014).

Table 3. Twenty-six-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

Note. Countries that are deemed to have a significantly noninvariant measurement parameter are shown in boldface within parentheses.

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 were introduced in Conventional Multiple-group CFA of the 26-country Example section. 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 (2014).

Table 3 shows the (non)invariance results for the measurement intercepts and factor loadings. The countries that are deemed to have a significantly noninvariant measurement parameter are shown as bolded within parentheses. As seen in Table 3, most of the items show a large degree of measurement noninvariance for the measurement intercepts and, to a lesser extent, the loadings. The large degree of noninvariance 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 noninvariance, 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

Table 4. Twenty-six-country Example: Alignment Fit Statistics.

	Intercepts			Loadings		
Items	Fit Function Contribution	R^2	Variance	Fit Function Contribution	R^2	Variance
IPMODST	−229.849	.203	.105	−158.121	.000	.020
IMPTRAD	−199.831	.566	.058	−134.042	.000	.014
IPFRULE	−213.806	.198	.103	−113.305	.263	.008
IPBHPRP	−32.836	1.000	.000	−33.941	.999	.000

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 noninvariance. 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 (The Alignment Method), 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 Simulations Based on the 26-country Data section. 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 percent 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 .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,

Table 5. Twenty-six-country Example: Factor Mean Comparisons of Countries.

Ranking	Group	Value	Groups With Significantly Smaller Factor Mean
1	26	.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	.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	.391	30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
4	7	.357	30 8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
5	11	.342	8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
6	4	.331	8 6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
7	12	.310	6 17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
8	30	.247	17 9 2 13 22 25 15 23 28 16 18 10 3 14 27 5
9	8	.200	13 22 25 15 23 28 16 18 10 3 14 27 5
10	6	.161	22 25 15 23 28 16 18 10 3 14 27 5
11	17	.130	22 25 15 23 28 16 18 10 3 14 27 5
12	9	.121	22 25 15 23 28 16 18 10 3 14 27 5
13	2	.114	22 25 15 23 28 16 18 10 3 14 27 5
14	13	.100	25 15 23 28 16 18 10 3 14 27 5
15	22	.007	15 23 28 16 18 10 3 14 27 5
16	25	.000	15 23 28 16 18 10 3 14 27 5
17	15	−.114	18 10 3 14 27 5
18	23	−.145	10 3 14 27 5
19	28	−.185	3 14 27 5
20	16	−.190	3 14 27 5
21	18	−.214	14 27 5
22	10	−.234	14 27 5
23	3	−.288	5
24	14	−.314	5
25	27	−.327	5
26	5	−.478	

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.

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:

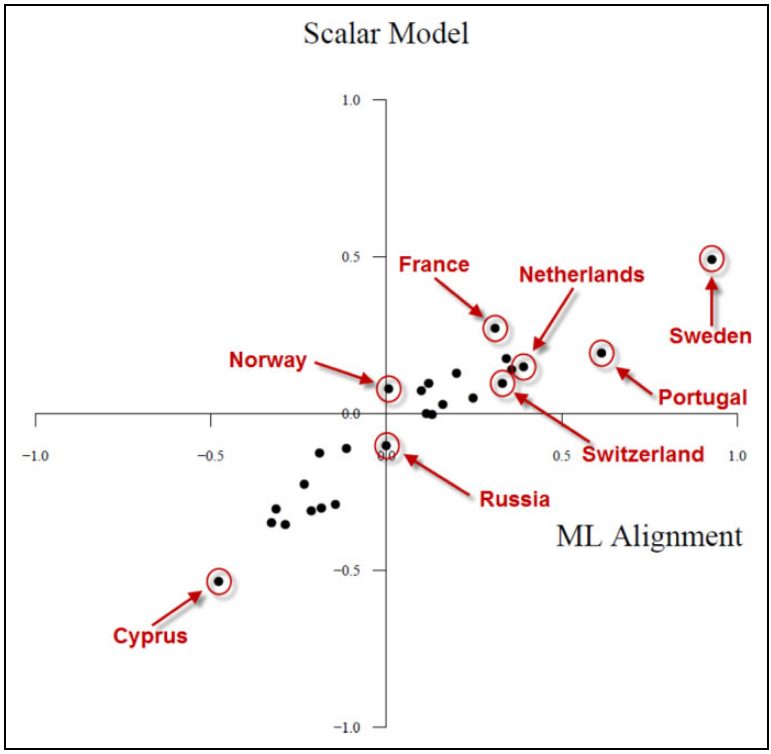

Figure 1. Twenty-six-country example: Factor means for alignment method versus scalar model.

1. Random intercepts, nonrandom (invariant) loadings: Different within- and between-level factor loadings.
2. Measurement invariance (nonrandom intercepts and loadings): Same within- and between-level factor loadings and zero between-level residual variances.
3. Random intercepts and random loadings.

Model Type 1: Random Intercepts, Nonrandom Loadings, and Different Within- and Between-level Factor Loadings

As a background for model type 1, recall random effect analysis of variance for individual i in cluster j,

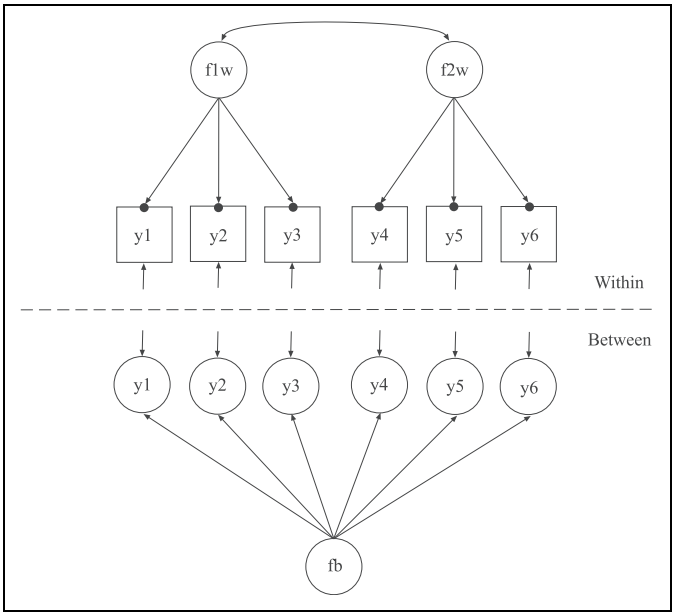

Figure 2. Random intercept two-level factor analysis in figure form.

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

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

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

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 equation (8) can be equivalently expressed as a random intercept model:

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

$$\text{Level 2}: \; v_j = v + \lambda_B f_{B_j} + \epsilon_{B_j}. \tag{10}$$

The variation in the random intercept v_j is expressed in terms of variation in a between-level factor f_{B_j} and a between-level residual ϵ_{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, labeled ϵ_{B_j} on the between level in equation (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).

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 2005, 2010; Fox and Glas 2001) considers the special case of $\lambda_W = \lambda_B = \lambda$ and $V(\epsilon_{B_j}) = 0$, so that equations (9) and (10) become

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

$$\text{Level 2}: \; v_j = v + \lambda f_{B_j} + 0, \tag{12}$$

so that v_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 rewritten as

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

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

which shows that the model assumes invariance of the intercept v and the loading λ across clusters and that the same λ multiplies 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, Oort, and Dolan (2013, 2014). This involves testing the general model of equations (9) and (10) against the model with $\lambda_W = \lambda_B = \lambda$ and $V(\epsilon_{B_j}) = 0$ using likelihood ratio χ^2. Modification indices (Lagrange multipliers; Sörbom 1989) are used to reveal model misfit due to nonzero $V(\epsilon_{B_j})$, pointing to factor indicators that have significant between-level residual variance and therefore noninvariant intercepts. This approach is illustrated in Random Intercept Analysis section.

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 et al. (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 (2015). Bayesian estimation is needed because random loadings with maximum likelihood estimation give 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 for the measurement parameter. As with the alignment method, only approximate measurement invariance is presumed. Different groups/clusters have different random deviations from the common mean. For example, 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 percent of the groups/clusters have a factor loading between 0.8 and 1.2.

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

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

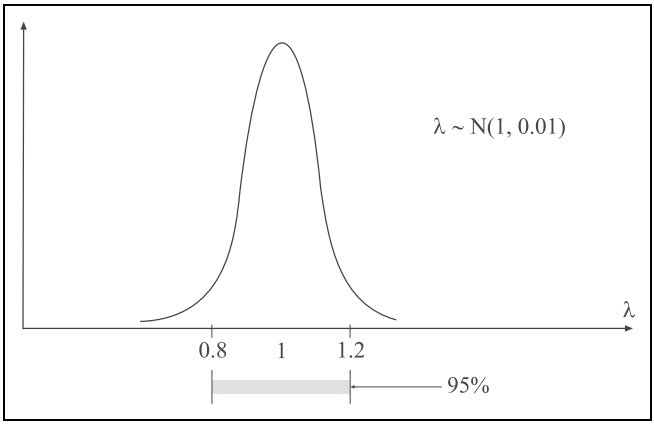

Figure 3. Random measurement parameter.

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

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

where Φ is the standard normal distribution function, θ_{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 two-parameter probit IRT model where both discrimination (a) and difficulty (b) vary across groups/clusters. The θ 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 noninvariance. 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 (2015) 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

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

$$v_j = v + \lambda f_{B_j} + \epsilon_{v_j}, \tag{23}$$

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

where $f_{W_{ij}} \sim N(0,1)$, $\epsilon_{ij} \sim N(0,\theta)$, $\epsilon_{v_j} \sim N(0,\sigma_v^2)$, $\epsilon_{\lambda.j} \sim N(0,\sigma_\lambda^2)$, $f_B \sim N(0,\psi)$, and $f_{\psi_j} \sim N(0,\sigma^2)$. The variation in intercepts is captured by σ_v^2, the variation in the loadings is captured by σ_λ^2, the variation in factor means is captured by ψ, and the variation in the factor variance is captured by σ^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 equations (22) to (24). Model type 2 of equation (13), equation (14) is obtained when setting $\sigma^2 = 0$, $\sigma_\lambda^2 = 0$, and $\sigma_v^2 = 0$, 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 $v_j = v + \lambda\, f_{B_j}$. Model type 1 of equation (9), equation (10) is obtained when setting $\sigma^2 = 0$, $\sigma_\lambda^2 = 0$, 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 σ^2 be freely estimated.

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.

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

Model 1 fits rather well, given the large sample size of 49,894 subjects. The χ^2 p value is .000, but good fit is indicated by RMSEA = .011 and CFI = .999. A test of model 2 against model 1 leads to a χ^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 χ^2 of 6,703 with 7 degrees of freedom.

Table 6. Twenty-six-country Example: Two-level Random Intercept Analysis.

Models	χ^2	*df*	RMSEA	CFI
1	Different loadings across levels, residual variances free on between			
	28.010	4	.011	.999
2	Equal loadings across levels, residual variances free on between			
	31.868	7	.008	.999
3	Equal loadings across levels, residual variances fixed at 0			
	6,731.072	11	.111	.723

Table 7. Twenty-six-country Example: Two-level Random Intercept Predicted and Actual χ^2 Improvement for Model 3 Between-level Residual Variances.

	χ^2 Improvement	
Items	Predicted by Modification Index	Actual
IPMODST	201,293	3,549
IMPTRAD	29,726	1,201
IPFRULE	161,347	2,924
IPBHPRP	14,347	852

The influence on model misfit for model 3 due to nonzero residual variances on the between level is shown in Table 7. In addition to modification indices, the actual χ^2 improvements (the drop in χ^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 χ^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 noninvariance of the measurement intercepts.

Random intercept and random loading analysis. Bayesian analysis was applied to the random-intercept, random-loading model of equations (22) to (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.

Table 8. Twenty-six-country Example: Two-level Random Intercept and Random Loading Variance Estimates and 95 Percent Credibility Intervals.

	Intercept		Loading	
Items	Estimate	CI	Estimate	CI
IPMODST	.122	[.070, .240]	.022	[.012, .042]
IMPTRAD	.056	[.031, .116]	.010	[.005, .021]
IPFRULE	.100	[.055, .196]	.008	[.003, .019]
IPBHPRP	.008	[.000, .038]	.006	[.002, .016]

Note. CI = confidence interval.

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 and 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 .987. Figure 4 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 .980.

Simulation Studies Comparing Fixed Versus Random-mode Analysis

This section compares the alignment method and the random-intercept, random-loading method using a Monte Carlo study with simulated data. Due to having known measurement noninvariance with known factor means and variances, a Monte Carlo study makes it possible to gauge the success of each method in finding the correct parameter values.

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

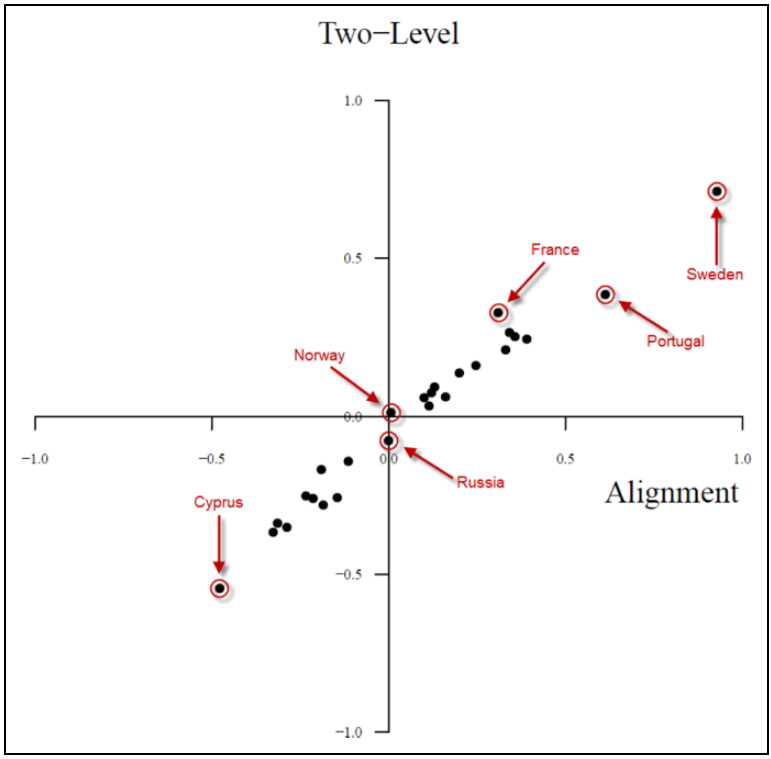

Figure 4. Twenty-six-country example: Factor means for two-level versus alignment analysis.

the alignment approach and using the two-level approach. The results for the model with full measurement invariance are also shown as a comparison.

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

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

Correlation	.95	.96	.97	.98	.99	.995	.999
Error	.620	.554	.480	.392	.277	.196	.088

do a good job for the extent of noninvariance studied, whereas the former value shows the performance of the alignment method for the extent of noninvariance 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. Online Appendix A 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 percent 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 percent of the groups, a correlation of .99 is required. A correlation of .95 gives a large error of .620. Figures 1 and 4 exemplify the difference in ordering of the countries for a correlation of .943 and .987, respectively. A correlation of at least .99 has also shown to be a good requirement for low bias in estimating each group's factor mean.

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.

Simulations Based on the 26-country Data

As discussed in Asparouhov and Muthén (2014), 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. In this section, data for each of the two methods are generated by the model assumed for that method. In the Online Appendix B, however, data for each method are generated by the model assumed for the other method. 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 are 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 .990 for the factor means. According to Table 9, the high-factor mean correlation corresponds to a relatively small absolute error of .277. The correlation between the true factor means and the average estimated factor means, where the average is over the replications, is .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 .940 and .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 .950 corresponding to an absolute error of .620. The correlation using averages is not applicable in this case, given that average scores are 0. 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 .977, .982, .985, and .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 .874 is obtained. This is akin to using the scalar model in the fixed-mode case. Applying model type 1, a correlation of .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 nonrandom loadings that are equal across the two levels obtains a correlation of .951, that is, the same as when also letting the loadings to 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. Online Appendix B 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 that describes in one statistic both the bias and variability of the estimates. The reader is referred to Online Appendix B for the results.

Conclusions

This article 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 differences between the two methods discussed in this article are 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 noninvariant. The assumption of the random intercept and loading method is that all parameters are approximately the same, that is, 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 noninvariance and produces a model that has clear interpretation in terms of invariance and noninvariance. The random intercept and slope method are not as detailed or focused on the actual parameter variations across the groups but instead look 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.

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

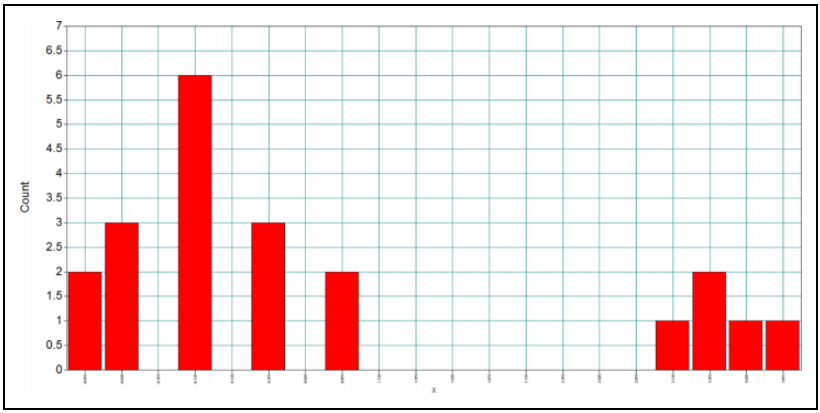

Figure 5. Group-varying intercepts.

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. Nonnormality 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 5. 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.

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. 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 noninvariance	+	−
Not requiring normality of measurement parameters	+	−
Ability to relate noninvariance to other variables	−	+
Complex survey data	+	−
Computational speed	+	−

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 are sufficient in principle.

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 are 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.

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 (2014) 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.

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

Information about groups contributing to noninvariance. Measurement invariance studies benefit from information on which groups contribute to noninvariance. 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.

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 nonparametric.

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

Complex survey data. Comparisons of many groups often arise in surveys of many countries where a complex survey design is used. For instance, with Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and other surveys of school

children, sampling of schools is carried out using probability proportional to size, 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 cannot accommodate complex survey features.

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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Asparouhov, T. and B. Muthén. 2009. "Exploratory Structural Equation Modeling." *Structural Equation Modeling* 16:397-438.

Asparouhov, T. and B. Muthén. 2014. "Multiple-group Factor Analysis Alignment." *Structural Equation Modeling* 21:495-508. doi:10.1080/10705511.2014.919210.

Asparouhov, T. and B. Muthén. 2015. "General random effect latent variable modeling: Random subjects, items, contexts, and parameters." in *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*, edited by J. R. Harring, L. M. Stapleton, and S. N. Beretvas. Charlotte, NC: Information Age Publishing, Inc.

Beierlein, C., E. Davidov, P. Schmidt, and S. H. Schwartz. 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.

Bou, J. C. and A. Satorra. 2010. "A Multigroup Structural Equation Approach: A Demonstration by Testing Variation of Firm Profitability across EU Samples." *Organizational Research Methods* 13:738-66. First published on January 26, 2010.

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, Stanford, CA.

De Boeck, P. 2008. "Random Item IRT Models." *Psychometrika* 73:533-59.

de Jong, M. G. and J. B. E. M. Steenkamp. 2010. "Finite Mixture Multilevel Multidimensional Ordinal IRT Models for Large-scale Cross-cultural Research." *Psychometrika* 75:3-32.

De Jong, M. G., J. B. E. M. Steenkamp, and J. P. Fox. 2007. "Relaxing Measurement Invariance in Cross-national Consumer Research Using a Hierarchical IRT Model." *Journal of Consumer Research* 34:260-78.

Fox, J. P. 2005. "Multilevel IRT Using Dichotomous and Polytomous Response Data." *British Journal of Mathematical and Statistical Psychology* 58:145-72.

Fox, J. P. 2010. *Bayesian Item Response Theory*. New York: Springer.

Fox, J. P. and C. A. W. Glas. 2001. "Bayesian Estimation of a Multilevel IRT Model Using Gibbs." *Psychometrika* 66:269-86.

Fox, J. P. and J. Verhagen. 2011. "Random Item Effects Modeling for Cross-national Survey Data." Pp. 461-82 in *Cross-cultural Analysis: Methods and Applications*, edited by E. Davidov, P. Schmidt, and J. Billiet. New York: Routledge.

Frederickx, S., F. Tuerlinckx, P. De Boeck, and D. Magis. 2010. "RIM: A Random Item Mixture Model to Detect Differential Item Functioning." *Journal of Educational Measurement* 47:432-57.

Goldstein, H. and R. P. McDonald. 1988. "A General Model for the Analysis of Multilevel Data." *Psychometrika* 53:455-67.

Härnqvist, K. 1978. "Primary Mental Abilities of Collective and Individual Levels." *Journal of Educational Psychology* 70:706-16.

Härnqvist, K., J. E. Gustafsson, B. Muthén, and G. Nelson. 1994. "Hierarchical Models of Ability at Class and Individual Levels." *Intelligence* 18:165-87.

Jak, S., F. J. Oort, and C. V. Dolan. 2013. "A Test for Cluster Bias: Detecting Violations of Measurement Invariance across Clusters in Multilevel Data." *Structural Equation Modeling* 20:265-82.

Jak, S., F. J. Oort, and C. V. Dolan. 2014. "Measurement Bias in Multilevel Data." *Structural Equation Modeling: A Multidisciplinary Journal* 21:31-39. doi:10.1080/10705511.2014.856694.

Jennrich, R. I. 2006. "Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case." *Psychometrika* 71:173-91.

Kim, E. S. and M. Yoon. 2011. "Testing Measurement Invariance: A Comparison of Multiple-group Categorical CFA and IRT." *Structural Equation Modeling* 18:212-28.

Lee, J., T. D. Little, and K. J. Preacher. 2010. "Methodological Issues in Using Structural Equation Models for Testing Differential Item Functioning." Pp.55-84 in *Cross-cultural Analysis. Methods and Applications*, edited by E. Davidov, P. Schmidt, and J. Billiet. New York: Routledge.

Longford, N. T. and B. Muthén. 1992. "Factor Analysis for Clustered Observations." *Psychometrika* 57:581-97.

McDonald, R. P. and H. Goldstein. 1989. "Balanced versus Unbalanced Designs for Linear Structural Relations in Two-level Data." *British Journal of Mathematical and Statistical Psychology* 42:215-32.

Millsap, R. E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Taylor & Francis.

Muthén, B. 1994. "Multilevel Covariance Structure Analysis." Pp. 376-98, Vol. 22 in *Multilevel Modeling, A Special Issue of Sociological Methods & Research*, edited by J. Hox and I. Kreft. Thousand Oakes, CA: Sage Publications.

Sörbom, D. 1989. "Model Modification." *Psychometrika* 54:371-84.

Stark, S., O. S. Chernyshenko, and F. Drasgow. 2006. "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology* 91:1292-306.

Thissen, D., L. Steinberg, and H. Wainer. 1993. "Detection of Differential Item Functioning Using the Parameters of Item Response Models." Pp. 67-113 in *Differential Item Functioning*, edited by P. W. Holland and H. Wainer. Hillsdale, NJ: Lawrence Erlbaum.

Verhagen, A. J. 2013. "Bayesian Item Response Theory Models for Measurement Invariance." Doctoral dissertation, University of Twente, the Netherlands.

Verhagen, A. J. and J. P. Fox. 2013. "Bayesian Tests of Measurement Invariance." *The British Journal of Mathematical and Statistical Psychology* 66:383-401.

Author Biographies

Bengt Muthén obtained his PhD in Statistics at the University of Uppsala, Sweden and is a professor Emeritus at UCLA. He was the 1988-89 president of the Psychometric Society and the 2011 recipient of the Psychometric Society's Lifetime Achievement Award. He has published extensively on latent variable modeling and many of his procedures are implemented in the Mplus program.

Tihomir Asparouhov obtained his PhD in Mathematics at the California Institute of Technology. He is a senior statistician with the Mplus group. His research interests are in the areas of structural equation modeling, dynamic modeling, complex survey analysis, multilevel modeling, survival analysis, and Bayesian analysis.