# 3

# Methodological Issues in Using Structural Equation Models for Testing Differential Item Functioning

*Jaehoon Lee, Todd D. Little, and Kristopher J. Preacher*
University of Kansas

## 3.1 INTRODUCTION

In cross-cultural studies, groups often differ in various characteristics (e.g., demographics, socioeconomic status, language, culture, etc.) and these characteristics may not be relevant to the goals of a particular study. Even when reasonable precautions have been taken to prepare a test or survey that is equivalent across cultural groups, it is possible that the attribute being measured has different conceptual meanings in different groups (de Beuckelaer, Lievens, & Swinnen, 2007) or that some items have different importance for one group more than another (Cheung & Rensvold, 1999). In such cases, observed group differences may represent measurement artifacts related to the instrument rather than true differences on a relevant construct. This disparity between observed and true group differences, in turn, adversely affects the comparability of their scores (Byrne & Stewart, 2006; de Beuckelaer et al., 2007; Raju, Laffitte, & Byrne, 2002; Vandenberg & Lance, 2000; van de Vijver & Poortinga, 1992). Thus, researchers have highlighted the importance of measurement equivalence as a prerequisite for meaningful group comparisons (Drasgow, 1984; Little, 1997; Reise, Widaman, & Pugh, 1993). Accordingly, standards established by both the American Psychological Association (APA) and the International Test Commission (ITC) have

emphasized evaluation of measurement equivalence for fair use of a scale (1999).

Structural equation modeling (SEM) has been highlighted as a useful and powerful tool for assessing measurement equivalence, or equivalently factorial invariance, across different cultural groups. For example, researchers have successfully evaluated factorial invariance in questionnaires for physical and mental health (Liang, 2001; Wang, Liu, Biddle, & Spray, 2005), mood and depression (Bagozzi, 1994; Byrne & Stewart, 2006; Gregorich, 2006; Reise et al., 1993), self-concept and personality (Katsuya, 2007; Leone, van der Zee, van Oudenhoven, Perugini, & Ercolani, 2005; Marsh, Tracey, & Craven, 2006), and consumer and organizational behavior (Dholakia, Firat, & Bagozzi, 1980; Raju et al., 2002; Riordan & Vandenberg, 1994; Schaffer & Riordan, 2003; Steenkamp & Baumgartner, 1998). In this chapter, we address some methodological issues that may arise when researchers conduct the SEM analysis of factorial invariance. This chapter consists of two parts. In Part I, we (a) introduce the concept of factorial invariance, (b) review the levels of invariance, and (c) introduce the concept of differential item functioning (DIF), which refers a lack of invariance at the item level. In Part II, we (a) describe two SEM-based DIF analyses, (b) summarize two Monte Carlo studies that examine the effects of employing different scaling designs, analytic strategies, and test statistics, and (c) provide general procedural guidelines for evaluating invariance of a scale. This chapter contributes to the cross-cultural measurement literature by cautioning researchers against the use of the conventional analytic approach in DIF analysis. Throughout this chapter, we will show that an innocuous choice of identification condition in the conventional approach involves the danger of inflating Type I error for tests of DIF, and therefore any cross-cultural group comparisons can be jeopardized by falsely identified item bias.

## 3.2 PART I

### 3.2.1 Factorial Invariance

Factorial invariance, which originated from the factor analytic and SEM literatures (Meredith, 1993), has a long history in the study of group



differences. The key question that factorial invariance addresses is "Are the underlying (latent) constructs measured in a comparable manner across two or more groups?" If the answer is "Yes," the indicators or items of the constructs behave similarly (psychometrically speaking) in each group. When items behave similarly, any observed differences represent "true" differences in the constructs, but not artifactual differences stemming from any differential functioning of the items. We will explain more precisely what we mean by the phrase "behave similarly" later. For now, the idea of similar behavior implies that key item parameters are statistically equivalent across two or more groups.

#### *3.2.1.1 Mean and Covariance Structure Analysis for Factorial Invariance*

Currently, mean and covariance structure (MACS; Sörbom, 1974) analysis is preferable for evaluating factorial invariance for several reasons (Little, 1997). In MACS analysis, a hypothesized factor structure is fitted simultaneously in two or more groups. Between-group equality of all parameters can be assessed, and "strong" tests for factorial invariance are possible. MACS analysis can be thought of as an extension of standard confirmatory factor analysis (CFA). That is, CFA attempts to reproduce the covariance structure that underlies a set of measured variables, while MACS analysis considers their mean structure as well. Thus, both CFA and MACS analysis are special cases of SEM.

The basic equations for MACS analysis are presented in Table 3.1. In Equation 3.1, observed examinee scores are depicted using a typical regression layout, where $\mathbf{T}$ is a $(p \times 1)$ vector of regression intercepts; $\Lambda$ is a $(p \times m)$ matrix of regression slopes, or loadings, which define the associations between items and latent constructs; $\eta$ is an $(m \times 1)$ vector of latent scores; and $\Theta$ is a $(p \times 1)$ vector of residual or unique factor scores. A key feature of this equation (and Equations 3.2 and 3.3) is that the parameters in each matrix are estimated uniquely in each group (denoted by the subscript $g$).

#### *3.2.1.2 Levels of Factorial Invariance*

Vandenberg and Lance (2000) extensively reviewed different levels of factorial invariance proposed in the literature and recommended a number of

**TABLE 3.1**

The Basic Equations for MACS Analysis and Levels of Invariance

| Invariance Level | Equation | |
|---|---|---|
| Configural | $y_g = \mathbf{T}_g + \Lambda_g \eta_g + \Theta_g$ | (3.1) |
| | $E(y_g) = \mu_{y_g} = \mathbf{T}_g + \Lambda_g \mathbf{A}_g$ | (3.2) |
| | $\Sigma_g = \Lambda_g \Psi_g \Lambda_g' + \Theta_g$ | (3.3) |
| Metric (weak factorial) | $y_g = \mathbf{T}_g + \Lambda \eta_g + \Theta_g$ | (3.4) |
| | $E(y_g) = \mu_{y_g} = \mathbf{T}_g + \Lambda \mathbf{A}_g$ | (3.5) |
| | $\Sigma_g = \Lambda \Psi_g \Lambda' + \Theta_g$ | (3.6) |
| Scalarb (strong factorial) | $y_g = \mathbf{T} + \Lambda \eta_g + \Theta_g$ | (3.7) |
| | $E(y_g) = \mu_{y_g} = \mathbf{T} + \Lambda \mathbf{A}_g$ | (3.8) |
| | $\Sigma_g = \Lambda \Psi_g \Lambda' + \Theta_g$ | (3.9) |

*Note:* $y$ is a $p \times 1$ vector of observed responses on the $p$ items and $g$ is an index that refers to the group. When $g$ is present, the parameters in the associated matrix are freely estimated across groups. $E(\,)$ is the expectation operator and $\mu$ is a $p \times 1$ vector of item means. A is an $m \times 1$ vector of latent construct means. $\Sigma$ is the model implied variance–covariance matrix of $y$.

invariance levels that could be evaluated in empirical research.* *Configural invariance* describes the situation when the parameters are estimated uniquely in each group but the pattern of free and fixed parameters is the same (or very similar). Configural invariance is determined by overall model fit and simple judgment regarding the adequacy of the hypothesized model in each group.

Different levels of factorial invariance require the parameter estimates in different matrices to be constrained across groups. In the model of *metric invariance*, the loading estimates in $\Lambda$ are constrained to be equal

* The invariance levels that are not discussed here include invariance of unique factor variances, invariance of latent construct variances/covariances, and invariance of latent construct means. For a detailed discussion on factorial invariance, see Meredith (1993) and Vandenberg and Lance (2000).



across groups and therefore common values are generated that are optimal for all groups (see Equation 3.4). In order to determine whether or not metric invariance holds, we evaluate the reasonableness of the metric invariance model (i.e., the imposed equality constraints) relative to the configural invariance model. Although metric invariance suggests that two or more groups share the same unit of measurement, it does not necessarily indicate that the origins (i.e., intercepts) of the scale are equivalent across groups. Thus, this invariance level is often called *weak* factorial invariance (Meredith, 1993).

Similarly, in the model of *scalar invariance*, the intercept estimates in $\mathbf{T}$ are also constrained to be equal across groups (see Equation 3.7). We evaluate the reasonableness of the scalar invariance model by assessing the fit change from the metric invariance model. Given scalar invariance, the scale is considered to have the same unit of measurement as well as the same origin, and therefore group mean comparisons become tenable (Widaman & Reise, 1997). Thus, this invariance level is often called *strong* factorial invariance (Meredith, 1993).

#### 3.2.1.3 Testing Factorial Invariance

As mentioned previously, invariance testing involves judging the reasonableness of the sequentially added constraints. Although one could evaluate the imposed equality constraints by assessing the $\chi^2$ differences between two nested models (i.e., likelihood ratio [LR] test), $\Delta\chi^2$ value may not a practical test statistic because of its dependency on sample size (Brannick, 1995; Kelloway, 1995). Instead, the set of equality constraints can be evaluated by assessing the change in key global fit indices. Most recently, Meade, Johnson, and Braddy (2008) conducted a conservative simulation study (i.e., 0.01 Type I error, 0.90 power) and concluded that the optimal criterion for rejecting a hypothesized invariance model is the change in the comparative fit index (CFI) of greater than −0.002. This recommendation represents a more stringent criterion than a previous simulation study conducted by Cheung and Rensvold (2002). They recommended that a $\Delta$CFI value less than 0.01 was sufficient evidence that a hypothesized invariance model holds with regard to a more conventional Type I error (i.e., 0.05). In addition, Chen (2007) recommended assessing changes in the root mean square error of approximation (RMSEA) and standardized root mean square residual ($\Delta$SRMR) as well. Nevertheless, she also concluded that

ΔCFI should be the main criterion because ΔRMSEA and ΔSRMR tests tend to overreject an invariant model when sample size is small.*

Taken together, if imposing equality constraints leads to a loss in CFI that is greater than 0.002 or 0.01, then one or more of the constraints are not tenable. In such cases, a set of "offending" (i.e., noninvariant) items must be located in a scale. A variety of analyses have been proposed for this purpose in the SEM literature (e.g., Chan, 2000; Ferrando, 1996; Muthén, 1988). As we detail later, most of the analyses are influenced by how we scale the latent constructs.

### 3.2.2 Methods of Scaling

In any structural model, the scale of the construct needs to be identified in order to obtain a unique solution for every parameter (Bollen, 1989). There are now three useful, statistically equivalent scaling methods (Little, Slegers, & Card, 2006). When three or more items are used to measure a construct, each scaling method provides the necessary condition for identifying the scale of the construct. Because using fewer than three indicators risks underidentification and increases the probability of obtaining an infeasible solution (Bollen, 1989), our discussion focuses on situations when a researcher has three or more items for each construct.

The most common scaling method is the *marker-variable method.* This method constrains one of the loadings and a corresponding intercept, by which the other parameters are estimated (see Little et al., 2006; see also Little, in press, Chapter 3). Generally, the loading is fixed to 1 and the intercept is fixed to 0. The second common method involves fixing the variance of the construct to 1 and the mean to 0. This method is termed the *fixed-factor* or *reference-factor method.* The third method is the recently introduced *effects-coded method.* This method involves placing a set of constraints so that the loadings average 1 and the intercepts average 0. Unlike the other two methods that provide an arbitrary scale of the construct, the effects-coded method provides a scale of the construct that directly reflects the scale of its indicators (see Little et al., 2006; Little, in press).

* Information-theoretic measures of fit (e.g., Akaike information criterion, Bayesian information criterion) are also suitable for evaluating factorial invariance, but they have not been supported in the literature as being informative beyond the CFI and the other fit measures discussed here.



### 3.2.3 Differential Item Functioning

The concept of factorial invariance also underlies the concept of DIF; a concept originated from the item response theory (IRT) literature. The IRT includes the statistical models specialized for different types of categorical responses (e.g., binary, ordinal), and the models for binary responses can be viewed as special cases of graded response model (Samejima, 1969, 1972). Before we detail in Part II how the concept of DIF is integrated into the SEM framework, we will briefly introduce the key terms and parameters in IRT. More detailed discussion on IRT can be found in Chapters 15 through 17 in this book.

The basic assumptions in IRT are that a set of items assesses a single ability dimension (unidimensionality) but they are pairwise uncorrelated at a given value of latent ability (local independence). That is, residual variances are uncorrelated when conditioned on the common ability variance—similar to the conditional independence assumption in CFA. In the graded response model, the relationship between the latent ability (θ) and the probability of choosing a progressively increasing response category is depicted by a series of boundary response functions, $p^*_{ik}(\theta_j) = e^{a_i(\theta_j - b_{ik})} / 1 + e^{a_i(\theta_j - b_{ik})}$, where $p^*_{ik}(\theta_j)$ is the probability that an examinee $j$ with a certain value of θ will respond to an item $i$ at or above a response category $k$. For a particular item with $k$ response categories, $k - 1$ boundary response functions are present. As observed in this equation, the boundary response functions depend on the θ parameter as well as the $b$ and $a$ parameters. The latter two parameters are usually termed the *attractiveness* (or *difficulty*) and *discrimination* of an item, respectively. For each item, $k - 1$ attractiveness/difficulty estimates are possible and a common value is generated for the discrimination estimates. As noted previously, if $k - 2$, the graded response model simplifies to a two-parameter model (Birnbaum, 1968; Lord & Novick, 1968; McDonald, 1967). Furthermore, if $a - 1$, this model becomes a one-parameter model (Hambleton, Swaminathan, & Rogers, 1991; Rasch, 1960).

In IRT, a lack of factorial invariance is referred to as *differential item functioning* or DIF. More specifically, DIF represents group differences in the probability of an item response after their ability scores are placed on a common scale, or "statistically matched" (Mellenbergh, 1994). When present, DIF indicates either *impact* or *item bias* depending on the source or nature of DIF (Camilli & Shepard, 1994; Zumbo, 1999). When group

**TABLE 3.2**

The IRT and SEM Parameters that Determine the Type of DIF or Item Bias

| Type of DIF | IRT Parameter | SEM Parameter |
|---|---|---|
| Uniform (i.e., a failure of scalar invariance) | Attractiveness/ difficulty ($b$) | Intercept ($\tau$) |
| Nonuniform (i.e., a failure of metric invariance) | Discrimination ($a$) | Loading ($\lambda$) |

truly differ in a latent ability being measured, they are expected to provide different responses on the same items. In such case, the parameter estimates of these items accurately reflect group differences in the ability or impact. In contrast, item bias occurs when different item responses are caused by factors that are irrelevant to the ability being measured. Because the conditional probability of an item response depends on the item parameters (see boundary response function described above), DIF (either impact or item bias) is present when the item parameter estimates are not invariant across groups (Raju et al., 2002).

DIF can be either uniform or nonuniform depending on what item parameter differs across groups. *Uniform DIF* is present when only the attractiveness/difficulty parameter estimates differ across groups. *Nonuniform DIF* exists when the discrimination parameter estimates differ across groups regardless of whether the attractiveness/difficulty parameter estimates are different or not. As detailed later, uniform DIF corresponds to group differences in intercepts, whereas nonuniform DIF corresponds to group differences in loadings. Table 3.2 presents the corresponding IRT and SEM parameters that determine the two types of DIF.

## 3.3 PART II

The links between IRT and SEM have been well demonstrated in the literature (e.g., Brown, 2006; Fleishman, Spector, & Altman, 2002; Goldstein & Wood, 1989; Kamata & Bauer, 2008; MacIntosh & Hashim, 2003; McDonald, 1999; Mellenbergh, 1994; Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002; Muthén & Christoffersson, 1981; Muthén & Lehman, 1988; Takane & de Leeuw, 1987). Researchers have extended these links, allowing us to test DIF within the SEM framework (e.g., Chan, 2000; Ferrando, 1996; Muthén, 1988). The most common SEM techniques employ either MACS (Sörbom, 1974) analysis or multiple indicators multiple causes (MIMIC; Jöreskog & Goldberger, 1975) analysis. The model specification and analytic strategies for each analysis are demonstrated here.



### 3.3.1 MACS Analysis for DIF

MACS analysis for DIF detection assumes that the responses on a given set of items reflect a single common construct. Further assuming that the covariances among the unique factor scores of this set of items are 0 in the population, the mean of an item is equal to the intercept when the construct score is zero and the covariances between each item and construct are equal to the loading (Jöreskog, 1971). The assumptions that (a) a single latent construct underlies the correlations among a set of items and that (b) off-diagonal elements in $\Theta_g$ are zero and are the analogs of, respectively, the unidimensionality and local independence assumptions in IRT. Note that in MACS DIF analysis, the local independence assumption can be violated and estimated (i.e., true population correlated residuals can be specified). In addition, multiple constructs—each with a unique set of indicators—can be included in simultaneous tests for DIF across all constructs.

The intercept parameters correspond to the attractiveness/difficulty parameters in IRT; the higher the intercept, the more attractive/difficult the item is (i.e., a higher mean response is obtained). The loading parameters correspond to the discrimination parameters; the higher the loading, the more discriminating the item is (i.e., this item better differentiates examinees of different construct scores; see Ferrando, 1996; Grayson & Marsh, 1994; Mellenbergh, 1994). As noted previously, uniform DIF exists when only the attractiveness/difficulty parameter estimates differ across groups; nonuniform DIF is present when the discrimination parameter estimates differ across groups regardless of whether or not the attractiveness parameter estimates are invariant. Thus, a lack of invariance in $\mathbf{T}$ implies uniform DIF, whereas a lack of invariance in $\Lambda$ implies nonuniform DIF regardless of whether or not $\mathbf{T}$ is invariant (Chan, 2000).

### 3.3.2 MIMIC Analysis for DIF

Multiple indicators multiple causes (MIMIC) analysis extends the standard MACS analysis by regressing latent constructs on measured grouping variables (covariates). Muthén and colleagues (Gallo, Anthony, & Muthén, 1994; Muthén, 1988) further extended MIMIC analysis such that item responses are also regressed on the covariates (MIMIC DIF analysis). Table 3.3 shows the basic equations for MIMIC and MIMIC DIF analyses. In Equations 3.10 through 3.15, B is a $(p \times q)$ matrix of regression slopes of the responses on the covariates and $\Gamma$ is an $(m \times q)$ matrix of regression slopes of the constructs on the covariates.

The regression slopes in $\Gamma$ are termed the *indirect effects* and they account for (latent) group mean differences across groups. The regression slopes in B are termed the *direct effects* because they influence the responses, unmediated by the latent constructs (Bollen, 1989; Dorans & Holland, 1993; Dorans & Kulick, 1986; Jones, 2006). The direct effects indicate whether item responses differ across groups after controlling for any latent mean differences, which is the definition of DIF (Fleishman, 2005; Fleishman & Lawrence, 2003; Fleishman et al., 2002). Accordingly, DIF is evident when the direct effects are statistically significant (Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Jones, 2006). The direct and indirect effects are conceptually illustrated in Figure 3.1. Note that because common loading parameters are assumed for different groups, MIMIC DIF analysis is limited to tests for uniform DIF.

### 3.3.3 Analytic Strategies

There are two analytic strategies useful for SEM DIF analysis. The first strategy tests DIF one item at a time, assuming that other items are DIF-free

**TABLE 3.3**

The Basic Equations for MIMIC Analysis

| Analysis | Equation | |
|---|---|---|
| MIMIC | $y = \mathbf{T} + \Lambda\eta + \Theta$ | (3.10) |
| | $E(y) = \mathbf{T} + \Lambda\mathbf{A}$ | (3.11) |
| | $\eta = \mathbf{A} + \Gamma x_k + \zeta$ | (3.12) |
| MIMIC DIF | $y = \mathbf{T} + \Lambda\eta + \mathrm{B}x_k + \Theta$ | (3.13) |
| | $E(y) = \mathbf{T} + \Lambda\mathbf{A} + \mathrm{B}x_k$ | (3.14) |
| | $\eta = \mathbf{A} + \Gamma x_k + \zeta$ | (3.15) |



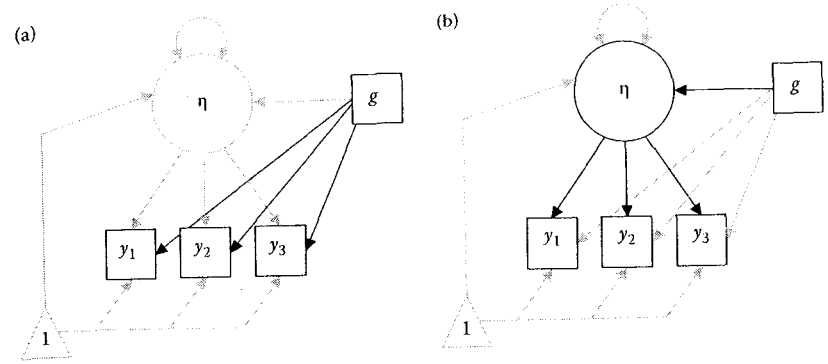



**FIGURE 3.1**
Direct and indirect effects in MIMIC DIF analysis. (a) Direct effects of a covariate on item responses. (b) Indirect effects of a covariate on item responses via construct.

anchors (e.g., Chan, 2000; Chen & Anthony, 2003; Finch, 2005; Gelin, 2005; Muthén & Asparouhov, 2002; Oishi, 2006; Stark, Chernyshenko, & Drasgow, 2006). In MIMIC DIF analysis, this strategy involves starting with a baseline model in which no direct effects are specified. After the fit of this baseline model is established, it is then statistically compared with each of the $p$ models (where $p$ = number of items), where a direct effect is allowed for only one item at a time.

To test nonuniform DIF using MACS DIF analysis, a baseline model constrains each item's loading and intercept to be equal across groups. Then, this model is compared with each of the $p$ models in which one respective loading is freely estimated in each group. Uniform DIF is usually examined only for those items whose loadings have been found to be invariant (Steenkamp & Baumgartner, 1998), although this is not a required condition (i.e., loading and intercept invariance can be evaluated simultaneously for an item; see Stark et al., 2006). A baseline model constrains the items' loadings (except for the nonuniform DIF items) and intercepts to be equal across groups. Then, this baseline model is then compared with each of $q$ models (where $q$ = number of items with invariant loadings) in which one respective intercept is freely estimated in each group. Because this analytic strategy starts by constraining the parameters of interest across groups, it is termed the *constrained-baseline strategy.*

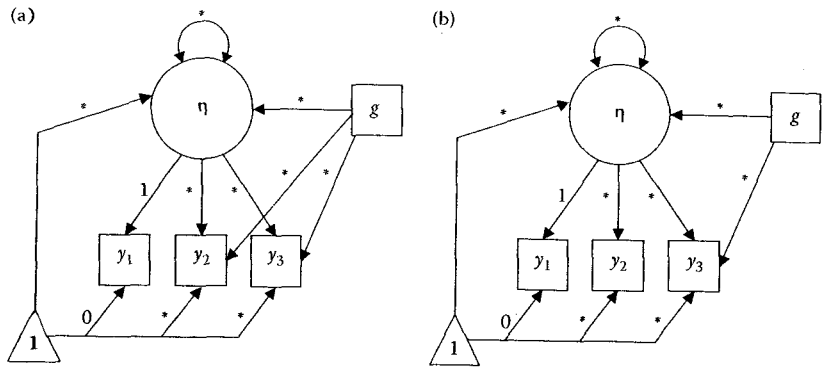



**FIGURE 3.2**
The free-baseline strategy for MIMIC DIF analysis. (a) Free-baseline model. (b) Model of a single restrictive direct effect. 3.2A and 3B depicts two nested MIMIC models. This example illustrates a simple case in which (a) a test includes three items and (b) the marker-variable scaling method was used for scaling. For simplicity, unique factor variances are omitted. Free parameters are marked by "*."

The second strategy tests DIF one item at a time, assuming that other items are not free from DIF (e.g., Fleishman et al., 2002; Stark et al., 2006; Woods, 2009; Woods, Oltmanns, & Turkheimer, 2009). Accordingly, this strategy starts with the baseline model in which all the parameters are freely estimated except those needed for scaling. This *free-baseline strategy* is depicted in Figures 3.2 (MIMIC DIF analysis) and 3.3 (MACS DIF analysis). For MIMIC DIF analysis, all possible direct effects except for at least one anchor are freely estimated in the baseline model (Figure 3.2a). Then, this baseline model is compared with each of the $p$ models that remove one respective direct effect (Figure 3.2b).

To test nonuniform DIF, the MACS baseline model freely estimates all the loadings and intercepts in each group (Figure 3.3a). This model is compared with each of the $p$ models that constrain one respective loading to be equal across groups (Figure 3.3b). Then, to test uniform DIF, the invariant loadings are constrained to be equal and this model is compared with each of the $q$ models in which one respective intercept is constrained to be equal across groups (Figure 3.3c).*

* The constrained-baseline strategy is similar to the itop-downi approach for assessing scale-level invariance in that it starts with a model that imposes the most restrictive (or full) metric or scalar invariance. In contrast, the free-baseline strategy has similarities with the "bottom-up" approach in that it starts with the least restrictive (or partial) metric or scalar invariance model. For more details on these two approaches, see Welkenhuysen-Gybels and van de Vijver (2001).



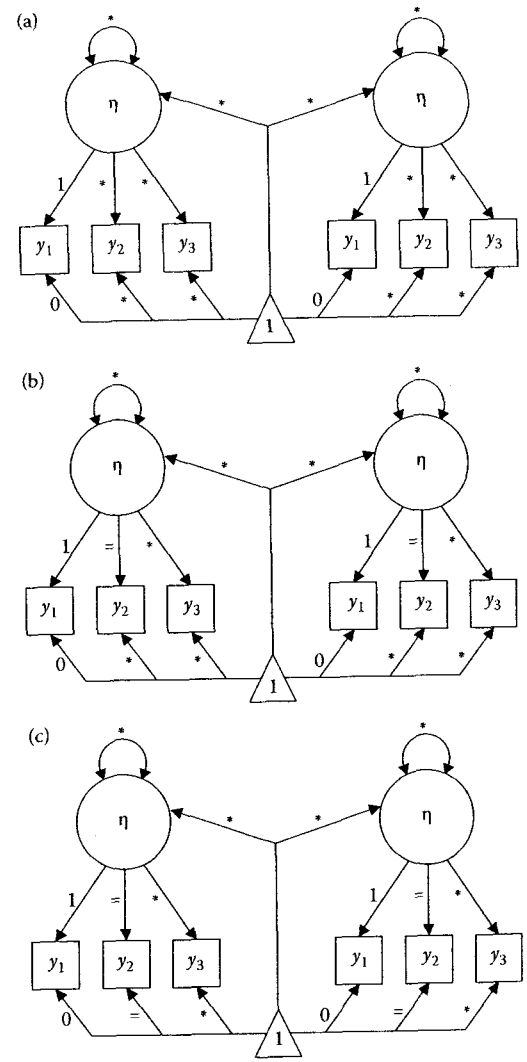



**FIGURE 3.3**
The free-baseline strategy for MACS DIF analysis. (a) Free-baseline model. (b) Model of a single restrictive loading. (c) Model of a set of restrictive loading and intercept. 3.3A and C depicts three nested MACS models. This example illustrates a simple case, in which (a) the scale includes three items, (b) only the second item exhibits nonuniform DIF, and (c) the marker-variable scaling method is used.

Regardless of which analytic strategy is used, the LR test is most frequently used to test DIF. Although empirical sampling distributions for other global fit indices have been provided (e.g., Cheung & Rensvold, 2002; Meade et al., 2008), there is no standard against which a researcher can compare the changes in global fit indices in order to test factorial invariance at the item level.

### 3.3.4 Previous Simulation Studies

Empirical evaluations of the SEM DIF analyses are scant. Recently, Stark et al. (2006) found that the constrained-baseline strategy is suitable for testing DIF only when there is no DIF item in a scale; Type I error was considerably inflated especially in uniform DIF cases. In contrast, the free-baseline strategy works fairly well; power was high enough, while Type I error was near or below the nominal alpha value. They also found that Type I error could be decreased substantially by using the Bonferroni-corrected LR test in large sample, large DIF cases. Similarly, Hernández and González-Romá (2003) showed that the constrained-baseline MACS analysis provided reasonable Type I error and power for detecting uniform DIF but power was not acceptable in the nonuniform DIF case.

For MIMIC DIF analysis, Finch (2005) reported that using the constrained-baseline strategy, Type I error was acceptable and power was very close to 1 unless a scale was relatively short and three-parameter logistic IRT model underlay the data. Using the free-baseline strategy, Woods (2009) also found that Type I error was controlled at the nominal alpha level, and power was reasonable when the focal group's sample size was equal to or greater than 100 (with reference group's sample size equal to or greater than 500).

### 3.3.5 Methodological Issues

The SEM DIF analyses, which are simple variations of the idea of partial factorial invariance (Raju et al., 2002), involve some methodological issues to be resolved in practice. Generally, the scaling method does not change the conclusions about overall model fit or the tests for omnibus scale-level invariance. However, when a researcher locates DIF in a scale after metric or scalar invariance has been rejected, a potential problem arises. Specifically, different scaling methods can lead to different conclusions of



DIF analysis because this post-hoc analysis relies on an examination of individual parameters.

To scale a latent construct in MACS DIF analysis, researchers conventionally fix an item's loading and intercept to be equal across groups (i.e., marker-variable method). This marker variable is termed an *anchor* in the DIF literature. In the case of MIMIC DIF analysis, researchers use an anchor or anchor set to which no direct effect is estimated. These scaling approaches essentially assume that the anchor set is truly invariant. If an invariant anchor set cannot be guaranteed or a researcher arbitrarily chooses an anchor set, other parameter estimates may be biased against invariance (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2005). Indeed, Stark et al. (2006) found that a biased anchor set severely inflated Type I error of the MACS DIF analysis. Finch (2005) and Navas-Ara and Gómez-Benito (2002) also reported that MIMIC DIF analysis was adversely affected by a biased anchor set.

A variety of empirical solutions for choosing an invariant anchor set have been proposed in the literature (e.g., Cheung & Rensvold, 1999; Christensen, MacKinnon, Korten, & Jorm, 2001; Fleishman et al., 2002; González-Romá, Tomás, Ferreres, & Hernández, 2005; MacKinnon et al., 1999; Stark et al., 2006). However, such solutions necessarily increase the number of nested-model comparisons, which requires setting a more conservative alpha level. Even when the alpha level has been adjusted, Type I error is severely inflated if no item appears to be invariant.

### 3.3.6 Current Simulation Studies

In a series of simulation studies, we included scaling method as a study condition, along with other conditions commonly examined in the DIF literature. For MACS DIF analysis, three different scaling methods (i.e., marker-variable, fixed-factor,* effects-coded; see *Method of Scaling* in Part I of this chapter) were examined using the free-baseline strategy as we illustrated previously (see Figure 3.3). Because different analytic strategies have not been empirically compared for MIMIC DIF analysis, the three scaling methods were combined with two analytic strategies

* When evaluating scale-level invariance, the variance and/or mean of a construct are freely estimated in one group. In contrast, when evaluating item-level invariance, they are constrained to equality across groups. In other words, the configural invariance model is used as the baseline model when examining nonuniform DIF and uniform DIF.

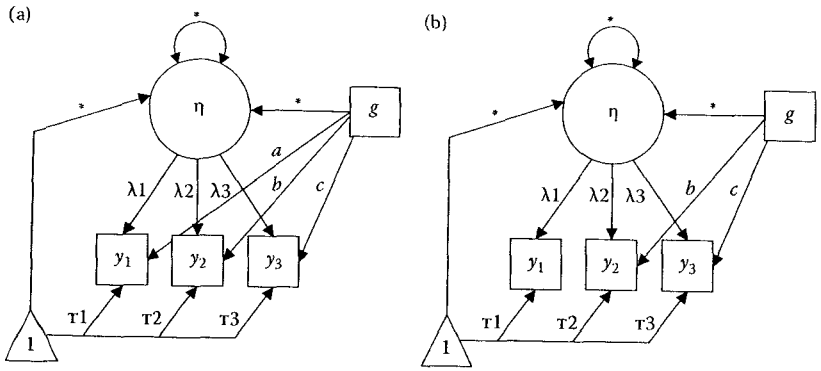



**FIGURE 3.4**
The effects-baseline strategy for MIMIC DIF analysis. (a) Free-baseline model. (b) Model of a single restrictive direct effect. 3.4A and B depicts two nested MIMIC models. This example illustrates a simple case, in which (a) the scale includes three items and (b) the effects-coded scaling method is used ($\Sigma_i^p \lambda_i = p$, $\Sigma_i^p \tau_i = 0$). A set of regression paths from the covariate $g$ to the items ($a$, $b$, $c$ or $b$, $c$) averages 0 ($\Sigma_i^p b_{ik} = 0$) in the effects-baseline strategy. Note that no anchor set is required for the purpose of scaling.

(i.e., constrained-baseline, free-baseline; see *Analytic Strategies for Testing DIF* in this chapter). We also tested a variant of the free-baseline strategy in which all possible direct effects are estimated as an optimal balance around a certain value (i.e., $\Sigma B_k = 0$). This *effects-baseline strategy* is theoretically appealing because there is no need for an anchor set. Figure 3.4 depicts the effects-baseline strategy used for MIMIC DIF analysis.

For both MACS and MIMIC DIF analyses, we also considered the biased anchor as a study condition. In addition, we used four different criteria for rejecting the assumption of partial factorial invariance; uncorrected and corrected critical $p$ values for the LR test ($p = .05/n$, where $n$ is the number of nested-model comparisons; Stark et al., 2006)* and ΔCFI values of −0.01 (Cheung & Rensvold, 2002) and −0.002 (Meade et al., 2008). The study conditions considered in the current simulation studies are presented in Table 3.4. More details on the current simulation studies and outcomes are available in Lee (2009, in preparation).

* For example, in Figure 3.3, two nested-model comparisons are possible for testing nonuniform or uniform DIF. Thus, the corrected critical $p$ value equals .025 (=.05/2).



**TABLE 3.4**
The Study Conditions Examined in the Current Simulation Studies

| Item Response (Scale Point) | Test Length | Sample Size ($N_F : N_R$) | Type of DIF in Anchor Item | Type of DIF in Target Item | DIF Amount | Scaling Method | Analytic Strategy[a] |
|---|---|---|---|---|---|---|---|
| Binary (2) | 6 items | 100 : 900 | Uniform | Uniform | No | Marker-variable | Constrained-baseline |
| Ordinal (5) | 12 items | 250 : 750 | Nonuniform | Nonuniform | Small | Fixed-factor | Free-baseline |
| | | 500 : 500 | | | Large | Effects-coded | Effects-baseline |

*Note:* The MACS study had a 2 × 2 × 2 × 3 × 2 × 2 × 3 × 3 × 1[a] factorial design (432 cells) and the MIMIC study had a 2 × 2 × 3 × 2 × 2 × 2 × 3 × 3 × 3[a] factorial design (1296 cells). Five hundred replications were made within each cell.

#### 3.3.6.1 Simulation Results for MACS DIF Analysis

In terms of Type I error and power, the three scaling methods used for MACS DIF analysis yielded different outcomes of testing (large) DIF. The conventional marker-variable method tested DIF effectively in most conditions. Some exceptions were the conditions when nonuniform DIF (loading invariance) was located in a relatively short test (e.g., 6 items). Similarly, the fixed-factor method detected uniform DIF (intercept invariance) quite well especially when DIF was introduced in the ordinal responses. When nonuniform DIF was present in the ordinal responses, this scaling method performed marginally well unless groups differed largely in size (e.g., $N_F = 100$ vs. $N_R = 900$). In contrast, the effects-coded method was tenable only for testing uniform DIF in a short test.

We found that a biased anchor greatly deteriorated the accuracy of testing nonuniform DIF for the marker-variable and effects-coded methods. That is, neither of these scaling methods was suitable for detecting DIF in almost all conditions.

Generally, the use of ΔCFI test greatly decreased Type I error as well as power. When used with the ΔCFI value of −0.002, the fixed-factor method worked well for testing uniform DIF unless groups differed greatly in size. With the same criterion, however, the marker-variable method was tenable for testing uniform DIF only when the anchor was not contaminated by the same type of DIF. In contrast, Bonferroni-correction on the LR test statistic did reduce Type I error, while retaining reasonable power (i.e., > 0.80) to detect both nonuniform and uniform DIF. Thus, the use of Bonferroni-correction is strongly recommended when using MACS analysis for DIF detection. For example, the fixed-factor method detected uniform DIF reasonably well in almost all conditions, including the biased anchor item.

#### 3.3.6.2 Simulation Results for MIMIC Analysis

As mentioned previously, MIMIC DIF analysis is not applicable to tests for nonuniform DIF because it presumes equal loadings across groups. Supporting this limiting assumption, power for detecting nonuniform DIF was not satisfactory in all conditions. Thus, our discussion is limited to the cases that, if present, only uniform DIF appears in a target.



Contrary to the case of MACS DIF analysis, scaling method had no impact on the accuracy of the MIMIC analysis for DIF detection. This finding was not surprising because DIF is determined by the significance of the direct effect estimate, not the invariance of the loading or intercept estimates as in the MIMIC DIF analysis.

Generally, each of three analytic strategies effectively detected uniform DIF. Type I error was below or near the nominal alpha value except in a few conditions (e.g., the constrained-baseline strategy with binary responses). Unless group sizes were largely different, power was satisfactory (i.e., > 0.80) in all conditions.

We found that the accuracy of MIMIC DIF analysis was considerably degraded by the presence of DIF in the anchor item. That is, none of the three analytic strategies was tenable for testing DIF when the anchor was biased by uniform DIF. Nevertheless, Type I error was substantially reduced by using the Bonferroni-corrected LR test. Consequently, the constrained-baseline strategy performed marginally well even with the biased, uniform DIF anchor. When the anchor had nonuniform DIF, the free-baseline strategy performed fairly well regardless of whether the Bonferroni-correction was used or not. In contrast, the ΔCFI tests markedly decreased power, making MIMIC DIF analysis useful in only a few conditions. Thus, if one cannot guarantee an anchor set devoid of uniform DIF, we recommend using the constrained-baseline strategy with the Bonferroni-corrected LR test for uniform DIF.

#### 3.3.6.3 Summary of Simulation Results

Our simulation results indicate that MACS analysis for DIF detection should be conducted using the fixed-factor scaling method. This method consistently outperformed the marker-variable and effects-coded scaling methods. For MIMIC DIF analysis, the scaling method had no impact, but the analytic strategy did. That is, either the free-baseline or the effects-baseline strategy effectively identified uniform DIF when the anchor set was DIF-free. When the anchor set had nonuniform DIF, the free-baseline strategy outperformed the other two strategies. In contrast, when the anchor set had uniform DIF, the constrained-baseline strategy performed the best. Finally, for both MACS and MIMIC DIF analyses, the Bonferroni-correction for nested-model comparisons should be considered to improve

the accuracy of these analyses, particularly when a DIF-free anchor set has not been established.

#### 3.3.6.4 Limitations

There are several weaknesses that require readers to interpret and generalize our simulation results with caution. First, maximum likelihood (ML) estimation, which assumes normality of the measured variables, was used with binary and ordinal responses. Lubke and Muthén (2004) noted that ML estimation can lead to erroneous invariance detection when used with categorical measured variables without accounting for their nonnormality. Second, no missing values were assumed in the responses although conclusions of any DIF analysis likely depend on the amounts and the patterns of missing values. Finally, sample sizes were selected so as to represent those often observed in psychological assessment. However, in some cases, smaller samples (i.e., less than 100) may be encountered especially with low-incidence groups.

## 3.4 CONCLUSIONS

Factorial invariance is a critical concern in cross-cultural research. Although researchers in this field have applied different methodologies to this issue, SEM has offered an integrative framework in which factorial invariance can be evaluated at both scale and item level. For example, MACS and MIMIC analyses reflect general IRT concepts, still accounting for measurement error in the responses and offering a variety of flexible options (e.g., multiple latent constructs, more than two groups, categorical or continuous covariates). The empirical findings illustrated in this chapter bring up some methodological issues and recommendations to be considered when a researcher conducts DIF analysis using SEM.

In a series of simulation studies, we found that statistically equivalent scaling methods did not provide identical outcomes when MACS analysis was used for testing DIF. We recommend using the fixed-factor scaling method (see *Methods of Scaling* in Part I of this chapter). If a test to be analyzed is relatively short, the effects-coded method may be considered for



testing uniform DIF. The scaling method does not impact the accuracy of MIMIC DIF analysis, but its analytic strategies may lead to different conclusions about DIF. Either the free-baseline or effects-baseline strategy is recommended for testing uniform DIF under favorable conditions such as comparable group sizes and a DIF-free anchor set (see *Analytic Strategies for Testing DIF* in this chapter). In less than favorable conditions, we recommend using the MACS DIF analysis.

An important issue in testing factorial invariance is the presence of DIF in the anchor set. Researchers have shown that having bias in the anchor set adversely affects invariance testing (Cheung & Rensvold, 1999; Finch, 2005; Navas-Ara & Gómez-Benito, 2002; Stark et al., 2006). Supporting this literature, we found that a biased anchor severely degraded Type I error and power of SEM DIF analyses. Nevertheless, our empirical results suggest a couple of possibilities to ameliorate these problems. That is, if used with the fixed-factor scaling method, Bonferroni-corrected LR test, and comparable large samples, the MACS DIF analysis would be a nearly fail-safe methodology for testing DIF, even when a designated anchor set is not readily available. Similarly, if accompanied by the Bonferroni-corrected LR test, constrained-baseline MIMIC analysis would produce accurate conclusions about DIF.

The critical ΔCFI values that have been suggested for scale-level invariance testing (i.e., −0.01, −0.002) were not optimal criteria for testing item-level invariance or DIF. In our simulation studies, they markedly reduced power of MACS and MIMIC DIF analyses. This is consistent with the results from the previous simulation study conducted by French and Finch (2006). They found that despite the fact that the ΔCFI test (of −0.01) has comparable power to the LR test (at .01 alpha level) for testing metric invariance in some conditions, this criterion rarely performs as well for detecting noninvariance (i.e., nonuniform DIF) of a single item.* Taken together, future efforts are needed to empirically examine the potential criterion values under a variety of DIF conditions (e.g., sample size, number of items, proportion of DIF in a scale, etc.). These criterion values should be independent from the overall fit of the baseline model, should not be influenced by model complexity, and should not be redundant with

* In supplementary analyses, we found that the critical values of RMSEA and SRMR suggested by Chen (2007) are not suitable for DIF analysis as well. When used to detect nonuniform and uniform DIF under our simulated conditions, generally they inflated Type I error above the nominal alpha level and/or provided very low power.

other fit indices (Cheung & Rensvold, 2002). Until optimal criterion values become available, we recommend using the Bonferroni-corrected LR test for testing DIF.

With regard to categorical measured variables, the use of alternative test statistics and estimation methods would provide more reliable DIF analyses. For example, the Satorra–Bentler (SB) $\chi^2$ incorporates a scale correction, taking into account the hypothesized model and the distributional characteristics of the data (1988). Satorra and Bentler (2001) further demonstrated how to calculate SB $\Delta\chi^2$ and corresponding degrees of freedom suitable for nested-model comparisons. Another alternative is to use a robust estimation method such as weighted least square (WLS) and robust WLS (RWLS).* These methods use polychoric correlations, item means, and weight matrices to produce an asymptotic covariance matrix of measured variables, which in turn is used to estimate the loading and intercept parameters (Muthén & Asparouhov, 2002).

Combining these methodological issues and recommendations, we suggest general procedural guidelines for testing factorial invariance in a flowchart (see Figure 3.5). This testing procedure proceeds in three stages. In the first stage, omnibus metric invariance of a scale is evaluated (see *Testing Factorial Invariance* in Part I of this chapter). If metric invariance holds, then omnibus scalar invariance of the scale is evaluated in the second stage. The ΔCFI test of −0.01 (or −0.002 for high-stakes testing environments) is recommended for assessing scale-level invariance. If it is appropriate to use ML estimation, the conventional LR test will be a comparable, or better, choice (see French & Finch, 2006). Because the scaling method generally does not change conclusions about scale-level invariance, any scaling method is applicable.

If metric invariance is rejected, locating the source of noninvariance would occur within the first stage. Nonuniform DIF is examined in each item, one at a time by conducting the free-baseline MACS analysis with the fixed-factor scaling method and Bonferroni-corrected LR test (see Figure 3.3). After flagging nonuniform DIF items, loading parameters for

* In fact, WLS and RWLS are not recommended in some cases. For example, Flora and Curran (2004) noted that the WLS $\chi^2$ is inflated, as are the parameter estimates, whereas their standard errors are negatively biased. Also, French and Finch (2006) found that the RWLS LR test provides very low power for testing scale-level metric invariance.



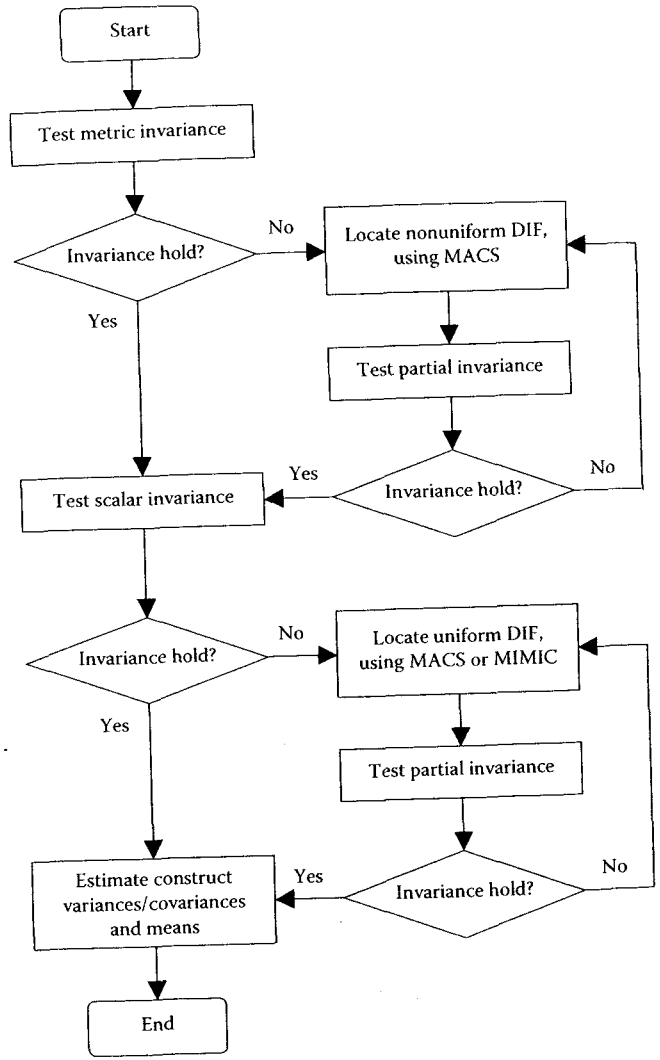



**FIGURE 3.5**
The procedure for testing factorial invariance in the SEM framework.

the DIF items are allowed to be freely estimated in each group and remain unconstrained throughout the next stages (i.e., partial metric invariance).

In the second stage, either MACS or MIMIC analysis is used to identify the items exhibiting uniform DIF. It should be noted that, because MIMIC analysis presumes equal loadings across groups, this analysis should be

avoided if scale-level metric invariance has not been established in the first stage. If metric invariance holds, one can use the constrained-baseline MIMIC analysis with the Bonferroni-corrected LR test to detect uniform DIF in each item, one at a time. Regardless of whether full or partial metric invariance holds, one can also use the free-baseline MACS analysis with the fixed-factor scaling method and the Bonferroni-corrected LR test.

There is some debate as to what minimum number of items should be invariant. For example, as a conservative approach to employing partial metric invariance, Vandenberg and Lance (2000) recommended that loading constraints should be relaxed for only a minority of items. In contrast, Steenkamp and Baumgartner (1998) suggested that at least two (loading- and intercept-) invariant items are sufficient for meaningful group comparisons. Although we agree with the recommendation of Steenkamp and Baumgartner (1998), the choice of a minimum number of invariant items must remain the prerogative of a researcher. This choice should be based on empirical evidence as well as practical considerations (Vandenberg & Lance, 2000).

After locating DIF items, further invariance tests (e.g., construct variances/covariances, construct means) may continue in the third stage. A baseline model should maintain the constraints of partial metric and scalar invariance that have been supported in the first two stages. When used in the multiple-group case, the effects-coded scaling method provides some preferable features (see Little et al., 2006). For example, because the scale of a construct is optimally weighted by all of its indicators in the effects-coded method, this method would be more useful in practice than the fixed-factor method in which the scale is defined by a single, arbitrarily chosen anchor. Furthermore, in MACS analysis, when invariance constraints are placed on the loadings and intercepts, the effects-coded method provides the scale of a construct within each group, which is not the case with the fixed-factor method. Accordingly, we recommend using the effects-coded scaling method when testing invariance of construct parameters.

Upon completing the illustrated testing procedure, a researcher may determine the "biasedness" of the DIF items through subsequent empirical and content analyses (Zumbo, 1999). As noted previously (see *Differential Item Functioning* in Part I of this chapter), only when observed group differences are attributable to the construct-irrelevant group characteristics



can DIF be considered item bias. Because groups in cross-cultural research often differ in various demographic and socioeconomic characteristics, these post-hoc analyses are strongly recommended to accomplish valid, meaningful group comparisons.

## REFERENCES


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bagozzi, R. P. (1994). The effects of arousal on the organization of positive and negative affect and cognitions: Application to attitude theory. *Structural Equation Modeling, 1*, 222–252.

Birnbaum. A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287–321.

Camilli, G., & Shepard, L. A. (1994). *Measurement methods for the social sciences series: Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaptation-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169–199.

Chen, C., & Anthony, J. C. (2003). Possible age-associated bias in reporting of clinical features of drug dependence: Epidemiological evidence on adolescent-onset marijuana use. *Addiction, 98*, 71–82.

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

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.

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

Christensen, H., MacKinnon, A. J., Korten, A., & Jorm, A. F. (2001). The "common cause hypothesis" of cognitive aging: Evidence for not only a common factor but also specific associations of age with vision and grip strength in a cross-sectional analysis. *Psychology and Aging, 16*, 588–599.

de Beuckelaer, A., Lievens, F., & Swinnen, G. (2007). Measurement equivalence in the conduct of a global organizational survey across countries in six cultural regions. *Journal of Occupational and Organizational Psychology, 80*, 575–600.

Dholakia, N., Firat, A. F., & Bagozzi, R. (1980). The de-Americanization of marketing thought: In search of a universal basis. In C. Lamb & P. Dunne (Eds.), *Theoretical developments in marketing* (pp. 25–29). Chicago, IL: American Marketing Association.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95*, 134–135.

Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research, 31*, 419–439.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278–295.

Fleishman, J. A. (2005). Using MIMIC models to assess the influence of differential item functioning. Retrieved from http://outcomes.cancer.gov/conference/irt/fleishman.pdf

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning. *Medical Care, 41*, 75–86.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences, 57*, 275–283.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491.

French, B. F., & Finch, H. (2006). Confirmatory factor analytic procedures for determination of measurement invariance. *Structural Equation Modeling, 13*, 378–402.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences, 49*, 251–264.

Gelin, M. N. (2005). *Type I error rates of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation*. Unpublished doctoral dissertation, University of British Columbia, Canada.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42*, 139–167.

González-Romá, V., Tomás, I., Ferreres, D., & Hernández, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups of adolescents? An application of the MACS model. *Structural Equation Modeling, 12*, 157–171.



Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiological Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences, 55B*, 273–282.

Grayson, D. A., & Marsh, H. W. (1994). Identification with deficient rank loading matrices in Confirmatory Factor Analysis: Multitrait-multimethod models. *Psychometrika, 59*, 121–134.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*, 78–94.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hernández, A., & González-Romá, V. (2003). Evaluating the multiple-group Mean and Covariance Structure analysis model for the detection of Differential Item Functioning in polytomous ordered items. *Psichotema, 15*, 322–327.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care, 44*, 124–133.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10*, 631–639.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–153.

Katsuya, T. (2007). Cross-cultural validity of self-construal scales: An investigation of differential item functioning using multigroup mean and covariance structure. *Japanese Journal of Behaviormetrics, 34*, 79–89.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16*, 215–224.

Lee, J. (2009). *Type I error and power of the MACS CFA for DIF detection: Methodological issues and resolutions*. Unpublished doctoral dissertation, University of Kansas.

Lee, J. *Type I Error and power of the MIMIC technique for DIF detection: Problem of biased anchor set and a recommended procedure*. Manuscript in preparation.

Leone, L., van der Zee, K., van Oudenhoven, J. P., Perugini, M., & Ercolani, A. P. (2005). The cross-cultural generalizability and validity of the Multicultural Personality Questionnaire. *Personality and Individual Differences, 38*, 1449–1462.

Liang, J. (2001). Assessing cross-cultural comparability in mental health among older adults. *Journal of Mental Health and Aging, 7*, 21–30.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53–76.

Little, T. D. (in press). Longitudinal structural equation modeling: Individual-difference panel models. New York: Guilford press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*, 59–72.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514–534.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27*, 372–379.

Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Rodgers, B. (1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences, 27*, 405–416.

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disability: A hybrid multigroup-MIMIC approach to factorial invariance and latent mean differences. *Educational and Psychological Measurement, 66*, 795–818

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15. New York, NY: Springer.

McDonald, R. P. (1999). *Test theory: Unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in test of measurement invariance. *Journal of Applied Psychology, 93*, 568–592.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223–237.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153–172). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479–515.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainder & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Los Angeles, CA: University of California and Muthén & Muthén.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407–419.

Muthén, B. O., & Lehman, J. (1988). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133–142.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment, 18*, 9–15.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*, 411–423.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen. (Expanded edition, 1980. Chicago: The University Chicago Press).



Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory Factor Analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Riordan, C. R., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*, 643–671.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, No. 17. New York, NY: Springer.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monographs*, No. 18. New York, NY: Springer.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 proceedings of the business and economics section* (pp. 308–313). Alexandria VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.

Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods, 6*, 169–215.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

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*, 1202–1306.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*, 78–90.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

van de Vijver, F., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17–24.

Wang, C. K. J., Liu, W. C., Biddle, S. J. H., & Spray, C. M. (2005). Cross-cultural validation of the Conceptions of the Nature of Athletic Ability Questionnaire Version 2. *Personality and Individual Differences, 38*, 1245–1256.

Welkenhuysen-Gybels, J., & van de Vijver, F. (2001). A comparison of methods for the evaluation of construct equivalence in a multigroup setting. *Proceedings of the annual meeting of the American Statistical Association*. Alexandria VA: American Statistical Association.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1–27.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*, 320–330.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

# 4

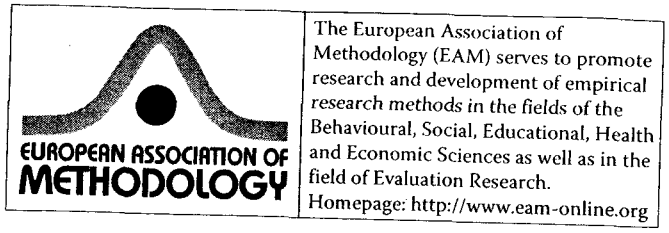
# Estimation and Comparison of Latent Means Across Cultures

*Holger Steinmetz*
University of Giessen

## 4.1 INTRODUCTION

One of the most often conducted kind of analysis in cross-cultural research is to compare the mean of some construct across two or more cultural populations. Although one of the long-term goals of cross-cultural research may be to understand cultural functioning with regard to underlying cultural dimensions (Hofstede, 1980; House, Javidan, Hanges, & Dorfman, 2002) or contextual factors, mean comparisons are an important first way to generate knowledge about cross-cultural differences.

Although the use of structural equation modeling has increased in the last decades, researchers still rely on traditional methods (e.g., analysis of covariance (ANOVA) and *t*-test) when comparing means. The typical procedure consists in aggregating items, for instance, from a questionnaire or telephone interview, to a composite score and comparing the composites' mean across the cultural samples. Therefore, although researchers are well aware that observed variables differ from latent variables (Borsboom, 2008), traditional analyses treat observed *means* as if they were equal to latent means. However, observed means cannot simply be equated with the latent mean of the underlying construct. As I will show later in more detail, the relationship between an observed mean and the latent mean is a function that contains two other important parameters, that is, the indicator intercept and the factor loading. Consequently, group differences on an observed composite can be solely attributed to a latent mean difference when the intercepts and loadings of the indicators are invariant (i.e., equal) across the groups.




The European Association of Methodology (EAM) serves to promote research and development of empirical research methods in the fields of the Behavioural, Social, Educational, Health and Economic Sciences as well as in the field of Evaluation Research. Homepage: http://www.eam-online.org

The purpose of the EAM book series is to advance the development and application of methodological and statistical research techniques in social and behavioral research. Each volume in the series presents cutting-edge methodological developments in a way that is accessible to a broad audience. Such books can be authored, monographs, or edited volumes.

Sponsored by the European Association of Methodology, the EAM book series is open to contributions from the Behavioral, Social, Educational, Health and Economic Sciences. Proposals for volumes in the EAM series should include the following: (1) Title; (2) authors/editors; (3) a brief description of the volume's focus and intended audience; (4) a table of contents; (5) a timeline including planned completion date. Proposals are invited from all interested authors. Feel free to submit a proposal to one of the members of the EAM book series editorial board, by visiting the EAM website http://eam-online.org. Members of the EAM editorial board are Manuel Ato (University of Murcia), Pamela Campanelli (Survey Consultant, UK), Edith de Leeuw (Utrecht University) and Vasja Vehovar (University of Ljubljana).

Volumes in the series include

**Davidov/Schmidt/Billiet:** Cross-Cultural Analysis: Methods and Applications, 2011

**Das/Ester/Kaczmirek:** Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies, 2011

**Hox/Roberts:** Handbook of Advanced Multilevel Analysis, 2011

**De Leeuw/Hox/Dillman:** International Handbook of Survey Methodology, 2008

**Van Montfort/Oud/Satorra:** Longitudinal Models in the Behavioral and Related Sciences, 2007

# Cross-Cultural Analysis
## Methods and Applications

Edited by

Eldad Davidov
University of Zurich, Switzerland

Peter Schmidt
University of Marburg, Germany
Professor Emeritus, University of Giessen, Germany

Jaak Billiet
University of Leuven, Belgium

Routledge
Taylor & Francis Group
New York London