5

Biased Latent Variable Mean Comparisons due to Measurement Noninvariance: A Simulation Study

Alain De Beuckelaer
Ghent University and Radboud University Nijmegen

Gilbert Swinnen
Hasselt University

5.1 INTRODUCTION

Making valid comparisons of latent variable (LV) mean scores[1] across groups is anything but a trivial task as it relies on stringent measurement invariance conditions. Many authors (e.g., Little, 1997; Meredith, 1993; Van de Vijver & Leung, 1997) have firmly stated that in order for such comparisons to be valid, *all* factor loadings as well as *all* indicator intercepts need to be invariant across groups (i.e., they are required to exhibit *scalar* invariance, also referred to as *full score* invariance across groups; Van de Vijver & Leung, 1997, p. 144).

In contrast, some authors have promoted the use of less stringent measurement invariance conditions. For instance, according to Alwin and Jackson (1981) the equality of *all* factor loadings (i.e., an invariance condition referred to as "metric invariance" across groups) would

[1] In this study, we deal only with the comparison of LV means across groups. Cross-group comparisons of structural relationships between observed or LVs are not considered. Such comparisons require only the factor loadings of LV indicators to be identical across the groups (i.e., metric invariance across groups) involved in the comparison (see, for instance, Van de Vijver & Leung, 1997, De Beuckelaer et al., 2007).

be a sufficient condition. Still others (Byrne, 1989; Byrne, Shavelson, & Muthén, 1989; Marsh & Hocevar, 1985; Muthén & Christofferson, 1981; Reise, Widaman, & Pugh, 1993) have argued that only a subset of all factor loadings (i.e., "partial metric invariance" across groups) would be a sufficient condition. The willingness to accept less stringent measurement invariance conditions (when comparing latent variable mean [LV] mean scores) may be caused by a growing belief that (survey) measurement instruments can hardly ever be totally invariant across groups (e.g., Byrne & Watkins, 2003; Horn, McArdle, & Mason, 1983). As a matter of fact, many large-scale international studies (e.g., Davidov, 2008; Davidov, Schmidt, & Schwartz, 2008; De Beuckelaer, Lievens, & Swinnen, 2007) have shown that survey instruments typically do not exhibit scalar invariance across a large number of nations—a group delimiter that is very frequently used in cross-cultural comparative research (e.g., Schaffer & Riordan, 2003).

A major point of concern is that much supportive evidence exists to support the claim that many researchers doing cross-cultural comparative research fail to test formally for possible sources of noninvariance of measurement parameters across groups (for instance, see Cheung & Rensvold, 1999; He, Merz, & Alden, 2008; Schaffer & Riordan, 2003; Vandenberg & Lance, 2000; Williams, Edwards, & Vandenberg, 2003). As such, many researchers run a high risk of drawing erroneous conclusions from a cross-group comparison of LV mean scores. For this reason it makes sense to examine the consequences of *falsely* assuming that the measurement instrument used exhibits a high level of measurement invariance, namely, scalar invariance across groups.

As argued by Vandenberg (2002), there is a growing need for research to help understand the sensitivity of the analytical procedures when making cross-cultural comparisons (e.g., multigroup mean and covariance structure (i.e., multigroup MACS) analysis, and the extent to which less stringent measurement invariance conditions (if unnoticed) can affect LV mean comparisons across groups. One way to adequately explore the extent to which the requirement of scalar invariance can be relaxed (without threatening the validity of cross-group comparisons) is to make use of carefully designed simulation research. Such simulation research may (at least potentially) provide valuable empirical evidence to justify the adoption of less stringent measurement invariance conditions. Except for an older study by Kaplan and George (1995) that did not include invariance conditions relating to indicator intercepts, we did not find any published

AU: Please check the shortened running head.

simulation study that enables us to infer to what degree the measurement invariance conditions across groups can be relaxed without running the risk of drawing erroneous conclusions from cross-group comparisons of LV mean scores.

In this chapter, we will try to fill this gap in the literature by presenting a simulation study that investigates the extent to which noninvariance of factor loadings and indicator intercepts may lead to false statistical conclusions in terms of (the reported significance of) LV mean differences across groups. The next sections will elaborate on the method of research and the analysis strategy, the presentation of some detailed results, and the general conclusions.

5.2 METHOD

AU: Please add a reference with full publication information for Chou & Bentler, 1995.

AU: Please add a reference for Curran et al., 1996 with full publication information.

In this simulation study, Satorra and Bentler's scaled chi-square statistic[2] was used because of its superior performance in earlier simulation studies (Chou & Bentler, 1995; Curran et al., 1996; Hu et al. 1992; Olsson et al. 1995). The design characteristics of this simulation study resemble, at least to some extent, the design characteristics of the simulation study by Kaplan and George (1995). As will be explained in the next paragraphs however, there are also some important differences.

AU: Please add a reference with full publication information for Olsson et al., 1995.

AU: Please add a reference with full publication information for Hu et al., 1992.

5.2.1 Experimental Design Factors Included

In this simulation study, noninvariance conditions were represented by just one LV indicator (i.e., always the second LV indicator) with a noninvariant factor loading applied to the LV indicator and/or a noninvariant indicator intercept.

The following design factors were used in the simulation study:

- Number of indicators for the LV (design factor 0)
- Type of distribution of the indicators (design factor 1)
- Sample size in the different groups (design factor 2)

[2] More precisely, it is the mean- and variance-adjusted chi-square statistic with robust standard errors (see Muthén & Muthén, 1999), which is used in this simulation study.

- LV mean difference between groups at population level (design factor 3)
- Noninvariance of factor loadings and indicator intercepts (design factor 4 and design factor 5, respectively)

The simulation was set up using a full-factorial experimental design so that all possible combinations are represented in the simulation study.

We make a distinction between two groups of design factors: "side design factors" and (measurement) "noninvariance design factors." Side design factors relate to characteristics of the data that are not related to the measurement (non)invariance of the data across groups. They include (see Table 5.1):

- A fixed number of two groups to be involved in the LV mean comparison
- The number of indicators used to measure the LV under study (either 3 or 4)
- The statistical or empirical distribution underlying LV indicator scores (standard normal distribution)
- Discrete 5-point response scales showing either a unimodal left-skewed distribution or a symmetric bimodal distribution[3]
- The size of the sample in the two groups involved in the LV mean comparison.

Next, measurement noninvariance design factors create noninvariance conditions as they determine the degree of measurement noninvariance of the noninvariant indicator (i.e., always the second indicator of the LV). Specifications regarding the levels of individual design factors are presented in Table 5.1. This table contains essential information about the simulations conducted. As the Results section often makes reference to abbreviations indicating simulated conditions that are specified in Table 5.1 (see further on in this chapter), a technically interested reader may benefit greatly from copying this table prior to examining in-depth the results discussed in that section. By doing so, technical

[3] Especially because of these scales, the present simulation study is to be conceived as more realistic than the study by Kaplan and George (1995).

TABLE 5.1

Values Specified for the Design Factors

Design Factors	Specification of Levels/Values Assigned	Further Remarks
Side design factors		
Number of groups	Always 2 groups to be compared	
F0: Number of indicators for the LV (i.e., design factor 0, shortly F0)	3 (F0 = 1) or 4 (F0 = 2)	
F1: Distribution of indicators (i.e., design factor 1, shortly F1)	*Continuous data following a standard normal distribution* (F1 = 1) as well as two discrete response scales with five ordinally ordered response categories following either a *unimodal, left-skewed distribution* (F1 = 2) or a *symmetric bimodal distribution* (F1 = 3); thresholds have been specified such that the distribution (in % of cases) across category points is as follows: unimodal, left-skewed: 8% [1]; 12% [2] 15% [3]; 40% [4]; 25% [5], and symmetric bimodal: 15% [1 and 5], 30% [2 and 4], 10% [3].	The study by Kaplan and George (1995) made use of the standard normal distribution only (also replicated in our study if F1 = 1); the two empirical distributions (e.g., unimodal, left-skewed [F1 = 2] and bimodal symmetrical [F1 = 3] based on five-ordered response categories) as used in our simulation study are often encountered in applied research making use of Likert-type of scales.
F2: Sample sizes (i.e., design factor 2, shortly F2)	Group sample size varies between 200 and 750; $F_2 = 1$: $N_1 = N_2 = 200$; $F_2 = 2$: $N_1 = N_2 = 300$; $F_2 = 3$: $N_1 = N_2 = 400$; $F_2 = 4$: $N_1 = N_2 = 500$; $F_2 = 5$: $N_1 = N_2 = 750$; $F_2 = 6$: $N_1 = 200$, $N_2 = 400$ (N_i means sample size in group i; i = 1,2; there are always 2 groups to be compared)	Except for one condition (i.e., (F2 = 6) all experimental conditions are balanced (i.e., having an equal number of observations in both groups).

(*Continued*)

TABLE 5.1 (Continued)

Values Specified for the Design Factors

Design Factors	Specification of Levels/Values Assigned	Further Remarks
Side design factors		
F3: Latent variable mean difference (i.e., design factor 3, shortly F3)	The LV mean in group 1 (G1) is always 0.00; the LV mean in group 2 (G2) is –0.30 (F3 = 1); –0.15 (F3 = 2); 0.00 (F3 = 3); 0.15 (F3 = 4); and 0.30 (F3 = 5); conditions 4 and 5 (i.e., F3 = 4 and F3 = 5) represent positive discrepancy cases (i.e., conditions in which the LV mean in the second group is larger than the LV mean in the first group). In negative discrepancy cases (F3 = 1 and F3 = 2), the reverse condition applies; an absolute difference of 0.15 (F3 = 2 and F3 = 4) may be considered to be small, whereas an absolute difference of 0.30 (F3 = 1 and F3 = 5) may be considered to be large (adequate empirical evidence for the use of the qualifiers small and large can be provided by the first author using detailed descriptive results of correct statistical conclusions obtained in the simulation study; see Results section).	Condition 3 (F3 = 3) serves as a means to assess the test's actual control over the type I error rate (i.e., falsely concluding that LV means are different); all other conditions help determine the test's statistical power.

Noninvariance design factors

F4: Noninvariance of factor loadings across groups (i.e., design factor 4, shortly F4)	In group 1 (G1) the factor loadings are always fixed as follows: $\lambda\_1 = 0.7$; $\lambda\_2 = 0.6$; $\lambda\_3 = 0.5$ (in 3-indicator conditions; see F0) or $\lambda\_1 = 0.7$; $\lambda\_2 = \lambda\_3 = 0.6$; $\lambda\_4 = 0.5$ (in 4-indicator conditions; see F0); in group 2 (G2) the factor loading of the second indicator (i.e., λ_2) varies as follows: F4 = 1: $\lambda\_2 = 0.4$ (indicator reliability = 0.24); F4 = 2: $\lambda\_2 = 0.6$ (indicator reliability = 0.41); F4 = 3: $\lambda\_2 = 0.8$ (indicator reliability = 0.56) The factor loadings of all other indicators are invariant across both groups.	Either the second indicator (i.e., λ_2) and/or its intercept are noninvariant across groups in noninvariance conditions. Condition 2 (i.e., F4 = 2) represents a condition in which no violation of invariance (across groups) occurs, at least not as far as factor loadings are concerned
F5: Noninvariance of indicator intercepts across groups (i.e., design factor 5, shortly F5)	The four experimental conditions are as follows: F5 = 1: Indicator 2 has an intercept in G2, which is identical to the intercept of indicator 2 in G1; F5 = 2: Indicator 2 has an intercept in G2, which is 0.15 higher than the intercept of indicator 2 in G1; F5 = 3: Indicator 2 has an intercept in G2, which is 0.30 higher than the intercept of indicator 2 in G1; F5 = 4: Indicator 2 has an intercept in G2, which is 0.45 higher than the intercept of indicator 2 in G1 The indicator intercepts of all other indicators are invariant across both groups.	Only the intercept of the second indicator may but does not have to be noninvariant. Condition 1 (i.e., F5 = 1) represents a condition in which no violation of invariance (across groups) occurs, at least not as far as indicator intercepts are concerned.

Notes: The notation Fx = k is used to refer to level number k for design factor x (e.g., F1 = 1); In all replications the variance of the LV is always fixed to one.

details concerning the nature of the simulated experimental conditions as displayed in Table 5.1 are readily available to the reader at all times.

5.2.1.1 Noninvariance and Invariance Conditions

As explained before, (measurement) noninvariance is caused by only one indicator, namely, the second LV indicator out of three or four. Depending on the particular condition, the second LV indicator may exhibit noninvariance due to a noninvariant factor loading across groups (F4 = 1 or F4 = 3 [i.e., level 1 or level 3 is specified for design factor 4]; see F4 in Table 5.1) and/or a noninvariant indicator intercept across groups (F5 = 2, 3 or 4; see F5 in Table 5.1). Alternatively, the second indicator may exhibit measurement invariance (i.e., if F4 = 2 and F5 = 1; see F4 and F5 in Table 5.1). Further on in this chapter we will use the notation "λ_2" and "Int_2" to refer to the factor loading and the indicator intercept of the second LV indicator, respectively. Whenever we refer to a group, we will add a suffix labeled "Gi" (i = 1 or 2).

The settings for factor loadings used in our study resemble the settings specified by Kaplan and George (1995). As indicated in Table 5.1, indicator reliabilities ranged between 0.24 (with a factor loading equal to 0.4) and 0.56 (with a factor loading equal to 0.8).[4] Differences in the indicator intercepts across groups varied between 0.00 and 0.45. The latter value of the indicator intercept represents a distance of nearly one-tenth of the length of the total scale consisting of five response categories (see Table 5.1).

AU: Please review footnote 4. Is this set correctly? There is a low line between lambda and number, is something missing?

One may reasonably expect that differences in indicator intercepts across groups are more harmful than differences in factor loadings when (estimated) LV mean scores are to be compared across groups. Differences in indicator intercepts will bias estimated LV mean scores equally for each observation (or person), whereas the bias resulting from differences in factor loadings really depends on the observation's (or person's) score on the underlying construct.

As mentioned before, we specified (corresponding) measurement invariance conditions in addition to noninvariance conditions. The major advantage of doing so is that measurement invariance conditions may serve as a natural benchmark (condition) against which the

[4] The reliability of the *i*th LV indicator is calculated as follows: 1 – (error variance / [λ_i^2 + error variance]). The error variance is always fixed to 0.51 in the simulation study.

(statistical) performance of the LV mean difference test may be evaluated in noninvariance conditions. Further details concerning exactly how this will be done are provided in the Analysis section.

5.2.1.2 Asymmetrical Structure of the Simulation Design

The specific noninvariance conditions, as specified in Table 5.1, demonstrate that the experimental design used has a structure that is asymmetrical. The factor loading of the (possibly noninvariant) second LV indicator is specified to be *smaller* than, equal to, or *larger* than the corresponding factor loading in group one. In contrast, the indicator intercept of this LV indicator in the second group (G2) is specified to be either equal to or *larger* than the corresponding indicator intercept in group one. As the smaller than condition is missing here, the asymmetric structure is entirely due to the experimental settings specified for the intercept of the (possibly noninvariant) second indicator of the LV.

Because of this asymmetry, the effect of unequal indicator intercepts across groups on the estimated size of the (absolute) difference in LV means across groups were different for positive and negative discrepancy cases (i.e., conditions in which the LV mean in group 2 [G2] is higher [*positive* discrepancy] or lower [*negative* discrepancy] compared to the LV mean in group 1 [G1]). In positive discrepancy cases (i.e., $\mu_{LV,G2} > \mu_{LV,G1}$), unequal indicator intercepts increase the estimated discrepancy between LV means. In negative discrepancy cases (i.e., $\mu_{LV,G2} < \mu_{LV,G1}$), the estimated discrepancy between LV means decreases due to the inequality of indicator intercepts across groups. Therefore, the inclusion of negative indicator intercepts in the simulation design (in addition to positive indicator intercepts) would only lead to duplicate information as some conditions with a positive discrepancy between LV means would be identical to some other conditions with a negative discrepancy between LV means.

5.2.2 Data Generation and Analysis Strategy

5.2.2.1 Data Generation

Multiple data files (i.e., 50) were generated for each experimental condition. Several software programs were written to run the simulations. These software programs took care of the data preparation and data extraction

tasks. The actual parameter estimations were provided by a dedicated software program, namely, Mplus (Muthén & Muthén, 1999). A more detailed description of all programs used and their functionality in the simulation process is available from the first author.

5.2.2.2 Data Analysis Strategy

The results from the simulation study were analyzed in two consecutive steps. These two steps are explained below.

Step 1: Correct and Incorrect Statistical Conclusions

Using the simulated data files, LV means were estimated for both groups. The estimation was carried out under the (possibly false) assumption that scalar invariance holds for all indicators across groups (i.e., imposing constraints regarding the equality of factor loadings and indicator intercepts across groups; i.e., imposing the scalar invariance model across groups onto the data).

The robust maximum likelihood procedure as implemented in the software Mplus (Muthén & Muthén 1999) was used to estimate the model parameters. To test whether the LV mean in G2 was identical to the LV mean in G1 (the latter one is fixed to zero in all simulations; see Table 5.1), a simple *z*-statistic (i.e., the estimated LV mean in the second group divided by its standard error) was used. Provided that the estimated LV mean in G2 is zero (i.e., that the null-hypothesis holds), the *z*-statistic follows a standard normal distribution, asymptotically. The correctness of the LV mean difference test (i.e., reject or don't reject the null-hypothesis) was then assessed using information about the true difference (if not zero) in the LV means (see Table 5.1, design factor 3). For each of the 50 replications of all experimental conditions the correctness of the statistical conclusion was flagged by a "not correct" [0]/"correct" [1] indicator. Incorrect decisions in the opposite direction (e.g., finding a significant positive discrepancy when the actual difference between LV means was negative) were not produced in this simulation study.

Next, the influence of the individual design parameters (see Table 5.1) on the *correctness* of statistical conclusions regarding the LV mean difference test across groups was assessed. Previous research (i.e., Kaplan & George 1995) has shown that the effect of the difference between LV means at population level (i.e., design factor 3) is dominant when compared to other

effects. This is quite obvious as the probability of finding a significant difference between LV means in two independent samples is directly related to the size of the difference between LV means at population level. This effect is, however, not relevant for the research problem at hand. The main research question is to evaluate the extent to which measurement noninvariance conditions (i.e., design factor 4 and 5) and certain side conditions (such as number of indicators, distribution of indicators, and sample sizes) bias LV mean comparisons across groups. Therefore, the LV mean difference at population level may be regarded as an extraneous factor. Consequently, the effects of all other design parameters were assessed separately for various levels of the LV mean difference at population level.

The design parameters were indicated by means of binary variables (i.e., 0/1 variables). The following notation was used: Fi_Dj with i representing the number identifying the design factor and j representing the number identifying the level specified for that design factor (see Table 5.1). So, F5_D1 means that the first level applies to design factor 5 (i.e. indicator intercept of indicator 2 is identical in G2 and G1; see Table 5.1).

We made use of the classification and regression tree (C&RTree) technique by Breiman, Friedman, Olshen, and Stone (1984) to identify the impact of the design parameters (i.e., levels of design factors) on the correctness of the statistical conclusion regarding the LV mean difference between groups. In these C&RTree analyses, the particular values of design parameters (as represented by a series of binary Fi_Dj variables; see above) serve as independent variables, while the binary indicator showing the correctness of the LV mean difference test across groups (i.e., not correct [0]/correct [1]) acts as the dependent variable. By optimizing a statistical criterion (i.e., the statistical significance of difference in percentage correct statistical conclusions), the C&RTree technique successively splits the entire sample into (sub)samples until at least one convergence criterion is reached. Just as in our study, a convergence or stop criterion is reached if statistical significance of the next candidate sample-split drops below a minimum level or, alternatively, the number of observations (or, in our study, replications of simulated conditions) in the sample to be split has dropped below a minimum value. Sample-splits can be made using a particular main effect (e.g., F2_D5, i.e., replications for which design factor 2 has value 5 versus replications for which design factor 2 has values other than 5; for details regarding the design factors see Table 5.1), or a particular interaction effect (e.g., F2_D5 * F1_D3, i.e., replications for which design

factor 2 has value 5 and design factor 1 has value 3 versus replications for which other combinations of the design factors 2 and 1 apply).

The results of a C&RTree analysis are graphically depicted in a tree-based structure, which is referred to as a C&RTree. An example of a C&RTree is provided in Figure 5.1.

Provided that none of the convergence criteria are met, a first sample-split is made based on one main effect or one interaction effect between (levels of) the design factors. The first sample-split in Figure 5.1 differentiates between replications using very large datasets (i.e., 5th level of design factor 5 applies [F2_D5 = 1]; see Table 5.1), and replications using smaller data sets (F2_D5 = 0). So, a main effect (F2_D5 = 0 or 1) is used to split the overall sample (with 48% correct statistical conclusions) into two (sub) samples with 31% (F2_D5 = 0) and 57% (F2_D5 = 1) correct statistical conclusions regarding the LV mean difference test, respectively. Among all candidate sample-splits (i.e., all main and interaction effects of levels of design parameters), the main effect F2_D5 = 0 or 1 leads to the largest possible difference between the percentage of correct statistical conclusions as obtained for both subsamples.

Next, a C&RTree analysis will determine whether or not further sample-splits can be made. To this end, some statistical evaluations are made for each of the two (sub)samples resulting from the first sample-split. In fact, for each subsample one statistical evaluation is made for all levels (or combinations of levels) of the design factors that have not been used in the first sample-split (or, in more general terms, higher up in the C&RTree). As far as the example is concerned (see Fig. 5.1), this means that the main effect

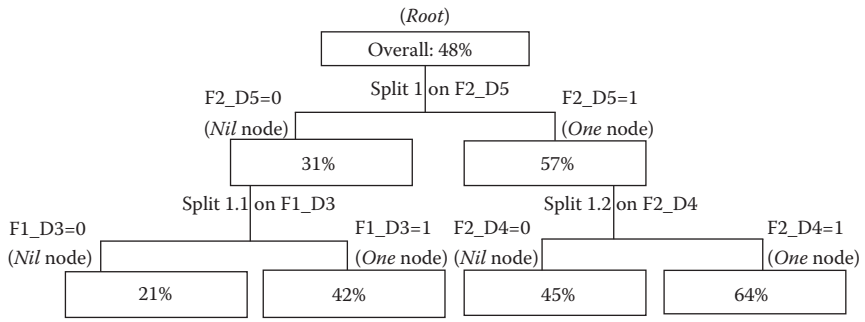

FIGURE 5.1
Percentage correct statistical conclusions (4-indicator cases; only large negative discrepancy cases [i.e., F3 = 1]).

F2_D5 = 0 or 1 cannot be used anymore as a splitting variable after it has been used as the variable determining the very first sample split. Provided none of the convergence criteria are met, each (sub)sample emerging from the first sample-split may be split again in two new (sub)samples, producing up to four different (sub)samples. In Figure 5.1 further sample-splits are made using the following two main effects: F1_D3 = 0 or 1, and F2_D4 = 0 or 1. Once again, these sample-splits lead to the largest possible difference between the percentages of correct statistical conclusions as obtained for both subsamples. This process of splitting (sub)samples continues until no further sample-splits can be made (i.e., when at least one of the convergence criteria is met).

In order to correctly interpret the results of C&RTree analyses, one may therefore rely on the following two basic principles:

Principle #1: The importance of the individual design parameters (i.e., levels of the design factors) in terms of predicting the correctness of the LV mean difference test is reflected by the sequence in which these sample-splits are made. In other words, sample-splits that are positioned lower down the C&RTree are, from a statistical point of view, somewhat less important than sample-splits that are positioned higher up the C&RTree.
Principle #2: All design parameters that have not been used anywhere in the C&RTree are, from a statistical point of view, of a lesser importance (i.e., they do not help very much in discriminating in terms of percentage of correct statistical conclusions regarding the LV mean difference test).

Step 2: Robust and Nonrobust Conditions

So far, the unit of analysis has been a replication of an experimental condition (i.e., 50 replications per experimental condition). To assess the robustness of the experimental conditions against violations of the scalar invariance assumption across groups (i.e., step 2), aggregated data are needed. In particular, the data of all replications need to be aggregated for every experimental condition.

The analysis strategy is to use the total number of correct statistical conclusions of invariance conditions as a reference against which the robustness of all (related[5]) noninvariance conditions is evaluated. Based on the

[5] Related noninvariance conditions are characterised by an identical LV mean difference between both groups and a noninvariant indicator having an unequal factor loading and/or indicator intercept across groups.

binomial distribution, a 99% confidence interval[6] is specified around the number of correct statistical conclusions of invariance conditions. If the number of correct statistical conclusions of a related noninvariance condition falls within this interval, the noninvariance condition is considered to be robust against violations of the scalar invariance assumption across groups. Otherwise, it is not considered to be robust. Based on such an analysis, all noninvariance conditions are flagged with a "not-robust" [0]/"robust" [1] indicator. In summary, the idea is to examine the decrease (or increase) in the number of correct statistical conclusions of noninvariance conditions using the number of correct statistical conclusions of invariance conditions as a benchmark (or reference condition).

C&RTree analyses are also used in this second step to determine the influence of the individual design parameters on the robustness of the experimental condition against violations of the scalar invariance assumption across groups. So, instead of using the statistical correctness of the LV mean difference test (in each replication of an experimental condition), the robustness of the experimental condition (i.e., across all replications of that experimental condition) is used as the dependent or criterion variable in these C&RTree analyses. As the unit of analysis is an experimental condition rather than a replication (of a particular experimental condition), a relatively small number of observations is available for these C&RTree analyses.

5.3 RESULTS

Step 1: Correct and Incorrect Statistical Conclusions

5.3.1 Descriptive Results

The percentage of correct conclusions regarding the LV mean difference test varies around 66% across all simulated conditions, regardless of the

[6] The specification of a confidence interval (CI) is always a somewhat arbitrary decision. Changing from a 99% CI to a 95% CI would not have had a substantial impact on the decisions regarding robustness/nonrobustness of the noninvariance condition (i.e., on average across all noninvariance conditions, less than one replication [i.e., 0.80; SD = 0.60] would have been classified differently).

TABLE 5.2

Percentage of Correct Conclusions Regarding the LV Mean Difference Test

Percentage of Correct Statistical Conclusions	Negative Discrepancy Cases		No Discrepancy Cases	Positive Discrepancy Cases	
	LV Mean in Group 2 (G2)				
	= –0.30 (F3 = 1)	**= –0.15 (F3 = 2)**	**= 0.00 (F3 = 3)**	**= 0.15 (F3 = 4)**	**= 0.30 (F3 = 5)**
3 indicators	62.5%	32.9%	55.5%	79.4%	95.5%
Overall: 65.2%					
4 indicators	69.5%	28.7%	67.3%	76.1%	95.0%
Overall: 67.3%	[+]	[–]	[+]	[–]	

Note: A plus or minus sign between square brackets indicates the direction of significant increases [+] or decreases [–] in terms of the percentage of correct statistical conclusions (when comparing 4-indicator conditions to 3-indicator conditions).

number of indicators used in the LV indicator model (i.e., 3 or 4 indicators). This is shown in Table 5.2. The table also shows percentages of correct statistical conclusions tabulated for different levels of the LV mean difference at population level (i.e., the different levels for F3).

Table 5.2 shows that the percentage of correct statistical conclusions increases with an increasing positive difference in LV means. This finding is in line with our expectations as the asymmetrical structure of the experimental design (in particular the larger indicator intercept in G2 in noninvariance conditions including one unequal indicator intercept across groups) artificially increases the estimate of the LV mean in G2. The artificial increase works in favor of a rejection of the null-hypothesis in positive discrepancy cases (see Table 5.2) in which the LV mean at population level is higher in G2 than in G1. However, in negative discrepancy cases in which the LV mean at population level is lower in G2 than in G1, the (negative) difference in the LV means across both groups is underestimated (i.e., pushed up toward zero, the LV mean of G1) because of a larger indicator intercept of one LV indicator in G2. This is particularly true in those noninvariance conditions that include one unequal indicator intercept across groups (see F5 in Table 5.1).

When mutually comparing 4-indicator conditions with 3-indicator conditions, significantly different percentages of correct statistical conclusions were obtained (see plus or minus signs indicated between square brackets in Table 5.1). Taking into account the large number of (simulated)

replications included in each cell (*N*/cell = 10,800) in Table 5.2, significant differences in the percentage of correct statistical conclusions should not come as a surprise. A further inspection of the sign of the significant differences across 4- and 3-indicator conditions shows that there is no winner. The 4-indicator conditions report higher percentages of correct statistical conclusions for F3 = 1 and F3 = 3, whereas the 3-indicator conditions report higher percentages for F3 = 2 and F3 = 4. More detailed tables showing aggregated results for different noninvariance conditions as well as scalar invariance conditions may be obtained from the first author.[7]

5.3.2 Inferential Results

As expected, the LV mean difference at population level (i.e., design factor 3) turned out to be the most influential design factor determining the correctness of the statistical conclusion regarding the LV mean difference test. In the C&RTree analyses for both 3- and 4-indicator conditions (not listed in Appendix A to save book space), all sample-splits involved different levels of design factor 3.

Inferential results for large negative discrepancy cases: In large negative discrepancy cases (i.e., F3 = 1), the noninvariant factor loading (i.e., design factor 4) was selected as the first variable to split all (simulated) replications in two subsamples. This is shown in the C&RTrees CT_01 (3-indicator condition) and CT_02 (4-indicator condition) in Appendix A. This sample-split indicated that a *factor loading* of 0.4 for the noninvariant indicator in G2 (versus 0.6 in G1) substantially lowered the probability of making correct conclusions regarding the LV mean difference test. Further sample-splits (in both subsamples) were based on the noninvariant indicator intercept (i.e., design factor 5). Furthermore, the analyses showed that the larger the discrepancy in the noninvariant indicator *intercept*, the

[7] These additional tables were not included in this chapter to save space. These tables do however show that the test of equality of LV means in both groups does provide reasonable control over the type I error rate (i.e., as assessed in scalar invariance conditions). As type I error rates occur when one *falsely* rejects the hypothesis of equal LV means at population level (i.e., the null-hypothesis), the test's control over the type I error rate is evaluated by examining statistical results obtained in "no discrepancy conditions" (i.e., those conditions in which the null-hypothesis truly holds at population level). In addition, such tables provide useful information on the power of the statistical test (i.e., as assessed in noninvariance conditions).

lower the probability of drawing the correct statistical conclusion based on the difference in LV means (i.e., as the noninvariant indicator intercept is opposite to the direction of the difference between LV means at population level). In summary, the results showed that in large negative discrepancy cases, both a noninvariant factor loading and a noninvariant indicator intercept were factors that had a strong influence on the correctness of the statistical conclusion regarding the LV mean difference test.

Inferential results for small negative discrepancy cases: In small negative discrepancy cases (i.e., F3 = 2), the sample of simulated replications was first split using the size of differences in the noninvariant indicator intercept (i.e., design factor 5) as a variable on which to split the sample. This is shown in the C&RTrees CT_03 and CT_04 in Appendix A. Further down in both trees, more splits were made using other levels of design factor 5 as splitting variables. The implication is (once again) that a larger noninvariant indicator intercept has a strong negative impact on the percentage of correct statistical conclusions in negative discrepancy cases. Further inspection of C&RTrees CT_03 and CT_04 revealed that additional sample-splits were made using the degree of noninvariance of the factor loading as a variable on which to split the sample (e.g., F4_D1 and F4_D3). These findings support the conclusion that noninvariance conditions (i.e., design factors 4 and 5) have a strong impact on the percentage of correct statistical conclusions.

Inferential results for no discrepancy cases: In the no difference cases (i.e., F3 = 3) successive splits were made using various levels of design factor 5 as splitting variables. This is shown in C&RTrees CT_05 and CT_06. The smaller the difference in the noninvariant indicator intercept, the higher the probability of drawing the right statistical conclusion with respect to the difference in LV means across groups. Since noninvariant indicator intercepts exert an upward bias on the estimated LV mean in G2, this will be reflected on the LV mean difference between G2 and G1. As a consequence, the probability of rejecting the hypothesis of equal LV means at population level (i.e., in this case, the correct statistical conclusion) decreases.

Inferential results for small positive discrepancy cases: In small positive discrepancy cases (i.e., F3 = 4), the difference in the noninvariant indicator intercept was successively used as the design factor on which sample-splits were made (see C&RTrees CT_07 and CT_08 in Appendix A). Larger differences in the noninvariant indicator intercept enlarge the

(estimated) difference between LV means at population level. As a result, the probability of drawing the correct statistical conclusion (namely, a significant difference between the LV means in both groups) increased artificially because of the upward bias on the LV mean in G2 caused by the noninvariant indicator intercept.

Inferential results for large positive discrepancy cases: In large positive discrepancy cases (i.e., F3 = 5), the percentage of correct statistical conclusions regarding the LV mean difference test turned out to be very high (i.e., around 95% in both 3- and 4-indicator conditions). The C&RTrees CT_09 and CT_10 (see Appendix A) showed that the sample was first split using the first level of design factor 4 (i.e., a noninvariant factor loading of 0.4 in G2 versus a factor loading of 0.6 in G1) as the variable on which to split the sample. The small difference in the percentage of correct statistical conclusions reported for both subsamples (as well as the size of the calculated measure of improvement) showed that this sample-split was only marginally relevant. In conclusion, the difference in LV means at population level (+0.30) was large enough to ensure a very high proportion of correct statistical conclusions (i.e., close to 95%). Obviously, the bias caused by a noninvariant indicator intercept (as present in many simulated conditions) was to a large extent responsible for this high percentage in correct statistical conclusions.

Summary of inferential results on statistical correctness: Overall, the C&RT analyses showed that measurement noninvariance exerted a strong influence on the percentage of correct statistical conclusions regarding the LV mean difference test. In particular, a difference in the noninvariant indicator intercept as large as (approximately) one-tenth of the total length of the scale (a difference of 0.45 on a 5-point scale), or even smaller, was found to have a strong effect on the correctness of the outcome of the LV mean difference test. The effect could either be positive (in positive discrepancy cases [i.e., F3 = 4 and F3 = 5]) or negative (in negative discrepancy cases [i.e., F3 = 1 and F3 = 2] and the no difference cases [F3 = 3]). Next, a noninvariant factor loading showing a difference of 0.2 (factor loading in G1 is 0.6; factor loading in G2 is 0.4) was also found to have a substantial effect on the correctness of the outcome of the LV mean difference test. These findings were largely consistent across 3- and 4-indicator conditions.

Step 2: Robust and Nonrobust Conditions

5.3.3 Descriptive Results

Robustness is only a relevant concept in noninvariance conditions. Of all noninvariance conditions, (only) about 35% was found to be robust. This conclusion applies to both the 3- and 4-indicator conditions. The percentage of robust noninvariance conditions was relatively high in large positive discrepancy cases (i.e., F3 = 5) as this percentage varied between 65% and 75%. In all other cases the percentage of robust noninvariance conditions was much smaller (see Table 5.3). When comparing the percentage of robust conditions across 3- and 4-indicator conditions, no significant differences were found. As a consequence, we had to conclude—once more—that neither of the two LV indicator models outperformed the other and that the number of indicators (at least when 3 and 4 indicators are compared) does not have an effect of the percentage of correct conclusions of the LV mean difference test.

5.3.4 Inferential Results

In an overall C&RT analysis (i.e., across all levels of design factor 3) for both 3- and 4-indicator conditions (C&RTrees are not included in Appendix A), design factor 3 popped up as the first design factor on which to split the sample. Large positive discrepancy cases (i.e., F3 = 5) were separated

TABLE 5.3

Percentage of Robust Noninvariance Conditions

	Negative Discrepancy Cases		**No Discrepancy Cases**	**Positive Discrepancy Cases**	
	LV Mean in Group 2 (G2)				
Percentage of Robust Cases	**= –0.30 (F3 = 1)**	**= –0.15 (F3 = 2)**	**= 0.00 (F3 = 3)**	**= +0.15 (F3 = 4)**	**= +0.30 (F3 = 5)**
K = 3 indicators	19.7%	39.4%	30.3%	15.2%	72.7%
Overall: 35.5%					
K = 4 indicators	28.3%	26.3%	34.9%	18.7%	66.2%
Overall: 34.9%					

Note: The percentage of robust noninvariance conditions is not significantly different across 3- and 4-indicator conditions.

from all other conditions (i.e., F3 different from 5) in the first sample-split. Consistent with the results presented in Table 5.3, the percentage of robust noninvariance conditions were relatively large in conditions representing large positive discrepancy cases. Further down in the C&RTrees, subsamples were formed based on the degree of noninvariance of the indicator intercept (i.e., design factor 5). Larger differences in the noninvariant indicator intercept decreased the probability that the noninvariance condition was robust. In the next paragraphs C&RTrees will be presented for each level of design factor 3.

Inferential results for large negative discrepancy cases: As far as large negative discrepancy cases (i.e., F3 = 1) are concerned, C&RTrees RT_01 and RT_02 in Appendix A reveal that the first important sample-split was made using the third level of design factor 4 (i.e., a noninvariant factor loading equal to 0.8 in G2 versus 0.6 in G1) as the variable on which to split the sample. Actually, in C&RTree RT_01 the very first sample-split was made using an interaction effect as the splitting variable (i.e., interaction effect: F4 = 3 and F5 = 2). This sample-split may be considered to be relatively unimportant because of the limited number of observations in the right branch of the tree ($N = 18$, a detail that is not listed in Appendix A). In C&RTree RT_02, the very first sample-split was made using the third level of design factor 4 (i.e., F4 = 3) as the splitting variable. In the same tree, further sample-splits were made using various degrees of noninvariance of the indicator intercept as splitting variables. C&RTree RT_02 clearly shows that a large noninvariant factor loading combined with a large noninvariant indicator intercept may lead to a very small percentage of robust cases (i.e., 16.7%). C&RTree RT_01 shows different sample-splits, but they all turned out to be relatively unimportant as indicated by the small score obtained for the measure of improvement (a detail that is not listed in Appendix A).

Inferential results for small negative discrepancy cases: In small negative discrepancy cases (i.e., F3 = 2), the first sample-split distinguished between conditions with very small sample sizes ($N = 200$ per group) and all other conditions (see C&RTree RT_03 and RT_04 in Appendix A). Small sample sizes seem to have a positive effect on the robustness of the noninvariance condition. Further down the C&RTrees (C&RT RT_03 and RT_04), the sample was split using a pair of interaction effects between the noninvariant measurement parameters as splitting variables (i.e., the interaction effects: F4 = 1 & F5 = 4, and F4 = 3 & F5 = 2 in 3-indicator conditions,

and the interaction effect: F4 = 3 & F5 = 2 in 4-indicator conditions). Apparently, a smaller noninvariant factor loading (in G2 when compared to G1) can partially compensate for the decrease in robustness due to a larger noninvariant indicator intercept (in G2 when compared to G1). Still further down the same C&RTrees, most sample-splits were made using various levels of the noninvariant indicator intercept or the noninvariant factor loading as splitting variables.

Inferential results for no discrepancy cases: In no difference cases (F3 = 3), successive sample-splits were made using various degrees of noninvariance of the indicator intercept as splitting variables (see C&RTrees RT_05 and RT_06). The results actually show that the larger the difference in the noninvariant indicator intercepts becomes, the smaller the probability that the noninvariance condition is robust against violations of the scalar invariance assumption (across groups).

Inferential results for small positive discrepancy cases: In small positive discrepancy cases (i.e., F3 = 4), various levels of noninvariance of the indicator intercept were successively chosen as splitting variables (see C&RTrees RT_07 and RT_08). The results may be interpreted as follows: the higher the noninvariance of the indicator intercepts, the smaller the probability that the noninvariance condition is robust against violations of the scalar invariance assumption (across groups).

Inferential results for large positive discrepancy cases: In large positive discrepancy cases (i.e., F3 = 5), the first pair of sample-splits were made using different sample sizes per group (i.e., F2) as splitting variables (see C&RTrees RT_09 and RT_10). In contrast to small negative discrepancy cases, the first sample-split in C&RTrees RT_09 and RT_10 shows a negative rather than a positive impact of a small sample size per group (i.e., $N = 200$ in both groups) on the percentage of robust noninvariance conditions. Further sample-splits were made using the degree of noninvariance of the factor loading (i.e., F4) as the splitting variable. A substantially smaller percentage of robust noninvariance conditions were reported in conditions with a noninvariant factor loading equal to 0.4 in the G2 (and a corresponding factor loading of 0.6 in G1).

Summary of inferential results on robustness: Our inferential analyses have shown that violations of the scalar invariance assumption across groups may have a very strong impact on the robustness of (simulated) noninvariance conditions. The extent to which noninvariance conditions are nonrobust depends on which measurement parameters (i.e.,

factor loading and/or indicator intercept) fail to exhibit measurement noninvariance across groups. The influence of a noninvariant intercept is dominant when compared to a noninvariant factor loading in negative discrepancy cases and in small positive discrepancy cases. In large positive discrepancy cases, the effect of a noninvariant factor loading is more significant than in all other noninvariance conditions. In negative discrepancy cases, a smaller noninvariant factor loading (in G2) may partially compensate for the negative effect of a larger noninvariant indicator intercept on the robustness of the noninvariance condition (for instance, when sample size per group is small [i.e., F2 = 2]). The robustness of the noninvariance condition is also influenced by the size of the sample size in each group. This is true for small negative discrepancy cases and large positive discrepancy cases. The distribution of indicators does not affect the robustness of noninvariance conditions.

5.4 CONCLUSIONS

Our simulation study has shown that a noninvariant LV indicator (if not noticed by the researcher) may have a very strong impact on the percentage of correct statistical conclusions of a LV mean difference test. Of all simulated replications about 65% resulted in a correct (statistical) outcome for the LV mean difference test.

A difference in the noninvariant indicator intercept as large as (about) one-tenth of the total length of the scale (a difference of 0.45 on a 5-point scale)—or even smaller—strongly reduced the probability of drawing correct statistical conclusions based on a LV mean difference test. The same is true for a 0.2 difference in a noninvariant factor loading. In our study, neither sample size (per group) nor the underlying distribution of the LV indicator(s) was found to exert a substantial influence on the correctness of the LV mean difference test. This finding is important as it shows that treating ordinal data as if they were metric does not seem to be problematic. All of these conclusions apply equally well to 3- and 4-indicator conditions (i.e., conditions in which the LV is measured by 3 or 4 indicator variables, respectively). Obviously, these conclusions are contingent on the choices made with respect to the design parameters in our study (e.g., only 5-point Likert-type of scales with a left-skewed distribution or a symmetric

bimodal distribution; sample sizes exceeding 200 observations per group). However, these conditions are very common in survey research and thus reflect realistic conditions.

The main research question in this simulation study was to evaluate the extent to which noninvariance conditions are robust against violations of the scalar invariance assumption (across groups). Of all simulated noninvariance conditions, only about 35% turned out to be robust. The low overall percentage of robust noninvariance conditions shows that noninvariant measurement parameters (of one indicator across groups) have a very strong impact on the robustness of noninvariance conditions. In this simulation study, robust noninvariance conditions were rather exceptional.

Apart from a difference in LV means (at population level), the major determinant of the robustness of noninvariance conditions turned out to be the degree of noninvariance of the indicator intercept. This is true for all simulated noninvariance conditions, except for noninvariance conditions with a large *positive* discrepancy between LV means (i.e., the notion of *positive* and *negative* discrepancy is explained in detail in Table 5.1).

The effect of the noninvariant factor loading was somewhat more important in large positive discrepancy cases. In these cases, the percentage of robust noninvariance conditions was rather high (about 70%). The combination of: (1) a large difference in LV means at population level, and (2) the positive bias due to a noninvariant indicator intercept was responsible for a small difference in the percentage of correct statistical conclusions between noninvariance conditions and their corresponding scalar invariance condition. As a consequence, a high percentage of robust noninvariance conditions were obtained.

A smaller factor loading (in group 2) could partially compensate for the bias due to a larger indicator intercept (in the same group). In addition to the effect of noninvariant measurement parameters, there was also an effect of sample size per group on the robustness of the noninvariance condition. This effect was found in small negative discrepancy cases and large positive discrepancy cases. The distribution of the indicators did not exert an influence on the robustness of noninvariance conditions. All conclusions regarding the design factors determining the robustness of noninvariance conditions were consistent across 3- and 4-indicator cases.

Overall, this simulation study has shown that noninvariant measurement parameters form a serious threat to the correctness of a LV mean difference test between two groups (when making a *false* assumption that all indicators exhibit scalar invariance across groups). A noninvariant factor loading, and in particular a noninvariant indicator intercept, have a strong impact on the percentage of correct statistical conclusions regarding the LV mean difference test. The degree of noninvariance (as simulated in this study) was severe enough to seriously affect the robustness of the LV mean difference test against violations of the scalar invariance assumption (across groups). Furthermore, it does not seem to matter very much if one uses three or four indicators to measure the underlying (one-dimensional) LV. The results were highly consistent across 3- and 4-indicator conditions.

For these reasons, the general advice for researchers is to conduct formal tests on measurement invariance of construct indicators (e.g., running multigroup mean and covariance structure analyses) *prior* to conducting any LV mean comparisons across groups. It is crucial that indicators that do not exhibit measurement invariance across groups are either removed from the measurement model or, alternatively, adequate corrections are made to correct for the bias of (a) noninvariant LV indicator(s). Provided that two indicators of the same construct exhibit scalar invariance, such technical corrections are possible. This has been demonstrated by Steenkamp and Baumgartner (1998), and Scholderer, Grunert, and Brunsø (2005).

Even though such technical corrections (see Scholderer et al., 2005; Steenkamp & Baumgartner, 1998) have been introduced one never knows exactly what one is controlling for. For this reason, we recommend (whenever possible) to identify the causes of measurement noninvariance *prior* to making statistical corrections to the LV mean difference test. For instance, response styles such as acquiescence response style or extreme response style are generally known to form a serious threat to measurement (scalar) invariance of indicators across groups (Weijters, 2006).

A good overview of how response styles can be measured (and corrected for) in survey research is provided in Baumgartner and Steenkamp (2006). An even more recent paper by Weijters, Schillewaert, and Geuens (2008) introduces a sound but sophisticated approach to correct for

different types of response styles. Their approach is based on the inclusion of a separate, heterogeneous set of response style indicators drawn from a wide universe of multi-item survey measures. This approach enables a valid and reliable assessment of response styles (i.e., due to the heterogeneity of response style indicators), while avoiding a possible confound between questionnaire *content* and response *style* of the respondent (i.e., as response style indicators are not used for substantive purposes; see De Beuckelaer, Weijters, & Rutten, in press).

ACKNOWLEDGMENTS

We would like to thank Albert Satorra, Steffen Kühnel, and Gerrit K. Janssens for their valuable comments on an earlier version of this paper. In addition, we are also indebted to the critical reviewers Eldad Davidov and Holger Steinmetz, who provided us with valuable suggestions on how to improve this chapter.

REFERENCES

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 249–279). Beverly Hills, CA: Sage.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses, and future advances* (pp. 95–109). Thousand Oaks, CA: Sage.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models.* New York, NY: Springer-Verlag.

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

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155–175.

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

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods, 2,* 33–46.

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

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

AU: Please update publication information.

De Beuckelaer, A., Weijters, B., & Rutten, A. (in press). Using ad hoc measures for response styles. A cautionary note. Forthcoming in *Quality and Quantity*.

He, Y., Merz, M. A., & Alden, D. L. (2008). Diffusion of measurement invariance assessment in cross-national empirical marketing research: Perspectives from the literature and a survey of researchers. *Journal of International Marketing, 16,* 64–83.

Horn, J. L., McArdle, J. J., and Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1,* 179–188.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: An Interdisciplinary Journal, 2,* 101–118.

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

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

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

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

Muthén, L. K., & Muthén, B. O. (1999). *Mplus, the comprehensive modeling program for applied researchers. User's guide* (2nd printing). Los Angeles, CA: Muthén and Muthén.

Reise, S., 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.

Schaffer, B. S., & Riordan, C. H. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods, 6,* 169–216.

Scholderer, J., Grunert, K. G., & Brunsø, K. (2005). A procedure for eliminating additive bias from cross-cultural survey data. *Journal of Business Research, 58,* 72–78.

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

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organisational research. *Organizational Research Methods, 3,* 4–70.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* London: Sage.

Weijters, B. (2006). Response styles in consumer research. PhD Dissertation, Vlerick Leuven Gent Management School, Ghent, Belgium.
Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36,* 409–422.
Williams, L. J., Edwards, J. R., & Vandenberg, R. J. (2003). Recent advances in causal modelling methods for organizational and management research. *Journal of Management, 29,* 903–993.

APPENDIX A: CLASSIFICATION AND REGRESSION TREES (C&RTREES)

Percentage of correct statistical conclusions(latent variable [LV] mean difference test)
ILLUSTRATIVE EXAMPLE of a C&RTree (see also Figure 5.A.1)
Example_C&RTree (4 indicators; *N*=198) {[Overall: 22.3%];
[Split 1=F4_D3 [upper=root]: nil=7.9%; one=63.9%];
[Split 1.1=F5_D2 [upper=nil]: nil=1.1%; one=25.0%];
[Split 1.2=F5_D4 [upper=one]: nil=79.6%; one=16.7%];
[Split 1.1.1=F2_D1 [upper=nil]: nil=0.0%; one=6.7];
[Split 1.1.2=F4_D1 [upper=one]: nil=50.0%; one=0.0%];
[Split 1.2.1=F4_D3*F5_D3 [upper=nil]: nil=88.9%; one=61.1%]}

Notes: This example C&RTree is identical to RT_02 (see below); clarification of the notation used: F4_D3=1 means the factor loading of indicator 1, i.e. λ_2, equals 0.8 in Group 2 (see F4 in Table 5.1); F5_D2=1 means the intercept of indicator 2 is 0.15 higher for Group 2 (see F5 in Table 5.1); F5_D4 means the intercept of indicator 2 is 0.45 higher for Group 2 (see F5 in Table 5.1).

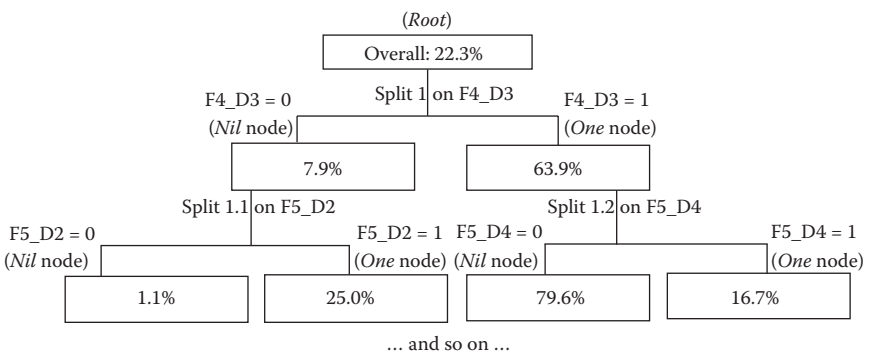

FIGURE 5.A.1
Percentage robust noninvariance conditions (F3 = 1; 4 indicators).

CT_01 (F3=1; 3 indicators; *N*=10,800):
{[Overall: 62.5%];
[Split 1=F4_D1 [upper=root]: nil=73.3%; one=40.9%];
[Split 1.1=F5_D4 [upper=nil]: nil=81.8%; one=47.9%];
[Split 1.2=F4_D4 [upper=one]: nil=47.0%; one=22.3%];
[Split 1.1.1=F5_D3 [upper=nil]: nil=89.5%; one=66.5%];
[Split 1.1.2=F4_D3*F5_D4 [upper=one]: nil=33.1%; one=62.7%];
[Split 1.2.1=F5_D3 [upper=nil]: nil=58.9%; one=23.2%];
[Split 1.1.1.1=F4_D3*F5_D3 [upper=one]: nil=53.6%; one=79.4%];
[Split 1.2.1.1=F5_D2 [upper=nil]: nil=71.5%; one=46.3%]}

CT_02 (F3=1; 4 indicators; *N*=10,800):
{[Overall: 69.5%];
[Split 1=F4_D1 [upper=root]: nil=81.6%; one=45.1%];
[Split 1.1=F5_D4 [upper=nil]: nil=87.3%; one=64.6%];
[Split 1.2=F4_D1*F5_D4 [upper=one]: nil=52.8%; one=21.9%];
[Split 1.1.1=F5_D3 [upper=nil]: nil=92.6%; one=76.7%];
[Split 1.2.1=F4_D1*F5_D3 [upper=nil]: nil=62.6%; one=33.2%];
[Split 1.1.1.1=F4_D3*F5_D3 [upper=one]: nil=64.1%; one=89.3%];
[Split 1.2.1.1=F5_D2 [upper=nil]: nil=73.3%; one=51.9%]}

CT_03 (F3=2; 3 indicators; *N*=10,800)
{[Overall: 32.9%];
[Split 1=F5_D3 [upper=root]: nil=37.5%; one=19.1%];
[Split 1.1=F5_D2 [upper=nil]: nil=43.3%; one=33.2%];
[Split 1.1.1=F5_D4 [upper=nil]: nil=51.7%; one=35.0%];
[Split 1.1.2=F5_D2 [upper=one]: nil=20.2%; one=37.0%];
[Split 1.1.1.1=F4_D3 [upper=nil]: nil=43.2%; one=68.9%];
[Split 1.1.1.2=F4_D1*F5D4 [upper=one]: nil=29.6%; one=45.8%]}

CT_04 (F3=2; 4 indicators; *N*=10,800)
{[Overall: 28.7%];
[Split 1=F5_D4 [upper=root]: nil=33.0%; one=15.8%];
[Split 1.1=F5_D3 [upper=nil]: nil=41.2%; one=16.5%];
[Split 1.2=F4_D3 [upper=one]: nil=17.8%; one=11.7%];
[Split 1.1.1=F4_D3 [upper=nil]: nil=33.5%; one=56.6%];
[Split 1.1.2=F4_D3 [upper=one]: nil=12.1%; one=25.2%];
[Split 1.1.1.1=F5_D2 [upper=nil]: nil=43.8%; one=23.2%];
[Split 1.1.1.2=F4_D3*F5_D2 [upper=one]: nil=12.1%; one=25.2%]}
[Split 1.1.1.1.1=F4_D1 [upper=nil]: nil=53.0%; one=34.6%];
[Split 1.1.1.1.2=F4_D1 [upper=one]: nil=28.9%; one=17.6%]}

CT_05 (F3=3; 3 indicators; *N*=10,800) {[Overall: 55.5%]; [Split 1=F5_D4 [upper=root]: nil=67.6%; one=19.2%]; [Split 1.1=F5_D3 [upper=nil]: nil=80.7%; one=41.2%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=87.7%; one=73.7%]}	CT_06 (F3=3; 4 indicators; *N*=10,800) {[Overall: 67.3%]; [Split 1=F5_D4 [upper=root]: nil=76.4%; one=39.7%]; [Split 1.1=F5_D3 [upper=nil]: nil=85.3%; one=58.6%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=90.0%; one=80.7%]}
CT_07 (F3=4; 3 indicators; *N*=10,800) {[Overall: 79.4%]; [Split 1=F5_D4 [upper=root]: nil=73.6%; one=97.0%]; [Split 1.1=F5_D3 [upper=nil]: nil=64.7%; one=91.5%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=50.0%; one=79.3%]; [Split 1.1.1.1=F4_D3 [upper=nil]: nil=42.8%; one=64.4%]}	CT_08 (F3=4; 4 indicators; *N*=10,800) {[Overall: 76.1%]; [Split 1=F5_D4 [upper=root]: nil=70.3%; one=93.6%]; [Split 1.1=F5_D3 [upper=nil]: nil=62.0%; one=86.8%]; [Split 1.2=F4_D1*F5D4 [upper=one]: nil=96.9%; one=87.0%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=51.1%; one=72.9%]; [Split 1.1.1.1=F4_D1 [upper=nil]: nil=59.5%; one=34.2%]; [Split 1.1.1.2=F4_D1*F5_D2 [upper=one]: nil=79.6%; one=59.4%]}
CT_09 (F3=5; 3 indicators; *N*=10,800) {[Overall: 95.5%]; [Split 1=F4_D1 [upper=root]: nil=97.8%; one=90.7%]}	CT_10 (F3=5; 4 indicators; *N*=10,800) {[Overall: 95.0%]; [Split 1=F4_D1 [upper=root]: nil=97.9%; one=89.1%]}

Notes: % indicate percentage of correct statistical conclusions (LV mean difference test); *nil* and *one* indicate the nodes of the tree where the condition (e.g., F5_D1) equals zero and one, respectively.

Percentage of robust noninvariance conditions

RT_01 (F3=1; 3 indicators; *N*=198) {[Overall: 19.7%];
[Split 1=F4_D3*F5_D2 [upper=root]: nil=12.8%; one=88.9%%];
[Split 1.1=F4_D3 [upper=nil]: nil=4.8%; one=31.5%];
[Split 1.1.1=F2_D1 [upper=nil]: nil=1.9%; one=19.0%];
[Split 1.1.2=F4_D3*F5_D4 [upper=one]: nil=47.2%; one=0.0%];
[Split 1.1.1.1=F5_D2 [upper=nil]: nil=0.0%; one=6.7%];
[Split 1.1.2.1=F4_D3*F5_D3 [upper=nil]: nil=72.2%; one=22.2%]}

RT_02 (F3=1; 4 indicators; *N*=198) {[Overall: 22.3%];
[Split 1=F4_D3 [upper=root]: nil=7.9%; one=63.9%];
[Split 1.1=F5_D2 [upper=nil]: nil=1.1%; one=25.0%];
[Split 1.2=F5_D4 [upper=one]: nil=79.6%; one=16.7%];
[Split 1.1.1=F2_D1 [upper=nil]: nil=0.0%; one=6.7];
[Split 1.1.2=F4_D1 [upper=one]: nil=50.0%; one=0.0%];
[Split 1.2.1=F4_D3*F5_D3 [upper=nil]: nil=88.9%; one=61.1%]}

RT_03 (F3=2; 3 indicators; *N*=198) {[Overall: 39.4%];
[Split 1=F2_D1 [upper=root]: nil=31.5%; one=78.8%];
[Split 1.1=F4_D1*F5_D4 [upper=nil]: nil=26.0%; one=86.7%];
[Split 1.1.1=F4_D3*F5_d2 [upper=nil]: nil=22.2%; one=60.0%];
[Split 1.1.1.1=F5_D3 [upper=nil]: nil=28.9%; one=8.9%];
[Split 1.1.1.1.1=F5_D2 [upper=nil]: nil=40.0%; one=6.7%];
[Split 1.1.1.1.2=F4_D1 [upper=one]: nil=3.3%; one=20.0%]}

RT_04 (F3=2; 4 indicators; *N*=198) {[Overall: 26.3%];
[Split 1=F2_D1 [upper=root]: nil=19.4%; one=60.6%];
[Split 1.1=F4_D3*F5_D2 [upper=nil]: nil=14.0%; one=73.3%];
[Split 1.1.1=F4_D3 [upper=nil]: nil=8.6%; one=26.7%];
[Split 1.1.1.1=F5_D3 [upper=nil]: nil=12.0%; one=0.0%];
[Split 1.1.1.2=F4_D3*F5_D4 [upper=one]: nil=40.0%; one=0.0%];
[Split 1.1.1.1.1=F2_D3 [upper=nil]: nil=8.3%; one=26.7%]}

RT_05 (F3=3; 3 indicators; *N*=198) {[Overall: 30.3%]; [Split 1=F5_D4 [upper=root]: nil=41.7%; one=0.0%]; [Split 1.1=F5_D3 [upper=nil]: nil=66.7%; one=0.0%]; [Split 1.1.1=F2_D4 [upper=nil]: nil=74.7%; one=26.7%]; [Split 1.1.1.1=F2_D3 [upper=nil]: nil=68.3%; one=100%]; [Split 1.1.1.1.1=F5_D2 [upper=nil]: nil=87.5%; one=55.6%]}

RT_06 (F3=3; 4 indicators; *N*=198) {[Overall: 34.9%]; [Split 1=F5_D4 [upper=root]: nil=47.9%; one=0.0%]; [Split 1.1=F5_D3 [upper=nil]: nil=71.1%; one=9.3%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=94.4%; one=55.6%]; [Split 1.1.2=F1_D2 [upper=one]: nil=2.8%; one=22.2%]; [Split 1.1.1.1=F1_D3 [upper=one]: nil=47.2%; one=72.2]}

RT_07 (F3=4; 3 indicators; *N*=198) {[Overall: 15.1%]; [Split 1=F5_D4 [upper=root]: nil=20.8%; one=0.0%]; [Split 1.1=F5_D3 [upper=nil]: nil=33.3%; one=0.0%]; [Split 1.1.1=F5_D2 [upper=nil]: nil=55.6%; one=18.5%]; [Split 1.1.1.1=F4_D1 [upper=one]: nil=8.3%; one=38.9%]; [Split 1.1.1.1.1=F4_D3*F5_D2 [upper=nil]: nil=16.7%; one=0.0%]}

RT_08 (F3=4; 4 indicators; *N*=198) {[Overall: 18.7%]; [Split 1=F4_D1*F5_D2 [upper=root]: nil=12.2%; one=83.3%]; [Split 1.1=F5_D4 [upper=nil]: nil=17.5%; one=0.0%]; [Split 1.1.1=F5_D3 [upper=nil]: nil=30.6%; one=0.0%]; [Split 1.1.1.1=F5_D2 [upper=nil]: nil=50.0%; one=11.1%]; [Split 1.1.1.1.1=F4_D3 [upper=one]: nil=22.2%; one=0.0%]}

RT_09 (F3=5; 3 indicators; *N*=198) {[Overall: 72.7%]; [Split 1=F2_D1 [upper=root]: nil=82.4%; one=24.2%]; [Split 1.1=F2_D6 [upper=nil]: nil=89.9%; one=60.6%]; [Split 1.1.1=F4_D1 [upper=nil]: nil=95.2%; one=75.0%]; [Split 1.1.1.1=F5_D2 [upper=nil]: nil=93.3%; one=100.0%]; [Split 1.1.1.1.1=F5_D4 [upper=nil]: nil=88.9%; one=100.0%]}

RT_10 (F3=5; 4 indicators; *N*=198) {[Overall: 66.2%]; [Split 1=F2_D1 [upper=root]: nil=73.3%; one=30.3%]; [Split 1.1=F2_D2 [upper=nil]: nil=79.6%; one=48.5%]; [Split 1.1.1=F2_D6 [upper=nil]: nil=85.9%; one=60.6%]; [Split 1.1.1.1=F4_D1 [upper=nil]: nil=95.2%; one=69.4%]; [Split 1.1.1.1.1=F4_D3 [upper=nil]: nil=100.0%; one=91.7%]}

Notes: % indicate percentage of robust noninvariance conditions; *nil* and *one* indicate the nodes of the tree where the condition (e.g., F5_D1) equals zero and one, respectively.