Testing Measurement Invariance Using Multigroup CFA: Differences between Educational Groups in Human Values Measurement

Holger Steinmetz, Peter Schmidt, Andrea Tina Booh, and Siegrid Wieczorek

University of Giessen / Germany

Shalom H. Schwartz

Hebrew University / Israel

Holger Steinmetz, Peter Schmidt, Andrea Tina Booh, and Siegrid Wieczorek, Institute for Political Science, University of Gießen / Germany; Shalom H. Schwartz, Department of Psychology, Hebrew University / Israel.

The research was supported by grants from the German Science Foundation (DFG) – SCHM 658 / 8 -1. The material provided in this study was presented at the 24th conference of the Society for Multivariate Analysis in the Behavioral Sciences (SMABS), July 2004, in Jena / Germany.

Correspondence concerning this article should be addressed to Holger Steinmetz, Institute for Political Science, University of Giessen, Karl-Glöckner-Str. 21 E, D-35394 Giessen, Germany. Telephone: 00496419923051 (Secr.), Fax: 00496419923059. Email: Holger.Steinmetz@web.de

Published in: Quality and Quantity

Link: http://www.springerlink.com/content/102985/

Abstract

This article applies the testing procedures for measurement invariance using multigroup confirmatory factor analysis (MGCFA). It illustrates these procedures by investigating the factorial structure and invariance of the Portraits Value Questionnaire (PVQ, Schwartz et al., 2001) across three education groups in a population sample ($N = 1,677$). The PVQ measures ten basic values that Schwartz postulates to comprehensively describe the human values recognized in all societies (achievement, hedonism, self-direction, benevolence, conformity, security, stimulation, power, tradition and universalism). We also estimate and compare the latent means of the three education groups. The analyses show partial invariance for most of the ten values and parameters. As expected, the latent means show that less educated respondents attribute more importance to security, tradition, and conformity values.

Keywords: *Measurement Invariance, Multigroup Analyses, Values, Cross-Cultural Psychology, Education, Survey*

Testing Measurement Invariance Using Multigroup CFA: Differences between Educational Groups in Human Values Measurement

The issue of measurement invariance is crucial for studies that investigate group differences. Cross-cultural methodologists have emphasized that group comparisons assume invariance of the elements of the measurement structure (i.e., factor loadings and measurement errors) and of response biases (Billiet, 2002; Little, 1997; van de Vijver and Leung, 1997). Less recognized is that group comparisons within a single culture also require measurement invariance to insure that potential differences (e.g., in means or regression coefficients) can be interpreted reliably (Vandenberg and Lance, 2000).

Sub-groups within populations are often heterogeneous with regard to the parameter values of a model. Nonetheless, most within-society research continues implicitly to assume homogeneity of the population (Muthén, 1989). This is especially so in field research with convenience samples of social, educational, or occupational sub-groups. These groups often differ from one another or from the overall population with regard to measurement or structural parameters. In the worst case, researchers measure different constructs in the groups. Hence within-society studies should assess possible lack of measurement invariance, when possible, to uncover potential population heterogeneity.

Multigroup confirmatory factor analyses (MGCFA) (Billiet, 2002; Jöreskog, 1971) is the most widely used method to test for measurement invariance. This method permits testing for invariance easily by setting cross-group constraints and comparing more restricted with less restricted models (e.g., Baumgartner and Steenkamp, 1998; Byrne, Shavelson, and Muthén, 1989).

This article illustrates the application of invariance testing to the value theory of Schwartz (2005a; 2005b), using this method with sub-groups within a single society. Schwartz postulates that ten human values comprehensively describe the basic values recognized in all societies (*achievement*, *hedonism*, *self-direction*, *benevolence*, *conformity*, *security*, *stimulation*, *power*, *tradition* and *universalism*). We test the factorial structure and measurement invariance of one of the instruments that operationalize the value theory, the Portrait Values Questionnaire (PVQ). The PVQ has not yet been tested for measurement invariance. The current study is the first to test the values theory for population homogeneity in a single society. Past studies have tested the theory cross-culturally (Davidov et al., in press; Schwartz & Boehnke, 2004; Spini, 2003).

We first provide an introduction to measurement invariance. We then briefly describe the Schwartz value theory. Finally, we test measurement invariance across three education groups. We expect these groups to differ in their responding behavior and their latent means.

1. Measurement Invariance

Researchers usually assume equivalence of the structure of the measures they compare across the groups. The validity of this assumption is critical for any conclusions about group related differences (see Vandenberg and Lance, 2000, for a review). Crucially, unless this assumption is true, one cannot even claim that the construct is the same in the different groups (Little, 1997). Thus, legitimate comparison of means or structural relations across groups requires equivalence of the measurement structures underlying the indicators (Ployhardt and Oswald, 2004; Thompson and Green, 2006). The manifest means in a comparison depend not only on the latent means but on the whole underlying measurement model (i.e., item intercepts and factor loadings).

Tests of measurement invariance address four questions: Are the measurement parameters (factor loadings, measurement errors etc.) the same across groups? Are there pronounced response biases in a particular group? Can one unambiguously interpret observed mean differences as latent mean differences? Is the same construct measured in all groups? In evaluation research (e.g., Millsap and Hartog, 1988) and field research with longitudinal data (e.g., Vandenberg and Lance, 2000), measurement invariance is also critical. Measurement parameters must be invariant across time.

1.1 The Measurement Structure

The following presentation refers to a case where a set of items (manifest indicators) measures an underlying (latent) construct ξ. Figure 1 shows the measurement models for two respective groups (A and B). Because the covariances between several latent constructs are important, we depict two latent variables (ξ_1 and ξ_2) with their respective indicators. Based on factor analytic tradition, we depict variation in a manifest indicator x_i as due to a construct ξ_j and an error δ_i. As a regression equation for a single indicator x_i^g, this causal influence is:

$$x_i^g = \tau^g{}_i + \lambda^g{}_i \xi_j^g + \delta^g{}_i \qquad (1)$$

Here x_i^g is the *ith* indicator in the set of indicators that measure ξ_j^g in the group *g*, τ_i^g is the intercept in the regression equation, $\lambda^g{}_i$ is the factor loading linking x_i^g and ξ_j^g and $\delta^g{}_i$ is the error of the indicator x_i^g.

The covariance equation – a matrix algebraic equation that links the measurement structure (see Figure 1) to the manifest covariance matrix – is:

$$\Sigma^g = \Lambda^g \Phi^g \Lambda^{g\prime} + \Theta^g{}_\delta \qquad (2)$$

Here Σ^g is the covariance matrix of the manifest indicators x_i^g in group *g*, Λ^g is the matrix containing the factor loadings ($\Lambda^{g\prime}$ is its transpose), Φ^g is the matrix of the variances and covariances of the latent constructs and $\Theta^g{}_\delta$ is typically a diagonal matrix containing the error variances of the indicators. In the common factor model, the intercept $\tau^g{}_i$ (see equation 1) is assumed to be 0 and therefore not estimated. Hence, the intercept does not appear in equation 2. However, it can be added to the model and estimated by including a vector of the manifest indicators' means in addition to the manifest covariance matrix (Bollen, 1989).

Each group has its own measurement model (Figure 1 and equation 2). Meeting the criteria for reliability and construct validity is, however, not enough for comparisons. The measurement structure must also be equivalent (invariant), albeit not perfectly (Byrne et al., 1989).

In the MGCFA framework, we test the invariance of the parameter matrices implied by equation (2) by constraining cross-group equality of these matrices. This is done in a stepwise approach; each step constrains a particular matrix (e.g., the Λ^g_x-matrix) to be equal across all groups. Each restricted model is nested within a less restricted one. Hence we can compare models statistically using the difference in the chi-square-statistics and degrees of freedom.

1.2 Tests of Measurement Invariance

We next describe the different types of measurement invariance. Byrne et al. (1989) and others distinguish two types of invariance: (a) 'Measurement invariance' (in a narrower sense) is invariance of item intercepts, factor loadings, and error variances; (b) 'structural invariance' is invariance of the variances and covariances of the latent variables. Table 1 depicts the invariance tests and their meanings.

Configural invariance implies the same number of factors in each group and the same pattern of fixed and free parameters. It is a prerequisite for the other tests.

Metric invariance implies equal factor loadings across groups. For instance, the parameter λ_{21} must be the same in groups A and B (see Figure 1). In terms of equation (2), this is tested by imposing equality constraints on the Λ-matrices that contain the factor loadings (i.e., $\Lambda^A = \Lambda^B = \dots \Lambda^G$; superscripts refer to groups A to G). Equal factor loadings indicate that the groups calibrate their measures in the same way. Hence, the values on the manifest scale have the same meaning across groups (Meredith, 1993; Vandenberg and Lance, 2000).

Metric invariance concerns construct comparability. Steenkamp and Baumgartner (1998) view configural invariance as sufficient for construct comparability across groups. We argue, in contrast, that metric invariance is a stricter condition of construct comparability. According to the common factor perspective, the factor loadings indicate the strength of the causal effect of the latent variable ξ_j on its indicators and can be interpreted as validity coefficients (Bollen, 1989). Significantly different factor loadings, hence, imply a difference in the validity coefficients. This raises concerns about whether the constructs are the same across groups. Hence, configural invariance, by providing evidence that the construct is related to the same set of indicators, is a prerequisite for inferring that the construct has *similar* meaning. However, metric invariance is necessary to infer that the construct has the *same* meaning, because it provides evidence about the equality of validity coefficients.

Scalar Invariance refers to invariance of the item intercepts in the regression equations that link the indicators x_i^g to their latent variable ξ_j^g (see equation 1). Hayduk (1989) notes that item intercepts can be interpreted as systematic biases in the responses of a group to an item. As a result, the manifest mean can be systematically higher or lower (upward or downward biased) than one would expect due to the groups' latent mean and the factor loading. Scalar invariance is present if the degree of up- or downward bias of the manifest variable is equal across groups. It is absent if one of the groups differs significantly in one or more of the item intercepts. The intercept also indicates the expected value of x_i when $\xi_j = 0$. To test for scalar invariance, one constrains the tau-vectors to be equal across groups ($\tau^A = \tau^B = \ldots = \tau^G$).

Invariance of factor variance exists when groups have the same variances in their respective latent variables. This is tested by constraining the diagonal of the phi-matrices ($\phi_{jj}^A = \phi_{jj}^B = \ldots = \phi_{jj}^G$) to be equal. This test assesses possible differences in homogeneity of the latent variables in the groups (Steenkamp & Baumgartner 1998).

Invariance of the factor covariances refers to equality of the associations among the latent variables across groups. It is tested by constraining the subdiagonal elements of the phi-matrices ($\phi_{jk}^A = \phi_{jk}^B = \ldots = \phi_{jk}^G$) to be equal. Covariances among constructs have implications for the constructs' meaning or validity (Cronbach and Meehl, 1955). Hence, unequal covariances raise concerns about equality of construct meanings (Cole and Maxwell, 1985).

In a similar vein, Millsap and Hartog (1988: 574) interpret changes in the covariances among constructs *over time* as ‘a shift in the meaning or conceptualization of the construct being measured’. In sum, the test of equal factor covariances has implications for ‘construct comparability’ (Little, 1997). As Marsh and Hocevar (1985) note, equal factor variances are required to interpret covariances as correlations.

Invariance of latent means. Most applications of structural equation modeling focus on the covariance part of the model. In such cases, the model assumes zero indicator intercepts and zero latent means. However, in some situations (mainly multigroup analyses and longitudinal designs) researchers are interested in the means and intercepts (Bollen, 1989; Hayduk, 1989; Sörbom, 1978). Analyses of invariance of the latent means test for differences between groups (or points of time) in the latent means. In contrast, traditional approaches to the analysis of mean differences use composite *manifest* scores and employ *t* tests, ANOVA, or MANOVA (Thompson and Green, 2006). The validity of testing group differences in manifest scores depends on whether the assumptions that underlie such comparisons are correct, specifically, that both the factor loadings and the item intercepts are equal (i.e., metric and scalar invariance).

Based on equation (1), the relationship between a latent and an observed mean or an expected observed value can be written as follows:

$$E(x_i^{\,g}) = \tau_i^{\,g} + \lambda_i^{\,g}\kappa_j^{\,g} \tag{3}$$

$E(x_i^g)$ is the expected value of the *ith* manifest indicator in group g, τ_i^g is the item intercept of the *ith* item in group g, λ_i^g is its factor loading and κ_j^g is the mean of factor j in group g. Equation (3) shows that a manifest mean depends not only on its latent mean but also on the factor loading and the item intercept. Thus, a manifest mean difference can be caused either by a latent mean difference or a difference in the loadings, intercepts, or both (Millsap and Everson, 1991). Therefore, a test of a latent mean difference requires the equality of both the factor loadings and item intercepts (Cole and Maxwell, 1985; Steenkamp and Baumgartner, 1998). The equality of the latent means is tested by constraining the kappa-matrices ($\kappa^A = \kappa^B = \ldots = \kappa^G$) to be equal across groups.

Invariance of error variances. The test of invariant error variances concerns the hypothesis that the measurement error in the manifest indicators (i.e., $\Theta^A = \Theta^B = \ldots = \Theta^G$) is the same in all groups. If the factor loadings and variances of the latent variables have been shown to be equal, then the error variances can be interpreted as equivalent to the reliability of the indicators (Cole and Maxwell, 1985; Steenkamp and Baumgartner, 1998).

In structural equation modeling, the test of invariance of the error variances is less important because the relationships between latent variables (correlations and regression coefficients) are corrected for measurement error. However, in analyses of manifest composite scales, unequal reliabilities lead to unequal biases in correlations or regression coefficients. Then, ‘pseudo-moderator-effects’ may occur. Relationships between variables may differ significantly across groups even though the latent construct correlations are in fact equal. This can occur because random measurement error attenuates observed correlations more in the group with the greater measurement error (Ployhardt and Oswald, 2004).

1.3 Full and Partial Invariance

Thus far we presented tests for measurement invariance that assess whether each element of the respective matrices is equal in all groups. This is *full* measurement invariance. It is widely acknowledged, however, that such a requirement may be too strict and unrealistic a goal for group comparisons. Consequently, Byrne et al. (1989) introduced the concept of *partial* invariance in which only a subset of parameters in each matrix must be invariant whereas others are allowed to vary between the groups. Byrne et al. argued that at least two indicators must be invariant to ensure the meaningfulness of latent mean comparisons. Baumgartner and Steenkamp (1998) compared two groups that shared a limited number of invariant indicators but other indicators that differed (e.g., groups A and B shared x_1 and x_2

but group A had x_3-x_5 whereas group B had x_6-x_8). Their findings provided evidence that two scalar and metric invariant indicators suffice to obtain estimates of latent mean differences that permit meaningful mean comparisons (De Beuckelaer, 2005).

Steenkamp and Baumgartner (1998) recommended the following order for tests of invariance: configural invariance, metric invariance, scalar invariance, invariance of the factor variances, invariance of the factor covariances, latent mean invariance, and invariance of the error variances.

1.4 Summary

MGCFA permits testing for full and partial invariance of the measurement (factor loadings, error variances) and structural parameters (variances and covariances). Both the intercepts of the indicators and the latent means can also be estimated and tested for invariance. Configural invariance of the whole factor structure and metric invariance of the factor loadings are critical for the interpretation of the constructs and are requisites for all other tests. Partial scalar invariance, at least, must be established before latent means can be compared. Moreover, some tests have implications for interpreting the results of subsequent tests (e.g., equal variances are necessary to interpret covariances as correlations). Before presenting the invariance tests, we briefly describe the theoretical foundations of our model.

2. The Theory of Basic Human Values

Schwartz (1992; 2005a) identifies five main features of values: (1) Values are beliefs linked to emotions. People for whom independence is an important value become aroused if their independence is threatened, for example, despair when they are helpless to protect it, and are happy when they can enjoy it. (2) Values refer to desirable goals that motivate action. People for whom social order, justice, and helpfulness are important values are motivated to pursue these goals. (3) Values are abstract goals that transcend specific actions and situations, a feature that distinguishes them from narrower concepts like norms and attitudes. (4) Values serve as standards or criteria that guide the selection or evaluation of actions, policies, people, and events.

(5) Values are ordered by importance, with each person characterized by his/her own distinctive system of value priorities.

The values theory defines ten broad values according to the motivation that underlies each. Presumably, these values encompass the range of motivationally distinct values recognized across cultures. Table 2 summarizes the defining goals of these broad values. Each value expresses a motivational goal that is either congruent or in conflict with the other values. The total set of relations of congruity and conflict among values yields a circular structure that organizes them. Motivationally congruent values are adjacent in the circle, conflicting values are opposed. Schwartz (1992; 2005a) posits that values form a motivational continuum: The closer two values around the circle, the more similar their motivational implications; the more distant around the circle, the more their motivational implications conflict.

The ten values are arrayed on two bipolar dimensions. The first dimension, *openness to change* vs. *conservation,* contrasts self-direction and stimulation values with security, conformity, and tradition values. The second dimension, *self-transcendence* vs. *self-enhancement,* contrasts universalism and benevolence values with power and achievement values. The configurations of values in smallest space analyses in over 200 samples from over 70 countries suggest that the theorized structure of value relations is near-universal.

The current study measured values with a variant of the PVQ (Schwartz et al., 2001). This is the first test of the factor structure and measurement invariance of this instrument. Schmitt, Schwartz, Steyer, and Schmitt (1993) used confirmatory factor analysis to analyze the Schwartz Value Survey (SVS), a predecessor of the PVQ, with data from a convenience sample. Schwartz and Boehnke (2004) used confirmatory factor analysis to test the structure of the SVS in two sets of 23 samples from 27 countries. However, they did not test for measurement invariance between or within countries. Thus, our research innovates in formally testing the underlying measurement theory in a representative sample from Germany.

3. Hypotheses

The study investigates a sample of the German working population. To assess population heterogeneity, we test the invariance of the measurement and structural part of the MGCFA across three educational groups. We expect differences between measurement errors, factor loadings, and latent means among educational groups because of two reasons.

First, highly educated individuals have more extensive and intense exposure to abstract verbal material and to testing situations. This should enable them to understand the items and instructions of the values questionnaire more easily. They should therefore provide more valid responses. Numerous studies reveal an association between education and the consistency of reported belief systems (Converse, 1964; Judd et al., 1981; Zaller, 1995). The greater validity of responses among more educated groups is likely to affect the factor loadings and measurement errors. We therefore predict lower measurement errors and higher factor loadings in the more educated sub-sample.

Second, Schwartz (2005b) reported substantial positive correlations of level of education with openness to change values (self-direction, stimulation, and hedonism) and substantial negative correlations with conservation values (tradition, conformity, and security). These correlations could replicated across eight countries. Education may therefore lead to differences in the latent means. Level of education did not relate substantially to self-enhancement (power, achievement) or self-transcendence (benevolence, universalism) values across countries. We therefore anticipate no differences between educational groups on the latent means of these values.

4. Methods

4.1 Sample

The sample included 1,677 respondents, 1,209 employed and 468 unemployed. The data were collected in April and May 2003 by a commercial survey institute as part of a study of flexible working time schedules and part-time work. Respondents were recruited by randomdialing and interviewed by telephone. The sample included 55.9% women and 44.1% men age 16 to 60 years. Of the respondents, 81.8% came from West-Germany and 18.2% from East-Germany. Missing data on the value measures averaged 4.7%, ranging from 4.5% to 5.1%

4.2 Measures

We measured the ten values with a German version of the PVQ (Bamberg et al., in press). The PVQ includes short verbal portraits of 40 different people, gender-matched with the respondent (Schwartz, 2005b; Schwartz, et al., 2001). Each portrait describes a person’s goals, aspirations, or wishes that point implicitly to the importance of a value, using two sentences. For example: “Thinking up new ideas and being creative is important to him. He likes to do things in his own original way” describes a person for whom self-direction values are important. Respondents report the similarity of the person described to themselves on a Likert-type rating scale. We infer respondents’ own values from their self-reported similarity to people who are described implicitly in terms of particular values.

Time limitations led us to reduce the number of items from 40 to 28 and the descriptions from two to one sentence each. For the first purpose, we performed an exploratory factor analysis on data from Hinz et al. (2005). After fixing the number of factors to ten, we selected the two or three items with the highest loadings on each factor. Before selecting single sentences for each item, we conducted a pilot study with two groups of undergraduate students to assess effects of this approach on reliability and validity. One group of students (n = 69) received the 28 two-sentence items, the second group (n = 68) received 56 single-sentence (i.e., 28 x 2) items. The two forms of the questionnaire differed only sligthly in internal consistency and correlations with external criteria. We therefore decided to construct a 28-item version of the PVQ with one sentence per item, selecting the three items with the highest item-total correlation for each factor, with the exception of universalism and tradition that were each measured with two items. Respondents rated their similarity to the person described in each item on a 4-point scale from 1 (*very dissimilar*) to 4 (*very similar*).

4.3 Modeling procedure

We used LISREL 8.54 (Jöreskog and Sörbom, 1993) to perform multigroup analyses. We compared three educational groups: individuals who had completed lower secondary school ("Hauptschulabschluss"; *low)*, secondary school ("Realschulabschluss"; *moderate*), and high school ("Allgemeine Hochschulreife"; *high*). The empirical covariance matrix of the items for each educational group served as the input. We used maximum likelihood as the estimation method. The sample size specified in the LISREL syntax was the median of the sample sizes in the various cells of each matrix. Because we intended to estimate the latent means, we added a vector of manifest means as input. With regard to the parameter matrices, we added the τ_x-vector and the κ-vector.

We conducted the analyses of invariance as follows. Each latent variable was measured with three items (indicators) (two for universalism and tradition, as noted above). We applied a new approach by Little, Slegers, and Card (2006) to scale the latent variables and to set their origins. Traditionally, this is done by fixing the first loading of a latent variable to one and by setting the first intercept to zero. If these parameters are not invariant across groups, however, this approach leads to a misfit of the model. In contrast, Little et al. (2006) propose estimating all factor loadings (and intercepts) but setting constraints that yield loading estimates which equal 1 *on average* and intercept estimates that *sum* to zero. When these constraints are imposed, all of the loadings together set the scale of the latent variable, although none is fixed to a specific value. Analogously, all of the intercepts together set the origin of the latent variable without fixing one intercept to zero.

We evaluated model fit with the root mean square error of approximation (RMSEA, Browne and Cudeck, 1993), the comparative fit index (CFI, Bentler, 1990), and the Akaike information criterion (Akaike, 1987). Values close to .95 for CFI and below.06 for RMSEA suggest a good fit (Hu and Bentler, 1999). Regarding the AIC, the model with the lowest value is preferred.

In a first step, we test for *full* parameter invariance, that is, we constrain the complete respective parameter matrices to be equal across the groups (e.g., $\Lambda_A = \Lambda_B = \Lambda_C$). If this step leads to a significant increase in chi-square ($\Delta\chi^2$), we use information from the modification indices and relax the constraints of the parameter with the highest modification index (cf. Byrne et al., 1989; Marsh and Hocevar, 1985; Steenkamp and Baumgartner, 1998). We then compare this *partially* invariant model with the initial reference model in which all of the respective parameters are unconstrained.

5. Results

5.1 Tests of Measurement Invariance in Educational Groups

Table 3 displays the fit indices for the models that tested measurement invariance. The initial model that assessed configural invariance (Model A) resulted in an acceptable fit ($\chi^2(915) = 1{,}808.36$, RMSEA = .044, CFI = .955, AIC = 2702.36).

The second step, testing *full metric invariance* (Model B) also yielded an acceptable fit: The chi-square increase was not significant ($\Delta\chi^2\,(36) = 41.05$, $p > .05$). The *full scalar invariant model* (Model C) failed as the chi-square increased significantly ($\Delta\chi^2(36) = 149.85$, $p < .001$). Relaxing the constraints for six intercepts in the high education group, one intercept in the moderate education group, and two in all of the groups (Model D) yielded a non-significant difference compared with the metrically invariant model (Model B) ($\Delta\chi^2(12) = 20.84$, $p > .05$). For some latent variables (tradition, self-direction, universalism, and hedonism), however, *partial scalar invariance* could not be established in the high education group. Therefore, any differences between the latent means of these latent variables when comparing this group to the others must be interpreted with caution.

The fully invariant model for *factor variances* (Model E) also failed ($\Delta\chi^2(20) = 46.82$, $p<.001$). However, after relaxing the equal factor variance constraint for security in the high education group (Model F), the increase was no longer significant ($\Delta\chi^2(9) = 9.55$, $p >.05$).

Constraining the *factor covariances* to be equal across the groups (Model G), significantly increased the chi-square ($\Delta\chi^2(90) = 134.22$, $p < .01$). To obtain a partially invariant model (Model H), we relaxed the constraints for three covariances in the high education group, two in the moderate education group, and one in all of the groups ($\Delta\chi^2(83) = 101.04$, $p > .05$).

We tested *full latent mean invariance* only for those latent means that had shown at least partial scalar invariance. Thus, we did not constrain the latent means of tradition, self-direction, universalism, and hedonism to be equal across the groups and let them be estimated freely in the high education group. Constraining the rest of the latent means across the groups impaired the model (Model I) significantly ($\Delta\chi^2(16) = 182.41$, $p < .001$). The partially invariant model (Model J) showed significant mean differences when comparing the high education to the low and moderate education groups for benevolence, security, power, and conformity. Moreover, the moderate education group had higher means for self-direction and lower means for tradition than the low education group. All three groups differed significantly from one another on tradition and self-direction values ($\Delta\chi^2(10) = 9.93$, $p > .05$).

The final analysis concerned *invariance of the error variances* (Model K). As in all of the other models, the fully invariant model failed ($\Delta\chi^2(56) = 248.84$, $p < .01$). Only after relaxing the constraints for eight error variances in the high education group, one in the moderate education group, and three in all three groups did we obtain a model (L) that did not differ significantly from model J ($\Delta\chi^2(40) = 44.17$, $p > .05$). Because only one factor variance, for security, was statistically different, these results can be interpreted in terms of reliability.

5.2 Differences among the Three Education Groups

Differences in the factor loadings, item intercepts and error variances. Table 4 displays the absolute values of all of the measurement parameters (factor loadings, item intercepts, and error variances) for the three groups. Where a parameter was invariant across all three groups, a single parameter value appears. Where a parameter varied significantly across groups, different parameter values are presented.

The factor loadings were equal across education groups. This indicates full metric invariance. It shows that the three groups use the same metric. The ten constructs also appear to have the same meaning across groups. The item intercepts and measurement errors, however, reveal a more diverse picture. The high education group differed from one or the other group on eight intercepts. The differences were not systematic: The intercept was lower in the high education group in five cases and higher in three cases. The low and moderate education groups differed on three intercepts, with two higher in the low education group.

Regarding measurement errors, 12 of the 28 items differed significantly. Eleven of these differences were between the high education group and the others. Seven error variances were lower in the high education group and four were higher.

Differences in latent means. For most values, we established invariance of the factor loadings and at least partial scalar invariance. In these cases it was possible to test mean differences. Because the high education group did not exhibit partial scalar invariance for self-direction, hedonism, universalism, and tradition, we did not test mean differences for these values. We permitted the latent means for these values to be freely estimated for the high education group, rather than constraining them to be equal. Consequently, differences between these latent means in the high vs. the low or moderate education groups were not tested for significance and must be interpreted with caution. We refer to these as *descriptive differences.*

Table 5 shows the latent means of the groups. In addition to the absolute means, we computed the effect size, Hedges' g, with the formula $g = (\kappa^1 - \kappa^2) / S_{pooled}$, where $S_{pooled} = \sqrt{(\phi_{11} + \phi_{22}) / 2}$. Because standardized effect sizes are easier to understand, we transformed Hedges' g into r with the formula $r = \sqrt{d^2 / (d^2 + 4)}$, where d = g.

Table 5 reveals that the high education group differed statistically on four of the ten latent means. They attributed significantly more importance than the others to benevolence and power values and less importance to conformity and security values. In addition, from a descriptive point of view, they attributed more importance to self-direction values and less to hedonism and tradition. In contrast, the low and moderate education groups differed statistically on only two latent means. The moderate education group attributed more importance to self-direction and tradition values. As expected, the lower education group attributed more importance than the high education group to the three conservation values (security, tradition, and conformity). Unlike Schwartz (2005b), we did not find a substantially greater emphasis on stimulation and hedonism values as a function of more education.

Differences in the factor covariances. Because all of the latent variables except security had invariant variances, we can regard differences in covariances among all other values as differences in correlations. Analogous to the treatment of descriptive mean differences mentioned before, differences in the correlations with security should be interpreted cautiously. Table 6 shows the correlations among the ten constructs. Most of these intercorrelations did not differ across groups. Only five differences were significant. In addition, these differences were of low magnitude.

6. Discussion

We investigated the factor structure of a modified form of the Portraits Values Questionnaire (PVQ, Schwartz, 2005a), assessing the assumption of population homogeneity across different levels of education. We employed multigroup confirmatory factor analysis to test cross-group equality constraints on the various parameters of the measurement model.

These tests confirmed that the modified measurement instrument for values based on the PVQ successfully measures the 10 types of values postulated by Schwartz. In contrast to most earlier studies, we used a population survey and confirmatory factor analysis rather than smallest space analysis to test the factorial structure of values. This allowed us to test the number of values and the factorial validity of the instrument formally.

We further investigated whether the common assumption of homogeneity of population surveys holds for the PVQ. Following Steenkamp and Baumgartner (1998), we tested whether all or some of the factor loadings, measurement errors, factor variances, covariances, intercepts, and latent means are equal across different educational groups.

We had expected less educated respondents to give more random answers, in keeping with the political attitudes literature (Converse, 1964; Judd et al., 1981; Zaller, 1995). This did not occur. The set of factor loadings was fully invariant and only nine of the 28 indicators showed different measurement errors. The less educated group had higher measurement errors in six indicators. These results suggest that individuals with different levels of education differ less in the thought they devote to values than to political beliefs (Saris and Sniderman, 2004).

The analysis of latent means presupposes partial invariance of loadings and intercepts. This held for most of the indicators. Although the tests of mean differences revealed eight significant differences among educational groups, only the conservation values (security, tradition, and conformity) exhibited substantial effect sizes. As hypothesized, less educated respondents attributed more importance to these values.

We employed a new scaling method from Little et al. (2006) to scale the factor loadings and origins of the latent variables. We constrained the factor loadings of each latent variable to equal 1 on *average* and the *sum* of the intercepts to equal zero. This method avoids the dangers of erroneously fixing a non-invariant loading to 1 or fixing a non-invariant intercept to zero.

Finally, we note some limitations of the current study. We performed the MGCFA only on two or three indicators per latent variable. This was due to time limits of the larger survey. Having two indicators for tradition and universalism values was the minimum necessary for identification and for testing the factorial structure. However, it led to problems in testing measurement invariance. Because partial invariance requires at least two indicators, even one non-invariant indicator obviates establishing partial invariance. This occurred with the test of scalar invariance for universalism, where one non-invariant item-intercept made it impossible to establish partial invariance and hence to test for mean invariance. Tests of invariance in the value inventory of the European Social Survey, where most values are measured with two items, suffered from the same problem (Davidov et al., in press). Therefore, if group comparisons are planned, we recommend including at least three indicators for each construct.

Tests of mean invariance are methodologically superior to the traditional tests which simply assume metric and scalar invariance. Nonetheless, there are some dangers. Like other simple mean comparisons or zero-order relationships, tests of mean invariance across groups cannot rule out the possibility of spurious relationships. A third variable that correlates with the group variable may cause significant mean differences among groups. Muthén (1989) proposed a MIMIC modeling approach in such cases. This approach includes several group variables in the model to predict differences in the latent variables. Tests of scalar invariance can then be performed by estimating direct structural effects from a group variable to the indicators of the latent variables. Significant direct effects on latent variables indicate latent mean differences. The MIMIC approach cannot test metric, variances, covariances, and error variances. Hence, the best solution may be to combine the MGCFA and MIMIC approaches. Future research should evaluate such a combined strategy.

7. Conclusions

Most research with instruments that measure the ten basic values in the Schwartz theory focuses on cross-cultural comparisons. Davidov et al. (in press) argue that measurement invariance is a prerequisite for cross-cultural or cross-national comparisons. But measurement parameters may also differ substantially within populations. Cross-cultural comparisons typically assume within-population invariance. This study demonstrates that the ten-factor model postulated by Schwartz holds across different educational groups in one society. The factor loadings were also invariant across educational groups. For most of the indicators, even the test of equal intercepts, a prerequisite for comparing latent means, produced no significant differences. This test ruled out only a minority of mean comparisons, those for self-direction, hedonism, universalism, and tradition. For these values, the intercepts in the high education group differed from those in one or both of the other groups.

Our findings should not be generalized to other constructs and groups. For example, education and interest in politics strongly affect factor loadings and measurement errors in the measurement of political attitudes (Saris and Sniderman, 2004; Zaller, 1995). This points to the importance of studying the effects of such variables as age, gender, social status, and salience of the survey topic in population surveys within-societies.

Measurement invariance should be added to the well-established criteria of reliability, homogeneity, and validity when constructing and validating a new scale. The goal is to construct scales with full invariance. A scale with partial invariance of the underlying measurement model may suffice in a structural equation model. If a researcher uses manifest composite scores, however, partial invariance is probably not sufficient because both invariant and non-invariant items are aggregated to form the composite.

8. References

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, *52*: 317-322.

Bamberg, S., Davidov, E., Herrmann, J., Schmidt, P., & Schwartz, S. H. (in press). Wie lassen sich Werte messen? Validierung des neuen Messinstruments "Portraits Value Questionnaire" [How can values be measured? Validation of the new measurement instrument "Portraits Value Questionnaire]. Zeitschrift für Sozialpsychologie.

Baumgartner, H., & Steenkamp, J.-B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters, 9* (1): 21-35.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin,, 107*: 238-246.

Billiet, J. (2002). Cross-cultural equivalence with structural equation modeling. In: P. P. Mohler (ed.), *Cross-Cultural Survey Methods*. New Jersey: John Wiley & Sons, Inc., pp. 247-264.

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In: K. A. Bollen & J. S. Long (eds.), *Testing structural equation models*. Newbury Park, CA: Sage, pp. 36-162.

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* (3): 456-466.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research, 20*: 389-417.

Converse, P. (1964). The nature of belief systems in mass publics. In: D. Apter (ed.), *Ideology and discontent*. New York: Free Press, pp. 206-261.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*: 281-302.

Davidov, E., Schmidt, P., & Schwartz, S. H. (in press). Bringing values back in: A multiple group comparison with 20 countries using the European Social Survey. *Public Opinion Quarterly*.

De Beuckelaer, A. (2005). Measurement invariance issues in international management research. Limburg: Limburgs Universitair Centrum.

Hayduk, L. A. (1989). *Structural equation modeling - essentials and advances*. Baltimore and London: The John Hopkins University Press.

Hinz, A., Brähler, E., Schmidt, P., & Albani, C. (2005). Investigating the circumplex structure of the Portraits Value Questionnaire (PVQ). *Journal of Individual Differences, 26* (4): 185-193.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*: 1-55.

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

Jöreskog, K. G., & Sörbom, D. (1993). *Lisrel 8 user's reference guide*. Chicago: Scientific Software International.

Judd, C., Milburn, M., & Krosnick, J. (1981). Political involvement and attitude structure in the general public. *American Sociological Review, 46*: 660-669.

Kruskal, J. B., & Wish, M. (1994). *Multidimensional scaling*. California: Newsbury Park.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32* (1): 53.

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* (1): 59-72.

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* (3): 562-582.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58* (4): 525-543.

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26* (3): 479-497.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research. *Journal of Applied Psychology, 73*: 564-574.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54* (4): 557-585.

Ployhardt, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7* (1): 27-65.

Saris, W. E., & Sniderman, P. M. (2004). *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*. Princeton: University Press.

Schmitt, M. J., Schwartz, S. H., Steyer, R., & Schmitt, T. (1993). Measurement models for the Schwartz values inventory. *European Journal of Psychological Assessment, 9* (2): 107-121.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: M. Zanna (ed.), *Advances in experimental social psychology* (Vol. 25). Orlando, FL: Academic Press, pp. 1-65.

Schwartz, S. H. (2005). Basic human values: Their content and structure across countries. In A. Tamayo & J. B. Porto (Eds.), *Valores e comportamento nas organizações* [Values and behavior in organizations] pp. 21-55. Petrópolis, Brazil: Vozes.

Schwartz, S. H. (2005b). Robustness and fruitfulness of a theory of universals in individual human values. In: A. Tamayo & J. B. Porto (Eds.), *op. cit.* pp. 56-95.

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38* (3): 230–255.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology, 32* (5): 519-542.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariances. *Psychometrika, 43*: 381-396.

Spini, D. (2003). Measurement equivalence of 10 value types from the Schwartz value survey across 21 countries. *Journal of Cross-Cultural Psychology, 34* (1): 3-23.

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

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent means. In: G. R. Hancock & R. O. Mueller (eds.), *Structural Equation Modeling: A Second Course*. Greenwich: CT: Information Age, pp. 119-169.

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

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* (1): 4-69.

Zaller, J. R. (1995). *The nature and origins of mass opinion*. Cambridge: University Press.

Table 1

Equality Constraints and Steps of Measurement Invariance

Constraints	Meaning	Label	Interpretation
No constraints	Same pattern of fixed and non-fixed parameters	Configural invariance	Same model structure in the groups
$\Lambda^{A} = \Lambda^{B} = \ldots = \Lambda^{G}$	Equally constrained matrices of factor loadings	Metric invariance	Same metric in the groups Implications for construct comparability
$\tau^{A} = \tau^{B} = \ldots = \tau^{G}$	Equally constrained vector with item intercepts	Scalar invariance	Same systematic response bias in the groups Prerequisite for latent mean comparison
${\phi_{jj}}^{A} = {\phi_{jj}}^{B} = \ldots = {\phi_{jj}}^{G}$	Equally constrained diagonal of the matrix with factor variances and covariances	Invariance of factor variances	Same heterogeneity of latent variables in the groups Prerequisite to interpret equal factor covariances as equal correlations and equal error variances as equal reliabilities
${\phi_{jk}}^{A} = {\phi_{jk}}^{B} = \ldots = {\phi_{jk}}^{G}$	Equally constrained sub-diagonal of the matrix with factor variances and covariances	Invariance of factor covariances	If equal factor variances, same correlations between factors Implications for construct comparability
$\kappa^{A} = \kappa^{B} = \ldots = \kappa^{G}$	Equally constrained vector with latent means	Invariance of latent means	If equal intercepts, same latent means in the groups
$\Theta^{A} = \Theta^{B} = \ldots = \Theta^{G}$	Equally constrained matrix with error variances and covariances	Invariance of error variances	If equal factor variances, same reliabilities in the groups

Table 2

Values and their Defining Goals

Value	Defining goal
Self-Direction	Independent thought and action
Stimulation	Excitement, novelty, challenge in life
Hedonism	Pleasure or sensuous gratification for oneself
Achievement	Personal success through demonstrating competence according to social standards
Power	Social status and prestige, control or dominance over people and resources
Security	Safety, harmony, and stability of society, of relationships, and of self
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms
Tradition	Respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides
Benevolence	Preserving and enhancing the welfare of those with whom one is in frequent personal contact (the "in-group")
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature

Table 3

Tests for Measurement Invariance across Three Education Groups

Model		*Compared Model*	χ2 (df)	Δχ2 (Δdf)	RMSEA	CFI	AIC
A	Configural invariance		1,808.36 (915)**		.044	.955	2,702.36
B	Full metric invariance	A	1,849.41 (951)**	+ 41.05 (36)	.043	.955	2,631.41
C	Full scalar invariance	B	1,999.26 (987)**	+ 149.85 (36)**	.045	.949	2,669.26
D	Partial scalar invariance	B	1,886.36 (976)**	+ 36.94 (25)	.043	.954	2,578.36
E	Full invariance of factor variances	D	1,933.18 (996)**	+ 46.82 (20)**	.043	.953	2,585.18
F	Partial invariance of factor variances	D	1,915.52 (995)**	+ 29.16 (19)	.043	.954	2,569.52
G	Full invariance of factor covariances	F	2,049.74 (1085)**	+ 134.22 (90)**	.042	.952	2,523.74
H	Partial invariance of factor covariances	F	2,016.56 (1078)**	+ 101.04 (83)	.041	.953	2,504.56
I	Full invariance of latent means[a]	H	2,198.97 (1094)**	+ 182.41 (16)**	.045	.944	2,654.97
J	Partial invariance of latent means	H	2,026.50 (1088)**	+ 9.93 (10)	.041	.953	2,494.50
K	Full invariance of error variances	J	2,275.34 (1144)**	+ 248.84 (56)**	.044	.945	2,631.34
L	Partial invariance of error variances	J	2,083.35 (1129)**	+ 44.17 (40)	.041	.953	2,469.35

Note. ***p* < .01; low education: *n* = 277, moderate education: *n* = 645; high education: *n* = 606; [a]with exception of the tradition, self-direction, hedonism, and universalism in high education.

Table 4

Invariant and Non-Invariant Factor Loadings, Item Intercepts, and Error Variances in Three Education Groups

		Factor loadings			Item intercepts			Error variances		
Latent Variable	Item	education			education			education		
		Low	medium	high	Low	medium	high	Low	medium	high
S-Dir	sd1		.94			.278		.254	.254	.191
	sd2		1.02		-.059	-.211	-.251		.348	
	sd3		1.03		-.218	-.218	-.117	.426	.308	.190
Stm	stm1		1.09			-.344			.359	
	stm2		0.81			.441			.463	
	stm3		1.10			-.097			.359	
Hed	hed1		1.02			-.129		.248	.200	.248
	hed2		0.98		.088	.088	.194	.184	.184	.155
	hed3		1.00		.040	-.044	-.071	.195	.195	.269
Ach	ach1		1.12			-0.246		.341	.341	.255
	ach2		0.95			-0.136			.455	
	ach3		0.93		.382	.382	.239		.448	
Pow	pow1		1.00			-.066			.314	
	pow2		0.86		.297	.397	.297		.388	
	pow3		1.14			-.230			.311	
Sec	sec1		1.15			-.936		.599	.599	.421
	sec2		1.01			.221			.205	
	sec3		0.85			.715		.266	.266	.343

(Table 4 continued)

Latent Variable	Item	Factor loadings			Item intercepts			Error variances		
		education			education			education		
		Low	medium	high	Low	medium	high	Low	medium	high
Con	con1		1.09			-.080			.283	
	con2		0.68			.684			.554	
	con3		1.23			-.604		.225	.225	.321
Trad	trad1		1.13			-.025			.443	
	trad2		0.87		.025	.025	-.071	.588	.407	.328
Ben	ben1		1.12			-.333			.129	
	ben2		1.08		-.201	-.201	-.095	.151	.151	.113
	ben3		0.80			.534			.257	
Uni	uni1		1.00			.053			.256	
	uni2		1.00		-.053	-.053	-.191	.247	.356	.405

Note. S-Dir = self-direction, Stim = stimulation, Hed = hedonism, Ach = achievement, Pow = power, Sec = security, Con = conformity, Trad = tradition, Ben = benevolence, Uni = universalism

Table 5

Latent Means of the Education Groups

	Means			Effect sizes (r)		
	Low	Moderate education	High	low vs. moderate	moderate vs. high	low vs. high
Self-direction	3.32*	3.42*	3.47$	.14	.07	.21
Stimulation	2.30	2.30	2.30	.00	.00	.00
Hedonism	3.49	3.49	3.41$	.00	.10	.10
Achievement	2.87	2.87	2.87	.00	.00	.00
Power	2.39	2.39	2.65*	.00	.22	.22
Security	3.37	3.37	3.10*	.00	.31	.31
Tradition	2.74*	2.56*	2.43$	.13	.16	.28
Conformity	3.16	3.16	2.93*	.00	.24	.24
Benevolence	3.47	3.47	3.36*	.00	.14	.14
Universalism	3.34	3.34	3.30$	.00	.04	.04

Notes. $Mean invariance not tested because of failure of scalar invariance; effect sizes of r = .00 indicate a non-significant difference in the latent means

Table 6

Correlations between the Latent Variables across the Three Education Groups

	Self-direction	Stimulation	Hedonism	Achievement	Power	Security	Tradition	Conformity	Benevolence
Self-direction									
Stimulation	.28								
Hedonism	.65	.45							
Achievement	.49	.41	.34						
Power	.56 .45 .56	.39	.16	.73					
Security	.32	-.21	.39	.34	.08				
Tradition	.12	-.01	.23	.26	.17	.61			
Conformity	.17	-.04	.28 .28 .18	.35	.12	.60 .60 .72	.48 .67 .63		
Benevolence	.42	.04	.49	.17	-.02	.56	.36	.53	
Universalism	.35	.02	.44	.08 .18 .08	-.06	.56	.38	.52	.59

Note. Low education: $n = 277$, moderate education: $n = 645$; high education: $n = 606$; three correlations in a cell reflect group specific correlations in the order low, moderate, and high education; correlations >.11 are significant (two-tailed), correlations with security should be interpreted with caution as the three education groups had significantly different variances in security

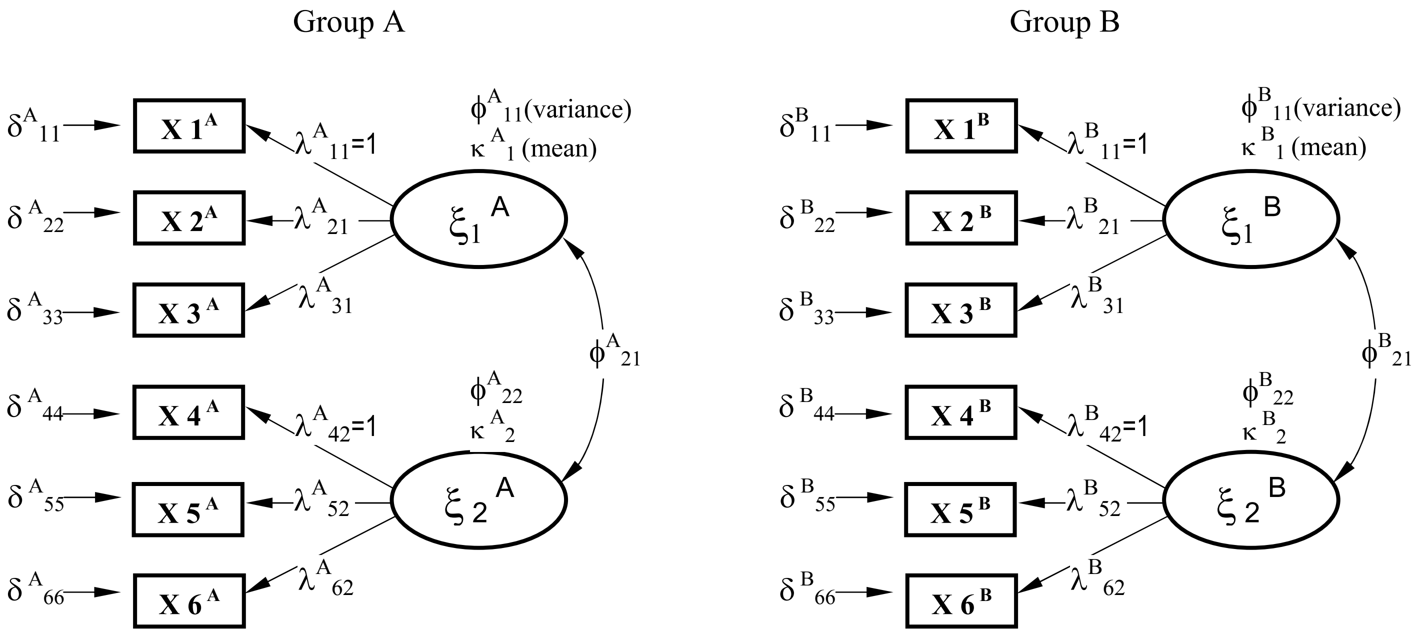

Figure 1. A two-group measurement model.