Analysis Models for Comparative Surveys[1]

Joop J. Hox, Edith D. de Leeuw, Matthieu J.S. Brinkhuis

Utrecht University, the Netherlands

21.1 Introduction

The aim of cross-national and cross-cultural surveys is to compare results across countries or cultural groups. Since the early international surveys in the 1970's (cf. Harkness, Mohler & Van de Vijver, 2003) the number of comparative survey programs have been growing and this trend is likely to continue (Lynn, Japec, & Lyberg, 2006). Over the years the methodological knowledge base on comparative studies has been growing and especially methods for questionnaire design (e.g., Harkness, Van de Vijver & Johnson, 2003, Smith, 2003), harmonization (e.g., Braun & Mohler, 2003), and questionnaire translation (e.g., Harkness, 2009) have been developed. Although cross-national surveys aim to use common data collection methods statistics, there generally remain a number of differences in the data collection process in the participating countries, in addition to the inevitable language differences. For instance, differences in the data collection methods (Sjak & Harkness, 2003) or mixed-mode strategies (De Leeuw, 2005), in details of the fieldwork of the data collection agencies (De Heer, 1999) and in survey climate, economic condition and culture (e.g., De Leeuw & De Heer, 2002). Even differences in the achieved response rates can affect the comparability (Couper & de Leeuw, 2003).

There are three main statistical issues in comparative research. Firstly, there is the issue of measurement equivalence. Can we assume that the instruments measure the same constructs in the same way, how can we assess whether we have measurement equivalence, and if not how can we correct measures in order to achieve valid comparisons? Secondly, if measurement equivalence is achieved, the analysis must deal with the issues of analyzing relationships within and between countries (or other contexts). That is, relationships can be established at the individual level within each country, but in comparative research the central issue is often the question whether such relationships are different between countries. Finally, the question is whether there are stable relationships between characteristics at the country level.

The classic statistical approach to deal with these questions is undoubtedly structural equation modeling (SEM) using a multi-group analysis. This analysis method makes it possible to test equivalence of measurement models and equivalence of structural (substantive) models. Using modern software that can model categorical data, modern measurement models like Item Response (IRT) models can be subsumed under SEM.

However, when the number of groups or countries becomes larger, multi-group

[1] Manuscript for Hox, J.J., de Leeuw, E.D. & Brinkhuis, M.J.S. (2010). Analysis models for comparative surveys. Pp. 395-418 in Harkness, J., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P., Pennell, B.E. and Smith, T.W. (Eds.) *Survey Methods in Multicultural, Multinational, and Multiregional Contexts.* Hoboken, NJ: Wiley.

SEM becomes unwieldy. The software setups become complicated, especially if subtle differences in measurement properties must be included. The statistical model also becomes complicated. Multi-group SEM is a fixed effects model, which means that it takes each group or country as given and the set of countries as the complete universe to generalize to. Unless a great many equality constraints are imposed, SEM estimates a unique set of parameter values for each country, which results in a large model. A random effects model, such as multilevel modeling (MLM) treats the countries as a sample from a larger population. Instead of estimating a different parameter value for each country, it assumes a (normal) distribution of parameter values and estimates its mean and variance (and covariances). This makes MLM much more parsimonious than SEM when the number of countries becomes large. A second advantage is that differences between countries can in turn be modeled using country-level characteristics. Simulations show that MLM can be used with second-level (group-level) samples as low as 20 (Maas & Hox, 2005), which means that the larger collaborative comparative surveys involve enough countries to consider employing multilevel modeling methods.

Recently, latent class modeling (LCM) has come into use, which does not model differences between countries as random effects, but attempts to identify latent classes of similar respondents. This approach combines some advantages of SEM and MLM: differences between countries are assumed to be differences between groups, but the groups are latent classes, and the number of latent classes is assumed to be much smaller than the number of countries or groups.

This chapter consists of three major sections. First, it compares the three statistical approaches outlined above (SEM, MLM, LCM). The underlying statistical models are explained at a general, non-technical level. The emphasis is on a comparison of the major characteristics: what is the structure of the model, what kind of questions can be answered using this approach and what are the important statistical assumptions underlying the model? The second section contains a small simulation study that compares the three approaches in a situation typical for comparative research: a relatively small number of groups (countries) but within groups relatively large sample sizes. This simulation addresses the question how accurate the estimates are with a small number of countries, and if there is sufficient power to detect anomalies in the measurement model. The third section applies and compares the three approaches on a real data set from a large scale comparative survey. A final section summarizes the findings and gives recommendations for model use and further methodological research.

21.2 Statistical Considerations in Comparing SEM, MLM, and LCM

21.2.1 Structural Equation Modeling (SEM)

The classic method for dealing with data from large cross-national surveys is to use multi-group structural equation modeling (SEM). This approach derives from the seminal work of Jöreskog (e.g., Jöreskog, 1971). Structural equation modeling provides a very general and convenient framework for statistical analysis that includes several traditional multivariate procedures as special cases, for example factor analysis, regression analysis, discriminant analysis, and canonical correlation. Structural equation models are often

visualized by a graphical *path diagram* (see for example Figure 21.1). In the path diagram, observed variables are represented by a square, and latent variables by a circle. The statistical model is usually represented in a set of matrix equations. Commonly, a distinction is made between the measurement part of the model and the structural part of the model. Figure 21.1 is a graphical presentation of the full structural equation model in the notation used by Bollen (1989). The diagram shows a measurement model for the latent factor ξ (ksi) and its associate observed indicators *x* and for the latent factor η_2 and its associated observed indicators *y*. The relationships between the latent variables ξ and η_1 and between η_1 and η_2 constitute the structural model. The latent variable ξ is denoted as exogenous, because there are only paths from it to other variables, and the latent variables η are denoted as endogenous because there are paths leading towards them. This distinction is important, because for the endogenous variables multivariate normality is assumed, but not for the exogenous variables (Bollen, 1989). Note that in this example the latent variable η_1 has no empirical indicators; it is solely defined by its role in the structural model. Such latent variables have their use in specialized models, but it is more common to have empirical indicators for all latent variables in the model. It is also possible to have a path model with only observed variables. In that particular case there is no explicit measurement model. Implicitly, it is then assumed that all variables are measured identically in all countries or groups, but it is not possible to test this important assumption.

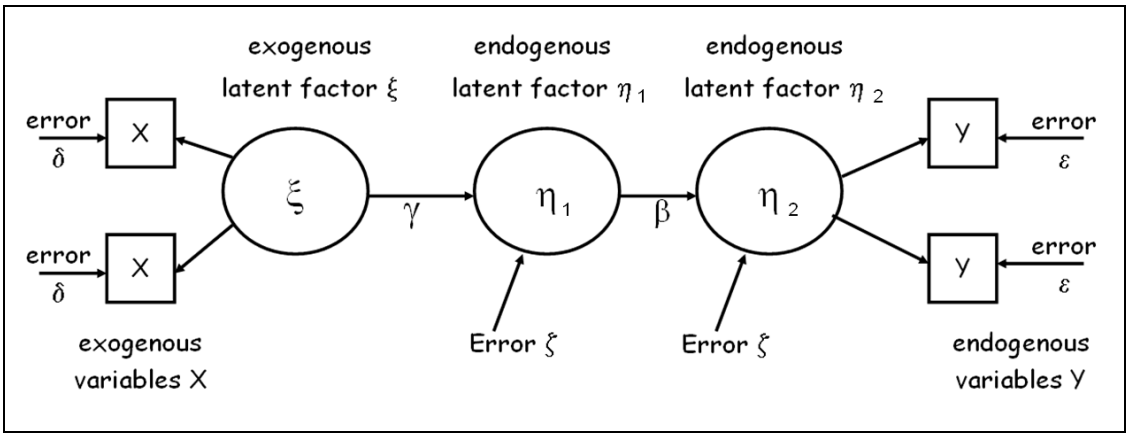

Figure 21.1. Diagram of full structural equation model

In SEM, it is usually assumed that the endogenous variables follow a multivariate normal distribution, which implies that the vector of means and the covariance matrix contain all the relevant information. The method most widely used for estimation is Maximum Likelihood (ML) estimation, assuming multivariate normal data and a reasonable sample size, e.g. about 200 observations. There are a variety of estimation procedures that can be used for non-normal continuous data. With non-normal data, including ordinal categorical data, the means and the covariance matrix do not represent all the information, and therefore these alternative estimation methods require raw data.

Comparative research generally employs large samples for each country.

Statistical tests for model fit have the general property that their power varies with the sample size. As a result, with large samples, we will almost always reject our model, even if the model actually describes the data very well. Conversely, with a very small sample, the model will always be accepted, even if it fits rather badly.

In comparative survey research, samples are typically large, e.g. above 1000 respondents in each country. As a consequence, even small discrepancies between the model and the data will lead to a significant test result and a formal rejection of the model. Given the sensitivity of the chi-square statistic to sample size, researchers have proposed a variety of alternative fit indices to assess model fit. Most of these fit indices not only consider the fit of the model, but also its simplicity. A saturated model, that specifies all possible paths between all variables, will always fit the data perfectly, but it is just as complex as the observed data itself. If two models have the same degree of fit, the principle of parsimony indicates that we should prefer the simpler model.

Modern SEM software computes a bewildering array of goodness-of-fit indices. For an overview and evaluation of a large number of fit indices, including those mentioned here, we refer to Gerbing and Anderson (1993). All fit indices are functions of the chi-square statistic, but some include a second function that penalizes complex models. For instance, Akaike's information criterion (AIC) is equal to the chi-square statistic plus twice the number of parameters in the model. Often used fit indices are the TLI (Tucker-Lewis Index) and the CFI (Comparative Fit Index), with values > 0.90 indicating a good fit, and 1.0 indicating a perfect fit (Bentler, 1990). A different approach to model fit is to accept that models are only approximations, and that perfect fit may be too much to ask for. Instead, the problem is to assess how well a given model approximates the true model. This view led to the development of an index called RMSEA (Root Mean Square Error of Approximation). If the approximation is good, the RMSEA should be small, with values <.05 indicating a good fit and values <.08 indicating an acceptable approximation (Browne & Cudeck, 1993).

If the fit of a SEM model is not adequate, it has become common practice to modify the model, by deleting parameters that are not significant and by adding parameters that improve the fit. To assist in this process, most SEM software can compute *modification indices* for each fixed parameter. The value of a given modification index is the minimum amount that the chi-square statistic is expected to decrease if the corresponding parameter is freed. Researchers often use this information to conduct a sequence of model modifications. At each step a parameter is freed that produces the largest improvement in fit, and this process is continued until an adequate fit is reached. For example, if in a confirmative factor model a loading that is fixed to zero shows a large modification index, we may free this parameter and estimate its value. This will improve the fit of the model at the cost of one degree of freedom.

The statistical model is usually described by separate equations for the measurement and the structural model. Thus, the equations for the measurement model in Figure 21.1 are in matrix format:

$$\mathbf{x} = \Lambda_x \xi + \delta \qquad (21.1)$$

$$\mathbf{y} = \Lambda_y \eta + \varepsilon\,, \qquad (21.2)$$

and for the structural model it is:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}\,. \qquad (21.3)$$

In this notation, $\boldsymbol{\Lambda}_x$ (lambda-x) is the factor matrix for the exogenous variables, and $\boldsymbol{\Lambda}_y$ (lambda-y) is the factor matrix for the endogenous variables. The notation used here distinguishes between independent variables *x* and ξ and dependent variables *y* and η. The paths from exogenous variables to endogenous variables are denoted γ (gamma), collected in the matrix Gamma (**Γ),** while the path coefficients among endogenous variables are denoted β (beta), collected in the matrix Beta (**B**). The distinction between measurement model and structural model is conceptual rather than statistical, and most SEM software hides these complications from the user. However, in comparative research this distinction between measurement and structural model is of particular interest, because equivalence of the measurement model across different groups is of central importance in order to achieve valid comparisons (cf. Bechger, van den Wittenboer, Hox & de Glopper, 1999). This means that before we can compare the structural (substantive) models for different countries, we must make sure that the factor matrices in the measurement model are in fact equal across countries.

This question whether measurement invariance may be assumed is generally investigated using multigroup SEM. Multigroup SEM makes it possible to test hypotheses concerning equivalence between groups, such as the hypothesis of measurement equivalence (cf. Vandenburg & Lance, 2000). The weakest form of measurement equivalence is *functional equivalence*, sometimes also denoted as configural equivalence, where the assumption holds that the different countries share a measurement model that has the same factor structure. This is a very weak form of measurement equivalence, because it only allows us to conclude that we are probably studying the same construct in each country, but there is no way to statistically compare the countries or to examine their differences. To analyze variation across countries, we need to prove first that the items that comprise a specific measuring instrument operate equivalently across the different populations or countries (Jöreskog, 1971, Meredith, 1964, 1993). In our SEM notation, different groups are denoted by superscripts. Thus, the hypothesis that there is measurement equivalence across two groups for the latent variable η would be written as:

$$\Lambda_{\mathbf{y}}^{(1)} = \Lambda_{\mathbf{y}}^{(2)}\,. \qquad (21.4)$$

If the factor loadings are invariant across all countries, we have a form of equivalence that is referred to as *scale equivalence* (Vanderburg & Lance, 2000). Scale equivalence means that the measurement scale is comparable across countries. Although the ideal is achieving complete measurement invariance across all countries, in practice a small amount of variation is often judged acceptable, which leads to partial measurement invariance (Byrne, Shavelson & Muthén, 1989; Steenkamp & Baumgartner, 1998).

When (partial) scale equivalence is achieved, it is possible to analyze differences between countries statistically. This includes the question whether paths in a specified causal structure are invariant across populations and the wider question whether the same

structural model holds in all countries. When comparisons of means of latent constructs are involved across countries, additional invariance restrictions are needed for the intercepts of the observed variables. If these intercepts can be considered invariant across countries, we have a form of equivalence that is referred to as *metric equivalence* (Vanderburg & Lance, 2000). When metric equivalence holds, the actual scores can be compared across countries. Again, the ideal is achieving complete invariance for all intercepts across all countries, but in practice a small amount of variation is judged acceptable, which leads again to partial measurement invariance. Regarding the minimal requirements for partial invariance, both Byrne et al. (1989) and Steenkamp and Baumgartner (1998) state that for each construct in addition to the marker item that defines the scale (marker item loading fixed at 1 and intercept fixed at 0) at least one more indicator must have invariant loadings and intercepts.

If (partial) scale equivalence has been established for the measurement model, we can use theoretical reasoning to specify substantive (structural) models for the relationships among constructs in different countries, and we can assess whether these relationships are the same across all countries. If (partial) metric equivalence has been established, we can also test if the countries differ on the means of the constructs. However, a major shortcoming of multigroup SEM in the context of comparative survey research is that there are no provisions to specify models that include country level variables to explain differences in these means.

It should be noted that the terminology used for the various forms of equivalence is not well standardized. For instance, the term scalar equivalence is sometimes used for metric equivalence (cf. Van de Vijver & Leung, 1998). Happily, there is consensus on the constraints that are needed to make specific comparisons valid. To compare relationships (regression, correlations) between countries we need equivalence constraints on the loadings, and to actually compare scores between different countries we need additional equivalence constraints on the intercepts.

21.2.2 Multilevel Modeling (MLM)

Multilevel models are specifically developed for the statistical analysis of data that have a hierarchical or clustered structure. Such data arise routinely in various fields, for instance in educational research where pupils are nested within schools, or in family studies with children nested within families. Clustered data may also arise as a result of a specific research design. An example is longitudinal designs; one way of viewing longitudinal data is as a series of repeated measurements nested within individual subjects. Comparative surveys also lead to a multilevel structure with respondents nested within countries. In comparative research there are besides respondent level variables also level variables measured on country level. In contrast to SEM, MLM can include these country level variables as explanatory variables in the model.

The most used multilevel model is the multilevel regression model (Raudenbush & Bryk, 2002; Goldstein, 2003; Hox, 2009). It assumes hierarchical data, with one response variable measured at the lowest level (e.g., respondents) and explanatory variables at all existing levels (e.g., respondent and country). Conceptually the model is often viewed as a hierarchical system of regression equations. For example, assume we have data in *J* groups, and a different number of individuals N_j in each group. On the individual (lowest) level we

have the dependent variable Y_{ij} and the explanatory variable X_{ij}, and on the group (higher) level we have the explanatory variable Z_j. Thus, we have a separate regression equation in each group:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}. \quad (21.5)$$

The β_j are *random coefficients*, assumed to vary across groups. They are modeled by explanatory variables at the group level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \quad (21.6)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \quad (21.7)$$

Substitution of (21.6) and (21.7) in (21.5) gives:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{1j} X_{ij} + u_{0j} + e_{ij} \quad (21.8)$$

In general, there will be more than one explanatory variable at each level. What these equations make clear is that there is a distinction between individual (respondent) level explanatory variables and group (country) level explanatory variables. Thus, in comparative surveys, we have respondent variables on the level of the respondents and country variables on the level of the countries. The outcome variable is always on the individual (respondent) level. For individual level predictors only, we can hypothesize that the regression coefficients of these variables differ between countries. If they do, we can attempt to explain the variation between countries using country level variables. As equation 21.8 makes clear, this is done by adding a cross-level interaction ($X_{ij}Z_j$) to the model.

The multilevel regression model is a univariate model, although it can be used to analyze multivariate outcome data by introducing an additional lowest level for the outcome variables. Multilevel structural equation modeling (MSEM) is more flexible. In multilevel SEM, we assume sampling at two levels, with both between group (group level) and within group (individual level) covariation. This approach includes a measurement and structural model at each level, with random slopes and intercepts (Mehta & Neale, 2005). Muthén and Muthén (2007) and Skrondal and Rabe Hesketh (2004) have suggested extensions of the conventional graphic path diagrams to represent multiple levels and random slopes. We use the notation proposed by Muthén and Muthén here, since it is close to the usual SEM path diagram (see Figure 21.2).

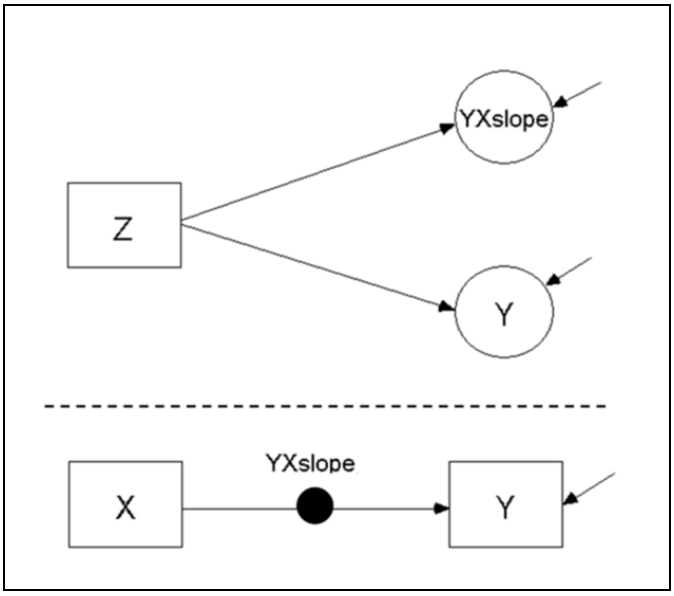

Figure 21.2 Path diagram for a two-level regression model

The two-level path diagram in figure 21.2 is based on the separate equations representation of the model, as provided by equations 21.5-21.7. The within part of the model in the lower area specifies that Y is regressed on 21. The between part of the model in the upper area specifies the existence of a group level variable Z. There are two latent variables represented by circles. The group level latent variable Y represents the group level variance of the intercept for Y. The group level latent variable XYslope represents the group level variance of the slope for X and Y, which is on the group level regressed on Z. The black circle in the within part is a new symbol, used to specify that this path coefficient is assumed to have random variation at the group level. This variation is modeled at the group level using group level variables.

Multilevel SEM provides a completely new approach to testing the equivalence of the measurement models across groups. Equivalent measurement models in this formulation means that the same factor model must fit in all groups, with no factor loading having a coefficient that varies across groups. In other words, testing whether factor loadings for the measurement model have significant variation across groups is a test on the scale equivalence of the measurement (c.f. 21.2.1).

Figure 21.3 shows the measurement model on the individual and the country level. There is no methodological requirement to constrain the corresponding loadings to be equal at the individual and the country level, but it is advantageous to do so. If there is scale equivalence across both levels, this indicates that the same process may be at work at both levels (cf. Van de Vijver & Poortinga, 2002). With respect to measurement invariance, it is interesting to note how this model handles unequal intercepts. At the within level, the observed variables are group-mean centered, and there are no intercepts (all intercepts are zero). Country-level variation in intercepts is modeled in the between (country level) model by allowing residual measurement errors on the latent variables η. As a result, countries are allowed to have different intercepts, but these are modeled

separately from the effect of the common latent variable η_B. Complete metric invariance can be imposed by restricting the residual measurement errors on the between level to zero, but there is no need to do so.

Figure 21.3 Measurement model for individuals (within) and countries (between)

In addition to testing measurement equivalence, using multilevel modeling to analyze substantive models and the potential differences between countries is an exciting new approach. In MLM, differences between countries can be modeled by explicitly including country level explanatory variables in the analysis. In this chapter, we focus on assessing measurement equivalence, but we return to the more general issue of modeling country differences in the discussion.

21.2.3 Latent Class Modeling (LCM)

Latent class modeling (LCM) was first described by Lazersfeld (1950, Lazersfeld & Henry, 1968) who introduced the concept of latent structure analysis to describe the use of mathematical models for the association between latent variables (McCutcheon, 1987). Classical structural equation modeling (SEM) is a form of latent structure analysis, with linear relations between continuous latent factors estimated on the basis of multivariate normal observed indicators. Latent class modeling assumes a discrete latent variable that represents latent classes, interpreted as subtypes of related cases in the population. The classical application of latent class analysis is to identify subtypes or segments in the population on the basis of categorical observed variables. In this application, latent classes are defined by the condition of local independence, which stipulates that conditional on the class membership, the observed variables are independent. In other words, all of the association between the observed variables is explained by the categorical latent class variable.

A typical procedure in classical latent class analysis is to determine the minimum number of latent classes needed to achieve local independence, and to interpret the latent classes in terms of the response probabilities on the observed categorical variables. Although the classes are latent, and class membership is therefore unobserved, cases can be classified into their most likely latent class using recruitment probabilities. An introduction in the classic latent class model is given by McCutcheon (1987). An application in the survey field is the study by Biemer *et al.*, who use latent classes to represent different types of misclassification in the U.S. census (Biemer, Woltmann, Raglin & Hill, 2001).

The classic latent class model has been extended to allow continuous variables, and to allow class membership to be predicted by observed covariates. In addition, latent class models have been formulated that relax the condition of local independence. Instead, a specific model is assumed to hold that produces the associations between the responses, which can be either continuous or categorical variables. Usual it is assumed that the same model holds in the distinct latent classes, while different latent classes are characterized by having different values for the model parameters. However, latent class models may also have different models in different classes. The extended latent class model is also referred to as a mixture model, because it assumes that the data are generated by a mixture of different distributions, generated by the corresponding models. An introduction in these extended latent class models is given by Rost and Langeheine (1997) and Magidson and Vermunt (2004).

Among the models that can be posed as generating the response distributions in the different classes are structural equation models (Magidson & Vermunt, 2005). When applied to assessing measurement equivalence in comparative survey data, we assume that the observed variables are generated by a measurement model such as a confirmative factor model. Following this assumption, the latent class analysis follows the same reasoning as the reasoning in SEM multiple group modeling. The relevant null hypothesis for measurement invariance across latent classes is

$$\Lambda_{\mathbf{y}}^{(1)} = \Lambda_{\mathbf{y}}^{(2)}, \qquad (21.13)$$

with the important difference that the superscripts in this case refer not to observed groups but to latent classes.

Although the methodological reasoning is the same, the actual testing procedure in latent class SEM is different than in classical SEM, due to differences in the underlying models. Searching for the correct number of latent classes is less straightforward than model testing in SEM, because there is no formal statistical test to test the 2-class model against the 1-class. In practice, decisions about the 'correct'number of classes are based on information criteria like Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). In addition, LCM has no global test for the fit of the model, and there are no modification indices to suggest model improvements. In classical SEM we can pose functional equivalence, the weakest form of measurement invariance (the different countries share a measurement model that has the same factor structure, cf. Vandenburg & Lance, 2000), as a reasonable starting model. If that model is rejected, modification indices will inform us how to proceed. In Latent Class SEM, we can follow the same approach, but without the guidance of global tests and modification indices. Here we can specify the same measurement model for each latent class, and search for the best number of classes. If there are two or more classes, we conclude that we have more than one subpopulation, but since the factor structure in all classes is the same we have functional equivalence. If the factor loadings can be constrained to be equal across all latent classes, we have scale equivalence (Vandenburg & Lance, 2000). Just as in multigroup SEM, some amount of invariance is judged acceptable, which leads to partial measurement invariance. When (partial) scale equivalence is achieved, the model can be extended to allow for different structural models across groups. When comparisons of means of latent constructs are involved, we need additional invariance

constraints across all latent classes on the intercepts of the observed variables to establish metric equivalence (Vanderburg & Lance, 2000). Again, a small amount of variation is judged acceptable. The minimal requirements for partial invariance given by Byrne et al. (1989) and Steenkamp and Baumgartner (1998) (for each construct in addition to the marker item at least one more indicator with invariant loadings and intercepts) appear to be reasonable also in the context of latent class SEM modeling.

The next section presents and discusses a small simulation study, aimed at clarifying how well the three methods work with varying number of sampling units at the group level. The section following the simulation study reports the results of applying multigroup SEM, multilevel SEM and LCM on a realistic data set.

21.3 A Comparison of SEM, MLM, and LCM by Simulation

Since measurement equivalence is of central importance, it is the topic of the simulation study. The simulation study uses two different data generating models, both simulating a simple measurement model. The data conform to the general structure of comparative studies, with a large sample within each country and a relatively small number of countries. We require that the latent variable is over-identified, which leads to four observed indicators for a single construct. For the model, there are two simulated conditions. In one condition, metric equivalence holds. The goal of simulating this condition is to investigate if the number of available countries permits accurate parameter estimates and standard errors. In the second condition, metric equivalence does not hold. The goal of simulating this condition is to investigate if the chosen analysis method has sufficient power to detect the violation of the equivalence of measurement.

21.3.1 Sample Sizes in Structural Equation and Multilevel Modeling

Simulation research on single level SEM has shown that with a good model and multivariate normal data a reasonable sample size is about 200 cases (cf. Boomsma, 1982), although there are examples in the literature that use smaller samples. Simulation studies (e.g., Chou & Bentler, 1995; Boomsma, 1982) show that with non-normal data, maximum likelihood estimation still produces good estimates in most cases, but that larger sample sizes are needed, typically at least 400 cases. Most surveys have sample sizes considerably larger than this, and carrying out a SEM analysis is feasible.

In comparative surveys, the sample size at the group or country level is often limited. Only a few simulation studies have investigated the sample size requirements for multilevel SEM. These studies typically report that at all simulated sample sizes the individual level coefficients and standard errors are estimated accurately. However, for the group level, it has been shown that multilevel SEM often results in good parameter estimates and reasonable but not highly accurate standard errors. For instance, Hox and Maas (2001) find that a group level sample size of 100 is required for sufficient accuracy of the model test and confidence intervals for the parameters. With the group level sample size set to 50, the parameters are still estimated accurately, but the standard errors are too small. In a later simulation using a new and more sophisticated estimation method (full maximum likelihood instead of the approximate limited likelihood), Hox, Maas and Brinkhuis (2009)

find accurate standard errors, even for group level factor loadings with only 50 groups of size 10 each. The coverage of the residual variances is not as good (90%). In the context of multilevel regression, Hox and Maas (2005) find similar results for multilevel regression modeling. Typically, regression coefficients can be accurate with higher level sample sizes as low as 20, standard errors require somewhat higher sample sizes, but accurate estimation and testing of variances requires group level sample sizes of at least 50 groups.

The simulations reported above are based on the usual multilevel designs, with group sizes reflecting applications in educational and organizational research. Comparative surveys typically use very large sample sizes within each country, and the number of countries in large scale comparative surveys is reaching the 20-40 range. Given the large samples within the countries, this smaller number of countries (the group level sample size) is likely to be sufficient to make multilevel SEM a serious analysis option. To gain a better understanding of how well the different analysis approaches fare under sample size conditions typical for comparative surveys, we report here the results of a small simulation study that investigates these issues.

21.3.2 A Small Simulation Study

To represent the number of countries as found in large scale surveys, three different values have been chosen for the Number of Countries in the simulation (NC=20, NC=30 and NC=40). Within each country, 1500 respondents are simulated. The model assuming metric equivalence is presented below. It should be noted that means are fixed at 0 and that all simulations are performed a 1000 times in each condition.

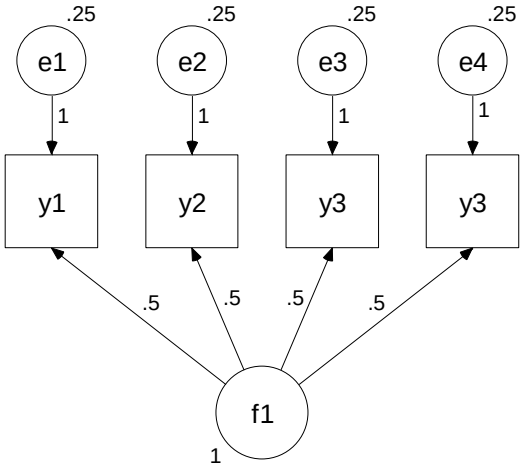

Figure 21.4 Path diagram for a factor model

For the model where metric equivalence is violated, data are simulated where the fourth factor loading is different from the others for half of the countries, namely 0.3 instead of 0.5. In terms of Cohen's effect size conventions, the loading in these countries changes from *medium* to *small*. Thus, half of the countries have different weights than the other half. The same simulated data are used for the multigroup model, the multilevel model, and the latent class model.

21.3.3 Simulation Results

When the three approaches are used to analyze the data where metric equivalence holds, the results are straightforward. Multigroup SEM performs very well in all simulated conditions. Latent class SEM with only one class ignores the country level, but performs just as well. Multilevel SEM performs well with 30 or 40 countries; with 20 countries the standard errors are too large, resulting in an operating coverage for the estimated 95% confidence interval that is actually between 92% and 93%. In all cases, the chi-square was significant in approximately five percent of the simulations, and the fit indices TLI, CFI and RMSEA all indicate a good fit.

When the three approaches are used to analyze the data that are simulated to violate metric equivalence by having one factor loading that varies across countries, the results are more comple21. One striking result is that only the chi-square test in the multigroup model is able to detect this model violation. The latent class model totally lacks power to detect the violation. The chi-square model test does not reject the model even once. The multilevel model fares somewhat better. The chi-square test does reject the model in the majority of cases, and the model improves significantly if the loading that violates the assumption of measurement invariance is allowed to vary across countries. The general fit measures CFI, TLI and RMSEA all perform very poorly. They look well in all simulated conditions, indicating that they lack power to detect the violation. Only the RMSEA in the multigroup model provides some indication of a non-perfect fit.

In sum, when the model is incorrect because metric equivalence is violated, the most striking result is a massive lack of power to detect this violation. Only in the classical multi-group analysis does the global chi-square test reject the model when the data that violate metric invariance are analyzed. But even in that case, the fit indices would indicate a very good fit, and most analysts would probably argue that given the large total sample size, the chi-square test is overly powerful and the model rejection therefore can be ignored. All the same, they would be ignoring a clear violation of measurement invariance. The conclusion is that a strong reliance on global fit indices is misleading. If most of the model is correctly specified, with only a few misspecifications in specific parts of the model, global fit tests and fit indices are overly optimistic. It is better to examine more specific indicators of lack of fit, such as modification indices and the corresponding estimated parameter change. In contrast to the chi-square test and associated fit indices, the modification indices are related to lack of fit for a specific parameter constraint. As such, when there is a specific fit problem in a model that fits well globally, the modification index has a much better power to indicate the source of this problem, and the estimated parameter change indicates how different the unconstrained parameter estimate is likely to be from the constrained estimate. This is the approach that is taken in the next section, where a realistic data set is analyzed.

21.4 A Comparison of SEM, MLM, and LCM on Existing Data

The substantive analyses presented here illustrate the three statistical approaches on a

realistic data set. We use data from the first round of the European Social Survey. The data collection took place in 22 countries between September 2002 and March 2003, the total number of subjects in these data is 41207. For details on the data collection we refer to the ESS website (www.europeansocialsurvey.org). We analyze a set of four items that measure "religious involvement". The items are (the C/E item codes refer to the question identification code in the ESS questionnaire):

- Regardless of whether you belong to a particular religion, how religious would you say you are? (C13)
- Apart from special occasions such as weddings and funerals, about how often do you attend religious services nowadays? (C14)
- Apart from when you are at religious services, how often, if at all, do you pray? (C15)
- How important is religion in your life? (E18)

Items C13 and E18 were measured on an 11-point scale ranging from 0=extremely unimportant to 10=extremely important. Items C14 and C15 were measured on a 7-point scale ranging from 1=every day to 7=never. The scores are reversed for C14 and C15, as a consequence, high scores are associated with high religiosity for all items.

These items have also been analyzed by Billiet and Welkenhuysen-Gybels (2004) who concluded that a single factor underlies them, and that partial measurement invariance exists. Our example data are somewhat different because we include France (not available for the 2004 analysis), and treat the data as continuous. On the country level, we have added the religious diversity index by Alesina et al. (2003). Since the data set is large and the amount of missing data on these four items is small, we use listwise deletion, which leaves 41207 subjects. All analyses were carried out using Mplus 5 (Muthén & Muthén, 2007).

21.4.1 Analysis Results, Multigroup SEM

The typical approach in investigating measurement invariance using multigroup SEM is to set up a series of models that specify a common factor model for all groups, starting with a model with no constraints (the functional equivalence model), and then adding constraints on factor loadings (the scale equivalence model) and intercepts (the metric equivalence model)) in two successive steps. Since these models are nested, both formal chi-square tests and inspection of fit indices can be used to decide whether measurement invariance holds. This approach has the advantage that if the functional equivalence model is rejected the analysis process may stop, because statistical analysis of the differences between countries is not valid. However, when the number of countries is large, it can take many analysis steps to find out which equality constraints are allowed. When the number of groups (countries) is large, we suggest using the opposite strategy: start by fitting a model with all constraints needed for metric invariance, and in addition for the purpose of comparison the functional equivalence model (no constraints). Again, if the functional equivalence model is rejected, the analysis should stop. If the functional equivalence model holds, but the metric invariance model does not fit, the modification indices for the metric invariance model and the differences between countries in their

parameter estimates in the functional equivalence model both provide information about the model modifications that are needed to achieve some form of equivalence.

The metric equivalence model is specified by constraining the latent variable mean to zero and its variance to one in the first country (Austria), and constraining all equivalent loadings and intercepts to be equal across all countries. A more common representation is to fix one intercept to zero and one loading to one for all countries, and allow the factor mean and variance of all countries to be estimated. The representation we have chosen here has the advantage that all intercepts and factor loadings are estimated, and therefore modification indices are available for all intercepts and factor loadings.

Since the sample sizes are very large, we do not use the chi-square statistic as the measure of model fit. Instead, we rely on the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI), with values > 0.90 indicating a good fit, and the Root Mean Square Error of Approximation (RMSEA), with values <.05 indicating a good fit and values <.08 indicating an acceptable approximation (c.f. 21.2.1). Given the simulation results discussed earlier, which show that a small violation of measurement invariance is easily masked when the remainder of the model fits well, we inspect all the modification indices that refer to the factor loadings and intercepts on a country-by-country basis, and follow up by comparing the constrained estimates to the unconstrained estimates in the functional equivalence model.

Table 21.1 Model selection in multigroup SEM

Model	Chi-square	df	CFI	TLI	RMSEA
1: Functional	1021.9	44	0.97	1.00	0.11
2: Metric	8783.5	170	0.90	0.92	0.16
3: Partial	1290.4	121	0.99	0.99	0.07

Table 21.1 presents the fit information for a sequence of models. Model 1, which states functional equivalence, shows a reasonable fit. The modification indices suggest correlated errors in some countries, mostly between C13 and E18 or between C14 and C15. Model 2, which states complete metric equivalence, has a much worse fit. The three countries with the largest contribution to the overall chi-square are Ireland (1603.8), Poland (1255.0) and Israel (991.7), perhaps not by accident countries where religion plays an important role.

Taking the metric equivalence model as a starting point, modification indices were inspected to discover the source of the lack of fit. This showed a striking result: the intercept of item C14 (attendance religious services) had a large modification index for almost all countries. When this intercept was set free the chi-square dropped by more than 3100. The CFI/TLI fit indices were acceptable for this solution (both 0.94) but the RMSEA was judged too large at 0.14. Following large values of the modification indices, intercepts and loadings were allowed to be freely estimated in specific countries, with the restriction that no country could have more than two intercepts or loadings free. This way, for all countries partial invariance is maintained. The result is the partial metric invariance model in row 3 of Table 21.1. This model fits well. The way this model is specified, the first country (Austria) has a mean constrained to zero and a variance

constrained to one. The means and variances in all other countries can be freely estimated. If the means are all constrained to be zero, the model is rejected again (chi-square 8876.3, *df*=142, CFI/TLI=0.90, RMSEA=0.18), and the difference between the two models is clearly significant (chi-square 77585.9, *df*=142, *p*=<.001). We conclude that the countries differ in the level of religious involvement. We present the country means and variances in section 21.4.4, where we compare the results from the three analysis approaches.

21.4.2 Analysis Results, Multilevel SEM

In multilevel SEM a measurement model is specified at both the individual level and the country level. The first model is a model with no varying slopes (implying scale invariance) and in addition equal loadings across both levels. The latter implies scale invariance across the individual and the country levels. This is not a necessary condition, but if it holds, it allows statistical comparisons of within and between factors. As explained earlier, metric invariance is not needed because the country-level variance of the intercepts is modeled by the country-level error variances (c.f. 21.2.2). The within model was identified by constraining the individual level factor mean at zero and the variance at one. All four loadings were estimated, and intercepts are zero by definition. Since the loadings are constrained to be equal across the two levels, the between model was identified by constraining the factor mean to zero, allowing all intercepts to be estimated. The variance of the between level factor can be estimated freely.

The model fits reasonably, but there is one very large modification index that suggests a covariance between the residuals of C13 (religiosity) and E18 (importance of religion). When this covariance is added, the model fits very well: chi-square=30.8 (df=6, p=0.00), CFI/TLI=1.00, RMSEA=0.01, and there are no large modification indices. In this model, the variance of the between factor is estimated at 0.25, which implies an intraclass correlation (ICC) of 0.20, meaning that 20% of the variance is at the country level. The ICCs for the observed variables are lower: they range from 0.12 to 0.17. Finding a larger ICC for the latent variables is typical, since measurement error in the observed variables ends up at the lowest level (Muthén, 1991).

Despite the good model fit, given the lack of power of the global tests in the simulations, four additional models were estimated that allowed each of the four loadings in turn to vary across countries. This shows a significant variance component for the loading of C14 (attendance of religious services). This variance component is small (0.094), but significant (S.E. is 0.03, *p*=.001). The loadings are a normally distributed random variable with a mean of 1.27 and a standard deviation of 0.31 ($=\sqrt{0.094}$) across countries. Since about 95% of a normal distribution is within two standard deviations from the mean, we can conclude that for 95% of all countries this loading is between 0.65 and 1.89. This is a sizeable difference, but we can also conclude that there are almost certainly no countries where this loading is negative.

The conclusion is that in the multilevel measurement model partial scale invariance appears to hold. For further modeling, we can include explanatory variables at either the individual or the country level to explain variation in religious involvement, or even to explain the variation in the loadings for C14 in the measurement model. Since the loadings can be constrained equal at both levels (for C14 this constraint refers to the

mean of the distribution across the countries), it is likely that the variation is caused by the same explanatory variables. Therefore, we may hypothesize that the difference between countries is not so much an effect of differences in country-level variables, but in differences in the composition of the countries' populace. This means that country-level variation is likely to be explained by aggregated individual variables. In principle, country variables can also be used to explain the variation in the slopes of C14, but this assumes a theory about the variations in measurement models across countries. Such analyses are not pursued here. To enable a comparison between the different statistical approaches, we estimate the factor scores for the between and the within factor. The mean factor scores across countries are presented in section 21.4.4, where we compare the results from the three analysis approaches.

21.4.3 Analysis Results, Latent Class SEM

In Latent Class SEM, a number of latent classes is postulated that are each characterized by their own structural model. The approach is analogous to multigroup SEM, with the latent classes replacing the groups. The difference is that the number of latent classes is assumed to be much smaller than the number of countries. In our case, we investigate both scale and metric invariance. Consequently, a confirmatory factor model was specified with equality constraints for the loadings and intercepts across all latent classes. Interestingly, when this model was specified for two classes, it did not converge. This could be solved by allowing the intercept of C14 (attendance of religious services) to be estimated freely in the two classes. The model for three classes did not converge, and a search for additional intercepts or loadings to be freed did not result in convergence. In the two-class model, C15 had a small negative residual variance in the first class; this was corrected by constraining it to zero. Table 21.2 presents the AIC and BIC measures of model fit for these models.

Table 21.2 Fit measures for different Latent Class Structural Equation Models

Model	AIC	BIC
1 class	675316.2	675428.4
2 class	632512.8	632702.7
2 class corrected	632564.7	632702.7
3 class	no convergence	no convergence

In the two-class model, the first class (about 44% of the respondents) constrains the latent religion variable mean to zero and the variance to one. The second class (about 56% of the respondents) estimates the mean as 8.9 and the variance as 7.7. The first class can be described as individuals for whom religious involvement is not part of their life. The second class can be described as individuals for whom religious involvement is important: they have a much higher mean but also display a larger variance in religious involvement.

Similar to the multilevel SEM approach, it is possible to add explanatory variables that explain variations in the latent variable religious involvement. In addition,

it is possible to add explanatory variables that explain class membership. This will not be pursued here. For comparison purposes, the factor scores and latent class probabilities are estimated, and discussed in the next section.

21.4.4 Analysis Results, Comparison of Approaches

The results from the three different approaches tend to converge on similar conclusions. All three approaches result in a verdict of partial metric invariance. All three approaches conclude that item C14 is problematic. There is also a covariance between the residuals of items C13 and E18 that is necessary in all approaches to achieve a good fit.

Table 21.3 Measurement models in different approaches (raw loadings)

Item	Multigroup	Multilevel	Latent Class
C13 (religious)	0.99	2.17	0.28
C14 (services)	2.35	1.04	0.45
C15 (pray)	1.79	1.83	0.45
E18 (importance)	2.80	2.52	0.54

Table 21.3 shows the factor loadings for the various measurement models. It is clear that the different approaches do not result in identical measurement models. The multigroup and the latent class model both analyze only within groups variation, and the variation between groups is modeled as differences in latent variable means and variances between observed or latent groups. In contrast, the multilevel approach estimates a common factor model for variation within and between countries. The multigroup and latent class approach both agree that C13 obtains the lowest loading and E18 the highest. As a result, the interpretation of the latent factor is subtly different in each approach. In each case, the interpretation of the latent factor is ‘religiosity’, but there are some subtle differences in the meaning of this construct. To investigate the correspondence between the different approaches we estimated factor scores for all latent variables in the final model in each of the three analysis approaches. The correlations between the multigroup factor score, the within factor score, and the latent class factor scores vary from 0.87 to 0.95. Thus, although the factor scores are not exactly the same, they are all very similar. When the factor scores are aggregated to the country level, the correlations between the multigroup factor score, the between factor score, and the latent class factor score vary from 0.96 to 0.99. Clearly, when the objective is to compare and analyze countries, all three approaches are effectively equivalent. Figure 21.5, which plots for all 22 countries their (Z-score transformed) multigroup means, country level factor scores from the multilevel analysis, and aggregated latent class factor scores, highlights the similarities at the country level.

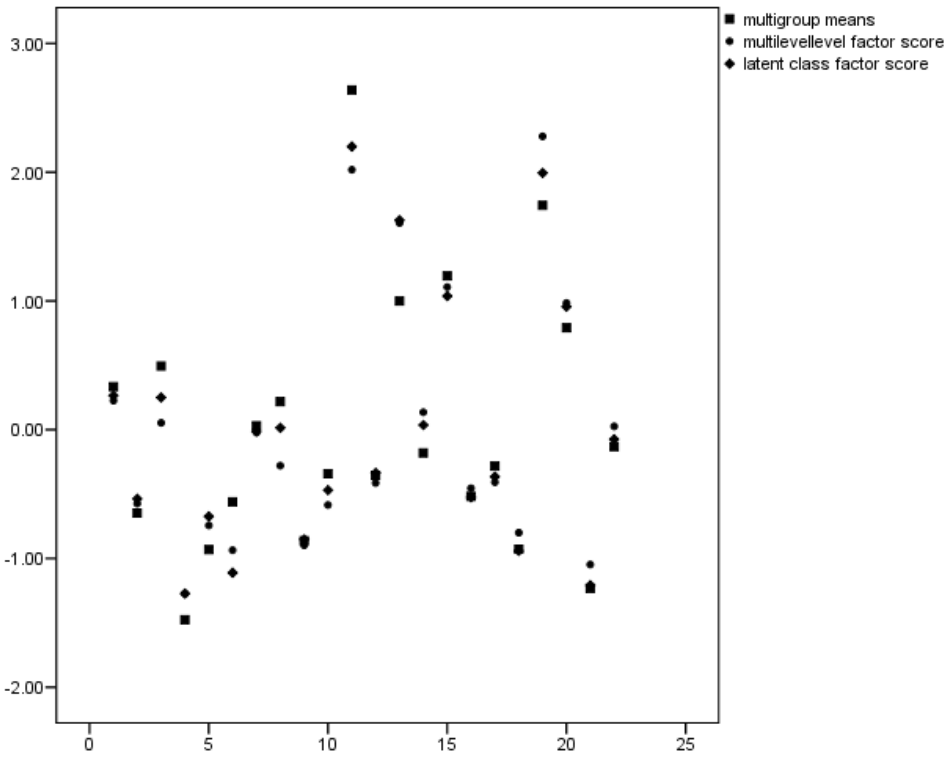

Figure 21.5 Standardized factor scores at the country level, three analysis approaches

The latent class provides additional information. The two class solution in the latent class approach suggests that the observed differences between the 22 countries can be the result of a different composition of their population in terms of these two classes. To investigate this possibility, a crosstabulation was made for estimated class membership against country. The correlation between these variables expressed as the phi coefficient is 0.39. The class membership per country is given in Table 21.4, in ascending order of membership in the first class.

Table 21.4 Estimated class membership (%) per country

Country	Class 1	Class 2
Greece	8.4%	91.6%
Poland	10.1%	89.9%
Ireland	15.4%	84.6%
Portugal	28.0%	72.0%
Italy	28.4%	71.6%
Switzerland	40.0%	60.0%
Austria	40.1%	59.9%
Israel	44.2%	55.8%
Spain	45.6%	54.4%

Slovenia	45.7%	54.3%
Total	46.8%	53.2%
Finland	50.2%	49.8%
Netherlands	55.1%	44.9%
Hungary	55.3%	44.7%
United Kingdom	55.9%	44.1%
Belgium	56.7%	43.3%
Luxembourg	57.1%	42.9%
Germany	60.5%	39.5%
France	62.5%	37.5%
Norway	69.0%	31.0%
Denmark	70.5%	29.5%
Czech Republic	73.2%	26.8%
Sweden	73.5%	26.5%

Estimated class membership indeed shows a large variation across the countries. To investigate to which degree differences between countries can be attributed to differences in the composition of the populace, mean factor scores for the factors in the multigroup solution are calculated, both for the total population and separately for the two classes. The results are presented in Table 21.5. The last column presents the factor means in the total population adjusted for a class membership indicator.

Table 21.5 Differences between countries

Country	Percent Class 1	Total Mean	Class 1 Mean	Class 2 Mean	Adjusted Total Mean
Greece	8.4%	1.17	-.47	1.32	.53
Poland	10.1%	.73	-1.05	.93	.12
Ireland	15.4%	.40	-1.03	.66	-.13
Portugal	28.0%	.33	-.72	.73	.00
Italy	28.4%	.52	-.54	.94	.20
Switzerland	40.0%	.18	-.65	.73	.05
Austria	40.1%	.13	-.74	.71	.01
Israel	44.2%	-.06	-.94	.63	-.10
Spain	45.6%	.03	-.78	.70	.02
Slovenia	45.7%	-.04	-.85	.64	-.07
Finland	50.2%	.10	-.57	.78	.13
Netherlands	55.1%	-.10	-.80	.77	.05
Hungary	55.3%	-.14	-.88	.78	.00
United Kingdom	55.9%	-.12	-.82	.76	.04
Belgium	56.7%	-.18	-.84	.68	-.02

Luxembourg	57.1%	-.19	-.90	.76	-.01
Germany	60.5%	-.34	-1.01	.68	-.10
France	62.5%	-.33	-.91	.65	-.06
Norway	69.0%	-.34	-.80	.68	.01
Denmark	70.5%	-.20	-.57	.68	.18
Czech Republic	73.2%	-.62	-1.13	.78	-.15
Sweden	73.5%	-.47	-.87	.65	-.01
R-squared		0.17	0.07	0.09	

Class membership accounts for approximately half of the variation in religious involvement between countries. The other half consists of variation in religious involvement within classes between countries. The effect of adjusting for differences in class membership are striking when we compare the total means with the covariate adjusted means. The total means vary from -0.62 to 1.17, while the adjusted means vary from -0.15 to 0.53.

To further analyze the differences between countries, we can use the country level religious diversity index (Alesina et al., 2003). In the multiple group approach, there is no formal way to include this variable in the model. We can correlate the countries' estimated means on the religiosity factor with the religious diversity index, which produces a correlation of -0.38 (p=.08). The same analysis can be done for the aggregated factor scores in the two-class latent class solution, this produces a correlation between the aggregated factor score and religious diversity of -0.40 (p=.07). In the multilevel model, we can directly predict the between countries religiosity by religious diversity, which produces a standardized regression coefficient of -0.18 (p=.78). In all approaches we conclude that there is a negative but nonsignificant relationship between religious diversity and religiosity. In the latent class model, we can predict class membership with religious diversity, this produces a raw regression coefficient of -1.29 (p<.001), indicating that high religious diversity is related to less membership in class 2. Although the analyses methods differ, the conclusions converge: high religious diversity in a country is (weakly) related to lower feelings of religiosity in general.

21.5 Discussion

21.5.1: SEM, MLM, and LCM

In the context of comparative research, we have a multilevel data structure consisting of respondents within countries. If we investigate the measurement model, we need to model respondent-level variation. Both the simulations and the example indicate that the classical analysis approach using multigroup SEM is still attractive. It is feasible with more than 20 countries, and the most powerful approach when violations of measurement equivalence are to be detected. The approach taken in the example data, which contrasts the functional equivalence model (same form, no constraints) with the metric equivalence model (all constraints on loadings and intercepts) has proven to be effective. Latent class

SEM is much like multigroup SEM: it is a fixed effects model, but is expected to include far fewer classes than we have countries, which makes it more parsimonious than multigroup SEM.

If we investigate differences between countries, we need to model country-level variation. Multilevel modeling (MLM) treats the observed countries as a sample from a larger population and can include country level variables in the analysis. It is more parsimonious than multigroup SEM, but assumes a reasonable sample size at the country level. Our simulations suggest 30 countries as a reasonable sample, but even at 20 countries the estimates and standard errors appear accurate enough.

The simulation results make clear that reliance on global model fit tests or indices can be dangerous. Such tests and indices take the entire model into account, and do not have sufficient power regarding very specific model violations, such as one factor loading that differs across countries. The simulations indicate that the multigroup SEM approach, although laborious, provides the most detailed information on the amount and the sources of lack of measurement invariance.

A comparison of the three approaches on the example data shows that some features of the data are identified in each of the three approaches. All approaches identify item C14 as problematic, and the residual correlation between items C13 and E18 is also present in each final model. Thus, the results of the different analyses do converge. There are also major differences between the final models, which are based on the elementary fact that we are considering models that are qualitatively very different. When issues of measurement invariance are involved, the multigroup SEM is clearly the best approach, because it provides very specific information on sources of misfit, and allows also very specific modifications of the strict invariance model. When the number of countries is large, latent class SEM offers a way to make the model more parsimonious. However, when we want to use country level variables to predict individual or country level outcomes, multilevel modeling is better suited, because the country level variables can be included in the model. In multigroup SEM, and latent class SEM, we need to use a two-step approach where country level variables and aggregated individual level variables are combined and analyzed in a separate analysis step.

It appears that the different analysis approaches should be seen as complementing each other, rather than competing. When measurement invariance has been established using multigroup SEM, subsequent analyses including country level variables can best be done using a multilevel model. Since the models that can be specified in the framework of multilevel regression are limited, the multilevel approach will likely involve multilevel SEM. Given the generally limited sample size at the country level, such models must at the country level be kept parsimonious. Latent Class SEM is an interesting addition to the statistical toolkit, but not a replacement for either multigroup or multilevel SEM.

21.5.2 Software Issues

The advent of powerful and user-friendly software for multilevel modeling has had a large impact in research fields as diverse as education, organizational research, demography, epidemiology, and medicine. Multilevel modeling has had less impact on the analysis of international and cross-cultural survey data. This is not caused by a lack of

interest in statistical techniques on the part of cross-cultural researchers; comparative researchers in intercultural and organizational research were among the first to use multigroup SEM to assess measurement comparability. The almost exclusive focus on multigroup SEM reflects specific challenges in comparative surveys. Until recently, few comparative surveys included a large enough number of countries to make multilevel modeling an attractive option. The sporadic use of latent class modeling in comparative survey research reflects mostly lack of powerful software. To be useful in the context of comparative surveys, latent class modeling must be able to include multiple and distinct observed groups such as countries in the model. That is, the latent class models and accompanying software must include either multigroup or multilevel capacities. Until recently, such software was simply not available.

In addition, most surveys result in complex data sets with weights and stratification that require specific analysis methods (Rabe-Hesketh & Skrondal, 2006). These analysis methods must be combined with the complex modeling techniques mentioned above. Few programs exist that can deal with these complexities. It should be noted that this software limitation also applies to multigroup SEM. The majority of SEM applications on survey data ignore the complex sampling design. It has been argued that when the focus is on estimating a multivariate model instead of estimating basic statistics for the population, ignoring the complex design is acceptable, since estimates of regression coefficients will still be unbiased. For a discussion of model-based versus design-based analysis we refer to Groves (1989). For the present, we point out that model-based analysis ignoring the complex structure of the data is only valid if the model is in fact correct and the assumptions are met. Since the kind of modeling discussed in this chapter is intricate and in general needs a fair amount of model specification searching to identify a good model, this strong reliance on the correctness of the model is undesirable. Therefore, we advise to take complex data structures into account whenever that is possible.

The most flexible software capable of multigroup SEM, multilevel analysis, and latent class analysis at the moment are M*plus* (Muthén & Muthén, 2007) and `GLLAMM` (Rabe-Hesketh, Skrondal & Pickles, 2004). These programs can both estimate models with varying combinations of multiple groups, multilevel, and (latent class) structural equation models. They can also deal with incomplete data and non-normal variables. A practical limitation for both programs is that when complex nonlinear models are estimated for nonnormal or incomplete data, the estimation methods that must be used are computationally very demanding, to the point where estimation is actually impossible.

Other, more limited software packages are Latent GOLD, which can analyze multilevel latent class models and has limited capacities for structural equation modeling, and the SEM packages LISREL and EQS, which have limited capacities for multilevel modeling and cannot estimate latent class models. The multilevel regression package HLM has limited capacity for multilevel structural equation models, which can be estimated provided they are recursive.

21.7 References

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. & Wacziarg, R. (2003).

Fractionalization. *Journal of Economic Growth, 8*, 155-194. Data accessible at http://www.stanford.edu/~wacziarg/papersum.html (accessed March 2008).
Bechger, T.M., van den Wittenboer, G., Hox, J.J. & de Glopper, C. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practice,* 18, 18-26.
Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.
Biemer, P.P., Woltmann, H., Raglin, D., & Hill, J. (2001). Enumeration accuracy in a population census: an evaluation using latent class analysis. *Journal of Official Statistics, 17*, 129-148.
Billiet, J., & Welkenhuysen-Gybels, J. (2004). Assessing cross-national construct equivalence in the ESS: the case of religious involvement. Paper presented at the Sixth International Conference on Social Science Methodology, August 17-20, 2004 Amsterdam. Accessed February 2008 on http://www.s3ri.soton.ac.uk/qmss/documents/BillietMainzpapermeasurementESSdefinite.pdf.
Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H.Wold (Eds.), *Systems Under Indirect Observation: Causality, Structure, Prediction* (Part I, pp. 149–173). Amsterdam: North-Holland.
Braun, M., & Mohler, P.Ph. (2003). Background variables. Pp. 101-115 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: 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* (pp.445-455). Newbury Park: Sage.
Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175.
Byrne, B.M., Shavelson, R.J. & Muthén, B.O. (1989). Testing for the equivalence of factor and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.
Chou, C.-P. & Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In: Hoyle, R.H. (ed.). *Structural equation modeling: concepts, issues and applications.* Thousand Oaks, CA: Sage.
Couper, M.P. & de Leeuw, E.D. (2003). Non-response in cross-cultural and cross-national surveys. Pp. 157-178 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.) *Cross-cultural survey methods.* Hoboken, NJ: Wiley.
Everitt, B.S. and Hand D.J. (1981). *Finite mixture distributions*. New York: Chapman & Hall.
Gerbing, D.W. & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. Sociological Methods & Research, 21. 132-161.
Goldstein, H. (2003).*Multilevel Statistical Models*. London: Arnold.
Groves, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.
Harkness, J, Mohler, P.,& Van de Vijver, F.J.R. (2003). Comparative research. Pp. 3-16 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.), *Cross-cultural survey methods.* Hoboken, NJ: Wiley.
Harkness, J, Van de Vijver, F.J.R., & Johnson, T.P. (2003). Questionnaire design in

comparative research. Pp. 19-34 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley.
Harkness, J. (2009). Comparative survey research: Goal and challenges. Pp. 56-77 in E.D. de Leeuw, J.J.Hox, & D.A. Dillman (Eds). *International handbook of survey methodology*. New York: Lawrence Erlbaum/psychology Press, Taylor & Francis group
Heer, W. de (1999). International response trends: Results of an international survey. *Journal of Official Statistics, JOS, 2*, 129-142.
Hox, J.J. *Multilevel Analysis; Techniques and Applications. 2nd Edition*. New York: The Psychology Press..
Hox (1998). Multilevel modeling: when and why. Pp. 147-154 in I. Balderjahn, R. Mathar & M. Schader (Eds.). *Classification, data analysis, and data highways*. New York: Springer.
Hox, J.J. & Maas, C.J.M. (2001). The Accuracy of Multilevel Structural Equation Modeling With Pseudobalanced Groups and Small Samples. *Structural Equation Modeling*, 8, 2, 157-174
Hox, J.J., Maas, C.J.M., & Brinkhuis, M. (2009). The effect of estimation method and sample size in multilevel SEM. Paper, 6th International Multilevel Conference, Amsterdam, 2007. *Submitted.*
Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.
Lazersfeld, P.F. (1950). The logic and foundation of latent structure analysis. Pp. 362-412 in S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazersfeld, S.A. Star & J.A. Clausen (eds.) (1950). *Studies in social psychology in world war II, Vol. IV*. Princeton, NJ: Princeton University Press.
Lazarsfeld P.F. & Henry N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
Leeuw, E. de, & Heer, W.de (2002). Trends in Household survey nonresponse: An international and longitudinal comparison. Pp.41-54 in R.M. Groves, D.A. Dillman, J.L. Eltinge, & R.J.A. Little (Eds). *Survey nonresponse.* New York: Wiley.
Lynn, P., Japec, L., & Lyberg, L. (2006). What's so special about cross-national surveys? Pp7-20 in J. Harkness (ed). Conducting cross-national and cross-cultural surveys ZUMA-Nachrichten Spezial, 12. Mannheim: ZUMA.
Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology. European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85-91.
Magidson, J. & Vermunt, J.K. (2004). Latent Class Models. Pp. 175-198 in D.Kaplan (Ed.). *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Thousand Oaks, CA: Sage.
Magidson, J. & Vermunt, J.K. (2005). Structural equation models: Mixture models. Pp. 1922-1927 in B. Everitt and D. Howell, (Eds.). *Encyclopedia of Statistics in Behavioral Science*. Chichester, UK: Wiley.
McCutcheon (1987). *Latent class analysis*. Thousand Oaks, CA: Sage.
Meredith, W. (1964). Notes On Factorial Invariance. *Psychometrika 29*, 177-185.
Meredith, W. (1993).Measurement invariance, factor-analysis and factorial invariance. *Psychometrika 58*, 525-543.

Munck, I.M.E. (1991). Path analysis of cross-national data taking measurement errors into account. Pp. 599-616 in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, & S. Sudman (eds.) *Measurement Errors in Surveys*. New York: Wiley.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, L.K. and Muthén, B.O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén

Rabe-Hesketh, S. & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society: Series A, 169*, 805-827

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). *GLLAMM Manual*. (www.gllamm.org).

Raudenbush, S.W & Bryk, A.W. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Raudenbush, S.W., Rowan, B., & Kang, S.J. (1991) A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 4, 295–330.

Rost J. & Langeheine, R. (1997). A guide through latent structure models for categorical data. Pp. 13-37 in J. Rost & R. Langeheine (Eds.) *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann.

Smith, T.W. (2003). Developing comparable questions in cross-national research. Pp. 69-91 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Skjak, K.K. & Harkness, J. (2003). Data collection methods. Pp. 179-193 in J.A. Harkness, F.J.R. Van de Vijver, & P.P. Mohler. (Eds.) *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 1, 78-90.

Vandenburg, R., & 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, 4-70.

Vijver, F. van de, & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks: Sage.

Vijver, F.J.R. van de, & Poortinga, Y.H. (2002). Structural equivalance in multilevel research. *Journal of Cross-Cultural Psychology*, *33*(2), 141-156

Vermunt JK (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-239.