An Introduction to Multilevel Modeling with SEM (Revised 2-2016)[1]

Ronald H. Heck
University of Hawai‘i at Mānoa

Over the past three decades, concerns in various fields with conceptual and methodological issues in conducting research with hierarchical (or clustered) data have led to the development of multilevel modeling techniques. For example, employees work in particular departments nested within organizations and countries. These individuals interact with their social contexts in a variety of ways. Individuals within successive units may share some common characteristics such as socialization patterns, traditions and values, and beliefs about work. This makes them more similar to each other than to individuals situated in other organizations and countries. Clustered data may also result from the specific research design. In survey research, for example, individuals are often selected to participate in a study from some type of stratified random sampling design (e.g., participants may be chosen from certain neighborhoods in particular cities and geographical areas). Longitudinal designs present another situation where a series of measurements is nested within the individuals who participate in the study and who may also be nested in higher-level units (e.g., students in classrooms and schools).

In the past, researchers had considerable difficulty analyzing data where individuals were nested within a series of hierarchical groupings due to various limitations of existing statistical software. This often led to a forced choice between either individuals or groups as the unit of analysis in order to conduct a single-level analysis. Ignoring the nested nature of data structures, however, can lead to false inferences about the relations among variables in a model as well as missed insights about the social processes being studied.

A variety of names are used to refer to methods for analyzing hierarchical data structures: multilevel regression models, hierarchical linear models, mixed-effects and random-effects models, random coefficients models, multilevel covariance structure models, and multilevel structural equation models. The statistical theory for multilevel models has developed out of several streams of methodological work in different fields of inquiry including biometric applications of mixed-effects analysis of variance (ANOVA), random coefficients regression models in econometrics, and developments in the statistical theory of covariance component models and Bayesian estimation of linear models (Bock, 1989; de Leeuw & Kreft, 1986; Efron & Morris, 1975; Fisher, 1918, 1925; Goldstein, 1987; Hartley & Rao, 1967; Laird & Ware, 1982; Lindley & Smith, 1972; Mehta & Neale, 2005; Morris, 1995; Muthén, 1989; Muthén & Satorra, 1989; Raudenbush, 1988; Raudenbush & Bryk, 1986; Rubin, 1950; Shigemasu, 1976; Smith, 1973; Wald, 1947; Wong & Mason, 1985). As multilevel analysis has become common, specialty software programs have been developed to address the research needs of an ever-widening group of users. Mainstream statistical packages (e.g., SAS, SPSS, Stata) have also developed analytic routines that can be used to examine hierarchical data structures.

The intent of this chapter is to provide an introduction to multilevel modeling using structural equation modeling (SEM). Several conceptual and methodological issues in multilevel modeling are discussed, followed by the mathematical models underlying multilevel SEM, and finally, examples of multilevel confirmatory factor analysis, multilevel path

[1]Originally in Marcoulides, G. & Schumacker, R. (2001). *New Developments and Techniques in Structural Equation Modeling* (pp. 89-128). Mahwah, NJ: Lawrence Erlbaum.

analysis, and multilevel structural models with latent variables are presented using the Mplus software program (Muthén & Muthén, 1998-2012). The examples are intended to show in simple terms how to set up and conduct analyses step by step. For this reason, substantive issues are kept at a minimum and the focus is placed on the methodological and practical issues in multilevel modeling with SEM.

Overview of Multilevel Modeling

Multilevel modeling is one of several approaches that can be used in analyzing clustered data. In studying organizations, for example, multilevel modeling is an attractive analytic approach because it allows the incorporation of substantive theory about individual and organizational processes into the clustered sampling designs of survey research. Despite the existence of hierarchical data structures in the social and behavioral sciences, past empirical studies generally did not address clustered data adequately, due primarily to software limitations, although substantive concerns such as choosing the proper unit of analysis and the precision of parameter estimates resulting from single-level analytic techniques such as multiple regression were periodically raised (e.g., Burstein, 1980; Cronbach & Webb, 1975; Goldstein, 1987; Lindley & Smith, 1972; Strenio, 1981; Walsh, 1947). Applying the single-level linear model to hierarchical data produced several analytic difficulties including the forced choice over the proper unit of analysis, trade-offs in measurement precision, and limitations of the analytic methods used to estimate the model's parameters properly (Raudenbush, 1995). In a way this represented a known "blind spot" in how researchers approached the analysis of hierarchical data.

For many years, therefore, empirical work lagged behind the substantive theory of multilevel modeling because of the known limitations of single-level analyses; that is, either individuals or groups were the unit of analysis. This precluded more thorough investigations concerning relationships between individual and group processes—one simple example in organizational settings being how various group structures might enhance or diminish individual job satisfaction and commitment to the organization. In hierarchical data sets, the lowest level of measurement is referred to as the *micro level*, with all higher level measurements referred to as the *macro level*. Before multilevel techniques were available, researchers had to either disaggregate or aggregate variables in a hierarchical structure to construct a data set that could be analyzed at either the micro level (individuals) or a specific macro level (departments, organizations) with available single-level techniques. Researchers did not always consider the implications of the assumptions they made about measuring variables at their natural level in a data hierarchy and subsequently moving them from one level to another in order to maintain a single-level analysis.

In the disaggregation approach to studying organizations, for example, the researcher would move variables conceptualized at the macro level to the micro level. Organizational-level variables like productivity and organizational size might be combined with information about departmental leadership and individual employees' workplace attitudes and motivation. The unit of analysis would then be individuals, and the analysis would be conducted using the number of individuals in the study as opposed to the number of departments or organizations.

Treating individuals as if they were independent of their various macro-level groupings, however, ignores the complexity inherent in the data and can introduce potentially

important biases into the analysis. Single-level analyses require the researcher to assume all observations are independent; that is, that individuals within similar subunits and organizations do not share any common characteristics (e.g., similar socialization processes or perceptions). As similarities among individuals within groups become more pronounced (i.e., more homogeneous), however, the model's regression coefficients, standard errors, and associated tests of parameter significance become more biased (Bryk & Raudenbush, 1992; Muthén & Satorra, 1995). For example, since hypothesis tests are often based on the ratio of a regression slope to its standard error (β/SE), the downward bias of standard errors in single-level analyses results in smaller estimated SEs and, hence, more findings of significant parameters in the model.

Moreover, efficient estimation based on ordinary least squares (OLS) regression requires that the random errors (or residuals) in the equation are independent, normally distributed, and have constant variance (Bryk & Raudenbush, 1992). This assumption regarding the model residuals will likely be violated in hierarchical data sets. More specifically, the random error components of multilevel data structures are more complex because the errors within each unit are dependent, since they are common to every individual within the unit. It is important to emphasize, therefore, that conducting an individual-level analysis implies that no systematic influence of macro-level variables is expected and, therefore, *all* macro influence is incorporated into the model's error term (Kreft & de Leeuw, 1998).

In contrast, if a researcher chose the aggregation approach, she or he would combine data from individuals and subunits within each organization to create an organizational-level set of measures and then investigate variation in the aggregated measures. Because organizations are the unit of analysis in this case, the individual and department data would be used to develop mean scores on the variables for each organization. Unfortunately, however, the aggregation approach also presents problems for single-level analyses. One is that differences at the aggregate level typically appear stronger than they would be if within-unit variation were also incorporated into the analysis because all the variability present within each unit (or subunit) is reduced to a single mean (Draper, 1995; Kaplan & Elliott, 1997). Ignoring individual variability and then making statements about individuals through conducting a group-level analysis is referred to as committing an *ecological fallacy* (Robinson, 1950).

The unit of analysis problem suggests that single-level OLS regression estimates are not robust to misspecification of the number of levels in the data structure (Raudenbush, 1995). Prior to the development of multilevel analytic techniques and their increased accessibility to researchers through emergent computer software, few satisfactory solutions to the analysis of clustered data emerged, although several approaches were laid out (e.g., Aitken & Longford, 1986; Cronbach, 1976; Cronbach & Webb, 1975; Dempster, Laird, & Rubin, 1977; Goldstein, 1987; Lindley & Smith, 1972; Muthén, 1989, 1991; Schmidt, 1969; Wong & Mason, 1985). Although analysis of variance (ANOVA) methods offered partial answers to some of the questions posed with nested data since it could address individual and group variation in one model (Draper, 1995), the general formulation of the multilevel linear model was not presented until the early 1970s (e.g., see Lindley & Smith, 1972).

Conceptual and methodological concerns in applying single-level analytic techniques to nested data structures has therefore led to the development of multilevel modeling techniques

over the past several decades, with advances in both theory and method currently continuing at an impressive rate. Most importantly, multilevel modeling provided an analytic framework that facilitates the investigation of theories about relationships among variables at each level of the data hierarchy, as well as how relationships at a higher level might moderate relationships at a lower level (i.e., referred to as a cross-level interaction). An example of this latter type of relationship might be where the size or hierarchical structure of the organization is proposed to influence the quality of interpersonal relationships at the departmental level. Multilevel analysis provides a means of partitioning an outcome's variance into different components (e.g., within and between units) and, within the analysis, facilitates assigning explanatory variables to their appropriate hierarchical levels (e.g., individual, department, organization). In this manner, the aggregation or disaggregation problem can be avoided by considering two or three hierarchical levels simultaneously in the analysis.

Currently, there are two basic classes of multilevel procedures typically used in investigating hierarchical data structures. One is multilevel modeling (MLM) using multiple regression-type techniques, where typically a single *Y* outcome is defined and the direct effects of several predictors specified at different hierarchical levels on the random *Y* intercept are examined. The regression slopes of lower level predictors can also be included as random effects in such models (Raudenbush & Bryk, 2002). The second class of procedures, utilizing SEM techniques, is characterized by models with latent variables which are defined by sets of observed indicators, and both direct and indirect effects can be specified between the latent variables at different levels of a data hierarchy. As it turns out, multilevel univariate regression models for both cross-sectional and longitudinal data can be conceptualized within the more general SEM multilevel framework, which provides a considerable advance in modeling capability (e.g., Curran, 2003; Mehta & Neale, 2005). In practice, however, there are some basic differences in the types of models that multilevel regression and multilevel SEM are optimally suited to examine, as well as some key differences in the capabilities of existing software to examine certain types of data structures (e.g., cross-classified data structures, data structures with a considerable number of hierarchical levels). Complex data structures may challenge the available syntax needed to specify the models, as well as the available estimation approaches and computer memory needed to estimate a proposed model's parameters.

Multilevel Regression Models

As the previous discussion implies, the choice of analytic paradigm requires the investigator to consider the research questions, theoretical model, and the structure of the data before considering the strengths and limitations of various multilevel techniques and software programs. Determining the extent to which clustering may be present in the data is generally the first step in deciding whether multilevel modeling will offer an improvement in the precision of estimates over single-level techniques (Longford, 1993). Often, single-level analyses will suffice quite well, depending on the structure and characteristics of specific data sets. Where variability due to clustering is present across levels, however, multilevel analyses yield better calibrated estimates of population parameters (intercepts, slopes, standard errors) than analyses conducted at a single level without adjustments for clustering effects.

The intraclass correlation describes the degree of correspondence within clusters or

groups regarding an outcome *Y* and may be expressed for a two-level model as

$$\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2) \quad (1)$$

where σ_B^2 represents the variability between groups and σ_W^2 represents the variability within groups. Therefore, ρ indicates the proportion of the total variability in *Y* that can be attributed to variability between the groups. The intraclass correlation should be zero when the data are independent--therefore, its magnitude depends on characteristics of the outcome measured and the attributes of the groups. The larger the intraclass correlation, the larger is the distortion in parameter estimation that results from ignoring this similarity. In studies of school outcomes, for example, intraclass correlations at the school level are often in the range of 0.10 to 0.25 (Hill & Rowe, 1996; Reynolds & Packer, 1992), suggesting considerable similarities due to clustering. Ignoring the similarity due to clustering will inflate hypothesis tests considerably as the intraclass correlation increases (Muthén & Satorra, 1995). In the absence of between-group variability (i.e., where the intraclass correlation is less generally less than 0.03-0.04), however, there is little need to perform a multilevel analysis.[i] In such cases where the observations are nearly independent, a single-level analysis would provide reasonable estimates of the parameters and standard errors.

It is instructive to begin with specifying a random coefficients, or multilevel regression, model. As readers may recall, in a single-level multiple regression analysis, the coefficients describing the model, such as the intercept and slopes, are considered as fixed values estimated from the sample data. For example, the regression coefficient describing the impact of employee job satisfaction on productivity would be fixed at some weight for the model. In contrast, the multilevel regression model can be viewed as a hierarchical system of regression equations (Hox, 2010; Raudenbush & Bryk, 2002). In this approach, the intercept for each group, as well as level-1 slopes of theoretical interest, can be treated as randomly varying rather than fixed. This implies that an outcome such as productivity can be specified as having different average levels across the sample of organizations. Similarly, the impact of job satisfaction on productivity within each organization in the sample can also be allowed to vary across the organizations and, therefore, can be treated as an outcome in the model. This specification facilitates the examination of the variability in intercepts and level-1 slopes that may be explained by individual- and group-level variables.

We can specify the level-1 model for individual *i* in group *j* as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} jobsat_{ij} + \varepsilon_{ij} \quad (2)$$

where β_{0j} is the intercept coefficient (representing the mean level of productivity adjusted for job satisfaction), β_{1j} is the slope coefficient describing the effect of a unit change in job satisfaction on productivity, and ε_{ij} represents error in predicting Y_{ij} from known values of X_{ij}. Between groups, the random intercept and slope parameters can be specified as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (3)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (4)$$

The γ_{00} and γ_{10} coefficients represent the average intercept and job satisfaction-productivity slope, respectively, at the group level, and u_{0j} and u_{1j} represent the random effects describing the group-level variability in the intercept and slope,

respectively. In this two-level formulation, the level of organizational productivity produced (i.e., the intercept) would be expected to be different for each organization in the study. Similarly, the relationship between job satisfaction and productivity might be larger in some organizations than in others.

When we combine the models using algebraic substitution, we can see the more complex error term associated with each level-1 unit:

$$y_{ij} = \gamma_{00} + \gamma_{10} jobsat_{ij} + u_{0j} + u_{1j} jobsat_{ij} + \varepsilon_{ij} \quad (5)$$

Given, the presence of variation in random intercepts and slopes across groups, the analyst might be interested in the average job satisfaction slope between organizations (i.e., the fixed effect), as well as how particular organizations deviate from the overall average slope (i.e., the random effect in Eq. 5). Organizational variables such as size and resource allocation might be included in the model to explain differences in productivity levels and differences in the effects of job satisfaction on productivity. These types of analysis can easily be conducted using Mplus and specifying a two-level (or three-level) model with level-1 random slope. For a two-level model, this type of analysis is referred to in Mplus as TYPE = TWOLEVEL RANDOM. An example Mplus input file for the job satisfaction example mentioned in this section is included in Appendix A.

Where there are not many groups in the sample, or where the within-group sample sizes are small, it can be more challenging to develop efficient parameter estimates at the group level. Although in most instances, the different sample sizes do not affect the precision of the estimated *Y* intercepts for each unit too much (unless they are quite extreme due to small numbers of individuals in the unit), the individual-level slopes (e.g., job satisfaction- productivity) are likely to be less reliably estimated, since their precision depends on both the sample size of the unit as well as the distribution of *Y* and *X* within the unit. This can result in less efficient prediction equations regarding the effect of job satisfaction on productivity (Raudenbush & Bryk, 2002).

The MLM framework can be extended to examine univariate or multivariate models with latent variables by defining a measurement model at level 1 that corrects observed indicators of the underlying constructs for measurement error. Individual (level 2) and group-level (level 3) predictors can be added to the model to explain variation in the latent constructs (Raudenbush & Bryk, 2002). This type of model is facilitated by restructuring the data set, so that the items defining the constructs at level 1 are in long format (i.e., where multiple lines are used to define the items comprising *Y* within each individual in the study). For example, if we had five items defining productivity, each individual would have five lines of data (assuming no missing data). The MLM framework can also address the estimation of mediating effects (see Raudenbush & Sampson, 1999, and MacKennon, 2008, for further discussion), as well as multilevel models with various types of categorical outcomes. It should be noted, however, that if several mediated effects were proposed, it would likely be easier to specify the model using multilevel SEM due to its facility in estimating several direct and indirect effects in one model.

Multilevel Structural Equation Modeling

The SEM approach is also well suited for examining multilevel data structures, although the range of this capability can vary considerably with respect to existing SEM software. The general SEM approach represents a synthesis of factor-analytic techniques developed in psychology with simultaneous equation models

from econometrics and sociology (Kaplan, 1998). Structural models are particularly designed to accommodate latent variables (underlying constructs defined by observed indicators), measurement errors in both dependent and independent variables, direct and indirect effects, reciprocal causation, simultaneity, and interdependence (Jöreskog & Sörbom, 1993).

Although the SEM approach and corresponding computer software have been widely accepted in the analysis of single-level multivariate data, the techniques have not been widely applied to the analysis of multilevel data structures (Hox, 2010; Muthén, 1994). In the past, there were considerable difficulties associated with using SEM techniques in examining multilevel data structures properly, since existing SEM software packages were designed to examine sample covariance matrices, which often utilize summary information about the sample rather than raw data on individuals, and SEM estimation techniques were utilized optimally with large samples (i.e., N = 200 or more). This made it very challenging to examine typical multilevel data sets where there might be only 10 or so individuals nested within each unit. In recent years, however, researchers have worked to integrate SEM techniques within MLM in order to provide an expanded general methodological approach that would account for clustered sampling, population heterogeneity, measurement error, and simultaneous equations (e.g., Arbuckle, 1996; Hox, 1993, 1995, 2002, 2010; Kaplan, 1998; Kaplan & Elliott, 1997; McArdle & Hamagami; 1996; Mehta & Neale, 2005; Muthén, 1989; 1990, 1991, 1992, 1994; Muthén & Muthén, 1998-2012; Muthén & Satorra, 1989, 1995; Raudenbush & Bryk, 2002; Willett & Sayer, 1996).

In particular, the application of full information maximum likelihood (FIML) estimation to missing data situations, which is based on individual raw data in the sample (Arbuckle, 1996), rather than the sample covariance matrix, has led to the wider application of SEM techniques to clustered data. Importantly, the FIML approach takes advantage of all available data on individuals, including individuals who may have only partial data. This "raw data" approach represents a sharp contrast to estimation procedures based on the sample covariance matrix which may be biased due to dropping incomplete cases. The FIML approach has proven to be quite general and can therefore be applied to a variety of analyses including the estimation of mean and covariance structures in the presence of missing data, unequal spacing of observations in studies of individual growth, and unbalanced level-2 group sizes in multilevel analyses (Enders, 2001; Mehta & Neale, 2005). Presently, therefore, the SEM approach can provide a flexible framework for investigating multilevel data structures, since it facilitates the specification and testing of a wide variety of theoretical models.

There is an expanding literature on the use of SEM techniques in defining and testing multilevel models (e.g., Asparouhov & Muthén, 2008; Curran, 2003; Goldstein & McDonald, 1988; Grilli & Rampichini, 2007; Heck & Thomas, 2009; Hox, 2010; Mehta & Neale, 2005; 1997; Muthén & Asparouhov, 2011; Muthén & Muthén, 1998-2012; Muthén & Satorra, 1989, 1995). Much of the methodological work on multilevel SEM is continuing at present, so there is still considerable discussion regarding specific statistical issues that have surfaced from initial multilevel modeling efforts (e.g., potential biases in parameter estimation, standard errors, and fit indices resulting from sampling issues, effects of missing data, violations of normality, statistical power, application of sample weights at multiple levels). As multilevel modeling with SEM is becoming more common, a number of these

issues have been resolved. One example, as noted earlier, was the application of FIML estimation to SEM, which alleviated a number of problems in examining relationships between individuals and groups using the general SEM framework.

SEM With Mplus

Mplus is one SEM software program that is designed specifically for examining multilevel data structures using a latent variable framework (Muthén & Muthén, 1998-2012). It represents a re-design, considerable extension, and replacement of Muthén's (1988) LISCOMP program and is designed for easy use by applied researchers. A defining feature of the program is its flexibility in handling numerous types of models with categorical observed and latent variables. Mplus can be used for analyzing single-level univariate and multivariate designs, multiple group designs with mean and threshold structures, designs with missing values, longitudinal (growth) designs, mixture model designs (i.e., where different individuals are hypothesized belong to different subpopulations whose membership must be inferred from the data), complex sample modeling using sampling weights and a cluster variable, and two- and three-level data structures obtained through cluster sampling. More recently, the program can also handle cross-classified data structures at two levels (e.g., individuals cross-classified in schools and neighborhoods).

Some of the general models include multilevel regression models (with random intercepts and slopes), multilevel path analysis, multilevel confirmatory factor analysis (used to examine the measurement properties of latent constructs within and across groups), multilevel structural models (which can include latent variables and separate sets of predictors within and between groups), and multilevel mixture models (i.e., models defined by categorical latent variables which focus identifying emergent groups of individuals within a population). The program is also capable of specifying and testing a wide variety of longitudinal multilevel models and two-level cross-classified models.

The general structural equation model used in Mplus consists of two interrelated submodels. Readers familiar with the general form of the structural equation model may notice sight differences in notation [see the *Mplus User's Guide* (Muthén & Muthén, 1998-2012) for further discussion]. The first is the measurement model that relates unobserved (latent) variables to their observed indicators:

$$y_i = \nu + \Lambda\eta_i + \varepsilon_i \qquad (6)$$

where y_i is a vector of observed dependent variables for individual i, ν is a vector of measurement intercepts, Λ is a factor loading matrix of measurement slopes (which link observed variables to their underlying factors), η_i is a vector of latent variables, and ε_i is a vector of measurement errors (contained in Θ) which are uncorrelated with other variables (Muthén & Muthén, 1998-2012). The covariance matrix is represented as

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \qquad (7)$$

The second submodel, the structural model, specifies the causal relationships between the latent variables. If the latent variables are outcomes or mediating variables, they are often referred to as endogenous variables. Endogenous variables are causally dependent on other endogenous variables or on exogenous (x) variables. In contrast, exogenous variables are determined by causes outside of the model and, therefore, are not explained by the model. The structural relationships in a model may be written as

$$\eta_i = \alpha + B\eta_i + \Gamma x_i + \zeta_i \quad (8)$$

where η_i is a vector of endogenous factors for individual i, α is a vector of latent variable intercepts, B is a matrix of regression coefficients relating the endogenous factors to other endogenous factors, Γ is a matrix of regression coefficients relating the exogenous variables (x_i) to the endogenous variables, and ζ_i is a vector of errors in the equations, indicating that the latent variables are not perfectly predicted by the structural equations. The covariance matrix of ζ is denoted ψ.

In cases where all variables are observed, the model with structural relationships reduces to a standard path analytic model:

$$y_i = \alpha + By_i + \Gamma x_i + \zeta_i \quad (9)$$

The approach was presented by Jöreskog in 1977. Since then, a great number of technical strides have been made in using structural equation modeling with real world data, including problems of statistical power, violations of normality, strategies for handling missing data, indices to assess the fit of models, and model modification strategies. For the interested reader, Byrne (2012) has provided one extended introduction to SEM using Mplus.

Options for Analyzing Multilevel Data

Researchers should be mindful of the multilevel structure of data present in many types of research, but even if the multilevel nature of the data is taken into account, there are a variety of modeling options that can be considered (de Leeuw & Kreft, 1995). Mplus offers several ways available to deal with clustering effects that result from the study's sampling design. For example, if a "design-based" approach is used, a single-level SEM analysis can be maintained using the conventional covariance matrix (i.e., based on the number of individuals in the study), after adjustments are made for design effects including the unequal subject selection probabilities and non-independence of observations (Muthén & Satorra, 1995). Equal weighting of the estimates (as would occur in simple random sampling) would bias the estimates of the model's parameters because of the over-sampling of certain subpopulations. With sample weights, however, the clustering effects in the data due to stratified random sampling are treated as "noise" that is filtered out of the analysis to provide more precise estimates of the population parameters. This approach corresponds to conducting a single-level analysis with sample weight (see Muthén & Satorra, 1995, for further discussion). In Mplus, this type of analysis is referred to as TYPE = COMPLEX. Mplus provides correct (robust) standard errors and chi-square test of model fit for stratified samplings designs with sample weights.

An example of the design-based approach to clustered data might be where participants are selected at random from a set of organizations of differing sizes with the intent of producing an analysis focusing on the job satisfaction of individuals. In this hypothetical analysis, the researcher would not be interested in examining how group-level variables that might account for variation in employee job satisfaction levels between the organizations, but she or he still wants to adjust the estimates for possible clustering effects (i.e., similarities among individuals within groups). Moreover, without applying sample weights that are included in the data set, the subjects in over-sampled units, perhaps from smaller organizations, would exert undue influence on the population estimates of individuals' work-related job satisfaction. Readers interested in this approach can consult the *Mplus User's Guide* for further discussion.

In contrast to design-based modeling, "model-based" approaches (or disaggregated modeling) tend to focus more on the effects due to clustering than the effects due to sampling design (Muthén & Muthén, 1998-2012; Muthén & Satorra, 1995), although design effects can also be incorporated into the model-based approach through the use of sample weights at two levels. MLM, for example, is well suited to partitioning univariate outcomes into within- and between-group variance components and accounting for the partitioned variance components through specifying sets of predictors at each level of the clustered data. As noted previously, prior to the application of FIML to the estimation of missing data in SEM, the general statistical model for multilevel SEM was complicated and difficult to implement as a practical matter because of the inherent complexities of computing separate covariance matrices for each unit (McArdle & Hamagami, 1996). This was because SEM techniques generally depended on large sample sizes for efficient estimation of multivariate outcomes using sample covariance matrices. An initial approach that simplified the application of SEM techniques to hierarchical data structures was to assume that there was one population of individuals clustered within groups. Instead of developing a separate covariance matrix for each unit (which would generally be problematic due to small within-group sample sizes), the individual data could be decomposed into two separate covariance matrices for the within-groups and between-groups structures (e.g., Cronbach & Webb, 1975; Muthén, 1991, 1994; Muthén & Muthén, 1998-2012; Muthén & Satorra, 1989, 1995). This general approach forms the basis of the application of SEM techniques to multilevel data.

In this approach, variation in a set of dependent variables can be decomposed into an individual component (i.e., the individual deviation from the group mean) and a group component (i.e., the disaggregated group means). This individual decomposition can be used to compute separate within-groups and between-groups covariance matrices. This approach facilitates the investigation of models with different sets of predictors at each level of the data hierarchy. As noted previously, however, the application of FIML to estimate multilevel data using SEM has expanded the capability to define multilevel models that include random slopes, unequal spacing of measurement occasions (for longitudinal analyses), and analyses where individuals with partial data can be included. For two-level models, in Mplus terminology this approach is referred to as TYPE = TWOLEVEL.

To represent the hierarchical nature of the data in a multilevel SEM analysis, the subscript *c* is added to represent the cluster (group) component, and *i* again represents the individual component. We can think of the basic SEM submodels representing the covariance structure specified in Eqs. 6-8 as being split into their within- and between-group covariance matrices. As an example, consider a number of items that are proposed to measure two underlying leadership factors. Following Eq. 6, the within-group and between-group portions of the measurement model can be specified as

$$y_{ci} = \nu_B + \Lambda_W \eta_{Wci} + \varepsilon_{Wci} + \Lambda_B \eta_{Bc} + \varepsilon_{Bc} \qquad (10)$$

where y_{ci} is a vector of observed leadership variables, ν_B is a vector of intercepts (which are specified between groups only), Λ_W and Λ_B are the respective factor loading matrices, η_{Wci} and η_{Bc} represent the vector of latent leadership factors, and ε_{Wci} and ε_{Bc} represent the vector of residuals (contained in Θ_W and Θ_B). Unlike conventional single-level analyses, where independence of observation is assumed over all *N* observations, in multilevel SEM independence

is only assumed over the *C* clusters (Mehta & Neale, 2005; Muthén, 1991, 1994). The sample covariance matrix can then be represented within and between groups as

$$\begin{aligned} \Sigma &= \Sigma_W + \Sigma_B \\ \Sigma_W &= \Lambda_W \Psi_W \Lambda_W' + \Theta_W \\ \Sigma_B &= \Lambda_B \Psi_B \Lambda_B' + \Theta_B \end{aligned} \quad (11)$$

We may also add observed and latent predictors at both the within- and between-group levels. This allows the specification of separate structural models at each level (Kaplan & Elliott, 1997). Following Muthén and Muthén's (1998-2012) discussion (see also Muthén & Satorra, 1995), the general two-level SEM model (without random slopes) considers a vector of observed variables which can contain cluster-specific, group-level variables z_c (c = 1, 2,...,C) and within-group variables (y_{ci} and x_{ci}) for individual i in cluster c, where

$$v_{ci} = \begin{pmatrix} z_c \\ y_{ci} \\ x_{ci} \end{pmatrix} = v_c^* + v_{ci}^* = \begin{pmatrix} v_{z_c}^* \\ v_{y_c}^* \\ v_{x_c}^* \end{pmatrix} + \begin{pmatrix} 0 \\ v_{y_{ci}}^* \\ v_{x_{ci}}^* \end{pmatrix} \quad (12)$$

Note that in this formulation the asterisks represent independent between- and within-group components of the respective variable vector (Muthén & Satorra, 1995). The between-group matrix contains the between-group predictors (z_c), group-level variation in intercepts (y_c), and group-level variation in the individual-level predictors (x_c). Note that the within-group matrix contains the intercepts and individual-level *x* and *y* predictors and zeros (0) for the group-level variables.

In this basic two-level framework, variation in dependent variables such as organizational outcomes can be explained by several sources. These sources could include between-group predictors (z_c) like organizational size, which are conceived of as affecting only the between-group variability in organizational outcomes; individual-level predictors (e.g., individual demographics) that may be considered in some models as varying only within groups (x_{ci}), that is, having no between-group variation; or individual-level predictors that may be decomposed into their own within- and between-group components (x_c and x_{ci}). An example of this latter formulation might be a predictor such as employee motivation, which could vary among individuals in an organization and also be defined as an aggregate measure across the set of organizations. In some cases, the researcher might want to consider certain individual background variables (e.g., minority by race/ethnicity, socioeconomic status) as having between-group components as well.

The basic two-level model can therefore be translated into a within-cluster model with latent variables, which is written as

$$\begin{aligned} v_{ci}^* &= \Lambda_W \eta_{Wci} + \varepsilon_{Wci} \\ \eta_{Wci} &= B_W \eta_{Wci} + \zeta_{Wci} \end{aligned} \quad (13)$$

and a between-cluster model with latent variables, which is written as

$$\begin{aligned} v_c^* &= v_B + \Lambda_B \eta_{Bc} + \varepsilon_{Bc} \\ \eta_{Bc} &= \alpha_B + B_B \eta_{Bc} + \zeta_{Bc} \end{aligned} \quad (14)$$

Equations 13 and 14 imply there can be separate structural models with underlying factors within and between groups (note with Mplus 7, the general two-level model can be extended to include a third level as well). The general mean and covariance structure model (i.e., consisting of within-group and between-group components) for two-level data (Muthén & Muthén, 1998-2012)

can be expressed as follows:

$$\mu = v_B + \Lambda_B (I - \mathrm{B}_B)^{-1} \alpha_B$$

$$\Sigma_B = \Lambda_B (\mathrm{I} - \mathrm{B}_B)^{-1} \Psi_B (\mathrm{I} - \mathrm{B}_B)^{-1\prime} \Lambda'_B + \Theta_B \qquad (15)$$

$$\Sigma_W = \Lambda_W (\mathrm{I} - \mathrm{B}_W)^{-1} \Psi_W (\mathrm{I} - \mathrm{B}_W)^{-1\prime} \Lambda'_W + \Theta_W$$

For the interested reader, structural equation models that are more general are also formulated in Schmidt and Wisenbaker (1986), McDonald and Goldstein (1989), and Muthén and Satorra (1995), Curran (2003), Mehta & Neale (2005), Muthén and Muthén (1998-2012), and Byrne, (2012).

Model Estimation

There are a number of different ways to estimate parameters in multilevel models. When sampling designs are unbalanced, iterative estimation procedures are needed to obtain efficient estimates. The most common approach used to estimate multilevel models is maximum likelihood (ML). ML estimation provides a means of dealing with the uncertainty present in the observed sample data by finding optimal values for the unknown parameters in a proposed model using a likelihood function that is based on the underlying sampling distribution of the outcome (e.g., normal, binomial, multinomial, Poisson). The likelihood function conveys information about the unknown quantities in the model.

One advantage of ML estimation techniques is that they are generally robust to departures from normality with sufficiently large samples, and they produce asymptotically efficient and consistent estimates (Hox, 2010). Of course, in real life settings, the sample data may depart from normality and, therefore, may not always represent the population accurately. Under less-than-ideal sampling conditions, there has been considerable debate among methodologists about the efficiency of ML estimation, given the non-normal features of the groups (e.g., Longford, 1993; Morris, 1995; Muthén, 1994). Where the number of groups available to study is small and the within-groups sample sizes are unbalanced, ML estimation may provide biased estimates that understate the between-groups variance and, hence, the group-level coefficients may be misleading (Morris, 1995; Raudenbush, 1995). This can occur because the group-level slope coefficients are conditional on estimates of the group-level variance. In actuality, in many cases, the group-level sample may only be a crude representation of some population, as in a convenience sample. Although researchers may believe there to be a close correspondence between their convenience sample and a real population of interest, this assumed correspondence can be difficult to quantify (Draper, 1995; Goldstein, 1995). One alternative for small level-2 samples available in Mplus is Bayes estimation.

A second caution is that ML estimation assumes that missing data are missing at random (MAR). MAR implies that the missing data can be a function of both the observed outcomes and covariates in the model, as long as the propensity for a data point to be missing is not related to standing on the missing data on *Y* (e.g., missing data due to high student absenteeism is not related to low math achievement). Missingness can, however, be related to some of the observed data. Mplus can provide ML estimation under various conditions of missing data including MCAR (missing completely at random), MAR, and NMAR (not missing at random) for continuous, censored, binary, ordered categorical (ordinal), nominal, counts, or combinations of these variable types (Little & Rubin, 1987). For data that do not support the standard of MAR, ML estimation can be used by defining categorical latent variables that are used to represent indicators of missingness, and the missingness

can be predicted by continuous and categorical variables such as in pattern-mixture models (Muthén, Jo, & Brown, 2003; Muthén, Asparouhov, Hunter, & Leuchter, 2011).

Obtaining a set of model estimates involves an iterative process that determines a set of weights for random parameters in the model in order to maximize the likelihood function. The likelihood function provides the probability of obtaining the observed data over a range of possible parameter values that may be almost as likely as the ML estimate (Pawitan, 2001). This allows us to compare the probability of the observed data under different parameter values. The likelihood function allows the parameters in a model to vary while holding everything else constant in order to find the set of ML estimates that maximizes it (Azen & Walker, 2011). This allows us to observe how the likelihood function changes for different values of the parameters while holding the data constant. For multi-parameter models, this is often expressed as the discrepancy between the sample covariance matrix and a model-implied covariance matrix, where a smaller discrepancy implies a stronger fit of the proposed model to the sample data.

In reality, since we do not know the population values, we attempt to find values from the sample data for the unknown model parameters that result in the smallest discrepancy in the likelihood function between the model-implied and the sample parameter values. For many types of analyses (e.g., SEM), the sample data are summarized as a sample covariance matrix and a vector of variable means. Model estimation then proceeds by specifying a set of restrictions on the sample covariance matrix and examining the difference between the original sample covariance matrix and the reproduced (or implied) covariance matrix with the restrictions imposed. If one considers the sample covariance matrix to represent the true population covariance matrix, then the difference between the sample covariance matrix (**S**) and a covariance matrix implied by the proposed model ($\hat{\boldsymbol{\Sigma}}$) should be small if the model fits the data.

Available approaches for estimating model parameters can be considered as different ways of weighting the discrepancy between the corresponding elements of the observed and implied covariance matrices:

$$F = (\boldsymbol{s} - \boldsymbol{c})'\boldsymbol{W}(s - \boldsymbol{c}) \qquad (16)$$

where $\boldsymbol{s}$ and $\boldsymbol{c}$ are the nonduplicated elements of the observed and implied covariance matrices $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$, respectively, arranged as vectors, and $\boldsymbol{W}$ is the weight matrix. For example, if $\boldsymbol{S}$ were a 3 x 3 covariance matrix, $\boldsymbol{s}$ would be a six-element vector and $(\boldsymbol{s} - \boldsymbol{c})'$ would contain the differences between these elements in their respective covariance matrices (Loehlin, 1992). If $\boldsymbol{W}$ is an identity matrix, the expression in Eq. 16 reduces to the sum of the squared differences between corresponding elements of the observed and implied matrix (i.e., which is the OLS criterion). Generalized least squares estimation (GLS) uses the inverse of the $\boldsymbol{S}$ covariance matrix ($\boldsymbol{S}^{-1}$). This only needs to be estimated once during the model iteration process. ML estimation uses the inverse of the model-implied covariance matrix ($\hat{\boldsymbol{\Sigma}}^{-1}$), which must be updated at each iteration of the model estimation process until the model converges (if it does converge) and an optimal set of estimates is obtained.

ML estimation entails finding an unknown parameter or parameters for which the probability of the observed data is greatest. Estimation starts with an educated guess of the parameter estimates and requires a computational algorithm to accomplish the iterations. The mathematical relationships implied in the proposed model are solved iteratively using the EM (expectation

maximization) algorithm (Dempster et al., 1977), by making guesses about the unknown (i.e., random) parameters, given increases in the observed data and the current estimates of the model during successive iterations used to solve the likelihood function. At each successive step, a new set of estimates is obtained and evaluated until the estimates are optimized.

Arriving at a unique solution, or a final set of model estimates, is known as model convergence (i.e., where the estimates no longer change and the likelihood is therefore at its maximum value). It is important that the model actually reaches convergence, as the resulting parameter estimates will not be trustworthy if the model does not converge. Sometimes increasing the number of iterations will result in a model that converges but, often, the failure of the model to converge is an indication that it needs to be changed and re-estimated. Even if a model converges, however, it does not mean the estimates are the *right* ones, given the sample data. This implies that proposed models have to be evaluated in terms of how well they fit the data by using various statistical and practical criteria and, ultimately, by making a judgment about whether they are theoretically sound.

For two-level models, there are actually several estimation options currently available in Mplus. The maximum likelihood estimator (referred to as ML in Mplus) is appropriate with balanced group sizes. As noted previously, one limitation of conventional analyses of covariance and mean structures using ML estimation is they depend on relatively large samples with complete data. Individuals with missing data have to be eliminated from analysis through listwise deletion. This can result in many individuals being excluded, which is likely to result in biasing the estimation of the model's parameters. Where we have missing data, or where there are random coefficients in a two-level model and unbalanced group sizes, we need to model the raw data instead of the estimated covariance matrices.

As noted, FIML represents an ML approach based on the raw data in the sample, rather than just the sample covariance matrix, or in the case of a two-level model, the within-group and between-group covariance matrices. This approach takes advantage of all *available* data on individuals (i.e., including those who may have only partial data). This is why FIML is often referred to as a raw-data estimation approach (Arbuckle, 1996). In Mplus, FIML is available for estimating single-level and multilevel models (referred to as MLR). MLR is maximum likelihood estimation with robust standard errors adjusted for nonnormality. It is appropriate for use with unbalanced group sizes. The MLR chi-square statistic is asymptotically equivalent to the Yuen-Bentler test statistic (Muthén & Muthén, 1998-2012). MLR facilitates the investigation of continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types; random intercepts and slopes; and missing data. It also facilitates the modeling of individually-varying times of observation and random slopes for time-varying covariates.

With respect to multilevel models, we can think of unbalanced group sizes as a type of missing data problem and use FIML to estimate the model parameters. Under the assumption of missing at random (MAR) and conditional on the data, we can assume that individuals with missing data are the same as those with complete data. FIML produces the model-implied mean and covariance matrices that are contributed for each individual response pattern, resulting in a covariance matrix whose dimensionality depends on the amount of actual data present for each individual; that is, the dimension of the covariance matrix can vary across individuals

(Mehta & Neale, 2005). Assuming multivariate normality, FIML minimizes the function for individual *i* cases as follows (Arbuckle, 1996):

$$F_{FIML} = \sum_{i=1}^{N} \log |\Sigma_i| + \log \sum_{i=1}^{N} (y_i - \mu_i)' \Sigma_i^{-1} (y_i - \mu_i) \quad (17)$$

where y_i is the vector of complete data for case *i* and μ_i contains the *p* x 1 vector of fitted population means observed for case *i* (Hox, 2010). The determinant and inverse of Σ are based on the variables observed for each individual *i* (Enders, 2001). The important point is that the notation simply suggests that the dimensions of $\boldsymbol{y}_i$ and of the associated mean and covariance structures can vary from observation to observation; that is, the approach allows for different amounts of data per individual. The SEM approach implies that the clusters are the units of observation and individuals within the units are treated as variables, which is consistent with the idea of having "incomplete data" in typical SEM analyses, since the cluster sizes are unbalanced (Mehta & Neale, 2005).

The likelihood, or probability, of the data can vary from 0.0 to 1.0 and, since this can be a very small number, it is often easier to work with its natural log, since it makes differentiating the likelihood function easier. It turns out that minimizing the log likelihood function amounts to maximizing the likelihood of the observed data (i.e., if the probability of the data = 1, the natural log of it = 0). Using the log of the likelihood, the overall fit function for the sample can be obtained by summing the *n* casewise likelihood functions across the sample:

$$\log L(\mu, \Sigma) = \sum_{i=1}^{N} \log L_i \quad (18)$$

Other estimation options available in Mplus include Muthén's limited-information approach, which is referred to as MUML. Given the usefulness of FIML for modeling multilevel data structures, however, MUML is not much used today, since it is limited to investigating only models with continuous variables, random intercepts, and no missing data (Muthén & Muthén, 1998-2012). There is also a limited-information weighted least squares (WLS) approach available (Asparouhov & Muthén, 2007), which allows continuous, binary, ordered categorical (ordinal), and combinations of these variable types; random intercepts; and missing data. WLS estimation also reduces estimation time for various models (e.g., categorical outcomes). Another recent option is Bayes estimation, which supports models with continuous, categorical, and combinations of these variable types; random intercepts and slopes; and missing data (Muthén & Muthén, 1998-2012). Interested readers can consult the *Mplus User's Guide* for further information about types of models that will support each type of estimation.

Once the proposed model is estimated and converges on a solution, we can assess how well it fits the data using various model fit indices. In general, confirmation of a proposed model relies on failure to reject the null hypothesis (H_0)—that is, concluding the data are consistent with the proposed model. The desired failure to reject the null hypothesis may be somewhat different for readers who are used to equating rejection of the null hypothesis as acceptance of an alternative research hypothesis (H_1). In contrast to this common use of the null hypothesis, we wish to conclude that the model cannot be rejected on statistical or practical grounds. Failure to reject the null hypothesis therefore implies that the proposed model is a plausible representation of the data, although it is important to note that it may not be the *only* plausible representation of the data. The value of the fit function based on the final parameter estimates is used to determine

how well the proposed model implied by $\hat{\boldsymbol{\Sigma}}$ fits the observed covariance matrix ***S***. As noted previously, if the value of the log likelihood function is 0, it suggests the proposed model fits the data perfectly.

The log likelihood estimate can be used to calculate a deviance statistic (i.e., also referred to as -2LL or -2*log likelihood) in order to compare models. We multiply the log likelihood by -2, so it can be expressed easily as a positive number. Models with lower deviance (i.e., indicating a smaller discrepancy function) fit better than models with higher deviance. If the observed and implied covariance matrices are identical, the value of the expression will be 0. We can also look at the model residuals, which describe the difference between the two matrices. Large residuals imply that some aspects of the proposed model do not fit the data well. Other criteria [e.g., comparative fit index (CFI), standardized root-mean-squared residual (SRMR, chi-square coefficient)] are also provided to help determine whether a particular model provides a plausible fit to the data. In addition, nested models (where a more specific model is formed from a more general one) can be compared by examining differences in deviance coefficients under specified conditions [e.g., see Muthén & Muthén (1998-2012) for more information about comparing models using Mplus].

Statistical Power

In investigating multilevel data structures, researchers should also consider issues surrounding statistical power and the sensitivity of their models to hypothesis testing. Since power is defined as the probability of finding a statistically significant effect if it indeed exists, power is closely tied to hypothesis testing. Tests of significance are designed to provide evidence with respect to an event having arisen because of sampling error. A *t*-test (or *z*-test) is a common statistical test that is often used to determine whether or not a model parameter is statistically significant (i.e., defined as the ratio of the estimate to its standard error). The test will therefore depend heavily on the accuracy of the standard error estimate. Unfortunately, in multilevel studies, when models are estimated with small numbers of groups (or unbalanced groups), the error variance is likely to be underestimated, resulting in standard error estimates that are too small, and a greater likelihood of committing *Type I* errors (i.e., falsely rejecting the null hypothesis).

Estimating power requires the researcher to consider the magnitudes of any anticipated effects (effect sizes), the sample size (i.e., number of clusters or groups needed and their within-group sizes), and the likely within- and between-group variance (intraclass correlation) associated with the observations (see Cohen, 1988; Hoyle & Panter, 1995; Kaplan, 1995; Kish, 1957; 1965; MacCallum, Roznowski, & Necowitz 1992; Muthén & Satorra; 1995; Saris & Satorra, 1993; Satorra & Saris, 1985 for further discussion). For example, larger anticipated effects and sample sizes are related to greater statistical power. Detecting smaller effects would of course require larger numbers of groups to be included in the study. The challenge for multilevel SEM analysis is that there may be several parameters of interest which have different anticipated effect sizes. This might imply different sample sizes that would be required to detect each effect. Of course, the best time to think about statistical power is in the design phase of studies. These issues and statistical power are so related that a small change in one can have a profound influence on power.

We should also consider the size of the interclass correlation in designing a multilevel study. Barcikowski (1981) demonstrated that ignoring intraclass correlations in analysis of variance studies can greatly inflate the chances of

making *Type I* errors. For example, in a study with 10 individuals in each group and an intraclass correlation of 0.20 (i.e., suggesting that 20% of the variance is between groups), the significance level of 0.05 is inflated to 0.28 (see also Muthén & Satorra, 1995). Of course, using this inflated alpha level would result in many more findings of significance. It is therefore important to note that in the presence of small intraclass correlations (e.g., as is typical in studies of school effects), it would be desirable to increase the number of groups included in the study in order to achieve more accurate estimates of parameters, standard errors, and error variances. This is especially important for obtaining accurate estimates of the model's between-group parameters, since the number of individuals contributing information to the calculation of the model's within-group parameters with ML estimation will generally be accurate if there are at least 200 or more individuals in the study (Boomsma, 1987; Chou & Bentler, 1995; Mok, 1995). Determining the required number of groups needed in a study, however, is more problematic and depends upon the anticipated effects and the complexity of the model being estimated. In samples that include a large number of groups, a change in the number of individuals within each group will have only a minimal impact on statistical power. On the other hand, when the sample is comprised of a small number of groups, a relatively small change in the within-group size can have a substantial impact on statistical power.

Most discussions of sample size and related issues are based on the use of probability samples. Although individuals may be chosen at random within units, they are seldom assigned at random to their existing units. When convenience samples are used for groups, it is unclear what the effects might be. As emphasized by others (e.g., Busing, 1993; Hox & Maas, 2001; Kreft & de Leeuw, 1998; Mok, 1995), caution should certainly be exercised in putting strong credibility in results where the number of groups is small (i.e., $N < 100$). Under these types of conditions, it is quite likely that the groups are not normally distributed. Accurately modeling the distribution of effects across a sample of groups that may not be randomly selected or that may depart from normality, therefore, is generally more of a problem in multilevel modeling than problems presented by the number of individuals sampled within each unit (e.g., Morris, 1995). Of course, the best data collection procedure is to have large numbers of observations per unit, relatively large numbers of units, and little or no missing data. Changes in any one of those conditions affect the completeness of our knowledge. In the real world, however, it is not always possible to utilize optimal sampling methods. Therefore, each individual study must be judged on its strengths and weaknesses, as well as how it contributes (whether flawed or not) to the development of research knowledge.

M**ultilevel CFA**

A first type of multilevel model to consider is where we wish to define and investigate underlying constructs and their observed indicators using confirmatory factor analysis (CFA). CFA is a useful general approach for investigating the relationships between constructs and their indicators because it provides a mathematical model that links the observations, or manifestations, of the underlying processes to the theories and constructs through which we interpret and understand them. The assumption is that an underlying construct such as job satisfaction may have a component that varies across individuals within an organization (i.e., individuals likely have different levels of job satisfaction) and a collective component that varies across organizations (i.e., organizations

vary in average levels of job satisfaction). Through CFA, the researcher can assess the reliability and validity of the measurements through the careful specification of constructs and their indicators prior to testing the model with the sample data. With multilevel factor analysis, we can examine the stability of factor models within and across groups. For example, we may find that three correlated factors consisting of several items each define leadership practices within organizations, while a general leadership factor is sufficient to describe the variation present between organizations. Moreover, we can examine the amount of measurement error in the observed items that define latent factors both within and between organizations. The unreliability of these measures affects the decomposition of variance, which can affect the intraclass correlations (Muthén, 1991). The individual-level error variance tends to inflate the contribution of within-level variation to the calculation of the intraclass correlation.

Multilevel factor analysis, therefore, gives results that correspond to those that would be obtained from perfectly reliable measures (Muthén, 1994). When we look at an error-free variance ratio for the intraclass correlation, we are gaining a more precise estimate of the within- and between-level contributions. Through this process, we can test the construct validity of our proposed multilevel CFA model and, hence, improve the estimation of proposed relationships between the latent factors by first paying more attention to the reliability and validity of constructs comprising the theoretical model.

Specifying a Multilevel CFA

In this example, 384 employees in 56 organizations rated their managers' leadership in 36 areas. We will concentrate on a subset of the data–six items that we propose define two latent leadership factors. It is important to note that ML estimation is best applied to interval data because of assumptions about the normality of the data. Its use with ordinal data has been open to debate (e.g., accuracy of parameter estimates, effects on fit indices). As the number of scale points increases, however, ordinal data behave more like interval data (Boomsma, 1987; Rigdon, 1998). One should, however, look at the measurement properties of ordinal data (and scales developed) closely before deciding which estimation method to use. We can also specify the items as continuous or ordinal and compare the results. Ordinal data in Mplus can also be examined using MLR estimation (which produces log odds coefficients observed variables by default and probit coefficients can be requested) or WLS estimation (which produces probit coefficients for observed dependent variables). Standardized coefficients are also available. Of course, if we define the items as ordinal, the resulting factor model has no residual errors for level-1 factor loadings.

The proposed multilevel factor model assuming continuous observed indicators is presented in Figure 1. The two latent factors of leadership (i.e., *Governance* and *Evaluation*) are enclosed in circles. Each factor is defined by arrows leading to the three observed indicators (enclosed in rectangles) and their corresponding unique factors (i.e., with short arrows representing measurement errors). The governance factor consists of the extent to which the manager involves employees in shared decision making (*shdec*), uses a team approach internally (*team*), and encourages client involvement in shaping the direction (*inclin*) of the organization. The evaluation factor consists of the extent to which the manager develops evaluation standards for assessing employee performance (*evstan*), uses systematic assessment procedures (*syasse*), and evaluates new programs that are implemented (*evprog*).

As summarized in the figure, the multilevel factor model suggests that there are four orthogonal sources of variation for each observed variable. The within-group sources of variation are (1) the individual variability common to the variables that load on each latent factor (shown with filled circles in the figure to indicate their intercepts vary at the group level), and (2) the individual variability that is specific to each observed variable (its residual). The between-group sources are (3) the group variability common to the variables that load on each factor (shown as ovals in the diagram since they are conceptualized as "unknowns" between groups), and (4) the group variability specific to each observed variable (the group-level residual).

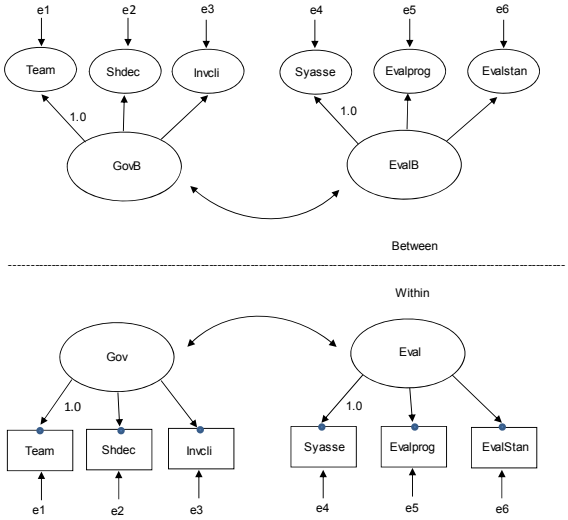

Figure 1. Proposed two-level CFA model.

The goal of the multilevel analysis is to summarize the within- and between-group variation in this leadership model and establish whether the same individual-level model holds at the organizational level. It is likely that there may be differences in the quality of measurement of items defining the factors that result from the multilevel nature of the data. For example, we may reasonably expect that employees within each organization differ to some extent in their assessments of a manager's performance with respect to these two leadership dimensions. The within-groups model addresses the portion of variance in the factors that results from variation among individuals.

Similarly, we can expect that there are also differences in managers' leadership performance between organizations (i.e., between-groups variance). The group-level model addresses the across-group variation rather than the across-individual variation (Muthén, 1991). Hence, we hypothesize that the same two-model factor holds across organizations, but that there may be likely differences in the measurement quality of the items used to define the two leadership factors. Some of this difference is also likely due to the differing amounts of variance contributed by each observed variable across levels (i.e., the intraclass correlation). Alternately, we could also hypothesize that one general leadership factor may be sufficient to account for the variation in leadership between organizations. Of course, there are many additional possibilities to consider when multilevel factor models are conceptualized and tested.

Steps in Testing the Model

It is sometimes useful to propose and test a conventional, single-level model (using the total sample covariance matrix $\mathbf{S}_T$) as a first step in specifying a two- (or three-) level CFA model. Of course, the conventional analysis using the total sample (N=384) will be to some extent biased (e.g., depending on the size of the intraclass correlations) because of its failure to consider the nested effects of the data, but this will likely give some indication of the observed variables that can be used to serve as indicators of the latent constructs. The model fit indices are likely to give rough estimates of the model's adequacy. Moreover, we may be able to spot obvious

misspecification, such as the presence of poor items or correlated error between indicators.

In this preliminary test, the proposed two-factor model fit the data well [χ^2 (8 df) =10.804, p = .21, and RMSEA = .03, p = .74]. Each set of three items loaded well on its hypothesized latent factor (with standardized loadings ranging from 0.61 to 0.81), and the errors were generally small (0.35 to 0.62). Because the single-level model fit the data adequately, we can specify the two-level model. We first focus on the size of the intraclass correlations for the observed indicators. The intraclass correlation summarizes the proportion of the total variation that lies between groups. They can be examined to check whether $\Sigma_B = 0$. If the intraclass correlations are very small, it would suggest a multilevel model may not be necessary. In this case, there was considerable variation between groups for the six observed measures of leadership, with intraclass correlations as follows:

```
SHDEC    0.225  INVCLI   0.155  TEAM   0.122
EVALSTAN 0.200  EVALPROG 0.176  SYASSE 0.201
```

Because there was sufficient between-groups variation in the observed variables, we can proceed to the simultaneous test of the within- and between-group models. Following are the Mplus input statements for this model. The DATA statement identifies the data file. The VARIABLE lines identify the variables in the data set, the variables used in the current analysis, and the variable used to form the clusters (group). The ANALYSIS command identifies the type of analysis (Type = Twolevel), and we can also specify the estimation method to use. With unbalanced group sizes, we can use MLR (which is the default estimation method), since it provides the correct the chi-square coefficient and standard errors. The Model statements for a two-level analysis require a separate model to be specified between (%Between%) and within (%Within%) groups. Each between-group leadership factor is measured by three observed variables. This is specified using the BY statement, which is short for "measured by." As part of the default specifications, Mplus automatically fixes the first observed indicator defined for each factor at 1.0 to provide a metric for measuring the latent factor. In this example, I fixed the error variance for *Team* to 0 (Team@0) between groups for model convergence. The covariance between factors is automatically estimated by the program. In the within-groups model, each leadership factor is similarly defined by three observed indicators. The OUTPUT statement requests sample statistics, standardized coefficients, modification indices (indicating parameters that could be freed to improve the model's fit), and TECH1 (which provides the vectors and matrices where the specified model parameters are contained). TECH1 output is useful to examine with respect to the within- and between-group specification of the factor model as defined in Eq. 11 (see Appendix B).

```
TITLE:     Two-Level CFA;
DATA:      FILE IS C:\Mplus\leadch.dat;
           Format is free;
VARIABLE:  Names are group shdec invcli
           team evalstan evalprog syasse;
           Usevariables are group
           shdec-syasse;
           cluster is group;
ANALYSIS:  Type = twolevel;
           Estimator is MLR;
  Model:   %Between%
           bgov by team shdec invcli;
           beval by syasse evalprog
           evalstan;
           team@0;
           %Within%
           gov by team shdec invcli;
           eval by syasse evalprog
           evalstan;
OUTPUT:    SAMPSTAT STANDARDIZED
           MODINDICES TECH1;
```

Output information includes the average cluster size, the intraclass correlations for the observed variables, the between-group and within-group covariance matrices, fit indices, the unstandardized and standardized parameter estimates, the intercepts and residuals, and the item squared multiple correlations, showing the variance accounted for by the factors.

Table 1. Model Fit Indices

Chi-Square Test of Model Fit	
Value	19.865*
Degrees of Freedom	17
P-Value	0.2812
Scaling Correction Factor for MLR	0.8914
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.021
CFI/TLI	
CFI	0.997
TLI	0.994
Chi-Square Test of Model Fit for the Baseline Model	
Value	860.860
Degrees of Freedom	30
P-Value	0.0000
SRMR (Standardized Root Mean Square Residual)	
Value for Within	0.021
Value for Between	0.059

In Table 1, we can first examine the fit of the proposed model to the data. The χ^2 (17 df) for the within- and between-groups models was 19.865 (p = .89), which suggests the model should not be rejected on statistical grounds alone. In addition, the comparative fit index 0.997 was above recommended levels (i.e., 0.95 or above), and the RMSEA was small (0.021, with values of 0.05 indicating a strong fit to the data). Finally, the standardized root-mean-squared residual (SRMR) was small 0.02 on the individual level but a bit larger (0.059) on the group level. Values of this index near 0.05 or smaller are generally considered evidence of a good fit. Modification indices (not tabled) indicated the error covariance between *team* and *evalstan* might be freed on the group level to improve the model fit; however, as this could not really be justified on theoretical grounds, we might just leave the covariance fixed at 0.

Once the fit of the model is determined to be adequate, it is important to assess the quality of the parameter estimates. The Mplus standardized estimates are summarized in Figure 2. There are a number of different ways to standardize estimates in multilevel modeling. Mplus standardizes the between-level parameters by the between-level variances for latent and observed variables and the within-group parameters by the within-group variances for latent and observed variables. This is helpful in determining how much variance is explained at each level separately (L. Muthén, personal communication, 12-17-98).

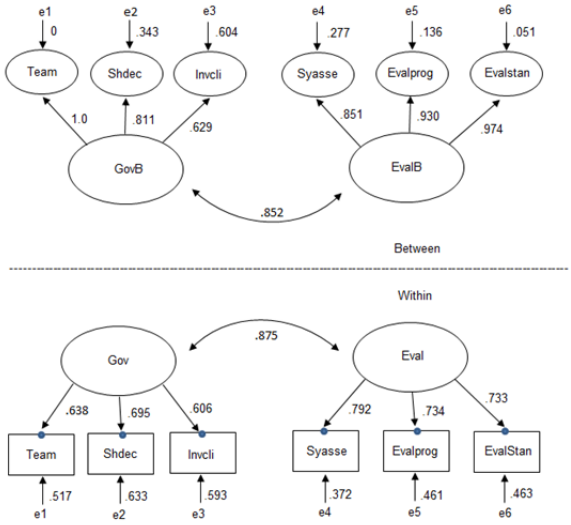

Figure 2. Standardized parameter estimates.

On the individual level, the standardized loadings on the governance factor ranged from 0.606 to 0.695 and were all substantial in size and statistically significant (i.e., tested as the ratio of the estimate to its standard error). This provides evidence that the observed variables serve as reliable indicators of the latent leadership variable. The evaluation factor was somewhat better measured (with loading ranging from 0.733

to 0.792). The correlation between the factors within groups was strong at 0.875.

Similarly, on the group level, the factor loadings were also relatively high (ranging from 0.629 to 1.00). We can also notice in the figure that the errors were generally smaller between groups (except for *invcli*). This suggests that measurement errors due to the observed indicators definitely affect the individual-level variance contributing to the intraclass correlations for each observed variable (Muthén, 1991). We would, of course, be a bit more cautious with this part of the model because of the relatively small number of groups in the study ($N = 56$). There was again a strong correlation between the factors (0.852), which can suggest that perhaps one leadership factor might be enough to capture the between-group variation. Subsequent testing of one general between-group leadership factor, however, did not result in an improved model fit. If we defined a common measurement scale for each item measuring the factors (by constraining each item loading to be invariant within and between groups), the variances of the factors at each level could be directly compared [see Mehta & Neale (2005) and Muthén & Muthén (1998-2012) for further discussion].

We can also estimate the model using ordinal data by adding the following line below the group identifier:

```
Categorical is shdec-syasse;
```

Latent variable models with categorical indicators using ML estimation require numerical integration in the computations (Muthén & Muthén, 1998-2012). In Figure 3, the results are summarized using (a) MLR estimation and the default logit link with Monte Carlo integration (which reduces the number of integration points in the model for faster estimation) and (b) WLSMV estimation (i.e., weighted least squares with mean and variance adjusted chi-square test statistic) with the default probit link. The figure suggests the results were similar, whether the items were defined as continuous or ordinal and either MLR or WLSMV estimation was used. Readers can assess for themselves the slight differences in presentation of the CFA model with items defined as continuous versus ordinal—one difference being with ordinal data there are no error terms for items at level 1. Of course, in some circumstances appropriate practice would dictate defining the items in the model as ordinal.

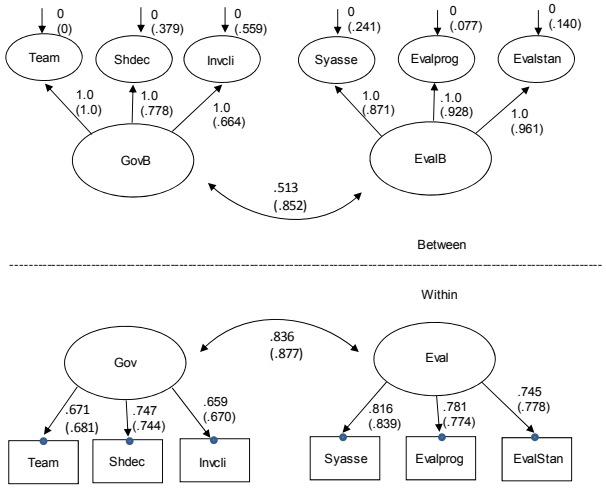

Figure 3. Standardized estimates using MLR estimation (WLSMV standardized estimates are in parentheses).

Multilevel Path Models

A second type of multilevel SEM model is the multilevel path model. This is a simplified model that uses only observed variables instead of latent constructs. Unlike the general multilevel regression model, however, multilevel path formulations encourage the investigation of more complex relationships that include multiple dependent variables, mediating variables and, therefore, indirect effects (i.e., combined effects

through two or more paths). This facilitates the specification of separate sets of structural relationships including both direct and indirect effects within groups and between groups. Indirect effects are overlooked within the typical multilevel regression model. Although we could define a series of separate multilevel regression models to estimate the parameters in a multilevel path model, one advantage of using the multilevel SEM framework is that we can specify all of the hypothesized relations within one model that provides simultaneous estimation of all proposed paths.

The primary disadvantage of a multilevel path analysis, however, is that it does not account for measurement error in defining the variables included in the model, which potentially introduces bias into the estimation of the model's parameters. This is an important limitation for several reasons. First, as we noted in developing the previous multilevel factor model, the unreliability of the observed measures affects, to some extent, the variance decomposition of the variables across organizational levels into their within- and between-group components (Muthén, 1991). Second, in the multilevel path model, measurement error in the endogenous variables (i.e., mediating or outcome variables) will affect the precision of the estimates (Kaplan, 1998). Where scales are used to define the variables in the model, it would be important to examine the reliability of the items comprising the scales preliminarily. When the specific research focus is not on measurement error, however, path models can be a useful approach to the multilevel modeling of individual and group processes. Mplus is able to formulate and test multilevel path models that contain random variation in intercepts and within-group regression slopes (Muthén & Muthén, 1998-2012).

Specifying a Multilevel Path Model

Consider an example where fourth grade students (N = 9,663) in 154 schools were measured on standardized tests of math skills at grades 3 and 4. We can define a separate path model within schools consisting of individual variables such as gender (with female coded 1), socioeconomic status (with low SES coded 1), and previous achievement (measured by the third grade standardized math test). It is hypothesized the student composition variables affect grade 3 achievement and also grade 4 achievement. Moreover, it is hypothesized that grade 3 achievement will affect grade 4 achievement.

In the between-school path model, we can specify a separate set of context and process variables proposed to impact grade 3 and grade 4 school math achievement. These variables include school SES composition (*schcomp*), the number of students enrolled (*zenroll*), the perceived effectiveness of the school's instructional program (*zinst*), and the average effectiveness of its teachers (*zteff*). It is also hypothesized that school grade 3 math scores will affect grade 4 math scores. Following is the final Mplus input file (after removing three paths that were not associated with the mediating and outcome math variables at $p < .10$ or $p < .05$).

In Mplus, structural relations between variables are specified with ON statements, which are short for "regressed on." As indicated in the model statements, there are no latent variables included in this model, since there are no BY statements included. In this example, we can begin by examining the intraclass correlations of the observed outcome variables. The intraclass correlations for *math1* and *math2* were 0.089 and 0.136, respectively (not tabled). This suggests that roughly 9-13 percent of the variance in math scores lies between schools. This is sufficient between-group variation to proceed with the multilevel analysis.

```
TITLE:     Two-level path model;
DATA:      FILE IS C:\Mplus\SEMpath.dat;
           Format is 6f8.0,5f8.2;
VARIABLE:  Names are schcode female lowses
           math1 var1 math2 zenroll ztexp
           zteff zinst schcont;
           Usevariables are schcode math1
           math2 female lowses  ztexp
           zteff zenroll zinst schcont;
           cluster = schcode;
           between =  ztexp zteff zenroll
           zinst schcont;
           within = female lowses;
ANALYSIS:  Type = twolevel;
           Estimator is MLR;
   Model:  %Between%
           math2 on math1 ztexp schcont;
           math1 on ztexp zenroll schcont
           zteff zinst;
           %Within%
           math2 on math1 lowses female;
           math1 on lowses female;
OUTPUT:    SAMPSTAT STANDARDIZED TECH1;
```

The initial model was specified as saturated (i.e., all possible path estimated), which resulted in a perfect fitting model (i.e., the chi-square coefficient for 0 degrees of freedom = 0). After removing three paths that were not statistically significant on outcomes at the school level, we obtained a chi-square estimate of 7.027 (3 df, p = .071). This suggests that the model should not be rejected on statistical grounds alone. Other indices were also strong, given there were only a few over-identifying constraints (i.e., 3 degrees of freedom in the model).

Because the reduced path model fit the data reasonably well, the parameter estimates can next be examined. We can test for indirect effects (*ind*) through specifying the following command before the OUTPUT command:

```
Model indirect:
           math2 ind lowses;
           math2 ind female;
           math2 ind schcont;
           math2 ind ztexp;
```

The standardized estimates are summarized in Figure 4. The within- and between-groups models indicate a variety of statistically significant direct and indirect effects on math outcomes. Within schools, third grade math (*Math 1*) scores had the strongest direct effect on fourth grade math (*Math 2*) scores (0.699, p < .05). In contrast, the direct effect of student low SES on Math2 outcomes was small but statistically significant (-0.147, p < .05), as was the effect of female on Math2 outcomes (0.110, p < .05). Low SES also produced a small direct effect on Math1 scores (-0.376, p < .05) as did female (-0.127, p < .05). In turn, low SES also exerted a small indirect effect on Math2 achievement through Math1 scores (-0.263, p < .05) as did female (0.089, p < .05).

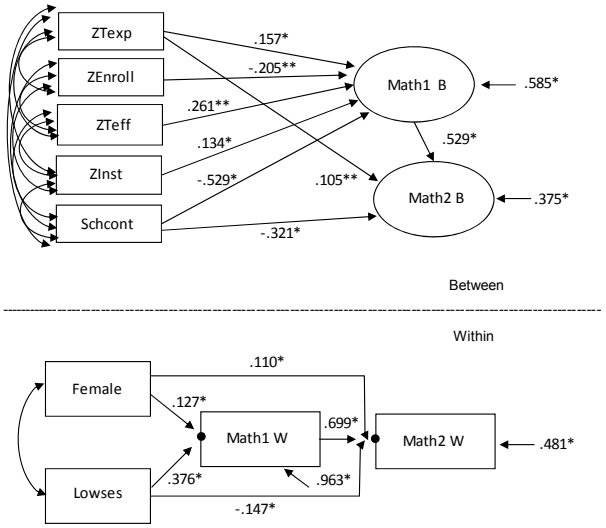

*Figure 4. Standardized path estimates (*p < .05, **p < .10)*

For the between-school model, the strongest predictor of Math2 achievement was Math1 achievement (0.529, p < .05). School composition was also associated with Math2 scores (-0.321, p < .05). There was also a weak direct effect associated with teacher experience on Math2 scores (0.105, p < .10). Regarding indirect effects

(through Math1), school context exerted a significant effect (-0.269, $p < .05$) as did teacher experience (0.083, $p < .05$) on Math2 outcomes. The figure also indicates that all five school predictors affected *Math1* at either $p < .10$ or $p < .05$. Other than school composition and teacher experience, however, the other predictors did not produce significant indirect effects on *Math2* outcomes.

We can also determine the variance in outcomes accounted for at each level. Between schools, the variables in the model accounted for 62.5% percent of the between-school variance in Math2 scores (with the 37.5% representing the errors in the equations, as summarized in the figure) and 41.5% percent of the between-school variance in Math1 scores (with 58.5% representing the errors in the equations). The within-school variables accounted for 51.9% of the within-school variance in Math2 scores and 3.7% of the within-school variance in Math1 scores. We would likely conclude that there are other variables that could be added to the model to improve its overall representation of variables explaining student learning.

Multilevel Structural Models with Latent Variables

We can also incorporate within- and between-group latent variables (with observed indicators) into multilevel analysis, which can bring several benefits to the measurement of variables in a model and, hence, improve the accuracy of its structural relations. To illustrate this type of model, we will again use the data set on organizational leadership ($N = 384$ individuals within 56 organizations), where the sampling design was unbalanced. The proposed multilevel structural model may be defined in manner that is very similar to the multilevel factor model.

Within groups, we again have two leadership factors, each comprised of three survey items with separate error terms. The three observed indicators defining *governance* practices are shared decision making, client involvement in setting company direction, and creating a team-oriented work environment. For *evaluation* practices, the observed indicators are using systematic assessment procedures, using of standards for personnel evaluation, and evaluating the implementation of new programs. In the within-group portion of the model, we propose that an individual's organizational *role* affects her or his perceptions of leadership practices. Managers (coded 1) were asked to complete a self-report of their leadership, and employees (coded 0) were asked to complete a performance assessment of their immediate supervisor. The hypothesis is that managers will systematically rate their leadership skills and activities more favorably than their employees will rate their leadership.

As in the multilevel factor model, the intercepts of the observed variables comprising the two leadership factors are hypothesized to vary across organizations. The group-level variation is modeled in terms of the two latent leadership factors (*GovB* and *EvalB*), each also defined by three observed indicators with separate error terms. In the between-group portion of the model, the predictors are organizational *effectiveness* in terms of the perceived quality of outputs produced (with effective coded 1 and ineffective coded 0) and organizational *type*, which was defined as product oriented (coded 1) and service oriented coded 0). The proposed structural model is summarized in Figure 5.

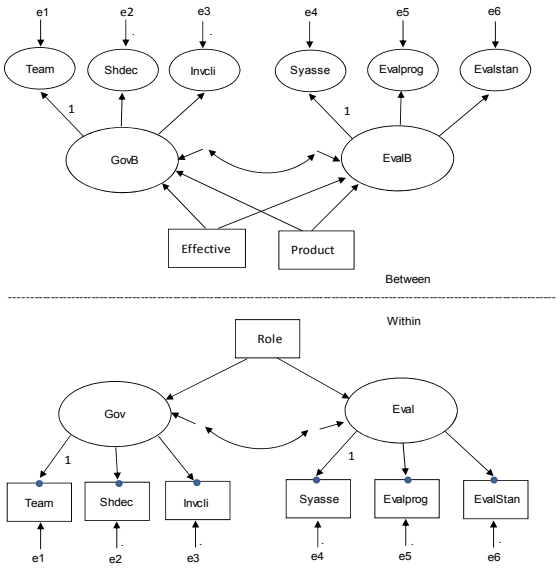

Figure 5. Proposed multilevel structural model.

The Mplus model input statements are presented next. Note again, I set the error term for *team* between groups to 0 (team@0;) to ensure the model would converge on a unique solution.

```
TITLE:      Two-Level lead SEM;
DATA:       FILE IS C:\Mplus\leadSEM.dat;
            Format is 10f8.0;
VARIABLE:   Names are group shdec invcli
            team evalstan evalprog syasse
            role level effect;
            Usevariables are group
            shdec-syasse role level effect;
            cluster is group;
            between = level effect;
            within = role;
ANALYSIS:   Type = twolevel;
            Estimator is MLR;
   Model:   %Between%
            bgov by team shdec invcli;
            team@0;
            beval by syasse evalprog
            evalstan;
            bgov beval on level effect;
            %Within%
            gov by team shdec invcli;
            eval by syasse evalprog
            evalstan;
            gov eval on role;
OUTPUT:     SAMPSTAT STANDARDIZED;
```

Output from the Analysis

The overall fit of the proposed structural model was judged as acceptable ($\chi^2 = 99.042$, df = 29, $p < .001$, RMSEA = .079, CFI = 0.93, results not tabled). Once again, the model fit could be improved by freeing error covariances (but making such changes could not be justified on theoretical grounds). In Figure 6, the standardized estimates are presented.

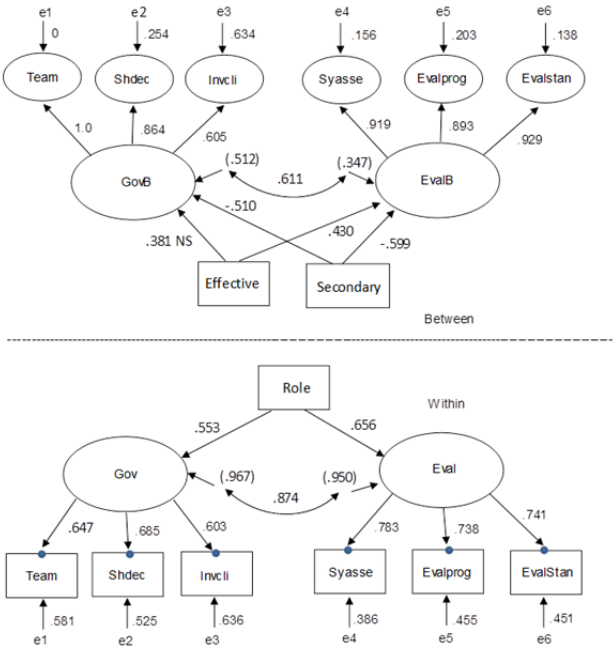

Figure 6. Standardized path estimates (note: NS = not significant)

All observed indicators of the two leadership latent variables loaded substantially within-groups and between-groups. This suggests the factors account for substantial within- and between-group variability in the observed variables. Organizational role exerted a moderate standardized effect on both leadership evaluation (0.656, $p < .05$) and governance (0.533, $p < .05$), suggesting that managers rated their leadership skills higher than their employees rated them.

At the group level, organizational effectiveness had a significant effect on between-group leadership ratings for evaluation (0.433, $p < .05$) but not governance (0.381, $p > .10$). Type of organization (i.e., product-oriented) had a negative effect on perceptions of manager evaluation (-0.599, $p < .05$) and governance (-0.510 $p < .05$) practices. The organizational-level variables accounted for 48.8% of the between-group variance in leadership governance and 65.3% of the variance in evaluation, while the one individual-level predictor accounted for only 3.3% of the within-group variance in leadership governance and 5.0% of the variance in leadership evaluation. Given the variety of information and the sensibility of the estimates, therefore, we can accept the model as a plausible representation of the data.

Summary

The goal of this chapter was to provide an interdiction to two-level modeling using SEM techniques. This is one approach that can be used to examine data obtained from cluster sampling, where measurement errors in defining constructs is considered an important issue and researchers may also be interested in defining models featuring separate structural models at two (or three) levels. The examples were intended to give readers a sense of a few of the many substantive problems that can be addressed using multilevel SEM techniques. The three basic models (CFA, path models, structural models with latent and observed variables) presented in this chapter can easily be extended to include three hierarchical levels, as well as various types of multilevel mixture models (i.e., where identifying subgroups within the population is the focus of the investigation).

It is important to keep in mind the role of theory in defining and testing multilevel models. Multilevel SEM using Mplus allows the investigation of a wide range of theoretical models with latent and observed variables defined with observed variables on diverse measurement scales. Despite some problems that may sometimes be encountered in specifying and estimating multilevel models with SEM, the approach can yield answers to variety of research questions in the social and behavioral sciences concerning individual processes, group processes, and outcomes.

References

Aitken, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of Royal Statistical Society*, *Series A, 149*, 1-43.

Arbuckle, J.J. (1996). Full information estimation in the presence of incomplete data. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-278).Mahwah, NJ: Lawrence Erlbaum.

Asparouhov, T. & Muthén, B. O. (2007). *Conceptually efficient estimation of multilevel high-dimensional latent variable models.* Proceedings of the 2007 JSM meeting in Salt Lake City.

Asparouhov, T. & Muthén, B. O. (2008). Multilevel mixture models. In G. Handcock & K. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing, Inc.

Azen, R. Walker, C. (2011). *Categorical data analysis for the behavioral and social sciences.* New York: Routledge.

Barcikowski, R. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6*(3), 267-285.

Bock, R.D. (1989). *Multilevel analysis of*

educational data. San Diego: Academic.

Boomsma, A. (1987). The robustness of maximum likelihood estimation in structural equation models. In P. Cuttance and R. Ecob (Eds.), *Structural Modeling by Example* (pp. 160-188). Cambridge: Cambridge University Press

Bryk, A.S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research in evaluation. *Review of Research in Education*, 8, 158-233.

Busing, F. M. (1993). *Distribution characteristics of variance estimates in two-level models*. Preprint PRM 93-04. Department of Psychometrics and Research Methodology, Leiden University, Netherlands.

Byrne, B. (2012). *Structural equation modeling with Mplus*. NY: Routledge.

Chou, C.P. & Bentler, P. (1995). Estimates and tests in structural equation modeling. In R. Hoyle (Ed.). *Structural equation modeling: Concepts, issues, and* applications (pp. 37-55). Newbury Park, CA: Sage

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd *Edition.* Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1976). *Research in classrooms and schools: formulation of questions, designs and analysis.* Occasional Paper, Stanford Evaluation Consortium, Palo Alto, CA.

Cronbach, L. J. & Webb, N. (1975). Between and within class effects in a reported aptitude-by-treatment interaction: Reanalysis of a study by G.L. Anderson, *Journal of Educational Psychology*, *6*, 717-724.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*, 529-569.

de Leeuw, J. & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*(1), 57-85.

de Leeuw, J. & Kreft, I. G. (1995). Questioning multilevel models. *Journal of Educational Statistics*, *20*(2), 171-189.

Dempster, A., Laird N., & Rubin, D. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *series B, 30*, 1-38.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational Statistics*, *20*(2), 115-148.

Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 74*, 311-319.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 121-141.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*, 399-433.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical models.* New York: Halsted.

Goldstein, H. & McDonald, R. (1988). A general model for the analysis of multilevel data.

Psychometrika, 53, 455-467.

Gustafsson, J. E. and Stahl, P. A. (1996). *STREAMS User's Guide*. *Version 1.6 for Windows*. Mölndal, Sweden: Multivariate Ware.

Heck, R. H. & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd Edition). New York: Routledge.

Hartley, H. O., & Rao, J. N. (1967). Maximum likelihood estimation fro the mixed analysis of variance model. *Biometrika, 54*, 93-108.

Hill, P. & Rowe, K. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7*(1), 1-34.

Hox, J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J. Oud & R. van Blokland-Vogelesang (Eds.) *Advances in longitudinal and multivariate analysis in the behavioural sciences*. Nijmegen, NL: ITS.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: T.T. Publikaties

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd Ed.). New York: Routledge.

Hox, J. J. & Maas, C.J.M. (2001). The accuracy of multilevel structural equation models with unbalanced groups and small samples. *Structural Equation Modeling, 8*, 157-174.

Hoyle, R. & Panter, A. (1995). Writing about structural equation models. In R. Hoyle (Ed.). *Structural equation modeling: Concepts, issues, and applications* (pp. 158-176). Newbury Park, CA: Sage.

Jöreskog, K.G. (1977). Structural equation modeling in the social sciences: specification, estimation, and testing. In P.R. Krishnaih (Ed.), *Applications of Statistics*. Amsterdam: North-Holland, 265-287.

Jöreskog, K.G. and Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago: Scientific Software.

Kaplan, D. (1995). Statistical power in SEM. In R. Hoyle (Ed.) *Structural equation modeling: Concepts, issues, and applications* (pp. 100-117). Newbury Park, CA: Sage.

Kaplan, D. (1998). Methods for multilevel data analysis. In G. A. Marcoulides (Ed.) *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum, 337-358.

Kaplan, D. & Elliott, P.R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling*, *4*(1), 1-23.

Kish, L. (1957). Confidence limits for cluster samples. *American Sociological Review*, *22*, 154-165.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kreft, I. & De Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Lindley, D. & Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, *B34*, 1-41.

Little, R. & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and*

structural analysis (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Longford, N. (1993). *Random coefficient models*. Oxford: Clarendon Press.

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis. The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490-504.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* New York: Taylor and Francis.

McArdle, J. & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides and R. Schumacker (Eds.) *Advanced Structural Equation Modeling: Issues and Techniques* (pp. 89-124). Mahwah, New Jersey: Lawrence Erlbaum.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research*, *22*, 399-413.

McDonald, R. P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*. *42*, 215-232.

Mehta, P.D. & Neale, M. C. (2005). People are variables too: Multilevel structural equations models. *Psychological Methods, 10*(3), 259-284.

Mok, M. (1995). *Sample size requirements for 2-level designs in educational research.* Macquarie University, Sydney, Australia.

Morris, C. (1995). Hierarchical models for educational data: An overview. *Journal of Educational Statistics*, *20*(2), 190-200.

Muthén, B. O. (1988). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Muthén, B.O. (1989). Latent variable modeling in heterogenous populations. *Psychometrika*, *54*, 557-585.

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Los Angeles: UCLA Statistics Series #62.

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

Muthén, B.O. (1992). *Latent variable modeling of growth with missing data and multilevel*. Paper presented at the Seventh International Conference on Multivariate analysis, Barcelona, Spain, September.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22(3), 376-398.

Muthén, B. O. (1997). Latent variable modeling with longitudinal and multilevel data. In Raftery (Ed.) *Sociological Methodology.* Boston: Blackwell Publishers, 453-480.

Muthén, L. (1998). Personal communication, December 17, 1998.

Muthén, B. O. & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel modeling in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15-40). New York: Taylor & Francis.

Muthén, B. O., Asparouhov, T., Hunter, A. & Leuchter, A. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STARD antidepressant trial. *Psychological Methods, 16*, 17-33.

Muthén, B. O., Jo, B., & Brown, H. (2003). Comment on the Barnard, Frangakis, Hill, & Rubin article. Principal stratification approach to broken randomized experiments: A case study of school

choice vouchers in New York City. *Journal of the American Statistical Association, 98,* 311-314.

Muthén, B. O. & Muthén, L. (1998) *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Muthén, B.O. & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.) *Multilevel analysis of educational data*. San Diego: Academic Press, 87-99.

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological methodology 1995* (pp. 267-316). Washington, DC: American Sociological Association.

Pawitan, Y. (2001). *In all likelihood: Statistical modeling and inference using likelihood.* Oxford: Clarendon Press.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear model: A Review. *Journal of Educational Statistics*, *13* (2), 85-116.

Raudenbush, S. W. (1995). Reexamining, reaffirming, and improving application of hierarchical models. *Journal of Educational Statistics*, *20*(2), 210-220.

Raudenbush, S. W. & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models* (*2nd Ed*.). Newbury Park, CA: Sage.

Raudenbush, S. & Sampson, R. (1999). Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods & Research, 28*(2), 123-153.

Reynolds, D. & Packer, A. (1992). School effectiveness and school improvement in the 1990s. In D. Reynolds & P. Cuttance (Eds.), *School effectiveness: Research, policy, and practice.* London: Cassell.

Rigdon, E. (1998). Structural equation models. In G. Marcoulides (Ed.) *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum, 251-294.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Sociological Review, 15*, 351-357.

Rubin, H. (1950). Note on random coefficients. In T. C. Koopmans (Ed.), *Statistical inference in dynamic economic models*. New York: Wiley.

Saris, W. E. & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen & J. S. Long (Eds.) *Testing structural equation models*. Newbury Park, CA: Sage, 181-204.

Satorra, A. & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-90.

Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model*. Unpublished doctoral dissertation, University of Chicago.

Schmidt, W. & Wisenbaker, J. (1986). *Hierarchical data analysis: An approach based on structural equations* (Technical Report No. 4). East Lansing, MI: Department of Counseling Psychology and Special Education.

Shigemasu, K. (1976). Development and validation of a simplified m-group regression model. *Journal of Educational Statistics, 1*(2), 157-180.

Smith, A. F. (1973). A general Bayesian linear model. *Journal of the Royal Statistical Society, Series B, 35*, 61-75.

Strenio, J. L. (1981). *Empirical Bayes estimation for a hierarchical linear model*. Unpublished doctoral dissertation, Department of Statistics, Harvard

University.

Wald, A. (1947). A note on regression analysis. *Annals of Mathematical Statistics, 18*, 586-589.

Walsh, J. E. (1947). Concerning the effect of the intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, *18*, 88-96.

Willett, J. & Sayer, A. (1996).Cross-domain analysis of change overtime: combining growth modeling and covariance structure analysis. In G. Marcoulides and R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 125-158) Mahwah, NJ: Lawrence Erlbaum Associates.

Wong, G. T. & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*(391), 513-524.

Appendix A

```
TITLE:      Explaining variation in a
            level-2 intercept and slope;
DATA:       FILE IS C:\mplus\slope.dat;
            Format is f8.0,f8.2,4f8.0;
VARIABLE:   Names are orgid jobsat product
            female orgvar1 orgvar2;
            Usevariables are orgid jobsat
            product female orgvar2;
            Between = orgvar2;
            Within = female jobsat;
            cluster = orgid;
  Define:   Center jobsat (grand);
ANALYSIS:   TYPE = Twolevel random;
  Model:    %Between%
            product S on orgvar2;
            S with product;
            %Within%
            product on female;
            S | product on jobsat;
OUTPUT:     SAMPSTAT TECH1;
```

Appendix B: Tech 1 Output for CFA Model

```
Within Groups
LAMBDA
              GOV           EVAL          BGOV          BEVAL
              ___           ____          ____          _____
SHDEC           1              0             0              0
INVCLI          2              0             0              0
TEAM            0              0             0              0
EVALSTAN        0              3             0              0
EVALPROG        0              4             0              0
SYASSE          0              0             0              0

THETA
              SHDEC    INVCLI    TEAM   EVALSTAN  EVALPROG    SYASSE
              _____    ______    ____   ________  ________    ______
SHDEC             5
INVCLI            0         6
TEAM              0         0        7
EVALSTAN          0         0        0         8
EVALPROG          0         0        0         0         9
SYASSE            0         0        0         0         0         10

PSI
              GOV           EVAL          BGOV          BEVAL
              ___           ____          ____          _____
GOV            11
EVAL           12            13
BGOV            0             0             0
BEVAL           0             0             0              0


Between Groups

NU (item intercepts)

SHDEC     INVCLI     TEAM     EVALSTAN   EVALPROG   SYASSE
_____     ______     ____     ________   ________   ______
   14         15        16         17         18         19

LAMBDA
              GOV           EVAL          BGOV          BEVAL
              ___           ____          ____          _____
SHDEC           0              0            20              0
INVCLI          0              0            21              0
TEAM            0              0             0              0
EVALSTAN        0              0             0             22
EVALPROG        0              0             0             23
SYASSE          0              0             0              0

THETA
              SHDEC    INVCLI     TEAM   EVALSTAN  EVALPROG    SYASSE
              _____    ______     ____   ________  ________    ______
SHDEC            24
INVCLI            0        25
TEAM              0         0         0
EVALSTAN          0         0         0        26
EVALPROG          0         0         0         0        27
SYASSE            0         0         0         0         0         28

PSI
              GOV           EVAL          BGOV          BEVAL
              ___           ____          ____          _____
GOV             0
EVAL            0             0
BGOV            0             0            29
BEVAL           0             0            30             31
```

[i] We can consider the *design effect*, which is a function of both the ICC and the average cluster size (Muthén & Satorra, 1995). The design effect quantifies the extent to which sampling error present in a sampling design departs from the sampling error that would be expected under simple random sampling. The design effect is approximately equal to [1+ (average cluster size – 1*ρ)], where ρ is the intraclass correlation. As Muthén and Satorra conclude, design effects of less than 2.0 do not appear to result in overly-exaggerated rejection proportions at $p = .05$ for using single-level analyses. For example, with average cluster size of 30 and ρ = 0.04, the design effect would be about 2.16. For the same average cluster size and a ρ = 0.03, the estimate would be about 1.87.