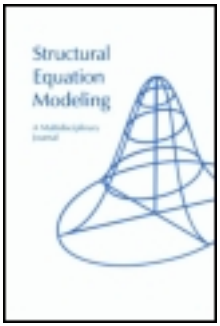

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons

Gitta H. Lubke & Bengt O. Muthén

Available online: 19 Nov 2009

To cite this article: Gitta H. Lubke & Bengt O. Muthén (2004): Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons, Structural Equation Modeling: A Multidisciplinary Journal, 11:4, 514-534

To link to this article: http://dx.doi.org/10.1207/s15328007sem1104_2

STRUCTURAL EQUATION MODELING, *11*(4), 514–534

Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons

Gitta H. Lubke
Virginia Commonwealth University

Bengt O. Muthén
University of California, Los Angeles

Treating Likert rating scale data as continuous outcomes in confirmatory factor analysis violates the assumption of multivariate normality. Given certain requirements pertaining to the number of categories, skewness, size of the factor loadings, and so forth, it seems nevertheless possible to recover true parameter values *if* the data stem from a single homogeneous population. It is shown that, in a multigroup context, an analysis of Likert data under the assumption of multivariate normality may distort the factor structure differently across groups. In that case, investigations of measurement invariance (MI), which are necessary for meaningful group comparisons, are problematic. Analyzing subscale scores computed from Likert items does not seem to solve the problem.

Questionnaires designed to measure latent variables such as personality factors or attitudes typically use Likert scales as a response format for the individual items. In response to statements such as "Does the student yell at others?" participants are asked to choose one of a given number of ordered response categories running for instance from *almost never* to *almost always.* Data arising from Likert-type items are often analyzed as multivariate normal outcomes, although the data are ordered categorical (Muthén & Kaplan, 1985). This article focuses on multigroup confir-

Requests for reprints should be sent to Gitta H. Lubke, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University USA, East Leigh Street, Suite 1–115, Richmond VA 23219–1534. E-mail: glubke@vcu.edu.

matory factor analysis of ordered categorical outcomes while incorrectly assuming multivariate normality of the data. Results from robustness studies in a single homogenous population concerning the analysis of Likert data while violating the normality assumption do not necessarily carry over to the multiple group situation, and group comparisons may have problems in addition to those encountered in single populations.

The multigroup confirmatory factor model is suitable to compare groups with respect to the latent variables underlying the individual items (Sörbom, 1974). In the analysis of longitudinal data a special type of factor model called growth curve model can be used to compare groups with respect to average latent growth trajectories (McArdle & Epstein, 1987; Muthén, 2001b). Although multigroup confirmatory factor models exist both for continuous outcomes and for ordered categorical outcomes, in practice the model for continuous outcomes is often applied to categorical outcomes. The commonly used maximum likelihood estimation of the model for continuous outcomes is based on the assumption of multivariate normality of the observed data, which is violated in case the outcomes are Likert scale data. The violation of the multivariate normality assumption may have especially serious consequences if groups are to be compared with respect to the factors a given test or questionnaire is designed to measure. To render these group comparisons meaningful, it is necessary to investigate MI (Meredith, 1993).

A test is measurement invariant if test takers belonging to different groups, who have the same score on the factor underlying the test, have on average the same score on an observed item. In other words, the distribution of observed scores conditional on the factor scores is the same for all groups (Mellenbergh, 1989). In the context of the factor model, this implies that the regression relations between observed items and underlying factors as specified by a given multigroup factor model have to be the same across groups (Meredith, 1993). In case of multivariate normal data, the parameters of interest are regression intercepts, factor loadings, and residual variances. If these parameters are known to be invariant across groups, the distribution of observed scores conditional on the factor scores is the same across groups, and groups may differ only with respect to the means and covariances of the factors. In case of ordered categorical data, the set of parameters required to be group invariant is different.

There are two approaches to model ordered categorical outcomes. One approach is an extension of the factor model for continuous data. The continuous outcomes are now assumed to be unobserved (e.g., latent response variables). A response category is chosen above a lower category if the latent response variable exceeds a certain threshold (Agresti, 1990; Jöreskog & Moustaki, 2001; Muthén, 1984). The thresholds are usually not known, that is, they are latent. In the other approach, the probability of choosing a certain response category conditional on the factor score is modeled directly (Agresti, 1990; Jöreskog & Moustaki, 2001; Muthén & Asparouhov, 2002). Independent of which of the two approaches is used

to model the relation of observed ordered categorical data to the underlying factors, the set of parameters that has to be group invariant for MI to hold includes the latent thresholds. Hence, the requirements for MI are different for multivariate normal data and ordered categorical outcomes.

In this study, it is investigated by means of simulated data whether tests of MI across groups (Dolan, 2000; Lubke, Dolan, Kelderman, & Mellenbergh, 2002) lead to meaningful results if the ordinal character of the data is ignored. When analyzing categorical data with models for continuous outcomes, the thresholds between categories are not estimated. Threshold differences between groups indicate that groups use a given Likert scale in a group-specific way and are a violation of MI (Millsap & Tein, 2003), whereas threshold differences between the observed indicators of a factor do not violate MI. A multigroup factor model for continuous outcomes restricted to represent MI may be rejected because of several different causes. First, the MI model may not fit adequately because threshold differences between groups are mistaken as structural differences between groups. This would lead to the correct conclusion that MI is violated. Second, the MI model may be rejected because threshold differences between observed indicators can lead to a distorted factor structure or because indexes of goodness of fit based on the assumption of normally distributed data do not work properly. The latter two cases would lead a researcher to believe that MI is violated when in fact it is not.

A substantial number of studies have focused on the robustness of factor analysis models with respect to nonnormality induced by ordered categorical outcomes (e.g., Bernstein & Teng, 1989; Dolan, 1994; Hoogland, 1999; Olsson, 1978, 1979). On the one hand, the studies show that small numbers of response categories, different thresholds across items, skewness, and high reliability of the items can all lead to distorted results. The distortion may result in the need for additional factors (e.g., difficulty factors; Bernstein & Teng, 1989; Carroll, 1945; Gorsuch, 1983), biased estimates of factor loadings, and inflated chi-square test statistics (Dolan, 1994). On the other hand, it has also been shown that given a sufficiently large number of response categories (at least seven), absence of skewness, and equal thresholds across items, it seems possible to obtain reasonable results (Dolan, 1994; Olsson, 1979). However, all these results concern data arising from a single, homogenous population and are, therefore, limited to the analysis of covariances of correlations. The results do not necessarily carry over to data arising from multiple groups. First, when comparing groups, the model of interest usually comprises a model for the means in addition to a model for the covariances. The two parts of the model are estimated simultaneously. Estimating regression intercepts and factor mean differences between groups may reveal additional distortions. Second, if data arise from a single homogenous population all possible response categories will be observed in a reasonably large sample. However, in data arising from a heterogenous population, not all response categories may be observable within all groups. If the groups are well separated with respect to their means, only the lower

categories may be observed in the group with the lower mean values. Hence, rules of the thumb with respect to the number of response categories required to obtain reasonable results may depend on the separation of the groups or classes. Third, group-specific thresholds can result in item distributions that differ across groups with respect to skewness even if the underlying factor is normally distributed in all groups.

In sum, the question arises to what extent groups can be compared in a meaningful way when assuming multivariate normality for Likert scale data. Distorted factor structures across items, or inadequate functioning of fit indexes based on multivariate normality of the data may lead a researcher to conclude that MI is absent when in fact it is not. Given the frequency with which Likert data are in practice analyzed with factor models for continuous outcomes, it is of interest to investigate the extent to which this practice leads to incorrect conclusions.

The multigroup models considered in this article are a single factor model and a linear growth model. The interest is to investigate through a simulation study whether MI is falsely rejected if the categorical character of the observed data is ignored. The simulation mirrors an empirical situation in which a researcher analyses categorical data with multigroup models for continuous outcomes and relies on measures of goodness of fit and parameter estimates obtained through normal theory maximum likelihood estimation. As mentioned previously, there are two approaches to generate categorical data. Although basic models corresponding to the two approaches are equivalent in terms of relating the observed scores to the underlying factors (Agresti, 1990; Muthén & Asparouhov, 2002), for our purpose the more convenient model is the extension of the factor model for continuous data in which an observed categorical response is related to the underlying factor(s) via a latent continuous response variable. We generate multivariate normal outcomes using a measurement invariant multigroup model and then categorize these data in different ways. The resulting categorized data are analyzed with measurement invariant multigroup models, and rejection rates, coverage of parameters of interest, and reliability of the observed indicators are computed.

We focus on the following issues. First, the number of response categories is investigated. Data are categorized with 5, 7, and 10 response categories. Second, the effect of the separation of groups is investigated. As mentioned previously, a larger separation may have a detrimental effect. Third, we vary the reliability of observed variables. On the one hand, high reliability is a positive quality; on the other hand, it might result in an increased power to reject a measurement invariant model because of the violation of the normality assumption. Fourth, we investigate the effect of threshold inequality across items. Threshold equality across items may be rather unlikely in practice. In the context of growth models, threshold differences across indicators are in fact threshold differences across time. The fifth and final issue is threshold inequality across groups.

Considering that analyzing Likert scale data with models for continuous data may be problematic, one might be tempted to compute sum or average scores for small subsets of items, and one might factor analyze the resulting, more continuous looking subscale scores instead of the individual item scores. Using the growth model as a data-generating model, it is also investigated whether averaging over different numbers of Likert items leads to improved results.

MODELS

The models used in this study have been broadly covered in the literature; therefore, their presentation in this article is brief. We first describe the confirmatory factor model for continuous outcomes in a single, homogenous population. The single population model is then extended to be applied to a heterogenous population consisting of a finite number of groups. A factor model is specified within each of the groups (Sörbom, 1974). Next, the linear growth model is described, which is a special case of the confirmatory factor model (Muthén, 2001a). The presentation of the models for continuous outcomes is followed by an outline of a factor model for ordered categorical outcomes (Jöreskog & Moustaki, 2001; Muthén & Muthén, 2001). Some attention is directed to the thresholds to illustrate their importance for the distribution of observed categorical scores. The section concludes with an explanation of the concept of MI in the context of the confirmatory factor model (Lubke et al., 2002; Meredith, 1993).

Confirmatory Factor Model for Continuous Outcomes: Single Population

The confirmatory factor model is a linear regression model in which a number of observed indicators are regressed on a smaller number of underlying latent variables called factors. When applied to a single population, we assume without loss of generality that all factor means are zero. Say, we have I participants, J observed indicators, and L factors, then the score of participant i on item j can be denoted as follows:

$$y_{ij} = \sum_{l=1}^{L} \lambda_{jl}\ \eta_{il} + \varepsilon_{ij}, \tag{1}$$

where y_{ij} is the score of participant i on indicator Y_j, λ_{jl} is the regression slope of the regression of Y_j on the factor score η_l and is called factor loading, and ε_{ij} is the residual score of participant i on indicator Y_j. The covariance matrix of the indicators, Σ, can be expressed as

$$\Sigma = \Lambda \Psi \Lambda' + \Theta \tag{2}$$

where Λ is the matrix of factor loadings, Ψ is the covariance matrix of factor scores, and Θ is the covariance matrix of the residual scores. Basically, the scores on the observed indicators are broken down into factor scores multiplied with the factor loadings and residual scores. Factor scores and residual scores are assumed to be uncorrelated, and the residual scores of different indicators are assumed to be uncorrelated. The factor loadings and factor covariance matrix account for the common content of the observed variables. The elements of the three matrices Λ, Ψ, and Θ are the parameters of the model and are in practice estimated from the covariance matrix of observed variables Σ. Methods to ensure that model parameters are uniquely defined are discussed in Bollen (1989).

Confirmatory Factor Model for Continuous Outcomes: Multiples Groups

The multigroup confirmatory factor model comprises a model for the means in addition to the model for the covariances described previously. For each of the groups, these two parts of the model are specified. The two parts of the model are then estimated simultaneously for all groups.

The observed score of participant i on item j where participant i is member of group g, $g = 1,\ldots,G$, is

$$y_{ijg} = \nu_{jg} + \sum_{l=1}^{L} \lambda_{jlg}\ \eta_{ilg} + \varepsilon_{ijg} \tag{3}$$

The model for the means and the covariances for group g, can be represented as

$$\mu_g = \nu_g + \Lambda_g \alpha_g \tag{4}$$

$$\Sigma_g = \Lambda_g \Psi_g \Lambda_g' + \Theta_g \tag{5}$$

In the single population model, the model parameters contained in the matrices Λ, Ψ, and Θ were used to impose a structure on the observed covariance matrix. In addition to these parameters, we now have regression intercepts ν and the factor means α_g to impose a structure on the observed means μ. The regression intercepts have to be equal across groups (i.e., $\nu_g = \nu$) . Furthermore, the factor means are fixed to zero in one of the groups. Here, we arbitrarily choose the first group such that $\mu_1 = \nu$. As a consequence, the elements of the vector α_g represent factor mean differences with respect to the first group.

The simultaneous estimation of the mean and the covariance model across groups offers the possibility to (a) restrict parameters to be equal across groups and (b) compare groups with respect to their factor means. Both are key issues in the context of MI (see later).

Linear Growth Curve Model

The linear growth curve model is a special case of a two-factor model. Consider a model with a single continuous indicator at each time point. The linear growth model can be specified as a single population model (G = 1), with the understanding that the observed outcomes, Y in Equation 1, now correspond to one indicator measured at different time points, and that the observed covariances, Σ in Equation 3, are the (co)variances of this indicator across time points. The linear growth curve model can be specified as a multigroup model using Equations 3 through 5.

The idea is to determine the average linear growth over time of each of the groups or classes to which the model is applied. Each participant within a group is assumed to follow an individual linear growth curve with a participant-specific intercept and slope. In the linear growth curve model, the intercept and the slope are represented by two factors. The factor scores on these two factors are the participant-specific intercept and slope and the means of the factors represent the average intercept and slope within a group. Using the linear growth model one can, therefore, estimate the average growth trajectories within group. In the simulation study presented later, we use the simplest case of the linear growth model in which all factor loadings are fixed. The loadings of the indicator on the intercept factor are fixed to one for all time points. The loadings on the slope factor represent the distances between time points and are fixed to 0, 1,…, $T - 1$ for T equally spaced time points. By fixing the loading on the slope factor to zero at the first time point, the intercept factor represents the participants' initial status. Let a and b indicate the intercept and slope factor, respectively. The linear growth model with a single indicator measured at four time points can be denoted for participant i in group g as

$$y_{1i} = a_i + \varepsilon_{1i} \tag{6}$$

$$y_{2i} = a_i + 1\ b_i + \varepsilon_{2i} \tag{7}$$

$$y_{3i} = a_i + 2\ b_i + \varepsilon_{3i} \tag{8}$$

$$y_{4i} = a_i + 3\ b_i + \varepsilon_{4i} \tag{9}$$

The model for the covariances and the means of the linear growth curve model are similar to Equations 4 and 5:

$$\mu_g = \Lambda\alpha_g \tag{10}$$

$$\Sigma_g = \Lambda\Psi_g\Lambda' + \Theta \tag{11}$$

with the difference that here the regression intercepts, ν in Equation 4, are here fixed to zero for all participants such that we can estimate the means of both factors in all groups. We have limited our description to the case of a single continuous in-

dicator, which is measured at each time point, although the model can be easily extended to more than one indicator. To accommodate for multiple time point indicators, a factor model is specified at each time point. Denote the factors of the model at a given time point as ξ. The difference with a single indicator per time point is that now, the factor scores ξ are regressed on *a* and *b* instead of regressing the observed single indicator on *a* and *b*. Hence, average growth curves of ξ are modeled. For an introduction to this specification of the growth model, the reader is referred to Muthén (2001b).

Ordered Categorical Outcomes With Underlying Confirmatory Factor Model

There is a variety of latent variable models for ordered categorical variables (Agresti, 1990; Jöreskog & Moustaki, 2001). Two different ways to derive latent variable models for ordered categorical variables can be distinguished. First, the conditional probability of choosing a response category given the score on the latent variable is modeled directly. A variety of item response models for polytomous outcomes fall into this category. The second approach is an extension of the confirmatory factor models for continuous outcomes, which we described in the previous section. The continuous outcomes are not observed and are called the latent response variable. A respondent chooses a response category above a lower category if the latent response variable exceeds a threshold. The two approaches emphasize different sets of parameters. Equivalences between parameters of a simple model derived via the first approach and the latent response variable model are shown in Agresti (1990) and Muthén and Asparouhov (2002). The interest of our study is in showing the effects of fitting multigroup models for continuous outcomes to categorical data. The latent response variable model is the most convenient choice, because we can generate multivariate normally distributed latent response variables with a multigroup factor model that represents MI (see later) and categorize these data in a second step. If fitting models for continuous outcomes to categorical data is unproblematic, then fitting the model that generated the latent response variables to the categorical outcomes would result in an acceptable model fit, and conclusions with respect to tenability of MI would be correct. Deviations from such results can demonstrate the problems at hand.

We indicate the observed categorized outcome as *Y* and the latent response variable as *Y**. The model can be conceptualized as follows. Suppose test takers respond to an item of an attitude test. The attitude (i.e., the factor scores) is assumed to be measured on a continuous scale. The same holds for the unobserved responses to the item *Y**. Due to the response format of the item, a test taker has to choose one of several possible response categories. Hence, the observed outcome

Y is ordered categorical. More formally, the model for participant *i* on item *j* can be represented as follows:

$$y_{ij}^{*} = \sum_{l=1}^{L} \lambda_{jl}\ \eta_{il} + \varepsilon_{ij} \tag{12}$$

$$y_{ij} = c \qquad \text{if} \quad \tau_{c-1} < y_{ij}^{*} \leq \tau_{c}. \tag{13}$$

The first part of the model is equal to Equation 1 with the only difference that the continuous outcome variable denoted as *Y** is not observed. The thresholds τ partition the range of *Y** into *C* categories, where $C = 1, \ldots, c$. The thresholds are therefore ordered, $\tau_0 < \tau_1 < \ldots, \tau_c$, where τ_0 and τ_c equal $-\infty$ and $+\infty$, respectively. The lowest category lies between $-\infty$ and τ_1, and the observed outcome *Y* falls in that category, if *Y** is smaller than τ_1.

In case of a single homogenous population, the distribution of the unobserved continuous outcome *Y** may be reasonably well reproduced by the categorical *Y,* if thresholds partition the range of *Y** into equidistant categories and if the number of categories is adequate (e.g., ≥7; see Dolan, 1994). This situation is depicted in the upper panel of Figure 1. The situation may be different in case the range of *Y** is not partitioned into intervals of equal length. The situation can also be different in case of several groups. The observed frequencies of the higher response categories may approach zero in the group with the lowest mean and vice versa. In other words, the number of *observed response categories within each group* may be smaller than the number of possible response categories of a given item (see shaded areas in the lower two panels of Figure 1 that represent the frequencies in the highest scoring group). The number of observed response categories within groups tends to be smaller with increasing separation of subpopulations (compare the middle and lower panel of Figure 1). In addition, skewness of *Y* within a group depends on the localization of the distribution of *Y** with respect to the thresholds.

In the lower two panels of Figure 1, the thresholds not only are equidistant but also are identical for all three subpopulations. Suppose that thresholds are not equidistant, that thresholds are not equal across groups or across items (i.e., time points in case of the growth model), or both. These conditions may all lead to a distortion of the estimated factor structure that relates *Y** to the factor scores η if *Y* is analyzed with models for continuous outcomes that neglects the thresholds.

Such a distortion of the factor structure may be especially critical if a study aims at comparing groups or classes with respect to the factors rather than with respect to the observed variables. These comparisons are only valid if it can be shown that the observed variables are measurement invariant, meaning that observed variables measure the same factors across groups or classes. Tests of MI may indicate a vio-

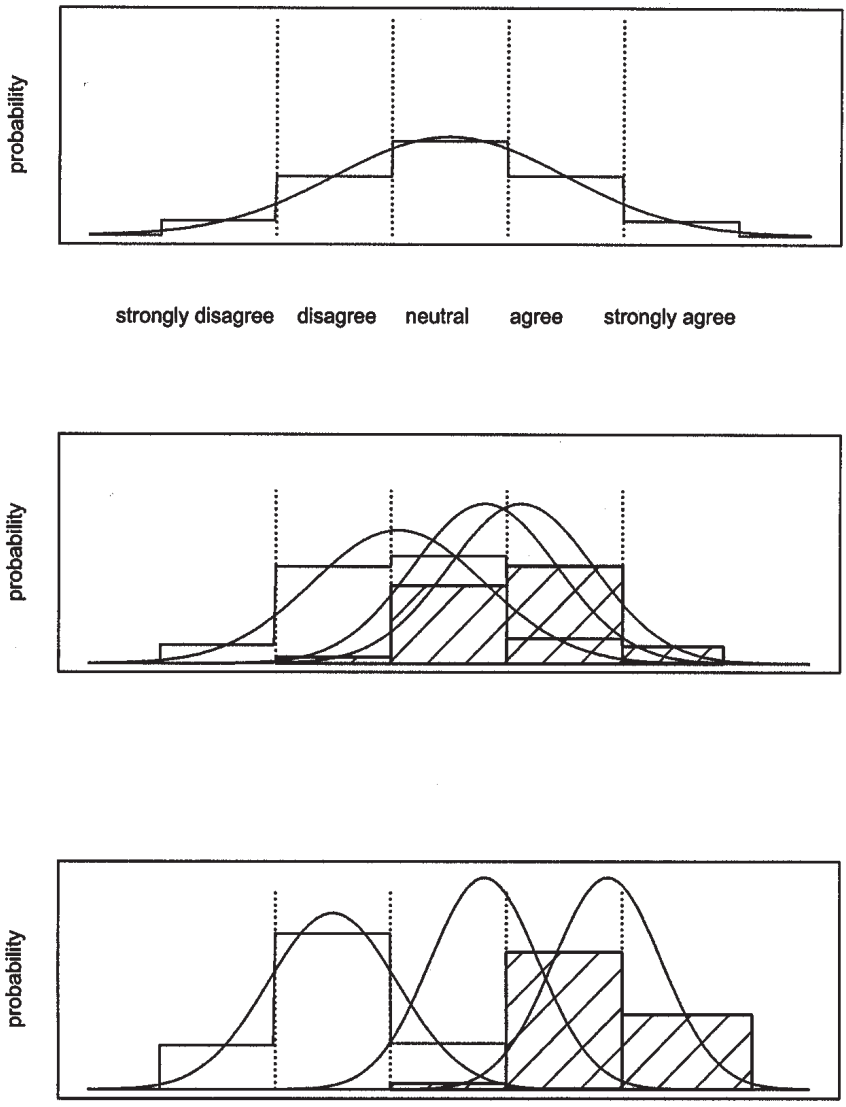

FIGURE 1 Single versus multiple subpopulations and separation.

lation of MI *even if the unobserved continuous outcomes Y* are measurement invariant across groups* because the factor structure in the observed ordered categorical data *Y* is distorted in a subpopulation specific way.

Measurement Invariance (MI)

The theoretical foundation of MI in the context of the confirmatory factor model is mainly due to Meredith (1964, 1993). It has been defined in a more general context by Mellenbergh (1989) as

$$f(\boldsymbol{Y} \mid \eta, s) = f(\boldsymbol{Y} \mid \eta) \tag{14}$$

Equation 14 shows that, given the factor score, the probability distribution of observed scores does not depend on subpopulation membership. Suppose a factor represents a certain attitude. MI means that, given the level of attitude, group or class membership has no additional effect on the observed scores of the attitude test. Groups or classes may differ only with respect to the means and (co)variances of the attitude factor(s).

MI is a hypothesis that can be tested using the multigroup confirmatory factor model (Meredith, 1993; for an application see Dolan, 2000). Meredith has shown that a factor model with intercepts, factor loadings, and residual variances restricted to be equal across groups or classes represents MI.[1] One strategy to test MI is to compare the fit of the fully restricted model (e.g., a model with intercepts, loadings, and residual variances held equal across groups) to the fit of a less restricted model in a likelihood ratio test. A significant increase in model fit when relaxing the MI restrictions is an indication that the questionnaire is not invariant across groups. In this way, testing MI may provide useful information concerning which of the observed questionnaire items is violating MI.

Another possibility is to fit the restricted model and evaluate measures of goodness of fit. In the simulation study, we use the second option.

SIMULATION STUDY

The general approach in the simulation study is to generate multivariate normal latent response variables *Y** under a measurement invariant multigroup confirmatory factor model, categorize the data, and fit the model that generated the latent response variables to the ordinal categorical outcomes. The measurement invariant model, although true for the continuous latent response variables, might be rejected when fitted to the categorical data because the factor structure may be dis-

[1]For some rather far-fetched exceptions see Meredith (1993) or Lubke et al. (2002).

torted differentially across groups or classes. We evaluate under which conditions MI would be falsely rejected if the decision is based on commonly used measures of goodness of fit. The measures of goodness of fit examined in this study are based on the assumption of multivariate normally distributed data and hence not adequate for categorical outcomes. However, researchers who treat categorical outcomes as if they were normally distributed would rely on these fit indexes. Our aim is to show under which conditions reliance on these fit measures leads to incorrect conclusions if the normality assumption is violated.

We use only a small number of different data-generating models. The aim of the simulation is not to quantify the effects of violating the normality assumption for different models. Instead, we aim at illustrating that there is a large number of characteristics of the data that may result in a distortion of the factor structure if factor models for continuous data are fitted to ordered categorical outcomes. By including the data characteristics as design factors in the simulation, we can show that certain combinations may be detrimental to a meaningful investigation of MI. We investigate three different ways of categorization (threshold schemes; see later), different numbers of response categories (5-, 7-, and 10-point items), two levels of reliability of observed data, two levels of group separation with respect to their factor means, and different total sample sizes ($N = 300$, $N = 500$, $N = 1{,}000$, and $N = 3{,}000$). Measurement invariant single factor models and linear growth models are used to generate the continuous Y^* data. The continuous data are categorized using different threshold schemes. The models that generated the continuous data are fitted to the categorized data. We also fit models for categorical outcomes to provide an indication of the results of fitting the correct model for categorical data. This is done only for the categorized linear growth data with time-invariant and time-varying thresholds due to current software limitations.[2] All simulations are carried out using the Monte Carlo feature in M*plus* 2.1 (Muthén & Muthén, 2002) if not mentioned otherwise. We report observed rejection rates at an expected 0.05 level and the coverage of parameters of interest.

Although we have carried out all meaningful combinations of the different levels of the design factors of our simulation study, the results of only some of the combinations are presented later in an attempt to avoid redundancy. The design factors and their levels are presented in Table 1. More specific explications are given in the next two paragraphs. The remainder of the section is divided in three parts covering the single factor model, the linear growth multigroup model, and the

[2]The current version of M*plus* requires all response categories to be observed in all groups. In the part of the simulation involving the single factor model, our aim is to show the effects of separation of groups, which means that low-response categories are not necessarily observed in a high-scoring group and vice versa. In addition, it is currently not possible to generate data with group-varying thresholds.

TABLE 1
Simulation Design: Manipulations of Data Characteristics and Type of Fitted Model

Data Characteristics	*Fitted Models*
Reliability: moderate (.55), good (.85)	SFMcont
Separation: moderate (1 SD), large (2 SD)	SFMcont
Threshold scheme: IT/IS, VT/IS, IT/VS	SFMcont (IT/IS, VT/IS)
	LGMcont (IT/IS,VT/IS, IT/VS)
	LGMcat (IT/IS, VT/IS)
Number of categories: 5, 7, 10	LGMcont and LGMcat
Total sample size: 300, 500, 1,000, 3,000	SFMcont

Note. SFM = single factor model for continuous data; LGMcont = linear growth model for continuous data; LGMcat = linear growth model for categorical data; IT/IS = invariant time or item/invariant subpopulation; VT/IS = variant time or item/invariant subpopulation; IT/VS = invariant time/variant subpopulation.

analysis of subscales created by averaging across multiple categorical items at each time point.

For each data type, 100 replications are generated that differ with respect to the factor scores and residual scores. Multivariate normal data are generated for three subpopulations throughout. Subpopulation sample sizes for total $N = 300$ are 100 for all three groups; for total $N = 500$ they are 100, 200, and 200; for total $N = 1{,}000$ they are 300, 300, and 400; and for $N = 3000$ they are again equal. All models for the continuous data are measurement invariant because the three subpopulations differ only with respect to the factor mean(s) and variances. The multivariate normal data are then categorized using three different schemes: (a) invariant indicator (or time point in the growth model) and invariant subpopulation thresholds (IT/IS), (b) variant indicator (time point) and invariant subpopulation thresholds (VT/IS), and (c) invariant indicator (time point) and variant subpopulation thresholds (IT/VS). Time-varying thresholds means that the observed variable is not measurement invariant over time. Group-varying thresholds represent a violation of MI across groups because the distribution of observed (ordered categorical) variables given the factor scores depends also on the thresholds and is therefore not the same for all subpopulations.

The different categorization schemes are carried out as follows. The continuous data generated by the single factor and the linear growth model, respectively, are categorized into 5-, 7-, and 10-point scale data. To obtain invariant indicator thresholds (or time invariant thresholds in case of the growth model), the average range of the continuous indicators taken over the whole sample (i.e., not separately per subpopulation) is divided into intervals of equal length. The number of intervals equals the number of response categories. The thresholds that separate the intervals are held constant across indicators (time points) when allocating a continu-

ous outcome to one of the categories using Equation 7. To obtain subpopulation invariant thresholds, the intervals and thresholds are held constant across subpopulations. Indicator-varying thresholds in the single factor model are obtained by shifting the thresholds linearly to the left for three of the five items and to the right for the other two (for details, see later). Time-varying threshold for the linear growth model correspond to slightly varying thresholds across the four time points, and group-varying thresholds mirror a low-, a medium-, and a high-scoring group (for details, see later).

Single Factor Model for Multiple Groups

Continuous data are generated for a single factor model with five indicators. The data are analyzed with a multigroup model for continuous outcomes. The single factor model is used to investigate the effects of (a) separation of groups, (b) average reliability of indicators.[3] (c) indicator invariant thresholds versus indicator-varying thresholds, and (d) sample size. Thresholds are group invariant throughout.

The continuous data are categorized in 5, 7, and 10 categories. Separation of groups refers to factor mean differences that equal 0, 1, –1 and 0, 2, –2 in the three groups in the continuous data. Factor variances equal 1 in all groups. Hence, standardized factor mean differences are 1 and 2 *SD*s between groups for the two levels of group separation. Average reliability of the indicators in the continuous data is 0.45 or 0.8. Invariant thresholds are obtained as described previously. The indicator-varying thresholds are obtained by subtracting a constant from the thresholds of two indicators and adding the same constant to the thresholds of the remaining three indicators. The constant equals $\frac{1}{5}$ of the range of the continuous outcomes. Hence, on a 5-point scale, participants would score on average one category higher on the first two indicators. It results in two indicators being skewed to the right and the remaining three being skewed to the left and represents a test with two "easy" items favoring the higher response categories and three more "difficult" items. Fitting a measurement invariant model may result in a rejection because of the necessity to add difficulty factors (Gorsuch, 1983). In this way, the work carried out by Bernstein and Teng (1989) in a single group context is extended to a multiple-group setting. Total sample sizes are 300, 500, 1,000, and 3,000.

Results of the single factor model for multiple groups. The effects investigated with the single factor model were consistent with our expectations (see Table 2). Increasing the separation between groups results in a slightly higher rejection rate when reliability is low; the effect is more pronounced when combined

[3]Here, reliability is understood as the percentage of variance of the indicators that is due to the factor.

TABLE 2
Observed Rejection Rates at an Expected 0.05 Level When Fitting the Single Factor Multi-Group Model for Continuous Outcomes to 5-Point Scale Data

Total N	*Low Rel/Small Sep*	*Low Rel/Large Sep*	*High Rel/Small Sep*	*High Rel/Large Sep*
Indicator invariant thresholds				
300	0.06	0.08	0.33	0.47
500	0.10	0.09	0.26	0.63
1,000	0.09	0.09	0.45	0.85
3,000	0.11	0.18	0.84	1.0
Indicator variant thresholds				
300	0.48	1.0	0.63	1.0
500	0.70	1.0	0.68	1.0
1,000	1.0	1.0	0.98	1.0
3,000	1.0	1.0	1.0	1.0

Note. Sep = factor mean separation between groups; Rel = reliability in the continuous data.

with high reliability. Introducing indicator-specific thresholds results in high rejection rates across the board even for small sample sizes. Results for fitting continuous models to 5-point Likert scale data are acceptable for larger sample sizes only given low reliability and indicator-invariant thresholds. Generally, higher reliability results in increased χ^2, which corroborates the findings of Bernstein and Teng (1989) in a single group setting.

In addition to the rejection rates that are based on the normal theory χ^2 fit statistic, we report the average coverage of parameters of interest. To obtain the coverage of parameters corresponding to the factor loadings and the latent mean differences the scale of the latent variable is fixed to the true value in the fitted model. The coverage of the estimates of the factor mean differences in this part of the simulation is larger than 0.85 even for small sample sizes. This finding does not seem to depend on any of the other design factors. However, the coverage corresponding to the factor loadings is smaller than 0.1. Hence, fitting the model for continuous outcomes seems to affect especially the estimated factor structure and to a much lesser extent the estimates of the latent mean differences.

Linear Growth Model for Multiple Groups

Continuous data are generated for linear growth models with a single indicator at each of four time points. The linear growth curve model is used to investigate the effects of (a) number of scale points, (b) time-varying thresholds, (c) group-varying thresholds, and again (d) total sample size. Average reliability in the

continuous data is held constant at approximately 0.7. The separation of groups is small. The growth curves for Group 1 is flat with zero intercept and slope. Group 2 has a 0.3 intercept and a 0.1 slope, whereas Group 3 has a 0.6 intercept and a negative 0.1 slope. Factor variances are constant across groups and are unity for the intercept factor and 0.25 for the slope factor. Hence, the differences in intercept correspond to standard deviations, and the slope differences have to be divided by 0.25. The covariance of the intercept and slope factor is zero. The number of response categories is 5, 7, and 10. The total sample size is 500, 1,000, or 3,000. Time-varying and group-varying thresholds are not combined in our design, although this may happen in empirical data. For the sake of interpretability of results, we keep group thresholds invariant when thresholds vary across time and vice versa. Time-varying threshold are obtained as follows. Taking the time-invariant thresholds as a starting point, we subtract a constant from all thresholds of the second time point and add the same constant to the thresholds of the third time point. The constant corresponds to 1/18 of the total range of observed continuous scores, which can be regarded as a very small average threshold differences across time. Subpopulation-varying thresholds are obtained by subtracting approximately 1/10 of the range from all thresholds of one of the three subpopulations. Note that group-varying thresholds violate MI across groups.

Results of the linear growth model for multiple groups. Given our choice of parameter values for the data generation, rejection rates when fitting a linear growth model for continuous data to categorical data are close to the expected 0.05 level only for sample size 500 and time-invariant and group-invariant thresholds (see Table 3). As can be expected, increasing sample size leads to increases in rejection rates. An increase in the number of response categories does not lead to lower rejection rates. All types of threshold variation chosen in this study result in unacceptable rejection rates. Note that the MI model should be rejected when fitted to the group- or time-variant threshold data because threshold variation represents a violation of MI across groups or time, respectively. However, when fitting a model for continuous data, the thresholds are not estimated and a researcher would not know whether unacceptable fit is due to group- or to time-varying thresholds. This lack of knowledge complicates group comparisons with respect to their latent growth trajectories.

The coverage of the parameters of the latent mean differences is again good. The factor loadings of the linear growth model are fixed; hence, no coverage is reported. However, the reliability (e.g., ratio of variance explained by the factors by total variance) is underestimated, and the underestimate is worse with smaller numbers of response categories.

Fitting the correct model, that is, a linear growth model for categorical outcomes, to the categorical data leads to rejection rates that do not deviate much from the expected 0.05. The range of observed rejection rates varies between 0.04 and

TABLE 3
Observed Rejection Rates at an Expected 0.05 Level for the Continuous Multi-Group Linear Growth Model When Fitted to 5-point Scale Data

Thresholds	*N = 500*	*N = 1,000*	*N = 3,000*
Number of categories = 5			
Equal	0.07	0.21	0.46
Time	0.92	1.0	1.0
Group	1.0	1.0	1.0
Number of categories = 7			
Equal	0.10	0.24	0.59
Time	0.98	1.0	1.0
Group	1.0	1.0	1.0
Number of categories = 10			
Equal	0.11	0.26	0.64
Time	1.0	1.0	1.0
Group	1.0	1.0	1.0

Note. Time and group refers to time and group varying thresholds. Rejection rates for the categorical linear groth model (LGM) with group varying threshold are not provided because the current version of the Monte Carlo utility does not provide this option.

0.13 across all sample sizes (500, 1,000, and 3,000) and is not affected by time-varying or time-invariant thresholds. The parameter coverage exceeds 95%. Hence, fitting the adequate model for categorical outcomes leads to very satisfying results.

Analyzing Subscale Scores

In an empirical situation, researchers often analyze more continuous looking sum or average scores instead of the individual ordered categorical indicators. Therefore, we construct subscale scores by generating multiple continuous indicators per time point (3, 5, 7, 9, or 15 indicators). The factor model used to generate the multiple indicator scores is a single factor model, which is invariant across time. On a conceptual level, this corresponds to indicators that measure the same single underlying construct at each time point, that is, the continuous items are unidimensional and measurement invariant across time. Next, the scores are categorized into 5-point scales using thresholds that are group invariant but that differ for the first time point as compared to the remaining time points. Finally, the categorized data for multiple indicators are averaged at each time point. This is done for 500 participants. The data are generated using Splus routines and analyzed with a modified RunAll utility.[4]

[4]Please see M*plus* Web page http://www.statmodel.com

Results of analyzing subscale scores. As before, the fitted model is restricted to represent MI. The main finding when analyzing subscale scores created by averaging over multiple indicators is that reliability of the observed average score at each time point increases as more indicators are added to the subscale. The reliability is 0.80 in the case of 3 indicators per subscale and increases to 0.90 in the case of 15 indicators. The increase might have been expected because reliability of the total test score increases with the length of a test because the measurement error cancels out. However, in the context of analyzing ordered categorical outcomes with models for continuous outcomes, increased reliability has a detrimental effect. It results in an increased power to detect the distortions of the linear growth model introduced here by time-varying thresholds. The average χ^2 over 100 replications equals 288.14 for 3 indicators and increases to 808.23 for 15 indicators per subscale. Note that the fitted model has 26 *df.* The detrimental effect of higher reliability is consistent with the results of the single factor model. Hence, unacceptable fit caused by threshold differences across time is not improved by creating and analyzing subscales instead of analyzing the individual outcomes.

We performed additional analyses with data obtained by averaging over multiple indicators that had been categorized with time- and group-invariant thresholds before averaging. Fit indexes remained approximately on the same acceptable level when increasing the number of averaged indicators. For both equal and unequal thresholds, parameter estimates were similar across different numbers of averaged indicators.

Discussion of the Simulation Results

The simulation illustrates some of the possible drawbacks that can be encountered when fitting factor models for continuous data to ordered categorical data. The magnitude of the effects of violating the normality assumption of course depends heavily on the parameter values that we have chosen for our data-generating models. However, the primary aim of this simulation is not to quantify the effects of violating the normality assumption but rather to illustrate the complexity of the interplay of the various effects. It is apparent that threshold differences across indicators or time points, threshold differences across groups, separation of groups, reliability of observed indicators, and sample size can all have impact on the model fit and hence on observed rejection rates.

Fitting the correct model, that is, a multigroup model for categorical outcomes to the categorical data leads to satisfying results. The observed rejection rates are close to the expected rates, and the parameter coverage exceeds 95%.

Interestingly, analyzing subscale scores obtained by averaging across categorical indicators does not necessarily improve the situation. Reliability increases by adding indicators to the subscale and so does the power to reject the measurement invariant model in case thresholds differ across time points.

CONCLUSION

Ordered categorical data are often analyzed with models for multivariate normal data. This practice can be problematic if the analysis aims at comparing groups on a latent level. Interpretations of factor mean differences between groups such as, for instance, differences in latent growth trajectories are valid only if the test or questionnaire can be shown to be measurement invariant across groups. Hence, investigation of MI is a necessary first step in this type of analysis.

The simulation study shows that fitting models for continuous data to ordered categorical data complicates a meaningful comparison of groups on a latent level. Not only different types of inequality of thresholds across groups, which represent a violation of MI, but also separation of groups, reliability of the observed variables, and sample size affect rejection rates. In an empirical setting, a researcher would not know whether unfavorable measures of goodness of fit are really due to a violation of MI, due to threshold differences across items that result in structural differences, or due to the fact that the data are categorical and measures of goodness of fit based on the assumption of normally distributed data do not function properly. No clear distinction can be made in case a latent growth model is rejected between MI across groups on the one hand and threshold changes across time on the other if thresholds are not estimated. Underestimation of factor variance further complicates the interpretation of group differences with respect to factor means. When applied to normally distributed data, tests of MI often provide useful information in case the MI model does *not* lead to acceptable fit because the results can reveal the sources of the violation. This, however, is not the case when categorical data are analyzed with models for continuous outcomes. The source of unacceptable fit remains obscure. Consequently, if a study aims at comparing groups or classes with respect to the factors that underlie a given test or questionnaire, it seems preferable to fit a model for ordered categorical outcomes. Fitting models for categorical outcomes provides the possibility to test hypotheses concerning threshold invariance across groups and indicators or time points.

ACKNOWLEDGMENTS

The research by the first author was supported through Grant MH65322 by NIMH, and through a subcontract to grant No. 5 R01 HD30995–07 by NICHD. The research of all authors was supported by both NIMH and NIDA under grant No. MH40859, and by NIMH under Grants No. MH01259, No. MH38725, and No. MH42968. Bengt Muthén was also supported by Grant K02 AA 00230 from NIAAA.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105,* 467–477.

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

Carroll, J. B. (1945). The effect of task difficulty and chance success on correlations between items or between tests. *Psychometrika, 10,* 1–20.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35,* 21–50.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance analysis.* Published doctoral dissertation, Groningen University, The Netherlands.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the context of the common factor model. *Intelligence, 31,* 543–566.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 111–133.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 121–143.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 521–543.

Millsap, R. E., & Tein, J. Y. (in press). Model specification and identification in multiple-group factor analysis of ordered-categorical measures. *Multivariate Behavioral Research*.

Muthén, B. O. (1984). A general structural equation model for dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. O. (2001a). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Muthén, B. O. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: American Psychological Association.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in* M*plus* [M*plus* Webnote #4]. Los Angeles: http://www.statmodel.com

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Muthén, L. K., & Muthén, B. O. (2001). M*plus user's guide.* Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). M*plus* 2.1. [Computer program]. Los Angeles: Muthén & Muthén.

Olsson, U. (1978). *Some data analytic problems in models with latent variables.* Unpublished doctoral dissertation, University of Uppsala, Sweden.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14,* 481–500.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 221–239.