Factor-analyzing Likert-scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes

Gitta Lubke
University of California, Los Angeles

Bengt Muthén
University of California, Los Angeles

Abstract

Treating Likert 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, etc., it seems nevertheless possible to recover true parameter values *if* the data stem from a single homogenous population. It is shown in a multi-group and a latent class context that analyzing Likert data under the assumption of multi-variate normality may distort the factor structure differently across groups or classes. Hence, investigating measurement invariance, which is a necessary requirement for a meaningful comparison of observed groups or latent classes, is problematic. Analyzing subscale scores computed from Likert items does not necessarily solve the problem. Based on a power study, some conditions are established to obtain acceptable results.

Questionnaires designed to measure latent variables such as personality factors or attitudes typically use Likert scales as a response format. 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 which run for instance from 'almost never' to 'almost always'. In case the interest of a study focuses on the latent variables underlying the items, data analysis will include fitting latent variable models such as confirmatory factor models. A special type of factor models, growth curve models, are used for the analysis of longitudinal data. Data arising from Likert-type items are often analyzed as multivariate normal outcomes in these models although the data are in fact ordered categorical. The present article focuses on analysis of ordered categorical outcomes from observed groups or latent classes while incorrectly assuming multivariate normality.

The main difference between multivariate normal and ordered categorical outcomes lies in parameters that govern the distribution of the items. The distribution of multivariate normal outcomes is completely specified by the item means and covariances. For ordered categorical items, information concerning the means and covariances is not sufficient.

Suppose the responses to the items of a questionnaire are multivariate normally distributed but, due to the response format of the items, the respondent has to choose one of several ordered categories for each of the items. Hence, the unobserved responses are multivariate normal, but the observed responses are ordered categorical. The distribution of the observed categorical items is determined by the underlying multivariate normal distribution *and* thresholds. A threshold is the level of the unobserved normal response at which, on average, respondents would choose say, category 3 instead of category 2. A correct model for Likert scale data would be one which includes estimation of threshold parameters.

A substantial number of studies has focused on the robustness of factor analysis models with respect to non-normality induced by ordered categorical outcomes. 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 can be manifest in terms of the need of additional factors (e.g., 'difficulty factors'), biased estimates of factor loadings, and inflated chi-square test statistics. On the other hand, it has also been shown that given a sufficiently large number of response categories (e.g., 7), and absence of skewness, and equal thresholds across items, it seems possible to obtain reasonable results. The latter results and the lack of choice concerning user-friendly software for the analysis of Likert scale data without assuming multivariate normality may explain the common practice of analyzing ordered categorical data using models for normally distributed variables.

It is important to note that the above mentioned studies all 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 or latent classes. First, when comparing groups or latent classes, 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 introduce an additional source of distortion. Second, if data arise from a single homogenous population, given a reasonably large sample, all possible response categories will be observed. However, in data arising from a heterogenous population, not all response categories may be observable in all groups or classes. If the groups or classes are well separated, for instance, only the lower categories may be observed in the group with the lower factor mean. 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. In addition, categorization can result in item distributions that differ across groups or classes with respect to skewness although the underlying variable is normal within all groups or classes.

Subpopulation specific distortion of the covariance and mean structure can complicate a meaningful comparison of groups or classes. Such comparisons are meaningful only in the absence of measurement bias. Consequently, measurement invariance has to be investigated. In the context of confirmatory factor models, this amounts to fitting a model with regression intercepts, factor loadings and residual variances restricted to be equal across groups or classes. Investigation of measurement invariance may fail due to a covariance and mean structure which is distorted differentially across groups or classes even though the underlying multivariate normally distributed outcome variables are measurement invariant. In addition, data may contain threshold differences across groups or classes which are in fact a form of measurement bias because the distribution of observed variables given the trait and group membership does not equal the distribution of observed variables given the trait. In sum, the question arises to what extend groups or classes can be compared in a meaningful way when assuming multivariate normality for Likert scale data. Differently distorted factor structures across groups or classes may lead a researcher to conclude that measurement invariance is absent when in fact it is not. Surely, one may advise to fit a more adequate model (i.e., for ordered categorical outcomes) under al circumstances, but given the frequency with which Likert data are in practice analyzed with factor models for continuous outcomes, it is of interest to investigate the extend with which this practice leads to incorrect conclusions with respect to group or latent class differences.

The models considered in the present paper are a single factor model and a linear growth model. Due to the growing popularity of growth mixture models, the growth model is investigated both in a multi-group and in a mixture context. Multivariate normal data are generated and subsequently categorized. We focus on the following issues. Firstly, the effect of the separation of groups or classes. As mentioned above, a larger separation may have a detrimental effect, although in a mixture setting it is well known that a larger separation facilitates model fitting. Secondly, we vary the reliability of observed variables. High reliability is usually aimed at, however, it might result in an increased power to reject the model due to violation of the normality assumption. Thirdly, we investigate the effect of threshold inequality across items. Threshold equality across items is rather unlikely in practice. Note that in a growth model setting inequality may result in a confounding between time effects and different use of a scale across time. The fourth and last issue is threshold inequality across groups or classes, which may result in a confounding between group differences in factor means and group differences in scale use.

The effects of the degree of separation of subpopulations and reliability are shown using the single factor model. The effect of inequality of thresholds across items is shown for the single factor model and for a linear growth model. Threshold inequality across groups is shown only for the growth model. Measurement invariant single factor or growth models with correct loading patterns are fitted to the categorized data and the power to reject the models is computed.

Considering the above one might be tempted to compute sum or average scores for small subsets of items, and factor analyze the resulting, more continuous looking subscales 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.