

## LATENT VARIABLE STRUCTURAL EQUATION MODELING WITH CATEGORICAL DATA\*

Bengt MUTHÉN

*University of California, Los Angeles, CA 90024, USA*

Structural equation modeling with latent variables is overviewed for situations involving a mixture of dichotomous, ordered polytomous, and continuous indicators of latent variables. Special emphasis is placed on categorical variables. Models in psychometrics, econometrics and biometrics are interrelated via a general model due to Muthén. Limited information least squares estimators and full information estimation are discussed. An example is estimated with a model for a four-wave longitudinal data set, where dichotomous responses are related to each other and a set of independent variables via latent variables with a variance component structure.

### 1. Introduction

This article gives a general overview of the specification and estimation of latent variable structural equation models, with particular emphasis on the case of dichotomous and ordered polytomous observed variables (indicators). With some recent exceptions, the methodology available to date is intended for the case of continuous indicators only. Developments for categorical indicators are important since in many applications, particularly in the social and behavioral sciences, observed variables frequently have a small number of categories with non-equidistant scale steps, and often they are dichotomous (binary). The categories of such variables may be scored for subsequent treatment as continuous, interval scale variables. Pearson product-moment correlations and covariances are, however, unsuited for these quasi-continuous variables, particularly when the variables are skewed. When such variables are forced into the mold of traditional structural equation models, a distorted analysis will result.

This article draws on new developments presented in Muthén (1981a), where a general structural equation model and its estimation was proposed. Muthén's model allows for both dichotomous, ordered polytomous, and continuous indicators of latent variables. With this general model, a large body of methodological contributions from psychometrics, biometrics, and

\*This research was supported by Grant 81-IJ-CX-0015 from the National Institute of Justice and by Grant DA 01070 from the U.S. Public Health Service.

econometrics can be conveniently interrelated. This is carried out with respect to modeling in section 3. Section 4 considers estimation approaches, while section 5 presents the estimation of a social-psychological longitudinal model with features that are relevant to many fields of application, including econometrics.

## 2. A general model

Muthén (1981a) considered the following model for  $G$  groups (populations) of observation units. The model is presented in a somewhat re-arranged way here. For each group  $g$  is observed a random dependent (endogenous) variable vector  $y^{(g)}$  ( $p \times 1$ ) and a random independent (exogenous) variable vector  $x^{(g)}$  ( $q \times 1$ ). Observations from different groups are assumed to be independent. In what follows the super-script  $g$  should be attached to each array of the model, but will be deleted for simplicity in cases where no confusion can arise. Each observed variable may be continuous or categorical with ordered categories. The observed variables are assumed to be generated by a set of underlying latent continuous variables in the following way. For each group, assume the linear structural equation system for a set of  $m$  latent dependent variables  $\eta$  and a set of  $n$  latent independent variables  $\xi$ ,

$$\eta = \alpha + B\eta + \Gamma\xi + \zeta, \quad (1)$$

where  $\alpha$  ( $m \times 1$ ) is a parameter vector of intercepts,  $B$  ( $m \times m$ ) is a parameter matrix of coefficients for the regressions among the  $\eta$ 's such that the diagonal elements of  $B$  are zero and  $I - B$  is non-singular,  $\Gamma$  ( $m \times n$ ) is a parameter matrix of coefficients for the regressions of  $\eta$ 's on  $\xi$ 's, and  $\zeta$  is a random vector of residuals (errors in the equations).

Also assume the linear 'inner' measurement relations for a set of  $p$  latent response variables  $y^*$  and a set of  $q$  latent response variables  $x^*$ ,

$$y^* = v_y + A_y\eta + \varepsilon, \quad (2)$$

$$x^* = v_x + A_x\xi + \delta, \quad (3)$$

where  $v_y$  ( $p \times 1$ ) and  $v_x$  ( $q \times 1$ ) are parameter vectors of intercepts,  $A_y$  ( $p \times m$ ) and  $A_x$  ( $q \times n$ ) are parameter matrices of coefficients (loadings) for the regressions of the latent response variables on the latent variables in the structural relations, and  $\varepsilon$  ( $p \times 1$ ) and  $\delta$  ( $q \times 1$ ) are random vectors of residuals (errors of measurement).

The observed variables are assumed to be related to the latent response variables by a set of  $p+q$  "outer" measurement relations. For a certain latent

response variable,  $z^*$  say, two alternative types of measurements,  $z$  say, are allowed. With a categorical  $z$  with, say,  $C$  categories we assume the monotonic relation,

$$\begin{aligned}
 z &= C-1 && \text{if } \tau_{C-1} < z^*, \\
 &= C-2 && \text{if } \tau_{C-2} < z^* \leq \tau_{C-1}, \\
 &\vdots && \\
 &= 1 && \text{if } \tau_1 < z^* \leq \tau_2, \\
 &= 0 && \text{if } z^* \leq \tau_1,
 \end{aligned} \tag{4}$$

where the  $C-1$   $\tau$ 's are threshold parameters defining category intervals on  $z^*$ . With a continuous  $z$  we simply have the identity

$$z = z^*. \tag{5}$$

The following specification is made regarding the first- and second-order moments of the random variables. In (1),  $E(\xi) = \kappa$  ( $n \times 1$ ),  $V(\xi) = \Phi$  ( $n \times n$ ),  $V(\zeta) = \Psi$  ( $m \times m$ ), and  $\zeta$  has zero expectation and is uncorrelated with  $\xi$ . In (2) and (3) the residual vectors have zero expectation,  $\varepsilon$  is uncorrelated with  $\eta$  and  $\zeta$ , and  $\delta$  is uncorrelated with  $\xi$ . Let  $\Theta_\varepsilon$  ( $p \times p$ ) and  $\Theta_\delta$  ( $q \times q$ ) be the covariance matrix of  $\varepsilon$  and  $\delta$ , respectively. Also,  $\zeta$ ,  $\varepsilon$  and  $\delta$  are mutually uncorrelated.

Muthén distinguished between two cases concerning the specification of the distribution of the observed variables. This distinction is particularly important with categorical dependent variables.

*Case A.* Specification of the density  $f(y^*, x^*)$  for the joint distribution of  $y^*$  and  $x^*$ . This determines the joint distribution of  $(y', x')$  by (4) and/or (5). For each group, the parameter arrays of Case A are  $\tau_y, \tau_x, \nu_y, \nu_x, A_y, A_x, \Theta_\varepsilon, \Theta_\delta, \alpha, B, \Gamma, \kappa, \Psi$ .

*Case B.* Specification of the density  $f(y^* | x)$  for the conditional distribution of  $y^*$  given  $x$ . This is of interest in the special case of the model where all independent variables are considered to be continuous, with  $q = n$  and

$$x = x^* = \xi, \tag{6}$$

so that (1) may be rewritten as

$$\eta = \alpha + B\eta + \Gamma x + \zeta. \tag{7}$$

This determines the conditional distribution of  $y$  given  $x$  by (4) and/or (5). Case B may be considered as the 'fixed  $x$ ' case, meaning that no structure is imposed on the marginal distribution of  $x$ . The parameter arrays of Case B are  $\tau_y, v_y, \Lambda_y, \Theta_\varepsilon, \alpha, B, \Gamma, \Psi$ .

Muthén considered the multivariate normal distribution for both Case A and Case B. For Case A, it is then sufficient to consider  $E(y^*, x^*)$  and  $V(y^*, x^*)$ , and for Case B,  $E(y^* | x^*)$  and  $V(y^* | x)$ ,

$$E \begin{bmatrix} y^* \\ x^* \end{bmatrix} = \begin{bmatrix} v_y + \Lambda_y(I - B)^{-1}(\alpha + \Gamma\kappa) \\ v_x + \Lambda_x\kappa \end{bmatrix}, \quad (8)$$

$$V \begin{bmatrix} y^* \\ x^* \end{bmatrix} = \begin{bmatrix} \Lambda_y \Sigma_{\eta\eta} \Lambda_y' + \Theta_\varepsilon & \text{(symmetric)} \\ \Lambda_x \Phi \Sigma_{\xi\xi} \Lambda_x' & \Lambda_x \Phi \Lambda_x' + \Theta_\delta \end{bmatrix}, \quad (9)$$

where

$$\Sigma_{\eta\eta} = (I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1}, \quad (10)$$

$$\Sigma_{\xi\xi} = \Phi\Gamma'(I - B)^{-1}, \quad (11)$$

and

$$E(y^* | x) = v_y + \Lambda_y(I - B)^{-1}\alpha + \Lambda_y(I - B)^{-1}\Gamma x, \quad (12)$$

$$V(y^* | x) = \Lambda_y(I - B)^{-1}\Psi(I - B)^{-1}\Lambda_y' + \Theta_\varepsilon. \quad (13)$$

### 3. Overview of related models

In its special cases, the general model reviewed above is related to several other models, used in different application areas. Modeling will be overviewed here utilizing this general model. Although the categorical case will be emphasized it is straightforward and convenient to also include in a condensed way the more familiar case of continuous variables.

#### 3.1. Continuous variables

A basic model is Jöreskog's so-called LISREL model, presented in Jöreskog (1973, 1977). In LISREL, all indicators are considered to be continuous, so that (5) holds for all outer measurement relations, i.e., the latent response variables are all observed. The original LISREL model was concerned with the special case of a single group ( $G=1$ ), and used the standardization  $\alpha=0, \kappa=0$ , so that  $E(\eta)=0, E(\xi)=0$ . Case A and Case B were both considered, using the normality assumptions. Case B, when further specialized to involve no measurement structure and no measurement errors

in (2), has  $p=m$  and  $y=\eta$ . The case of  $p=m$  (and  $q=n$ ) will be referred to as the single-indicator case, as opposed to the multiple-indicator case. It has been extensively studied by econometricians in the analysis of linear simultaneous equation systems [for familiar references, e.g. see the overview in Jöreskog (1973, pp. 93–95)]. LISREL is a hybrid modeling of linear factor analysis (inner) measurement relations [see e.g. Lawley and Maxwell (1971)], see (2) and (3), combined with a linear simultaneous equation system for the factors, see (1). This has proven very useful, particularly in social and behavioral science applications. For overviews with illustrations and additional detail, see e.g. Aigner and Goldberger (1977), Bentler (1980), Bentler and Weeks (1980), Bielby and Hauser (1977), Browne (1982) and Jöreskog (1978).

Retaining the requirement of continuous indicators, simultaneous analysis of several groups,  $g=1, 2, \dots, G$ , and the inclusion of structured means via the parameter arrays  $\alpha^{(g)}$  and  $\kappa^{(g)}$  has been incorporated in the LISREL framework more recently. The multiple-group factor analysis of Jöreskog (1971) was extended by Sörbom (1974) to study not only differences and similarities in covariance structure but also in factor means. Multiple-group analysis with structured means was developed into more general LISREL models in Sörbom (1982) with applications to latent variable ANCOVA [Sörbom (1978)] and the analysis of longitudinal data [Jöreskog and Sörbom (1980)]; see also Jöreskog and Sörbom (1981).

### 3.2. *Categorical variables: Single indicators*

Turning to situations with categorical response variables, consider first the single-indicator case. Here we find Case B models. The simplest situation is that of univariate and multivariate regression with categorical response variables. Methodology for this situation is well-known to econometricians and an excellent review with econometric applications covering dichotomous, ordered and unordered polytomous response is given in Amemiya (1981). These models originated in biometric work, notably probit/logit regression in bioassay [see e.g. Bliss (1935)]. Probit regression is a special case of the general model of section 2, while logit regression and related log-linear modeling fall outside this model. In the multivariate case the general model gives the multivariate probit model of Ashford and Sowden (1970). Multivariate logit models are not directly related to this model structure; there is no multivariate logistic distribution with logistic marginal distributions that have unconstrained correlation coefficients [see Gumbel (1961) and also Amemiya (1981, pp. 1525–1531) and Morimune (1979)]. As opposed to multivariate regression, simultaneous equation models generally place a structure on the reduced-form regression coefficients and possibly also the reduced-form error covariances/correlations. With categorical

response variables, such models have recently attracted a growing interest in econometrics, but do not seem to have been utilized in biometrics or psychometrics. Some important contributions are Amemiya (1978), Heckman (1974, 1978) and Maddala and Lee (1976).

### *3.3. Categorical variables: Multiple indicators*

We now consider the more complex situation of categorical response variables, where there are multiple indicators of latent variables. Developments here have mainly come from psychometric work. Consider first the measurement part of the general model. Here, the latent response variables for the observed response variables are related to the latent variable constructs by a factor analysis type measurement model. With dichotomous indicators, probit models have been considered also here, although the independent continuous variables are now latent. In item response (latent trait) theory language [see, e.g., Lord (1980)] the general model with dichotomous indicators implies the so-called two-parameter normal ogive item characteristic curve model of Lawley (1943, 1944), Lord and Novick (1968) and Bock and Lieberman (1970). For a set of items (dichotomous variables) designed to measure a certain trait (factor), conditional independence is assumed to hold, given the factor. In the general model the analogous assumption is the diagonality of the measurement error covariance matrix ( $\Theta_\epsilon$  or  $\Theta_\delta$ ). Note, however, that correlated errors can be handled. For related one-, two- and three-parameter logistic item response models, see e.g. Andersen (1980). The general multiple-factor model has been studied by Bock and Aitkin (1981), Christofferson (1975) and Muthén (1978), both for exploratory ('unrestricted') and confirmatory ('restricted') factor analysis. Muthén and Christofferson (1981) generalized the model to handle simultaneous multiple-group analysis, where various degrees of invariance over populations can be studied. As in the continuous variable case, modeling of factor mean differences over populations is then of interest, see e.g. Muthén (1981b).

The extension of the measurement model to more than two ordered categories by (4), in combination with both (2) and (3), is straightforward and natural. For special cases, this was first proposed by Edwards and Thurstone (1952), and later studied by e.g. Bock and Jones (1968), Samejima (1969) and Bartholomew (1980). [Note the biometric counterparts of Aitchison and Silvey (1957) and Gurland, Lee and Dahm (1960).] The unordered polytomous case, not covered by the general model above, was studied by Bock (1972). Further contributions are found in Samejima (1972).

The extension to structural equation modeling with categorical response variables as latent variable indicators was first brought forward in Muthén (1976a), and further developed in Muthén (1977, 1979, 1982a). Here, Case B

was considered with dichotomous observed variables for each latent response variable. Muthén (1979) considered a multiple-indicator–multiple-cause (MIMIC) model analogous to the MIMIC model discussed in Jöreskog and Goldberger (1975) for the case of continuous response variables, while Muthén (1976b) studied a model with reciprocal interaction between two dependent latent variable constructs.

The general model of section 2 covers not only Case A and Case B of the general structural equation model but also any combination of dichotomous, ordered categorical, and continuous indicators in the measurement part. Further generalizations of the measurement part are possible. One example is the inclusion of categorical–continuous or limited dependent observed variables [see, e.g., Tobin (1958) and Amemiya (1973, 1982)].

#### 4. Estimation

The general model of section 2 can be estimated in various ways. Two basically different approaches have been attempted for special cases of this model, limited information (univariate and bivariate) multi-stage weighted least-squares (WLS), and full information, maximum likelihood (ML) estimation. Limited information estimation has been motivated by the fact that when categorical response variables are involved, a straight-forward application of ML may lead to heavy computations.

##### 4.1. Limited information estimation

Muthén (1981a) proposed a three-stage limited information WLS estimator. Muthén summarized the structure of the general model in three parts, encompassing both Case A and Case B. The three parts are respectively a mean/threshold/reduced-form regression intercept structure, a reduced-form regression slope structure, and a covariance/correlation structure. Any of the three parts may be used alone or together with any of the other parts. A computer program LACCI [Muthén (1982b)] may be used for all computations (LACCI was utilized for the analyses of section 5). The model structure will first be presented in its full generality and then explained through a set of special cases. For each group, deleting the group index, consider the three population vectors  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$ :

*Part 1* (mean/threshold/reduced-form regression intercept structure)

$$\sigma_1 = \Delta^* [K_\tau \tau_z - K_v (v_z + \Lambda_z (I - B_z)^{-1} \alpha_z)], \quad (14)$$

*Part 2* (reduced-form regression slope structure)

$$\sigma_2 = \text{vec} \{ \Delta \Lambda_z (I - B_z)^{-1} \Gamma_z \}, \quad (15)$$

## Part 3 (covariance/correlation structure)

$$\sigma_3 = K \text{vec}\{\Delta[\Lambda_z(I - B_z)^{-1}\Psi_z(I - B_z)'\Lambda_z' + \Theta_z]\Delta\}. \quad (16)$$

Here,  $\Delta$  is a diagonal matrix of scaling factors particularly useful in multiple-group analyses with categorical variables,  $\Delta^*$  contains the same element as  $\Delta$  but diagonal elements are duplicated for categorical variables with more than one threshold (more than two categories),  $K_r$  and  $K_v$  similarly distributed elements from the vectors they pre-multiply, the  $\text{vec}$  operator strings out matrix elements row-wise into a column vector, and  $K$  selects lower-triangular elements from the symmetric matrix elements it pre-multiplies, where a diagonal element is only included if the corresponding observed variable is continuous.

For Case A, part 2 is not needed. We may stack the dependent variables followed by the independent variables into a single vector. Then, the arrays of the three-part model structure organize the parameters as

$$\begin{aligned} \tau_z &= \begin{bmatrix} \tau_y \\ \tau_x \end{bmatrix}, & v_z &= \begin{bmatrix} v_y \\ v_x \end{bmatrix}, \\ \Lambda_z &= \begin{bmatrix} \Lambda_y & 0 \\ 0 & \Lambda_x \end{bmatrix}, & \Theta_z &= \begin{bmatrix} \Theta_\epsilon & \text{(symmetric)} \\ 0 & \Theta_\delta \end{bmatrix}, \\ \alpha_z &= \begin{bmatrix} \alpha \\ \kappa \end{bmatrix}, & B_z &= \begin{bmatrix} B & \Gamma \\ 0 & 0 \end{bmatrix}, \\ \Gamma_z &\text{ has no counterpart, } & \Psi_z &= \begin{bmatrix} \Psi & \text{(symmetric)} \\ 0 & \Phi \end{bmatrix}. \end{aligned}$$

For Case B,

$$\begin{aligned} \tau_z &= \tau_y, & v_z &= v_y, & \Lambda_z &= \Lambda_y, & \Theta_z &= \Theta_\epsilon, \\ \alpha_z &= \alpha, & B_z &= B, & \Gamma_z &= \Gamma, & \Psi_z &= \Psi. \end{aligned}$$

With the normality specification on the latent response variables, any model that fits in the general framework is identified if and only if its parameters are identified in terms of  $\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(G)}$ , where  $\sigma^{(g)'} = (\sigma_1^{(g)'}, \sigma_2^{(g)'}, \sigma_3^{(g)'})$ . Muthén (1981a) utilized this fact in that statistics  $s^{(g)}$  were produced as consistent estimators of  $\sigma^{(g)}$ , in order to estimate the model parameters in a final estimation stage. Preceding estimation stages give  $s^{(g)}$ ,

where only limited information from bivariate sample distributions is needed. In the final estimation stage, a WLS fitting function with a general, full weight matrix is used,

$$F = \sum_{g=1}^G (s^{(g)} - \sigma^{(g)})' W^{(g)-1} (s^{(g)} - \sigma^{(g)}), \quad (17)$$

where the (limited information) generalized least squares (GLS) estimator is obtained when  $W^{(g)}$  is a consistent estimator of the asymptotic covariance matrix of  $s^{(g)}$ . For the estimator based on the minimization of (17) there is no requirement that the  $s_3^{(g)}$  elements form a positive definite matrix, although in large samples absence of this would indicate a mis-specified model. With GLS,  $F$  calculated at the minimum provides a large-sample chi-square test of model fit to the first- and second-order statistics. Large-sample standard errors of estimates are also readily available.

With continuous indicators only, the model structure in the single-group case can usually be encompassed by the covariance matrix structure alone, i.e., part 3 of Muthén's three-part structure. With  $\Delta = I$ , part 3 includes the LISREL model structure. In a multiple-group analysis, the model usually also implies a structure on the observed variable means, so that both part 1 and part 3 would be used, where part 1 in this case simplifies to  $v_z + A_z(I - B_z)^{-1} \alpha_z$ . The bivariate sample statistics vectors  $s_1^{(g)}$  and  $s_3^{(g)}$  have elements from the sample mean vector and the ordinary sample covariance matrix  $S^{(g)}$ . Part 2 is not needed here. Jöreskog (1973, 1977) considered the full information, maximum likelihood (ML) estimator in the single-group case, while Sörbom (1974, 1982) considered ML in the multiple-group case. GLS estimation was considered for various single-group cases of the model by Bentler and Weeks (1980), Browne (1974, 1982) and Jöreskog and Goldberger (1972), and for multiple-group cases by Lee and Tsui (1982) and Muthén (1981a). The ordinary GLS estimator requires that the fourth-order cumulants of the observed variables are zero, which is fulfilled by the normality specification for the latent response variables in the general model. In this case GLS is asymptotically equivalent to ML, and the limited information approach is equivalent to full information. Considering the single-group case, the GLS weight matrix for part 3 is here a function of  $S$  only, and the above fitting function  $F$  may be rewritten in the familiar way,

$$(N-1)^{-1} F = \frac{1}{2} \text{tr}(I - S^{-1} \Sigma)^2, \quad (18)$$

where  $N$  is the sample size and  $\Sigma$  is the population covariance matrix containing the  $\sigma_3$  elements. Going beyond the restriction of multivariate normal variables, Browne (1982) discussed asymptotically distribution-free

GLS estimation. The full weight matrix then includes third- and fourth-order moment information. This can also be handled by (17).

While obtaining  $s^{(g)}$  is straightforward with continuous indicators, its elements have to be computed iteratively when categorical indicators are involved. As opposed to the continuous case, the distinction between Case A and Case B becomes important to estimation. First consider Case A with dichotomous indicators. Christofferson (1975), Muthén (1978) and Muthén and Christofferson (1981) studied factor analysis by unweighted (ULS) and GLS estimators that use only first- and second-order sample information, i.e., univariate and bivariate proportions. With the normality specification on the latent response variables, this approach leads to the analysis of sample thresholds and sample tetrachoric correlations. The population tetrachoric correlation is the (latent) correlation between a pair of latent response variables; see, e.g., Brown and Benedetti (1977) and Pearson (1900). For each two-by-two table, one tetrachoric correlation element and two threshold elements are estimated by ML. The former produces an element for the  $s_3^{(g)}$  vector and the latter produces two elements for the  $s_1^{(g)}$  vector. Rapid and numerically efficient algorithms have been described by Divgi (1979) and Kirk (1973). Muthén (1978) developed the limited information GLS weight matrix corresponding to the tetrachoric correlation vector  $s_3^{(g)}$  and the threshold vector  $s_1^{(g)}$ . It uses sample information up to and including fourth-order moments.

For Case A with ordered polytomous variables and mixtures of ordered polytomous variables and continuous variables, similar statistics can be produced. The latent correlations are then called polychoric and polyserial respectively, and their estimation has been discussed in Jaspens (1946), Lancaster and Hamdan (1964), Martinson and Hamdan (1971), Olsson, (1979a), Olsson, Drasgow and Dorans (1981), Pearson (1904, 1913), Pearson and Pearson (1922) and Tallis (1962). Given bivariate information, ML estimation of the thresholds ( $C-1$ , for each categorical variable with  $C$  categories) and the latent correlation coefficient presents an overidentified model and is more time-consuming than in the tetrachoric case. Olsson (1979a) and Olsson, Drasgow and Dorans (1981) discuss a simpler two-stage estimator, where thresholds are estimated from the univariate distributions, and the correlation is estimated by conditional ML, given the threshold estimates from the first stage. The two estimators are equivalent in the tetrachoric case. Algorithms for computing tetrachoric, polychoric and polyserial correlations have recently also been included in the LISREL computer program; see Jöreskog and Sörbom (1981).

To summarize, the three stages of Muthén's estimation are for Case A with categorical variables: the estimation of population thresholds, the estimation of population latent correlations given estimated thresholds, and the estimation of model parameters given estimated correlations and thresholds.

For Case A with categorical indicators, only part 3 is needed in a single-group analysis; the threshold vector  $\tau$  would then not be estimated. With categorical indicators, diagonal elements are not included in part 3. In a multiple-group analysis, part 1 would also be included to identify and estimate level differences in latent variable constructs. With categorical indicators, the elements of part 1 correspond to the lower integral limits for the standardized latent response variables, when integrating to determine the distribution of the observed variables. Group differences in the variances of the latent response variables can be captured by  $\Delta^{(g)}(\Delta^{*(g)})$ .

The estimation approach presented for Case A with categorical indicators may be seen as producing more suitable coefficients of association than the ordinary Pearson product-moment correlations and covariances applied to scored categorical variables. In psychometrics it is well-known that product-moment correlations for the observed variables in these cases are not free to vary between  $-1$  and  $+1$ , are dependent on the marginal distribution for each variable, and underestimate the value of the latent correlations described above, particularly with markedly skewed categorical variables. General references, are, e.g., Carroll (1961) and Pearson (1913), and specific references on factor analysis are, e.g., Ferguson (1941) discussing 'difficulty factors' for dichotomous test items, and Mooijaart (1982) and Olsson (1979b) discussing skewed polytomous variables. Using latent correlations may be seen as robustifying correlations against categorized and skewed variables, 'stretching' correlations to be able to assume values throughout the  $-1$ ,  $+1$  range, independent of the univariate marginals. The use of latent correlations built on the normality specification is of course only one way of 'stretching'. Simpler well-fitting models may perhaps be found using other statistics  $s_3^{(g)}$ , such as the measures of association discussed in chapter 11 of Bishop, Fienberg and Holland (1975). Tetrachoric and biserial coefficients are used in psychometric item analysis work. For a discussion of the normality specification in this context, see e.g. Carroll (1961) and Lord and Novick (1968, ch. 15).

For Case B with categorical response variables, no distributional assumptions are made on the independent variable vector  $x^{(g)}$ . Particularly, normality is not required. Building on Muthén (1976b, 1982a), Muthén (1981a) utilized this fact in proposing univariate and bivariate regression to produce statistics for the final estimation stage. In the first estimation stage, reduced-form regression coefficients are estimated from univariate ML regressions to produce  $s_2^{(g)}$  for part 2. Given these estimates, stage two performs bivariate ML regression [see e.g. Ashford and Sowden (1970)] to estimate a single reduced-form residual correlation for each pair of variables. Computationally, this may be done very efficiently. The residual correlations correspond to part 3, where only lower-triangular elements are included in  $\sigma_3^{(g)}$ . Part 1 here structures reduced-form regression intercepts, which is of

particular interest in multiple-group analyses. This type of analysis has not yet been explored in practical work, but seems to hold great promise for both biometric, psychometric, and econometric applications. An application of this type is presented in section 5.

Related limited information, least squares estimation approaches for special cases of Case B have been presented in econometric work; see e.g. Amemiya (1978) and Lee (1979).

#### *4.2. Full information estimation*

The full information ML estimation of models fitting into the general framework above, is still largely undeveloped. Note that when all indicators are categorical the distribution of the observed variables ( $y$  and  $x$  for Case A,  $y$  for Case B) is multinomial, and that it will be determined by integration over the multivariate normal latent response variables. Since the multivariate normal distribution function does not exist in a closed form, computational difficulties may arise already for more than two variables. Bock and Lieberman (1970) developed the ML estimator for factor analysis of dichotomous variables. A small number of variables could be handled, since the normal multiple integral may in this case be replaced by a single integration over the assumed single factor. A computational breakthrough for this problem was presented in Bock and Aitkin (1981), building on the EM algorithm; see Dempster, Laird and Rubin (1977). In this approach, values for the factors are considered 'missing' and are estimated in the E-step, whereas univariate probit regressions may be carried out in the M-step to estimate model parameters. This avoids computation of multivariate normal distribution functions and models with many items and more than one factor can be handled. Bock and Aitkin (1981) and Gibbons (1981) studied exploratory factor analysis. However, it seems possible in principle to develop this EM approach to general Case A situations, although perhaps resulting in heavy computations with complex models involving correlated errors, a large structured latent variable construct covariance matrix and multiple groups. With continuous indicators, the EM approach has been studied by Rubin and Thayer (1982) [see also Bentler and Tanaka (1982)] for the factor analysis model, and by Chen (1981) for the MIMIC model.

Muthén (1979) discussed the ML estimator for Case B involving dichotomous response variables. No more than two response variables were, however, attempted. A numerically feasible approximation to the multivariate normal distribution function was presented in Clark (1961); see also Amemiya (1981, pp. 1523–1524). Gibbons and Bock (1982) discussed this approximation and the EM approach for full information estimation in a Case B model with dichotomous response variables.

Comparisons between limited information and full information estimators would be of interest. This has not yet been carried out rigorously for any general case of the model. Two empirical results are available for the factor analysis model with dichotomous indicators. The five-item LSAT6 and LSAT7 data sets were analyzed by one factor using both ML and limited information GLS; see Christoffersson (1975) or Muthén (1978). The model fit for the first data set was very good, while there was a marginal indication of misfit for the second. In both cases both the estimates and the estimated standard errors were close, with no indication of a significant loss of information by using only bivariate information.

### 5. A longitudinal model

The following example illustrates Case B of the section 2 general model, estimated by the Muthén (1981a) approach. The author is obliged to Paul Duncan-Jones at the Australian National University, Canberra, for providing the data and the basic ideas for the causal modeling. The data consist of a random sample of Canberra electors interviewed four times with four-month intervals, in 1977 and 1978. The data are fully described by Henderson, Byrne and Duncan-Jones (1981). We will analyze data for 231 individuals for whom complete data are available.

The response variables at each time point are four dichotomous items intended to indicate 'neurotic illness'. They arose from the following question:

'In the last month have you suffered from any of the following?'

---

Anxiety

Depression

---

Irritability

Nervousness

The response Yes was denoted by  $y=1$  and No by  $y=0$ . In the figure and tables below, the abbreviations  $A$ ,  $D$ ,  $I$ ,  $N$  will be used to denote these variables.

It was hypothesized that at each time point these four items are indicators of a latent construct of 'current level of neurotic illness' denoted by  $\eta_t$ ,  $t=1,2,3,4$ . The measurement structure of these items has also been investigated in Muthén and Christoffersson (1981). The four items will be related to a set of independent observed variables in the following way.

We will consider a longitudinal model for the relation between the  $\eta_t$  variables and the independent variables, which generalizes certain econometric variance component modeling to latent dependent variables; see e.g. Maddala (1971), Lillard and Willis (1978) and Jöreskog (1978b). Assume the model

$$\eta_t = \sum_{s=1}^t \gamma_{ts} \xi_s + \eta_0 + \zeta_t, \quad t = 1, 2, 3, 4, \quad (19)$$

where  $\gamma_{ts}$  is the coefficient for the time-specific effect of the latent independent variable  $\xi_s$  on  $\eta_t$ , allowing for lagged effects. Also,  $\eta_0$  is a latent individual component which does not vary over time and  $\zeta_t$  is an error component which is uncorrelated with the  $\xi_s$  variables and with  $\eta_0$ , but is possibly correlated over time. It is further assumed that the time-independent component is determined as

$$\eta_0 = \gamma_0 \xi_0 + \zeta_0, \quad (20)$$

where  $\xi_0$  is a latent independent variable and  $\zeta_0$  is a random error.

In this application, an indicator of  $\xi_0$  is the  $N$  scale from the Eysenck Personality Inventory, which is intended to measure long-term liability to neurosis. The variable  $N$  is here taken as the average score from occasions two and four. As indicator of each of the time-specific components  $\xi_s$ , we use the score on a variable measuring the number of 'life events' occurring to the respondent in the four months prior to each interview. These will be denoted by  $L1$ ,  $L2$ ,  $L3$  and  $L4$ . Since each of the five  $\xi$  variables above has only a single indicator, the corresponding continuous observed variables will be set identically equal to these latent variables. The resulting model is graphically depicted in a conventional path diagram as in fig. 1.

The model of fig. 1 may be fitted into the general framework as follows. Part 2 and part 3 of the three-part model structure will be used. Let  $y^*$  ( $16 \times 1$ ) be the latent response variables taken from left to right in fig. 1 and let  $\eta' = (\eta_0 \eta_1 \eta_2 \eta_3 \eta_4)$ . By standardizing the  $\nu$  parameters to zero, the (inner) measurement part is

$$y^* = A\eta + \varepsilon. \quad (21)$$

The first column of the  $16 \times 5$  matrix  $A$  consists of zeros. Since the same questionnaire was administered at all occasions, we let the  $A$  matrix have a time-invariant structure, with equality of loadings for the same item at different time points. Let  $\lambda_i$  be the loading for the  $i$ th item. To determine the scale of each  $\eta_i$ , we will set  $\lambda_1$  to unity. The mean of  $\eta$  is standardized to zero by fixing  $\alpha$  ( $5 \times 1$ ) to a vector of zeros. No time invariance is specified

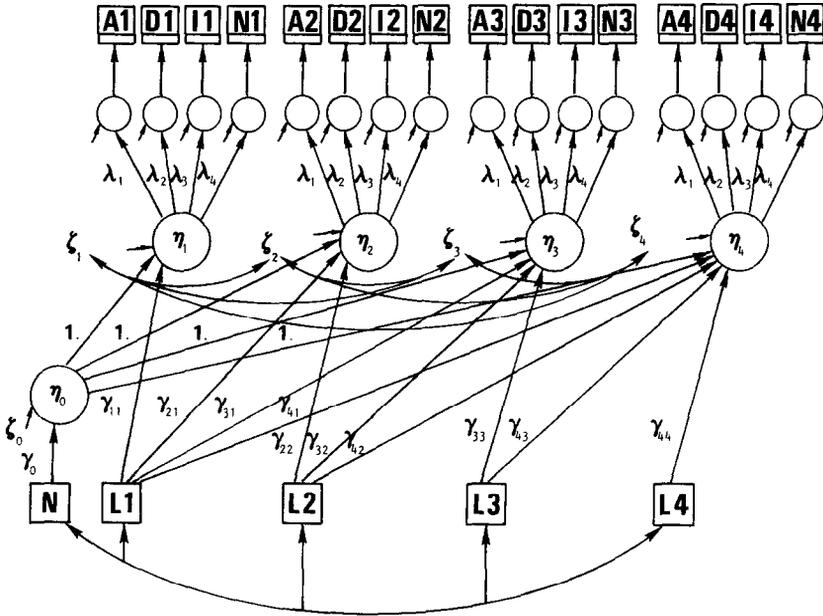


Fig. 1. A longitudinal model.

for the  $16 \times 1$  vector of thresholds  $\tau$ , allowing for differences in means of the same item over time. Modeling of these differences over time could be attempted, using the reduced-form regression intercepts of part 1. With time-invariant thresholds, differences in means of  $\eta_t$  over time are identifiable and may be estimated. This will not be carried out here.

Let  $x' = (N \ L1 \ L2 \ L3 \ L4)$  and  $\zeta' = (\zeta_0 \ \zeta_1 \ \zeta_2 \ \zeta_3 \ \zeta_4)$ . For the  $5 \times 1$  vector  $\eta$ , we have

$$(I - B)\eta = \Gamma x + \zeta, \tag{22}$$

where

$$B (5 \times 5) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\Gamma(5 \times 5) = \begin{bmatrix} \gamma_0 & 0 & 0 & 0 & 0 \\ 0 & \gamma_{11} & 0 & 0 & 0 \\ 0 & \gamma_{21} & \gamma_{22} & 0 & 0 \\ 0 & \gamma_{31} & \gamma_{32} & \gamma_{33} & 0 \\ 0 & \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{bmatrix}.$$

To complete the specification,  $\Theta$  ( $16 \times 16$ ) is a diagonal error covariance matrix restricted so that the variances of  $y^*$  given  $x$  are unity. Frequently in longitudinal data, errors for the same variables are correlated over time, but here there was no indication of this. Furthermore,  $\Psi$  ( $5 \times 5$ ) is here a diagonal covariance matrix with  $\zeta_t$  variances hypothesized to be equal. Also, we let  $\Delta = I$  ( $16 \times 16$ ), since a single group is analyzed.

Under Case B of the model, the first estimation stage consists of computing the sixteen univariate ML probit regression of each dichotomous item on the vector  $x$ . The correlations and variances of the  $x$  variables are given in table 1. The probit regressions give consistent estimates of the elements of the reduced-form regression matrix  $A(I-B)^{-1}\Gamma$ . The estimates are given in table 2. The eighty elements make up the sample vector  $s_2$ . The standard errors given in parentheses pertain to each univariate regression.

Table 1  
The longitudinal model: Correlations and variances for the independent variables.

<i>Correlations</i>					
<i>N</i>	1.0000				
<i>L1</i>	0.2162	1.0000			
<i>L2</i>	0.1595	0.5377	1.0000		
<i>L3</i>	0.1807	0.4967	0.4937	1.0000	
<i>L4</i>	0.2059	0.5270	0.4886	0.5144	1.0000
<i>Variances</i>					
	20.66	6.537	5.893	4.898	5.271

Given  $s_2$ , the second estimation stage involves the computation of the 120 bivariate probit regressions of all pairs of the dichotomous items on  $x$ , estimating [see, e.g., Ashford and Sowden, (1970)] each of the 120 lower-triangular elements of the reduced-form residual correlation matrix  $A(I-B)^{-1}\Psi(I-B)^{-1}A' + \Theta$ . These  $s_3$  values are given in table 3. Inspecting  $s_3$  shows that the elements correspond to a non-positive definite covariance matrix. The sub-set of elements corresponding to the covariance matrix for the first three occasions is positive definite as is the covariance matrix for the

Table 2

The longitudinal model:  $s_2$ , univariate regression standard errors (in parentheses), and estimated  $\sigma_2$  (underneath).

Dependent variable	$N$	$L1$	$L2$	$L3$	$L4$
$A1$	0.0722	0.0881	-0.0295	0.0063	0.0511
	(0.0222)	(0.0493)	(0.0518)	(0.0597)	(0.0524)
	0.0669	0.0397	0.0000	0.0000	0.0000
$D1$	0.1617	0.0841	0.0192	-0.0352	-0.0089
	(0.0269)	(0.0552)	(0.0557)	(0.0652)	(0.0560)
	0.1168	0.0694	0.0000	0.0000	0.0000
$I1$	0.0761	0.0773	0.0961	-0.1246	0.0217
	(0.0213)	(0.0444)	(0.0480)	(0.0558)	(0.0522)
	0.0801	0.0476	0.0000	0.0000	0.0000
$N1$	0.1612	0.0233	-0.0069	0.0600	-0.0296
	(0.0291)	(0.0584)	(0.0513)	(0.0628)	(0.0609)
	0.1545	0.0918	0.0000	0.0000	0.0000
$A2$	0.0521	0.0005	0.0329	0.0194	0.0671
	(0.0229)	(0.0559)	(0.0570)	(0.0699)	(0.0584)
	0.0669	0.0319	0.0123	0.0000	0.0000
$D2$	0.1153	0.0598	0.0194	0.0249	-0.0530
	(0.0231)	(0.0462)	(0.0541)	(0.0612)	(0.0653)
	0.1168	0.0558	0.0215	0.0000	0.0000
$I2$	0.1030	0.0745	0.0552	-0.0575	0.0413
	(0.0218)	(0.0470)	(0.0460)	(0.0496)	(0.0463)
	0.0801	0.0383	0.0147	0.0000	0.0000
$N2$	0.1642	0.0584	-0.0058	0.0419	-0.0534
	(0.0282)	(0.0525)	(0.0531)	(0.0677)	(0.0711)
	0.1545	0.0738	0.0284	0.0000	0.0000
$A3$	0.0998	0.0401	-0.0121	-0.0022	0.1090
	(0.0256)	(0.0594)	(0.0706)	(0.0708)	(0.0701)
	0.0669	-0.0139	-0.0131	0.0633	0.0000
$D3$	0.1342	-0.0865	0.0242	0.1419	0.0338
	(0.0291)	(0.0628)	(0.0584)	(0.0651)	(0.0557)
	0.1168	-0.0242	-0.0229	0.1106	0.0000
$I3$	0.0656	0.0300	-0.0086	0.0598	0.0593
	(0.0214)	(0.0477)	(0.0502)	(0.0552)	(0.0510)
	0.0801	-0.0166	-0.0157	0.0758	0.0000
$N3$	0.1371	-0.0560	-0.0811	0.1591	-0.0694
	(0.0292)	(0.0673)	(0.0650)	(0.0680)	(0.0749)
	0.1545	-0.0320	-0.0303	0.1462	0.0000
$A4$	0.0450	0.0387	-0.0645	0.0489	0.1070
	(0.0227)	(0.0598)	(0.0648)	(0.0657)	(0.0621)
	0.0669	0.0130	-0.0483	0.0201	0.0464
$D4$	0.0699	0.0309	-0.0875	0.0720	0.0771
	(0.0224)	(0.0682)	(0.0619)	(0.0690)	(0.0668)
	0.1168	0.0227	-0.0844	0.0351	0.0810
$I4$	0.0704	0.0857	-0.0254	-0.0182	0.0952
	(0.0215)	(0.0440)	(0.0469)	(0.0585)	(0.0512)
	0.0801	0.0156	-0.0579	0.0241	0.0555
$N4$	0.1837	-0.0379	-0.1471	0.0326	0.0357
	(0.0365)	(0.0541)	(0.0810)	(0.0767)	(0.0714)
	0.1545	0.0301	-0.1117	0.0464	0.1071

**Table 3**  
**The longitudinal model:  $s_3$  and estimated  $\sigma_3$  (underneath).**

<i>A1</i>	1.							
<i>D1</i>	0.5978 0.3950	1.						
<i>I1</i>	0.3315 0.3775	0.2213 0.2956	1.					
<i>N1</i>	0.5776 0.5817	0.5198 0.4555	0.2799 0.4353	1.				
<i>A2</i>	0.3765 0.2670	0.1076 0.2091	0.3389 0.1998	0.4432 0.3079	1.			
<i>D2</i>	-0.0981 0.2091	0.0616 0.1637	0.1220 0.1565	0.2942 0.2411	0.4555 0.4394	1.		
<i>I2</i>	0.2250 0.1998	0.0533 0.1565	0.3648 0.1496	0.1727 0.2305	0.2964 0.4199	0.4561 0.3288	1.	
<i>N2</i>	0.3371 0.3079	0.1989 0.2411	0.0855 0.2305	0.4198 0.3551	0.6554 0.6471	0.4686 0.5067	0.5193 0.4843	1.0
<i>A3</i>	0.4684 0.4480	0.3068 0.3508	0.2538 0.3352	0.4168 0.5106	0.6336 0.3578	0.1058 0.2802	0.1262 0.2678	0.1693 0.4127
<i>D3</i>	0.2820 0.3508	0.3988 0.2747	-0.1009 -0.2625	0.4460 0.4045	0.3229 0.2802	0.2472 0.2194	0.1722 0.2097	0.1978 0.3231
<i>I3</i>	0.4166 0.3352	0.2748 0.2625	0.4406 0.2509	0.3620 0.3866	0.5490 0.2678	0.3581 0.2097	0.4525 0.2004	0.2653 0.3088
<i>N3</i>	0.5636 0.5166	0.5507 0.4045	0.1180 0.3866	0.7732 0.5958	0.5419 0.4127	0.1284 0.3231	0.1981 0.3088	0.5161 0.4759
<i>A4</i>	0.2886 0.3643	0.1616 0.2853	0.2107 0.2727	0.2448 0.4202	0.5446 0.3365	0.0276 0.2635	0.2363 0.2518	0.4169 0.3881
<i>D4</i>	0.2805 0.2853	0.4201 0.2234	0.1104 0.2135	0.7030 0.3290	0.2833 0.2635	0.2144 0.2063	0.0255 0.1972	0.3793 0.3039
<i>I4</i>	0.2952 0.2727	0.4073 0.2135	0.3783 0.2041	0.2247 0.3144	0.1974 0.2518	-0.0564 0.1972	0.4267 0.1885	0.4103 0.2904
<i>N4</i>	0.3590 0.4202	0.2658 0.3290	0.1421 0.3144	0.5996 0.4845	0.4980 0.3881	0.1071 0.3039	0.2829 0.2904	0.4303 0.4475
<i>A3</i>	1.							
<i>D3</i>	0.3417 0.4280	1.						
<i>I3</i>	0.6480 0.4091	0.3922 0.3203	1.					
<i>N3</i>	0.5616 0.6304	0.4737 0.4936	0.4070 0.4718	1.				
<i>A4</i>	0.3455 0.4206	0.3009 0.3294	0.3900 0.3148	0.6026 0.4850	1.			
<i>D4</i>	0.2868 0.3294	0.3242 0.2579	0.2903 0.2465	0.4771 0.3798	0.5589 0.4810	1.		
<i>I4</i>	0.2642 0.3148	0.2084 0.2465	0.4013 0.2356	0.3048 0.3630	0.3637 0.4597	0.4821 0.3600	1.	
<i>N4</i>	0.2195 0.4850	0.3340 0.3798	0.3608 0.3630	0.6736 0.5594	0.7692 0.7084	0.6010 0.5547	0.3302 0.5302	1.

fourth occasion. The non-positive definiteness may be due to the relatively small sample size ( $N=231$ ) or it may indicate that normality for  $y^*$  given  $x$  does not hold exactly. This does not seem to be a serious problem, however, since the fit to  $s_3$ , as reported below, seems to be reasonably good. Also, for the parameters pertaining to the first three occasions the results are very similar when analyzing all four occasions as specified, compared to analyzing only the first three occasions.

In estimation stage three, we may fit  $\sigma'=(\sigma'_2 \sigma'_3)$  to  $s'=(s'_2 s'_3)$ . The  $\sigma_1$  part is not included since its 16 elements are not restricted with the 16 free parameters of  $\tau$ . The remaining total number of 16 free parameters are contained in  $\Lambda$ ,  $\Gamma$  and  $\Psi$  and give the 200 elements of  $\sigma_2$  and  $\sigma_3$ . The model is hence considerably overidentified. We note that the parameters of  $\Lambda$  and  $\Gamma$  are identified in terms of the elements of  $\sigma_2$ . Also, the parameters of  $\Lambda$  and  $\Psi$  are identified in terms of the elements of  $\sigma_3$ .

Differences in scales of the  $x$  variables and in variability of the elements of  $s_2$  and  $s_3$ , will here be taken into account in the following way. We may estimate  $\Lambda$  and  $\Gamma$  using the  $\sigma_2$  part with a diagonal weight matrix, with diagonal elements obtained from the standard errors reported in table 2 above. Furthermore,  $\Lambda$  and  $\Psi$  may be separately estimated by ULS, using the  $\sigma_3$  part only. Any differences in the estimates of the common parameters of  $\Lambda$  may be attributed to sampling variability and lack of model fit.

Table 4  
The longitudinal model: Parameter estimates.

Estimated loadings					
Alternative	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	
$\sigma_2$	1.*	1.747	1.198	2.309	
$\sigma_3$	1.*	0.763	0.746	1.139	
Estimated $\Gamma$					
	$N$	$L1$	$L2$	$L3$	$L4$
$\eta_0$	0.0669	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
$\eta_1$	0 <sup>a</sup>	0.0397	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>
$\eta_2$	0 <sup>a</sup>	0.0319	0.0123	0 <sup>a</sup>	0 <sup>a</sup>
$\eta_3$	0 <sup>a</sup>	-0.0139	-0.0131	0.0633	0 <sup>a</sup>
$\eta_4$	0 <sup>a</sup>	0.0130	-0.0483	0.0201	0.0464
Estimated $\Psi$					
$\zeta_0$	0.373				
$\zeta_1$	0 <sup>a</sup>	0.194			
$\zeta_2$	0 <sup>a</sup>	0 <sup>a</sup>	0.194		
$\zeta_3$	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0.194	
$\zeta_4$	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0 <sup>a</sup>	0.194

\*Fixed parameter.

The estimated  $\sigma_2$  and  $\sigma_3$  are reported above in table 2 and table 3, respectively, underneath the corresponding values of  $s_2$  and  $s_3$ . We note that in general the fit seems to be reasonably good. The parameter estimates are given in table 4. The  $\lambda$  coefficients do exhibit some differences between the two alternatives of estimation, using the  $\sigma_2$  part and using the  $\sigma_3$  part. The  $\Gamma$  coefficients for the lagged effects (elements below the diagonal) are in certain cases negative, but may be within the sampling variability.

From these results and the covariance matrix of  $x$  we may deduce the estimated covariance matrix for the latent variable constructs  $\eta_0, \eta_1, \eta_2, \eta_3$  and  $\eta_4$ . They are reported in table 5. Note that the variances of  $\eta_t$  ( $t = 1, 2, 3, 4$ ) are rather stable. It is also of interest to note that the variation in these constructs is to a large extent accounted for by the individual, time-invariant component of  $\eta_0$ . The percentage variation of  $\eta_0$  relative to the variation in  $\eta_t$  is estimated as 68.1, 68.1, 68.3 and 67.7, respectively.

Table 5  
The longitudinal model: Estimated covariance matrix  
for  $\eta_0, \eta_1, \eta_2, \eta_3$  and  $\eta_4$ .

$\eta_0$	0.465				
$\eta_1$	0.472	0.683			
$\eta_2$	0.472	0.489	0.683		
$\eta_3$	0.469	0.478	0.478	0.681	
$\eta_4$	0.471	0.483	0.481	0.482	0.687

## References

- Aigner, D.J. and A.S. Goldberger, 1977, Latent variables in socio-economic models (North-Holland, Amsterdam).
- Aitchison, J. and S.D. Silvey, 1957, The generalization of probit analysis to the case of multiple responses, *Biometrika* 44, 131–140.
- Amemiya, T., 1973, Regression analysis when the dependent variable is truncated normal, *Econometrica* 41, 997–1016.
- Amemiya, T., 1978, The estimation of a simultaneous equation generalized probit model, *Econometrica* 46, 1193–1206.
- Amemiya, T., 1981, Qualitative response models: A survey, *Journal of Economic Literature* 19, 1483–1536.
- Amemija, T., 1982, Tobit models: A survey, *CJRS* 4 (Rhodes Associates, Palo Alto, CA).
- Andersen, E.D., 1980, Discrete statistical models with social science applications (North-Holland, Amsterdam).
- Ashford, J.R. and R.R. Sowden, 1970, Multivariate probit analysis, *Biometrics* 26, 535–546.
- Bartholomew, D.J., 1980, Factor analysis for categorical data, *Journal of the Royal Statistical Society B* 42, 293–321.
- Bentler, P.M., 1980, Multivariate analysis with latent variables: Causal modeling, *Annual Review of Psychology* 31, 419–456.
- Bentler, P.M. and J. Tanaka, 1982, Problems with EM algorithms for ML factor analysis, *Psychometrika*, forthcoming.
- Bentler, P.M. and D.G. Weeks, 1980, Linear structural equations with latent variables, *Psychometrika* 45, 289–308.

- Bielby, W.T. and R.M. Hauser, 1977, Structural equation models, *Annual Review of Sociology* 3, 137–161.
- Bishop, Y.M.M., S.E. Fienberg and P.W. Holland, 1975, *Discrete multivariate analysis: Theory and practice* (MIT Press, Cambridge, MA).
- Bliss, C.I., 1935, The calculation of the dosage mortality curve (appendix by R.A. Fischer), *Annals of Applied Biology* 22, 134–167.
- Bock, R.D., 1972, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* 37, 29–51.
- Bock, R.D. and M. Aitkin, 1981, Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika* 46, 443–459.
- Bock, R.D. and L.V. Jones, 1968, *The measurement and prediction of judgement and choices* (Holden-Day, San Francisco, CA).
- Bock, R.D. and M. Lieberman, 1970, Fitting a response model for  $n$  dichotomously scored items, *Psychometrika* 35, 179–197.
- Brown, M.B. and J.K. Benedetti, 1977, On the mean and variance of the tetrachoric correlation coefficient, *Psychometrika* 42, 347–355.
- Browne, M.W., 1974, Generalized least squares estimates in the analysis of covariance structures, *South African Statistical Journal* 8, 1–24. Reprinted in: D.J. Aigner and A.S. Goldberger, eds., 1977, *Latent variables in socio-economic models* (North-Holland, Amsterdam).
- Browne, M.W., 1982, Covariance structures, in: D.M. Hawkins, ed., *Topics in applied multivariate analysis* (Cambridge University Press, Cambridge).
- Carroll, J.B., 1961, The nature of the data, or how to choose a correlation coefficient, *Psychometrika* 26, 347–372.
- Chen, C.-F., 1981, The EM approach to the multiple indicators and multiple causes model via the estimation of the latent variable, *Journal of the American Statistical Association* 76, 704–708.
- Christofferson, A., 1975, Factor analysis of dichotomized variables, *Psychometrika* 40, 5–32.
- Clark, C., 1961, The greatest of a finite set of random variables, *Operations Research* 145–162.
- Dempster, A.P., N.M. Laird and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B* 39, 1–38.
- Divgi, D.R., 1979, Calculation of the tetrachoric correlation coefficient, *Psychometrika* 44, 169–172.
- Edwards, A.L. and L.L. Thurstone, 1952, An internal consistency check for the methods of successive intervals and the method of graded dichotomies, *Psychometrika* 17, 169–180.
- Ferguson, G.A., 1941, The factorial interpretation of test difficulty, *Psychometrika* 6, 323–329.
- Gibbons, R.D., 1981, Full information factor analysis of dichotomous variables, Presented at the Psychometric Society meeting (Chapel Hill, NC).
- Gibbons, R.D. and R.D. Bock, 1982, A probit model for trends in correlated proportions, Presented at the Psychometric Society meeting (Montreal).
- Gumbel, E.J., 1961, Bivariate logistic distributions, *Journal of the American Statistical Association* 56, 335–349.
- Gurland, J., I. Lee and P.A. Dahm, 1960, Polychotomous quantal response in biological assay, *Biometrics* 16, 382–398.
- Heckman, J.J., 1974, Shadow prices, market wages, and labor supply, *Econometrica* 42, 679–694.
- Heckman, J., 1978, Dummy endogenous variables in a simultaneous equation system, *Econometrica* 46, 931–959.
- Henderson, A.S., D.G. Byrne and P. Duncan-Jones, 1981, *Neurosis and the social environment* (Academic Press, Sydney).
- Jaspén, N., 1946, Serial correlation, *Psychometrika* 11, 23–30.
- Jöreskog, K.G., 1971, Simultaneous factor analysis in several populations, *Psychometrika* 36, 409–426.
- Jöreskog, K.G., 1973, A general method for estimating a linear structural equation system, in: A.S. Goldberger and O.D. Duncan, eds., *Structural equation model in the social sciences* (Seminar Press, New York) 85–112.
- Jöreskog, K.G., 1977, Structural equation models in the social sciences: Specification, estimation and testing, in: P.R. Krishnaiah, ed., *Applications of statistics* (North-Holland, Amsterdam).
- Jöreskog, K.G., 1978a, Structural analysis of covariance and correlation matrices, *Psychometrika* 43, 443–477.

- Jöreskog, K.G., 1978b, An econometric model for multivariate panel data, *Annales de l'INSEE*, 30-31, 355-366.
- Jöreskog, K.G. and A.S. Goldberger, 1972, Factor analysis by generalized least squares, *Psychometrika* 37, 243-260.
- Jöreskog, K.G. and A.S. Goldberger, 1975, Estimation of a model with multiple indicators and multiple causes of a single latent variable, *Journal of the American Statistical Association* 70, 631-639.
- Jöreskog, K.G. and D. Sörbom, 1980, Simultaneous analysis of longitudinal data from several cohorts, Research report no. 80-5 (Department of Statistics, University of Uppsala, Uppsala).
- Jöreskog, K.G. and D. Sörbom, 1981, LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods, Research report no. 81-8 (Department of Statistics, University of Uppsala, Uppsala).
- Kirk, D.B., 1973, On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient, *psychometrika* 38, 259-268.
- Lancaster, H.B. and M.A. Hamdan, 1964, Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters, *Psychometrika* 29, 383-391.
- Lawley, D.N., 1943, On problems connected with item selection and test construction, *Proceedings of the Royal Society of Edinburgh* 61, 273-287.
- Lawley, D.N., 1944, The factorial analysis of multiple item tests, *Proceedings of the Royal Society of Edinburgh* 62-A, 74-82.
- Lawley, D.N. and A.E. Maxwell, 1971, Factor analysis as a statistical method (Butterworth, London).
- Lee, L.F., 1979, Health and wage: A simultaneous equation model with multiple discrete indicators, Discussion paper no. 79-127 (University of Minnesota, Minneapolis, MN).
- Lee, S.-Y. and K.-L. Tsui, 1982, Covariance structure analysis in several populations, *Psychometrika*, forthcoming.
- Lillard, L.A. and R.J. Willis, 1978, Dynamic aspects of earning mobility, *Econometrica* 46, 985-1012.
- Lord, F.M., 1980, Applications of item response theory to practical testing problems (Lawrence Erlbaum Ass., Hillsdale, NJ).
- Lord, F.M. and H. Novick, 1968, Statistical theories of mental test scores (Addison-Wesley, Reading, MA).
- Maddala, G.S., 1971, The use of variance components models in pooling cross section and time series data, *Econometrica* 39, 341-358.
- Maddala, G.S. and L.F. Lee, 1976, Recursive models with qualitative endogenous variables, *Annals of Economic and Social Measurement* 5, 525-545.
- Martinson, E.O. and M.A. Hamdan, 1971, Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables, *Journal of Statistical Computation and Simulation* 1, 45-54.
- Mooijart, A., 1982, Factor analysis for ordered categorical variables with discrete and continuous latent variables (Department of Methodology, Psychological Institute, University of Leyden, Leyden).
- Morimune, K., 1979, Comparisons of normal and logistic models in the bivariate dichotomous analysis, *Econometrica* 47, 957-976.
- Muthén, B., 1976a, Structural equation models with dichotomous dependent variables, Research report no. 76-17 (Department of Statistics, University of Uppsala, Uppsala).
- Muthén, B., 1976b, Structural equation models with dichotomous dependent variables: A sociological analysis problem formulated by O.D. Duncan, Research report no. 76-19 (Department of Statistics, University of Uppsala, Uppsala).
- Muthén, B., 1977, Statistical methodology for structural equation models involving latent variables with dichotomous indicators, Unpublished doctoral thesis (Department of Statistics, University of Uppsala, Uppsala).
- Muthén, B., 1978, Contributions to factor analysis of dichotomous variables, *Psychometrika* 43, 551-560.
- Muthén, B., 1979, A structural probit model with latent variables, *Journal of the American Statistical Association* 74, 807-811.
- Muthén, B., 1981a, A general structural equation model with ordered categorical and continuous latent variable indicators (Department of Psychology, University of California, Los Angeles, CA).

- Muthén, B., 1981b, Factor analysis of dichotomous variables: American attitudes toward abortion, in: D.J. Jackson and E.F. Borgatta, eds., *Factor analysis and measurement in sociological research: A multidimensional perspective* (Sage Publications, London).
- Muthén, B., 1982a, Some categorical response models with continuous latent variables, in: K.G. Jöreskog and H. Wold, eds., *Systems under indirect observation: Causality, structure, prediction* (North-Holland, Amsterdam).
- Muthén, B., 1982b, LACCI: Latent variable analysis with dichotomous, ordered categorical, and continuous indicators — An experimental program for researchers, Unpublished.
- Muthén, B. and A. Christofferson, 1981, Simultaneous factor analysis of dichotomous variables in several groups, *Psychometrika* 46, 407–419.
- Olsson, U., 1979a, Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika* 44, 443–460.
- Olsson, U., 1979b, On the robustness of factor analysis against crude classification of the observations, *Multivariate Behavioral Research* 14, 485–500.
- Olsson, U., F. Drasgow and N.J. Dorans, 1981, *The polyserial correlation coefficient* (University of Illinois, Urbana, IL).
- Pearson, K., 1900, Mathematical contributions to the theory of evolution, VII: On the correlation of characters not quantitatively measurable, *Philosophical Transactions of the Royal Society of London A* 195, 1–147.
- Pearson, K., 1904, Mathematical contributions to the theory of evolution, XIII: On the theory of contingency and its relation to association and normal correlation, *Drapers Company Research Memoirs, Biometric Series*, no. 1.
- Pearson, K., 1913, On the measurement of the influence of 'broad categories' on correlation, *Biometrika* 9, 116–139.
- Pearson, K. and E.S. Pearson, 1922, On polychoric coefficients of correlation, *Biometrika* 14, 127–156.
- Rubin, D.B. and D.T. Thayer, 1982, EM algorithms for ML factor analysis, *Psychometrika* 47, 69–76.
- Samejima, F., 1969, Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph* no. 17.
- Samejima, F., 1972, A general model for free-response data, *Psychometrika Monograph* no. 18.
- 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, 229–239.
- Sörbom, D., 1978, An alternative to the methodology for analysis of covariance, *Psychometrika* 43, 381–396.
- Sörbom, D., 1982, Structural equation models with structured means, in: K.G. Jöreskog and H. Wold, eds., *Systems under indirect observation: Causality, structure, prediction* (North-Holland Amsterdam).
- Tallis, G.M., 1962, The maximum likelihood estimation of correlation from contingency tables, *Biometrics* 18, 342–353.
- Tobin, J., 1958, Estimation of relationships for limited dependent variables, *Econometrica* 26, 24–36.