Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), Testing Structural Equation Models (pp. 205-243). Newbury Park, CA: Sage. (#45)

9

Goodness of Fit With Categorical and Other Nonnormal Variables

BENGT O. MUTHÉN

With continuous-normal variables, the testing of a structural model involves testing the fit of the restrictions imposed on the covariance matrix. Testing of structural models for categorical data involves additional considerations compared with those for normal data. Since these considerations are similar to testing with nonnormal continuous variables, the case of nonnormal continuous variables will also be discussed in this chapter and will thereby provide a convenient connection with the continuous-normal case discussed in other chapters.

For continuous-normal variables, the sample covariance matrix is a natural choice of statistic to analyze, and testing the model involves testing restrictions on the covariance matrix. Structural models for categorical data have no such natural choice of sample statistics, and testing the model can be done in more than one way. The structural models for categorical data considered here use a model specification that can be expressed as a certain model family for the observed variables, where the parameters of this model family are further restricted in terms of a smaller number of structural parameters. As an example, we may consider a model with four binary variables. With an assumption of nor-

AUTHOR'S NOTE: This research was supported by Grant AA 08651-01 from NIAAA for the project Psychometric Advances for Alcohol and Depression Studies and by National Science Foundation Grant SES-8821668. Albert Satorra provided stimulating discussions. I thank Ginger Nelson and Kathleen Wisnicki for helpful research assistance.

mality for continuous variables underlying the observed variables, the natural statistics to analyze are the correlations among the underlying variables, the tetrachoric correlations. As a structural model for the correlations among the underlying variables we may consider a standard one-factor model. In this example, the model family for the observed variables is one that specifies multivariate normality for the underlying variables. Only if this model family is accepted should tetrachorics be used. Having accepted the model family, the testing of the structural model proceeds with testing the restrictions it imposes on the tetrachorics. Since in this way variances are not considered, the usual residual variance parameters of covariance structure analysis are disregarded. In this way, the degrees of freedom for p variables are obtained as $p(p-1)/2$, the number of parameters in the unrestricted model, minus the number of free parameters. With four variables, the unrestricted model has as many parameters as there are correlations, in this case six, while a one-factor model uses four parameters to generate the correlations. In this way, the structural model implies that two restrictions are imposed on the correlations. Note, however, that the choice of model family, and hence of statistics to analyze, is not unambiguous. The use of the tetrachoric family implies accepting a model for the probabilities of the cells of the $2 \times 2 \times 2 \times 2$ table for the four binary variables. Testing of this model will be discussed below. It is interesting to contrast the incorrect acceptance of the normality assumption of the tetrachoric family with the incorrect use of normal theory estimators and tests for continuous, nonnormal variables. In the latter case, it is well known that the estimators will produce consistent estimates of structural model parameters and that only chi-square tests and standard errors are affected. In the former case, however, using the wrong model family results in using the wrong statistics, and inconsistent estimation of structural model parameters results.

In testing structural models for categorical data one may therefore consider two levels of goodness of fit: (a) the fit of the observed variable model family to the observed data where no further restrictions of model family parameters are made and (b) the fit of the structural model to the model family. As will be discussed, it is often hard to test the structural model directly against observed categorical data, and the two levels of fit may instead be considered in these two steps. The second level of testing corresponds most closely to the conventional testing for continuous variables. The first level of testing has no counterpart in normal theory analysis and it is unfortunately often ignored in structural analysis of categorical data. Examples of the first level of testing will be discussed in terms of logit and probit regression for binary dependent variables, regression with ordered polytomous response, analysis with polychoric correlation coefficients, analysis with tetrachoric correlations, and analysis with mixtures of continuous and categorical variables. Examples of the second level of testing will be discussed only briefly in terms of factor analysis and structural equation modeling of binary variables. Structural modeling with latent variables runs the risk of removing interpretations of the model too far from the observed data; this is perhaps particularly true for categorical data. A final section discusses how this can be avoided by complementing the two levels of testing by explicating predictions from the structural model for the observed data.

A general structural equation model will first be presented. Checking the appropriateness of model families against the data will then be discussed, followed by testing of underlying structure and how this structure can be tied back to the data.

The Structural Model

As in the LISCOMP program (Muthén, 1978, 1979, 1983, 1984, 1987, 1989b), a variety of response models for categorical and other nonnormal data can be put into a unifying framework by the use of latent continuous response variables. These y^* variables may be observed as y variables in a variety of forms. In the conventional case, the y^*s may be directly observed as continuous-unlimited y variables, whether they are normal or nonnormal (Muthén, 1989c). The y^* variables may also be observed as censored continuous variables (Muthén, 1987, 1989d) and as ordered categorical variables, including binary variables. Recently, such modeling has also been provided within the framework of Jöreskog and Sörbom's (1988, 1989) PRELIS and LISREL programs. Thus the estimation and testing procedures to be discussed are widely available.

This chapter will focus on the ordered categorical case,

$$y = c, \qquad \text{if } \tau_c < y^* < \tau_{c+1}, \tag{1}$$

where y^* is an underlying continuous variable with thresholds τ as parameters for the categories $c = 0, 1, 2, \ldots, C-1$, for a variable with C categories, and where $\tau_0 = -\infty$, $\tau_C = \infty$. In the binary case, there is a single τ parameter and it is equivalent to the negative of the intercept of conventional probit regression. If continuous variables y are present, this is taken to mean that the underlying variable is directly observed, $y = y^*$.

The LISCOMP model involves a measurement model

$$y^* = \Lambda\eta + \epsilon, \qquad [2]$$

and a structural model

$$\eta = B\eta + \Gamma x + \zeta. \qquad [3]$$

Here, ϵ is a random vector of measurement errors with zero means and covariance matrix θ_ϵ, η is a random vector of constructs with zero means, x is a random vector of observed background variables, and ζ is a random vector of residuals with zero means and covariance matrix Ψ. The vectors ϵ and ζ are assumed to be independent of each other and of η. For simplicity, arrays related to intercepts and means are left out in the present discussion and so are multiple-group specifications and scaling issues. The model parts of Equations 2 and 3 lead to the model structure for the conditional distribution of y^* given x,

$$E(y^* | x) = \Pi x, \qquad [4]$$

$$V(y^* | x) = \Omega, \qquad [5]$$

where

$$\Pi = \Lambda(I - B)^{-1}\Gamma, \qquad [6]$$

$$\Omega = \Lambda(I - B)^{-1}\Psi\Lambda'(I - B)^{-1\prime} + \Theta_\epsilon. \qquad [7]$$

The Π structure constitutes LISCOMP's model part 2, while the Ω structure constitutes LISCOMP's model part 3. Mean and threshold structures would be included in part 1. When categorical y variables are present, the corresponding diagonal elements of Ω are not identifiable and are not included in the LISCOMP analysis. In the modeling such diagonal elements are taken to be unity. For categorical variables, the PRELIS/LISREL system also involves the Ω structure, but the threshold structure of part 1 and the Π structure of part 2 have no counterparts in the current PRELIS/LISREL system.

For categorical and other nonnormal y variables, generalized least squares (GLS) estimation is used to estimate the model parameters in LISCOMP and PRELIS/LISREL (see also Jöreskog, 1991). LISCOMP has the special feature that when x variables are present, GLS estimation is done by fitting Π and Ω elements to the vector of sample statistics corresponding to the multivariate regression slopes and residual covariances (see Muthén, 1987). The advantage is that the calculation of these regression-based statistics draws only on the assumption of conditional normality for y^* given x, while the customary use of correlations would entail joint normality of y^* and x. When x variables are not present, the conditioning on x is vacuous, the Π matrix does not exist, and the Ω elements are fitted to a sample statistics matrix of underlying variable correlations and/or covariances. This includes tetrachoric, polyserial, and polychoric correlations as well as correlations and covariances for censored variables. In all cases, the sample statistics are calculated in two steps using univariate and bivariate response variable information and maximum likelihood (ML) estimation.

Testing of the First Level: The Fit to the Data

In this section the fitting of the model is related to the sampling scheme under which the data were observed. For categorical data, product-multinomial or multinomial sampling schemes are relevant (see, e.g. Agresti, 1990; Bock, 1975, chap. 8). The just-identified, or saturated, model then has the corresponding probabilities as parameters and the ML estimates are the observed proportions. Checking the model involves an investigation of how these estimated probabilities are reproduced by our model. The general point of this section is that this first-level model testing stage is often ignored, despite the fact that a host of model checking techniques are available. By means of a series of examples, this testing is shown to contribute crucially to the understanding of the data. Examples of the first level of testing will be discussed in terms of logit and probit regression for binary dependent variables, regression with ordered polytomous response, analysis with polychoric correlation coefficients, analysis with tetrachoric correlations, and analysis with mixtures of continuous and categorical variables.

Regression With a Binary Dependent Variable

The first two examples discuss regression analysis with categorical dependent variables. As a starting point, consider the simple case of regression of a binary dependent variable y on an x variable. In terms of the general model of the previous section, probit regression is obtained with $\boldsymbol{\Lambda} = \boldsymbol{I}$, $\boldsymbol{\Theta}_{\epsilon} = \boldsymbol{0}$, $\boldsymbol{B} = \boldsymbol{I}$, so that

$$y^{*} = \gamma x + \zeta\,, \qquad [8]$$

obtaining the ML-estimated probit slope of $\boldsymbol{\Pi}$ in Equation 6 as γ. The variance of the residual ζ is standardized to one so that $\boldsymbol{\Omega}$ of Equation 7 is the scalar 1. The model expresses the conditional probability of y given x as the probability that y^{*} exceeds the threshold τ_c in Equation 1,

$$P(y = 1 \mid x) = \int_{\tau - \gamma x}^{\infty} \varphi(t)dt\,, \qquad [9]$$

where φ denotes the univariate standard normal density. Equivalently, conventional probit regression parameterization expresses the negative of τ as an intercept, while γ is the conventional slope.

$$P(y = 1 \mid x) = \Phi(-\tau + \gamma x) \qquad [10]$$

$$= \Phi(\alpha + \beta x)\,,$$

where Φ is the standard normal distribution function.

Example 1: Probit Regression

Table 9.1 gives British coal miner data taken from Ashford and Sowden (1970). The x variable is age and the binary y variable is breathlessness. This is a case of grouped data in the sense that each distinct x value in the sample has more than a single observation. The sampling scheme may be considered as product-binomial so that the conditional probabilities of y given x are modeled. There are 9 different x values and for each x value a binomial variable is observed. Hence the unrestricted H_1 model for the data has one parameter, a probability, for each of 9 x values, giving a total of 9 parameters. In contrast, the linear probit

model has two parameters, τ and γ. The latter model is nested within the former. This can be seen by considering a transformation of the 9 probability parameters π_j, $j = 1, 2, \ldots, 9$, into 9 (probit) parameters, z_j, where $\pi_j = \Phi(z_j)$: The probit model restricts the 9 z_js to be a linear function of x. In this way, a Pearson or likelihood ratio chi-square test of fit has 7 degrees of freedom. If a multinomial sampling scheme is instead considered, the result is the same. The unrestricted model then has 17 parameters because there are 9 x 2 cells of probabilities and these have to add to one. As is the case in log-linear modeling, 8 of these parameters correspond to the marginal distribution of x and should be added to the H_0 model. However, in line with ordinary regression, the marginal distribution of x is not restricted here. The likelihood ratio chi-square value is 5.19 with 7 degrees of freedom and the model is not rejected despite the huge sample size. Note that this may be considered a test against the data of the probit model family for the relationship between y and x. The corresponding test of the logit family results in a likelihood ratio chi-square of 17.13, which is not significant on the 1% level. Table 9.1 also gives the fitted, or predicted, probabilities of breathlessness for each x value. It can be seen that the probit family captures the observed proportions better than the logit family at low and high x values. In Example 1 there is no further structure imposed on the linear probit model parameters, but this could be envisioned as a case

Table 9.1 Example 1: British Coal Miner Data

Age (x)	*N*	*N Yes*	*Proportion Yes*	*Probit Estimated Probability*	*Logit Estimated Probability*	*OLS Estimated Probability*
22	1,952	16	0.008	0.009	0.013	−0.053
27	1,791	32	0.018	0.018	0.022	−0.004
32	2,113	73	0.035	0.034	0.036	0.045
37	2,783	169	0.061	0.060	0.059	0.094
42	2,274	223	0.098	0.100	0.095	0.143
47	2,393	357	0.149	0.156	0.148	0.192
52	2,090	521	0.249	0.231	0.225	0.241
57	1,750	558	0.319	0.322	0.327	0.290
62	1,136	478	0.421	0.425	0.448	0.339
	18,282	2,427	0.130			

SOURCE: Data from Ashford and Sowden (1970).

of testing $\gamma = 0$, or equality of γ slopes with more than one x variable. Such a test can be performed using as the alternative hypothesis the probit/logit model with unrestricted slopes. In this way, the test is done on the second level without involving the unrestricted multinomial model.

If data are not grouped as in Example 1, model testing against the data is more difficult, because the chi-square approximation may be poor, with many cells having zero or very low expected frequencies. This is the more common case and illustrates the difficulty of testing the categorical variable model against the data directly. Standard computer packages offer a "model test" also in this case, but it refers to the H_0 hypothesis of γs all being zero tested against the H_1 hypothesis of the γs not being zero. Such a test is what is here termed a *second-level test.* Although it is interesting to know that your predictors have significant influence on y, this procedure does not offer the desired test of the probit/logit family against the data. For ungrouped data, Agresti (1990) discusses more suitable goodness-of-fit tests related to residuals.

Regression With an Ordered Polytomous Dependent Variable

Example 2: Ordered Polytomous Regression

Muthén (1987) considered an example of alcohol consumption where y corresponds to the number of drinks a person has per day on average and the xs are age and income. The y categories are 0 (nondrinker), 1 (1-2 drinks per day), 2 (3-4 drinks per day), and 3 (5 or more drinks per day). A U.S. general population sample of 713 males with regular physical activity levels was considered. In this example there are four ordered response categories, where

$$P(y=0 \mid x) = \Phi(\tau_1 - \gamma' x)\,, \qquad [11]$$

$$P(y=1 \mid x) = \Phi(\tau_2 - \gamma' x) - \Phi(\tau_1 - \gamma' x)\,,$$

$$P(y=2 \mid x) = \Phi(\tau_3 - \gamma' x) - \Phi(\tau_2 - \gamma' x)\,,$$

$$P(y=3 \mid x) = \Phi(-\tau_3 + \gamma' x)\,.$$

The arguments of Φ are called (population) probits and are linear in the xs. For example, $(-\tau_3 + \gamma' x)$ is the probit for $P(y = 3 \mid x)$. The conditional

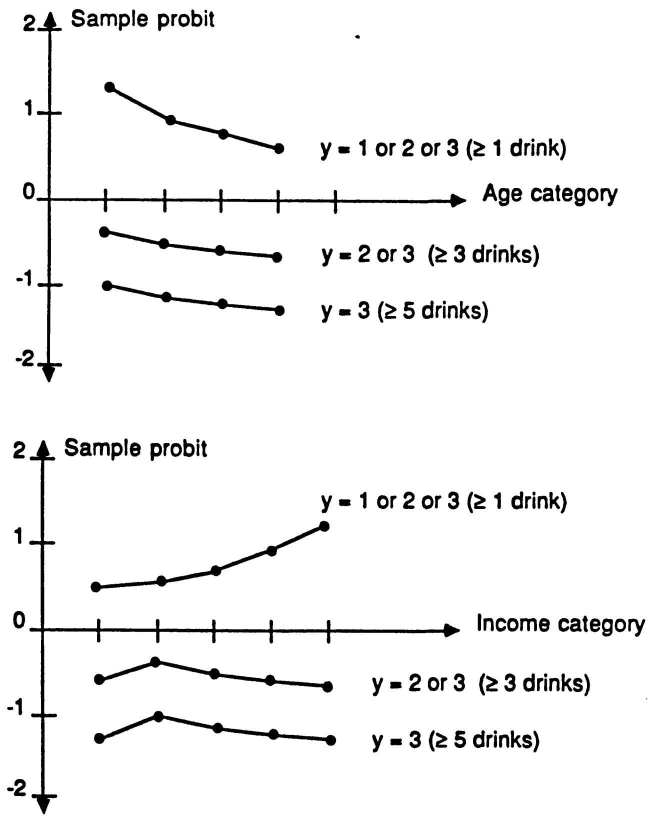

Figure 9.1. Example 2: Sample probit plots for alcohol data.

probabilities of the response categories imply that the probits are linear in the xs when the probabilities for the following three events are considered: $y = 3$, $y = 2$ or 3, $y = 1$ or 2 or 3. As an example, consider the event $y = 2$ or 3. Noting that $1 - \Phi(z) = \Phi(-z)$, the probit for this event is $-\tau_2 + \gamma' x$. The probits for these three events also have the same x slopes γ. These facts can be used to test the goodness of fit for the family of ordered four-category probit models against the data. Using grouped data, the corresponding sample probits based on observed proportions can be plotted against the x variables. Figure 9.1 shows the plots where age has been categorized into four categories and income has been categorized into five categories.

The top panel of Figure 9.1 shows three probit functions corresponding to the three events plotted against age. If the ordered probit model is correct, these functions should be linear and parallel, and this seems to hold to a reasonable degree of approximation. For income, however, the bottom panel indicates that the lines are not parallel. The probit for drinking at all appears to have a positive slope, while larger amounts of drinking appear to have zero probit slopes. These plots suggest that the ordered polytomous probit model is inappropriate for these data. Note that the plots should more properly be done for each variable while conditioning on the other, using, say, the modal category. This, however, will drastically reduce the sample sizes used for the plots, and the distortion may not be large if the xs are approximately normally distributed. ML estimation of the model results in a strongly significant negative slope for age and a weakly significant positive slope for income. If instead the y variable is dichotomized as nondrinker versus drinker, the age slope is about the same and the income slope becomes strongly significant, with a large positive value. The conclusion is that while drinking is strongly related to income, the amount of drinking is not. This outcome is predictable from the probit plots. The example shows the value of testing the model family against the data. If one goes ahead and uses the ordered polytomous probit model for these data, misleading conclusions can be drawn.

Regression With Several Dependent Variables

With multivariate categorical dependent variables, the testing and model checking becomes more complex. As an example of the increasing complexity, consider the case of bivariate, binary response. As a special case of the general model of Equations 2 and 3, Muthén (1979) studied a bivariate probit model with two indicators of a single factor, where the factor was regressed on a set of exogenous variables. Leaving the fit of the underlying latent variable structure aside, checking the appropriateness of the family of bivariate probit models against the data involves checking a 2 × 2 table for each distinct combination of values on the x variables. Unless data are grouped and a large sample is available, this is intractable. More informal model checking is, however, possible in line with the probit plots of the previous example. For bivariate binary responses, a probit plot is first carried out for each variable and checked for linearity in line with the previous example. In addition, a probit plot needs to be inspected for the joint event $y_1 = 1$, $y_2 = 1$. This requires transforming the sample proportion by use of a bivariate normal distribution function..

Correlations Between Ordered, Polytomous Variables

Multivariate response with a structural model is a special case of the general model of Equations 2 and 3 where xs are not present and $\boldsymbol{\Gamma} = \boldsymbol{0}$. Here, the assumption of conditional normality for y^*s given xs is replaced by the assumption of normality of the y^*s themselves. In particular, this occurs when the latent variables of η as well as the measurement errors ϵ are normal. Assuming normally distributed y^* variables and using only bivariate information from pairs of variables leads to the analysis of latent correlations such as tetrachorics, polychorics, and polyserials, as is done in LISCOMP and PRELIS/LISREL. The use of such correlations implies the acceptance of a model in itself, the model of underlying normality of y^*s. This is different from the case of continuous variables where Pearson product-moment (PPM) correlations are used. For categorical variables, underlying y^* variables have to be assigned a distribution, although not necessarily normal (see, e.g., Jöreskog, 1991), in order to enable the estimation of latent correlations. For continuous variables, a linear relationship between the two variables is assumed for a PPM correlation, but the full distribution of the two variables need not be given. If the two variables are not bivariate normal this does not invalidate the use of the PPM correlation as it does the latent correlation. The next two examples consider the goodness of fit of underlying normality models to data.

Example 3: Polychoric Correlations

Muthén, Huba, and Short (1985) analyzed quality of life data measured on seven-category Likert scales for 1,814 individuals. Questions referred to satisfaction with a person's house, leisure, family life, standard of living, and savings, among other factors. The scale steps ranged from "very satisfied" to "very dissatisfied," with the fourth category being neutral. Underlying normality for a pair of variables can be assessed by a chi-square test. With seven-category variables the pairwise chi-square tests have 35 degrees of freedom, obtained as the difference of 48 parameters in the unrestricted multinomial model for the 49 cells and 6 + 6 + 1 = 13 parameters in the normality model (6 thresholds for each variable and 1 correlation). The latter model may

Table 9.2 Example 3: Polychoric Versus Pearson Correlations: Quality of Life Data (N = 1,814; seven-category Likert)

	Pairwise Chi-Square Tests of Normality (35 df)				
	Neighborhood	*House*	*Leisure*	*Family*	*Standard*
House	160.7				
Leisure	117.8	145.7			
Family	73.7	88.2	197.5		
Standard	115.0	173.3	181.3	123.2	
Savings	93.6	139.8	120.1	92.0	173.7
		Correlations:	*Pearson* *Polychoric*		
House	.454 .515				
Leisure	.249 .309	.285 .335			
Family	.188 .249	.230 .291	.376 .437		
Standard	.323 .373	.384 .431	.376 .432	.305 .355	
Savings	.226 .261	.296 .336	.330 .373	.287 .330	.578 .632

SOURCE: Data from Muthén, Huba, and Short (1985).

be seen as nested within the former as follows. The unrestricted model parameters can be transformed into bivariate probit parameters. The normality model restricts these parameters so that they increase or decrease monotonically both horizontally and vertically in the table, reflecting the ordered nature of the pair of variables at hand.

The top panel of Table 9.2 shows the results of this chi-square testing using Pearson chi-squares. In my experience, the rejections of the normality model observed in this example are frequently found, but these rejections are often to a large extent caused by cells with low expected frequencies (see, e.g., Benson & Muthén, 1992). For cells with large enough numbers, however, there are often interesting information and ideas to be found in such testing. A particularly common outcome is one where "outlier" responses, or responses by a heterogeneous subpopulation, contribute to the rejection. In the quality of life example, certain individuals are very satisfied with one aspect of life and very dissatisfied with another, related, aspect. An example is given in Table 9.3 for the pair of variables Neighborhood and Savings. Consider first cell 1, 7, in the top panel. Here, 86 individuals are very satisfied with their neighborhoods and very dissatisfied with their savings. One may suspect either that certain individuals have made sloppy responses or that for these individuals the two matters have become negatively related, perhaps because of a recent house purchase. Consider next cell 3, 3, corresponding to a point on the response scale that is one step away from neutral toward the satisfied end. While 64 individuals respond in this way, the model estimates a lower number of 43. Perhaps the higher number of observed individuals can be explained by a noncommittal response style, where the individual perhaps responds to most questions

Table 9.3 Quality of Life, Neighborhood, Savings

				Observed Table			
	1	2	*3*	*4*	*5*	*6*	*7*
1	134	139	111	118	64	52	$\boxed{86}$
2	37	95	97	75	76	43	62
3	13	26	$\boxed{64}$	39	59	19	34
4	15	17	26	43	34	34	34
5	2	10	12	16	21	14	17
6	0	4	7	5	7	9	6
7	3	1	2	6	9	6	11
				Expected Table (rounded)			
1	115	139	135	113	90	52	$\boxed{59}$
2	49	77	87	84	75	48	64
3	20	35	$\boxed{43}$	44	42	29	42
4	13	24	32	34	34	25	40
5	4	9	13	15	16	12	22
6	2	3	5	6	7	5	10
7	1	3	4	6	7	6	12
				Chi-Square Elements (rounded)			
1	3	0	4	0	8	0	$\boxed{13}$
2	3	4	1	1	0	1	0
3	2	2	$\boxed{11}$	1	7	3	2
4	0	2	1	2	0	3	1
5	1	0	0	0	2	0	1
6	1	0	1	0	0	3	2
7	3	1	1	0	1	0	0

in a somewhat positive way although he or she has different true feelings. In any case, the bottom panel shows that these two cells contribute in a major way to the rejection of the underlying normality model. It is possible that retaining the model and using its polychoric correlation gives a smoothed estimate of association that is less influenced by outlier behavior than, say, PPM correlations. Such speculations can be made plausible only by in-depth analysis using additional information from other variables. A comparison of regular PPM correlations and polychoric correlations is given in Table 9.2. The testing of underlying normality in conjunction with polychoric correlations can be done in both LISCOMP and PRELIS/LISREL.

Correlations Between Binary Variables

While testing of underlying normality is possible for a pair of polytomous variables, the case of a 2 × 2 table results in a just-identified model that cannot be tested. As suggested in Muthén and Hofacker (1988), this fact may have contributed to the long-standing debate about whether or not to use tetrachoric correlations. Usually, parameter estimates, such as tetrachorics, are used only when the model from which they are derived fits the data well. But the customary way of computing tetrachorics from 2 × 2 tables does not provide such a test. In principle, one could attempt to use full information from the 2^p cells, but, as already noted, this leads to problems of small cell frequencies. Muthén and Hofacker propose a compromise using information from three variables at a time. The unrestricted model for a 2 × 2 × 2 table has 7 probability parameters. Since the trivariate normality model has three threshold and three correlation parameters, this gives a single-degree-of-freedom chi-square test of underlying normality. This triplet testing approach is suggested by Muthén and Hofacker (1988) for assessing the suitability of using tetrachoric correlations for dichotomous data. This technique is currently available only in LISCOMP.

Example 4: Triplet Testing of Tetrachorics

Consider panel data for a dichotomous variable observed at three time points. This example relates to attempts by Alwin (1992) to establish a quality declaration of attitudinal items by studying their correlation over time. This longitudinal model is shown in Figure 9.2 using the notation of the general framework of Equations 2 and 3. As in this framework, the paths involving the continuous variables of y^* and η are linear regression among latent variables. In line with Heise (1969), Alwin (1989) considered the case of standardized factors η and denoted the slope of η_2 on η_1 and the slope of η_3 on η_2 as stability coefficients. Standardizing the y^* variances to unity, the squared slopes (loadings) of the regressions of each y^* on the respective η then corresponds to the reliability of that y^* variable. Since y can be viewed as a crude measurement of y^*, this reliability may be viewed as the maximum attainable reliability under optimal measurement circumstances. In addition, the size of this reliability is directly related to how precisely the stability coefficients can be estimated. Following Heise's approach of assuming time-invariant reliabilities, the common reliability may be obtained as an estimate of the ratio $\rho_{21}\rho_{32}/\rho_{31}$, where in this case ρs denote correlations between the y^*s, estimated as the tetrachoric correlations for the ys. Under this structural model, there are three parameters: the common loading and the two stability coefficients. The structural model measurement error and residual variances are not free parameters, but are restricted to yield unit y^* and η variances. The structural model is therefore just-identified in terms of the three correlations. In this way, the second-level testing of the structure is not possible. The first-level

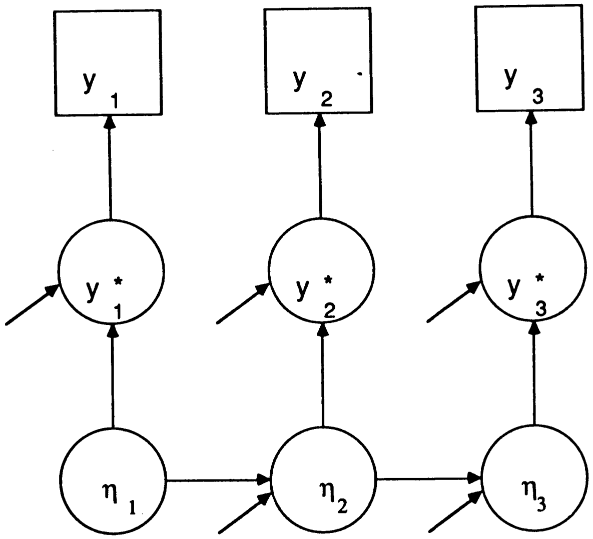

Figure 9.2. Example 4: Three-wave panel model for binary variables.

testing of underlying normality of the y^* variables, however, is possible using the Muthén-Hofacker triplet testing approach.

As an illustration, data were generated according to this model using stabilities of 0.7 and reliabilities of 0.64. A "population" was created by using LISCOMP to generate 100,000 trivariate normal y^* observations that were dichotomized at 0.25 to give 40% "1" responses at each time point. A random sample of 1,500 observations was then drawn from this population. For this sample, the one-degree-of-freedom likelihood ratio chi-square value was obtained as 0.003 with an estimated reliability of 0.64. Hence the sample reflects the population very well in that normality is not rejected and the true reliability value is obtained. The true situation is then distorted in order to violate underlying normality. This is done by assuming that after the first wave, 20% of those with true minority opinions ($y^* > \tau$) change their responses to the majority opinion ($y = 0$). In this way, 20% of the response pattern frequencies for 101, 110, and 111 are changed to response pattern 100. The triplet test for this new data set obtained a likelihood ratio chi-square value of 7.87, which exceeds the 1% critical value of 6.64 so that the tetrachoric model is rejected. The rejection is fortunate, since for these distorted data the estimated reliability is 0.93, which is a severe overestimation of the true value of 0.64. If only 10% are changed, the chi-square value is 2.24, so that the model is not rejected at the 5% level (critical value 3.84). In this case, however, the estimated reliability of 0.74 is closer to the true value.

This artificial example suggests that lack of first-level goodness of fit can have important consequences for inference on structural parameters. Very little is known about the interaction between first-level misfit and structural modeling, however. Muthén and Hofacker (1988) performed triplet testing of data on attitudes toward abortion. For six abortion items, there were 20 triplets to be tested. Four of the 20 triplets were rejected at the 5% level from a sample of 3,921. One item, RAPE, was involved in all four triplets. Factor analysis of tetrachoric correlations using all six items versus five items deleting RAPE gave very similar results, however. Muthén and Hofacker (1988) also provided an artificial example where response consistency effects were generated for the last three of a set of 12 items following a one-factor model. Triplet testing correctly identified the misfitting items. Incorrectly using tetrachoric correlations for all 12 items led to a two-factor model where the last three items were incorrectly taken to measure a second factor.

Correlations Between Ordered Polytomous and Continuous-Normal Variables

Given a mixture of ordered categorical variables and continuous variables, both polychoric and polyserial correlations are computed. For polyserial correlations, the underlying normality assumption pertains to the bivariate distribution of the y_j^* for the categorical variable and the continuous y_k variable. This case has been studied by Jöreskog (1985). Testing of underlying bivariate normality is available in PRELIS, but not in LISCOMP. Traditionally, polyserial correlations have been estimated by considering the product of the likelihood for the continuous y_k variable and the likelihood of the conditional distribution of y_j^* given y_k. Jöreskog (1985) points out that there is a strong computational advantage to considering instead the product of the marginal distribution of y_j and the conditional distribution of y_k. Approaching the estimation in this way, the individual data points on y_k are not needed because the means and variances of y_k for each category of y_j are sufficient statistics. This approach also provides a likelihood ratio chi-square test of underlying bivariate normality that is readily interpretable as follows.

The model of bivariate normality has $C + 2$ parameters: $C - 1$ threshold parameters (for a y_j with C categories), one mean and one variance parameter for the continuous y_k, and one correlation between y_j^* and y_k. This model expresses the C conditional means and C conditional variances for y_k given each y_j category in terms of only three parameters (the mean, the variance, and the correlation). In contrast, the unrestricted model uses the $2C$ mean and variance parameters, which together with the $C - 1$ probability parameters for the marginal distribution of y_j, gives a total of $3C - 1$ parameters. It follows that the chi-square test has $3C - 1 - (C + 2) = 2C - 3$ degrees of freedom (Jöreskog, 1985). The test is clearly applicable also when the categorical variable is binary, yielding 3 degrees of freedom.

To conclude the section on first-level testing of the observed variable model family against the data, one may safely say that much more research is warranted. First of all, not all cases have been covered. As an example, censored variables and mixtures of such variables and categorical variables have not been covered. Second, and perhaps more important, there is almost no research on how goodness-of-fit assessment on the first level interacts with structural inference. It is largely unknown how serious the consequences of lack of fit on the first level are for estimation and testing of structural models. Nevertheless, the techniques that are available today appear to be underutilized in terms of providing insights about the data.

Testing of the Second Level: The Structural Model Fit

This section discusses inferential procedures for structural models not only in cases with categorical variables, but also in those with continuous, nonnormal variables. It will be shown that testing techniques for nonnormal continuous variables are closely related to those for categorical data; this development ties the present discussion to that of other chapters for continuous variables. For simplicity, the general modeling framework of Equations 2 and 3 will be considered for the special case of no *x*s. The case of binary variables will be emphasized as an illustration of the ideas, focusing on the fitting of tetrachoric correlations. However, the general discussion carries over to polytomous and censored variables as well as to models with *x*s and fitting models to regression statistics.

Testing of a structural model in terms of summary sample statistics such as correlations involves the use of test statistics that under certain conditions are asymptotically distributed as chi-square. As is well known, such testing often leads to rejection of a hypothesized model. A primary suggested cause of this is overwhelming power due to large sample size. Many alternative fit indices have been proposed and some are discussed elsewhere in this book. In my opinion, however, there are good reasons for not discarding the chi-square approach for a test of overall model fit. Power issues can be directly addressed (see Saris & Satorra, Chapter 8, this volume). In situations where power is difficult to assess, there are practical ways of checking ill effects of large sample sizes. First, respecifications of the model as suggested by modification indices can be carried out until a nominally well-fitting model is approached. Second, the importance of these respecifications can be checked in terms of practical significance of the new parameter estimates and in terms of the change in the estimates of the original parameters of central interest. If, practically speaking, these changes are not large, then the original fit may be deemed sufficient. There are two important caveats to this approach. One is that the researcher must have started out in a modeling framework that is close enough to the true model for model modifications to be able to lead in the right direction. Second, the chi-square value itself must be trustworthy. It is this second point on which the following discussion centers.

In Equation 5, the covariance matrix of the y^* variables is expressed in terms of structural model parameters. In the continuous case, the y^* variables are directly observed as *y*s and the sample matrix is a conventional covariance or correlation matrix. In the binary case, the y^* variables are indirectly observed by the *y*s and the sample matrix is a tetrachoric correlation matrix. If the continuous variables are not multivariate normal, the use of the sample covariance matrix to fit the model represents a limited information estimation approach. This is also the case when using tetrachoric correlations. In both cases, the sample matrix is created using only bivariate information from pairs of variables.

Estimation and Testing for Continuous Variables: A Brief Overview

Under the conventional assumption of IID observations on a *p*-variate vector $\boldsymbol{y}$ and assuming normality for $\boldsymbol{y}$, the sample covariance matrix $\boldsymbol{S}$ contains sufficient statistics for estimating the structural model parameters of $\boldsymbol{\theta}$, say. In this case, two common fitting functions are normal theory maximum likelihood (NTML) and normal theory GLS (NTGLS),

$$F_{\mathrm{NTML}} = \ln|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - \ln|\boldsymbol{S}| - p\,, \qquad [12]$$

$$F_{\mathrm{NTGLS}} = \mathrm{tr}[(\boldsymbol{\Sigma} - \boldsymbol{S})\boldsymbol{S}^{-1}]^2\,, \qquad [13]$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix for $\boldsymbol{y}$. The expression for F_{NTGLS} is a special case of the general weighted least squares fitting function

$$F_{\mathrm{WLS}} = (\boldsymbol{s} - \boldsymbol{\sigma})' \boldsymbol{W}^{-1}(\boldsymbol{s} - \boldsymbol{\sigma})\,, \qquad [14]$$

where $\boldsymbol{s}$ and $\boldsymbol{\sigma}$ refer to the $p(p + 1)/2$ vectors of distinct elements of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$, respectively. If in Equation 14 $\boldsymbol{W}$ is taken as a consistent estimator of the asymptotic covariance matrix of $\boldsymbol{s}$, then F_{WLS} is referred to as a generalized least squares estimator or minimum chi-square analysis (Ferguson, 1958; Fuller, 1987, sec. 4.2). In the special case of multivariate normality for $\boldsymbol{y}$, the asymptotic covariance matrix of $\boldsymbol{s}$ has a particularly simple structure, depending only on second-order moments,

$$2\boldsymbol{K}_p'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{K}_p\,, \qquad [15]$$

where $\boldsymbol{K}_p'$ is given in Browne (1974, p. 210) and ⊗ denotes the Kronecker product. A consistent estimator of Equation 15 is obtained by replacing $\boldsymbol{\Sigma}$ with $\boldsymbol{S}$. This simplification of Equation 14 leads to Equation 13.

For arbitrary distributions, the asymptotic covariance matrix of $\boldsymbol{s}$, $\boldsymbol{\Gamma}$, say, has elements

$$\gamma_{ijkl}=\sigma_{ijkl}-\sigma_{ij}\sigma_{kl} \qquad [16]$$

(see, e.g., Browne, 1982), when these moments exist. (Note that this $\boldsymbol{\Gamma}$ matrix is not the same as in equation 3.) Define the $p(p+1)/2$ data vector $\boldsymbol{d}_i$ for observation i,

$$\boldsymbol{d}_i=\begin{pmatrix}(y_{i1}-\bar{y}_1)(y_{i1}-\bar{y}_1)\\(y_{i2}-\bar{y}_2)(y_{i1}-\bar{y}_1)\\(y_{i2}-\bar{y}_2)(y_{i2}-\bar{y}_2)\\\dots\\(y_{ip}-\bar{y}_p)(y_{ip}-\bar{y}_p)\end{pmatrix}, \qquad [17]$$

where y_{iv} is the ith observation on variable $v(v = 1, 2, \dots, p)$ and $\bar{y}_v$ is the sample mean for variable v, so that summing over the n sample units,

$$(n-1)^{-1}\sum_{i=1}^{n}\boldsymbol{d}_i=\boldsymbol{s}\,. \qquad [18]$$

A consistent estimator of $\boldsymbol{\Gamma}$ is obtained via the sample covariance matrix of $\boldsymbol{d}_i$, involving fourth-order moments (see, e.g., Browne, 1982, 1984; Chamberlain, 1982),

$$\hat{\boldsymbol{\Gamma}}=(n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{d}_i-\bar{\boldsymbol{d}})(\boldsymbol{d}_i-\bar{\boldsymbol{d}})' \qquad [19]$$

so that the estimator of the asymptotic covariance matrix of $\boldsymbol{s}$ is $n^{-1}\hat{\boldsymbol{\Gamma}}$. Taking $\hat{\boldsymbol{\Gamma}}$ as $\boldsymbol{W}$ in the weighted least squares fitting function of Equation 14 gives the asymptotically distribution free (ADF) estimator proposed by Browne (1982) for covariance structure analysis of nonnormal continuous variables.

Consider now standard errors of parameter estimates and tests of model fit. Let $p^*=p(p+1)/2$ and define the $p^*\times q$ derivative matrix

$$\boldsymbol{\Delta}=\partial\boldsymbol{\sigma}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\,. \qquad [20]$$

Estimating $\boldsymbol{\theta}$ with the weighted least squares fitting function of Equation 14, it is well known that a Taylor expansion gives the asymptotic covariance matrix,

$$nV(\hat{\boldsymbol{\theta}})=(\boldsymbol{\Delta}'\boldsymbol{W}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\boldsymbol{W}^{-1}\boldsymbol{\Gamma}\boldsymbol{W}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\boldsymbol{W}^{-1}\boldsymbol{\Delta})^{-1} \qquad [21]$$

(see, e.g., Browne, 1984; Ferguson, 1958; Fuller, 1987; Satorra, 1989). This matrix may be consistently estimated by inserting the estimated $\boldsymbol{\theta}$ in $\boldsymbol{\Delta}$ and using the estimated $\boldsymbol{\Gamma}$. It is common to set the weight matrix $\boldsymbol{W}=\hat{\boldsymbol{\Gamma}}$ so that the estimated asymptotic covariance matrix of the parameter estimates simplifies to

$$n\hat{V}(\hat{\boldsymbol{\theta}})=(\hat{\boldsymbol{\Delta}}'\boldsymbol{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}\,. \qquad [22]$$

Note that this expression is commonly used for both NTGLS and ADF. Satorra (1989) shows that asymptotically the same expressions hold for NTML. The expression in Equation 21 points to the fact that $\boldsymbol{W}$ need not be the same as $\hat{\boldsymbol{\Gamma}}$. For example, $\boldsymbol{W}$ may be calculated via the computationally simple normal theory expression of Equation 15 to give normal theory parameter estimates. Using the general $\hat{\boldsymbol{\Gamma}}$ expression of Equation 19 in Equation 21 provides the proper covariance matrix for these estimates even under nonnormality. Under nonnormality the estimates then have somewhat larger asymptotic variability than for ADF, but such an approach is strongly preferable to ADF from a computational point of view when the number of variables is large. In contrast to the ADF approach, Equation 21 shows that the large $\hat{\boldsymbol{\Gamma}}$ matrix need not be inverted to obtain the standard errors of the estimates.

Under normal theory, the conventional model test of fit of H_0 against an unrestricted covariance matrix is obtained as the likelihood ratio statistic nF, where F is the optimum value of the fitting functions in Equations 12 and 13. For Equation 14, a corresponding Wald statistic is obtained. This quantity is distributed as chi-square with p^*-q degrees of freedom. Relaxing the restriction of normality, Browne (1984) gives a more general expression for a chi-square test of model fit. Consider the residual expression,

$$nT = (s - \hat{\boldsymbol{\sigma}})'A(s - \hat{\boldsymbol{\sigma}})\,, \qquad [23]$$

where $\hat{\boldsymbol{\sigma}}$ is the estimated covariance structure and A is a consistent estimator of

$$\boldsymbol{V} - \boldsymbol{V\Delta}(\boldsymbol{\Delta}'\boldsymbol{V\Delta})^{-1}\boldsymbol{\Delta}'\boldsymbol{V} = \boldsymbol{\Delta}_{\perp}(\boldsymbol{\Delta}_{\perp}'\boldsymbol{V}^{-1}\boldsymbol{\Delta}_{\perp})^{-1}\boldsymbol{\Delta}_{\perp}'\,, \qquad [24]$$

where $\boldsymbol{\Delta}_{\perp}$ is an orthogonal complement of $\boldsymbol{\Delta}$ (see Satorra, 1992, p. 7). Robustness to nonnormality is obtained by using $\boldsymbol{V}^{-1} = \hat{\boldsymbol{\Gamma}}$ of Equation 19. Satorra (1989) shows that this also holds when the optimum of F is obtained via NTML. In line with Bartlett (1937), a simpler, mean-corrected expression is the scaled chi-square, nF/α, where

$$\alpha = \mathrm{tr}[(\boldsymbol{W} - \boldsymbol{W\Delta}(\boldsymbol{\Delta}'\boldsymbol{W\Delta})^{-1}\boldsymbol{\Delta}'\boldsymbol{W})\boldsymbol{\Gamma}]/r\,, \qquad [25]$$

where r is the degrees of freedom of the model. A mean- and variance-corrected chi-square may also be computed (see, e.g., Satorra & Bentler, 1990; Satterthwaite, 1941).

To summarize estimation and testing for continuous variables, three types of analysis approaches can be distinguished. First, *normal theory analysis* refers to obtaining model estimates by the NTML or NTGLS fitting functions and computing standard errors and chi-square test of model fit by the conventional formulas of Equation 22 and nF. Second, *ADF analysis* refers to obtaining parameter estimates by the weighted least squares fitting function of Equation 14, setting $\boldsymbol{W}$ to the ADF-type $\hat{\boldsymbol{\Gamma}}$ of Equation 19, computing standard errors via Equation 22, and a chi-square model test as nF. Third, *robust normal theory analysis* refers to obtaining estimates by the NTML or NTGLS estimators. Using the normal theory $\boldsymbol{W}$ and the ADF-type $\hat{\boldsymbol{\Gamma}}$, standard errors are computed via Equation 21, and the chi-square test of model fit is computed either as the residual chi-square of Equations 23 and 24 or as the scaled chi-square of Equation 25.

Muthén and Kaplan (1985, 1992) carried out Monte Carlo studies of normal theory analysis and ADF analysis using factor analysis on nonnormal data. They found that normal theory analysis gave good inference for small models (around five variables), but inflated chi-square values and a downward bias of standard errors for larger models (ten or more variables). ADF analysis gave good standard errors and chi-square tests for small models and large samples (at least 1,000), but larger models did not show good results. Larger models produced inflated chi-square values and a downward bias of standard errors that was comparable to or worse than that of normal theory analysis. Apparently the asymptotic properties of ADF are not realized for the type of models and the finite sample sizes often used in practice. The method is also computationally heavy with many variables. This means that while ADF analysis may be theoretically optimal, it is not a practical method. Robust normal theory analysis appears to be an attractive alternative. To date, however, there is very limited experience with this type of analysis. Muthén and Kaplan (1985, 1992) have demonstrated that normal theory estimates usually show very little bias, even under nonnormality. A few studies with small models have recently reported promising results with regard to the robust standard errors and robust chi-square for nonnormal data (see, e.g., Chou, Bentler, & Satorra, 1991; Satorra, 1992; Satorra & Bentler, 1990), but nothing has been reported for realistic-sized models. Robust normal theory analysis appears to warrant further study.

The previous discussion considered second-level testing using chi-square statistics to investigate the fit of a model to sample variances and covariances. At issue was the trustworthiness of the chi-square value itself, and it was noted that robust procedures seem to offer better results in this regard. Some investigations of robustness will now be presented in connection with binary data.

Correlation Structure Analysis With Dichotomous Variables

The robust standard errors and chi-square tests of model fit can be extended to analysis of statistics other than the sample covariance matrix $\boldsymbol{S}$ for continuous variables, for example, the use of tetrachoric correlations in the factor analysis of binary items. In Muthén (1978, 1984), a weighted least squares procedure was proposed for the estimation and calculation of standard errors and chi-square test of model fit. An estimated $\boldsymbol{\Gamma}$ matrix was computed as a consistent estimate of the asymptotic covariance matrix of the sample thresholds and tetrachoric correlations (see Muthén, 1978),

$$\boldsymbol{\Gamma} = (\partial\boldsymbol{\pi}/\partial\boldsymbol{\sigma})^{-1}V(\boldsymbol{p})(\partial\boldsymbol{\pi}/\partial\boldsymbol{\sigma})^{-1}\,, \qquad [26]$$

where $\boldsymbol{\pi}$ denotes the vector of univariate and bivariate marginal probabilities, $\boldsymbol{\sigma}$ denotes the vector of population thresholds and tetrachoric correlations, and $V(p)$ denotes the covariance matrix of the vector of univariate and bivariate marginal sample proportions, consisting of sample moments up to the fourth order. In this way, the Muthén (1978) estimator is analogous to the ADF analysis for continuous variables. In both cases, fourth-order moments are used in computing Γ. In practice, the method suffers from the same type of computational and statistical limitations for large models as does ADF, as described previously. Here, a counterpart to robust normal theory analysis will be discussed that avoids these limitations. I have recently proposed this method elsewhere (Muthén, 1992).

In the dichotomous case, correlations are analyzed and there is no issue of scale dependency. Because of this, a simple analogue to robust normal theory analysis for continuous variables is to obtain model parameter estimates by unweighted least squares, using $W = I$ for estimation, standard errors, and tests by Equations 14, 21, and 25. This new approach to the analysis of tetrachoric correlations shows promise for a computationally efficient way of obtaining more robust standard errors and chi-square tests of model fit. The approach has a further advantage as well. For strongly skewed dichotomous variables, the weight matrix is often singular, even with large samples. This was frequently found in analysis of symptom items for depression and alcohol, as in Muthén (1989a) and Muthén, Wisnicki, and Hasin (1991), owing to the fact that these items have small frequencies for $y = 1$ even with large samples. In these cases, the conventional weighted least squares estimator cannot be used, because the weight matrix cannot be inverted. The new approach, however, does not rely on the inversion of a weight matrix and provides standard errors and chi-square by Equations 21 and 25 even for a singular $\hat{\Gamma}$.

As shown in Muthén and Kaplan (1992), the size of the model is a crucial factor in testing structural models for categorical and other nonnormal data. The larger the model, the larger the sample size needed for the asymptotic properties of the chi-square approximations to hold to a reasonable degree. The importance of this may not be fully realized among practitioners. In particular, the fact that the chi-square value is inflated by a too small sample size may come as a surprise. Also, methodologists demonstrating the sampling behavior of new testing procedures are wise not to limit their study to the common choice of small-sized models of around five variables. Although such studies are valuable, they may seriously misrepresent results for larger models. These issues are illustrated in the next example.

Example 5: Size of the Model in Testing Skewed Likert Scale and Dichotomous Data

In this study, standard errors and chi-square tests of model fit are contrasted for a small model of 5 variables and a medium-sized model of 15 variables. Strongly skewed 5-category and 2-category scales were studied, each case generated as categorized multivariate normal data with a single-factor structure. The 5-category case was the same as "Case 4" in Muthén and Kaplan (1985), where category percentages for the variables were 5, 5, 5, 10, and 75. The 2-category case used 75%/25% splits for each variable, which may be viewed as dichotomizing the 5-category variables. The 5-category data were treated as continuous variables scored 0-4 and analyzed by regular sample covariance matrices, while the 2-category data were treated as binary and analyzed via tetrachoric correlation matrices. The different technologies for continuous and binary data are discussed together here because it turns out that they produce analogous results.

In both cases, the categorized data follow the postulated model of a single factor. Equal loadings were chosen to give correlations of 0.5 and reliabilities of 0.5. Two sample sizes were studied. Because such skewed variables were analyzed, these sizes were set at large values, using $N =$ 1,000 and 4,000, representing large-scale surveys of rare phenomena. The Monte Carlo study was carried out with 500 replications. For each data set, the conventional inference method as well as the robust inference method was used. The two methods will be referred to as MI and MII, respectively. In this way, the 5-category data were analyzed by conventional nonnormal variable GLS (ADF) using the estimator of Equation 14, standard error calculations as in Equation 22, and chi-square computed as nF (MI). The 5-category data were also analyzed by robust normal theory analysis of Equation 12, using W based on Equation 15, calculating the standard errors by Equation 21, and computing the mean-corrected chi-square as in Equation 25 (MII). The 2-category data were analyzed by the conventional GLS estimator of Equation 14 applied to tetrachorics, standard errors as in Equation 22, and chi-square as nF (MI). The 2-category data were also analyzed by ULS, using the standard error formula of Equation 21, $W = I$, and chi-square as in Equation 25 (MII; see Muthén, 1992, for technical details).

Table 9.4 presents results for both 5-category (labeled continuous) and 2-category (labeled dichotomous) data, contrasting the small model (5 variables) with the medium-sized model (15 variables), the smaller

Table 9.4 Example 5: Size of the Model in Testing Skewed Likert Scale and Dichotomous Data

N = 1,000				*N = 4,000*			
Continuous		*Dichotomous*		*Continuous*		*Dichotomous*	
MI	*MII*	*MI*	*MII*	*MI*	*MII*	*MI*	*MII*
5 Variables (5 df)							
Standard Error Bias %							
−7	−5	1	2	−6	−5	−4	−4
Chi-square (mean, variance, 5%, 1%)							
4.97	4.49	4.86	4.83	4.90	4.45	5.18	5.18
8.70	7.29	8.88	8.82	9.30	7.73	11.37	11.29
4.8	2.4	4.4	4.0	4.0	2.6	6.0	5.6
0.2	0.4	0.0	0.6	1.0	0.6	1.0	0.8
15 Variables (90 df)							
Standard Error Bias %							
−19	1	−13	−2	−5	−2	−8	−4
Chi-square (mean, variance, 5%, 1%)							
103.8	91.2	98.99	90.86	94.21	90.18	92.72	90.72
195.4	163.3	225.9	196.0	189.6	172.1	196.8	176.6
26.2	3.8	17.2	7.2	9.8	4.2	9.4	6.4
8.2	1.0	7.8	1.4	1.8	0.2	2.4	0.6

sample ($N = 1{,}000$) with the larger sample ($N = 4{,}000$), and the conventional inference method (MI) with the more robust inference method (MII). The table reports standard error bias percentage in the loadings, where bias is calculated as $(a - b)/b$, where a is the mean estimated standard error over the 500 replications and b is the standard deviation of the estimates over the replications. The table also reports chi-square test results. The four entries are mean value over the replications, variance over the replications, proportion of the replications leading to model rejection at the 5% level, and proportion of replications leading to model rejection at the 1% level.

Table 9.4 shows that for the smaller model of 5 variables the conventional method of MI works well at the lower sample size of 1,000 for both the continuous case and the dichotomous case. No adjustments seem needed. In fact, judging by the 5% rejection proportion, the more robust chi-square procedures of MII seem to overcorrect slightly in the continuous case. For the larger model of 15 variables the conventional methods of MI do not work well at the smaller sample size of 1,000,

and can be said to be only marginally acceptable at sample size 4,000. For this model size, the more robust procedures of MII give a marked improvement at sample size 1,000. It is interesting to note that the standard errors computed for MII by Equation 22 in all cases outperform the conventional standard errors.

The preceding analyses have shown that robust model testing methods are needed for models with categorical data whenever the model is not small. The new robust method for the analysis of dichotomous variables shows a marked improvement over the older method, particularly in terms of the chi-square test of model fit. In this way, a more trustworthy tool is provided for the second-level model testing.

Discussion

With continuous variables, the testing of a structural model involves testing the model's fit to its sample statistics, the sample covariance matrix. If techniques are used that are robust to deviations from normality, no further testing against the data seems essential. It is true, however, that checks of linear relations among variables should be carried out. Also, with strongly skewed variables it may be argued that a nonlinear model, using different sample statistics, may be more appropriate—for example, a "tobit" factor analysis model for censored variables (see Muthén, 1989d). For the categorical variables techniques that we have discussed, however, first-level testing of the model family against the data is an essential augmentation to second-level testing of the structural model. This is because first-level testing may affect the very choice of sample statistics.

If efficient estimates are obtained for the cell probabilities, first- and second-level testing can be replaced by a chi-square test of a structural model directly against the data by comparing it to the unrestricted multinomial alternative. To this end, full-information estimation via ML can be carried out, such as in the Bock and Aitkin (1981) approach to the exploratory factor analysis of dichotomous items. Such a direct test of a factor model against the unrestricted multinomial alternative is often problematic, however, because the chi-square approximation is poor in the typical situation of many cells having small or zero frequencies.

While informative, the combined use of first- and second-level testing proposed here poses the question of how to assess the seriousness of first-level misfits for inferences about structural models that have good second-level fits. Muthén and Hofacker's (1988) abortion example suggests a certain amount of robustness for first-level misfit, whereas the artificial three-wave panel example suggests that important structural parameters may be misestimated.

It is also important that a model be assessed in a way that is relevant for its ultimate use. To augment first-level and second-level testing, a useful approach would be to deduce predictions from the structural model to the observed data. Take as an example MIMIC modeling as discussed in Muthén (1989b). Here the MIMIC model was used to capture mean differences in factors and items for various sociodemographic groups by including group membership variables as dummies among the exogenous "multiple causes" variable set. This modeling was suggested as an alternative to multiple-group structural modeling with one group for each sociodemographic cross-classification. The multiple-group approach is unattractive because of small group sizes when there are many cells in the cross-classification, particularly when the *y* variables, the "multiple indicators," are strongly skewed, as in modeling of dichotomous alcohol dependence or depression symptom items. The MIMIC approach avoids these problems, but assumes instead that the factor variances and the loadings are invariant across groups. If the use of the model is to predict group differences in levels, a relevant test is to check how close the set of predicted item means are to the observed ones for as fine a sociodemographic grouping as possible. This sort of model check also ties the implications of the structural model back to the data.

References

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

Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality and Quantity, 23*, 277-331.

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and reliability of attitude measurement. In P. Marsden (Ed.), *Sociological methodology 1992*. Washington, DC: American Sociological Association.

Ashford, J. R., & Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics, 26*, 535-546.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A, 160*, 268-282.

Benson, J., & Muthén, B. (1992). *Testing the factor structure invariance of the Test Anxiety Inventory using categorical variable methodology*. Unpublished manuscript.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Browne, M. W. (1974). Generalised least squares estimates in the analysis of covariance structures. *South African Statistical Journal, 8*, 1-24.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis*. Cambridge: Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics, 18*, 5-46.

Chou, D., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Ferguson, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Annals of Mathematical Statistics, 29*, 1046-1062.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlations. *American Sociological Review, 34*, 93-101.

Jöreskog, K. G. (1985, July 2-5). *Estimation of the polyserial correlation from summary statistics*. Paper presented at the Fourth European Meeting of the Psychometric Society, Cambridge, England.

Jöreskog, K. G. (1991, July 15-17). *Latent variable modeling with ordinal variables*. Paper presented at the International Workshop on Statistical Modeling and Latent Variables, Trento, Italy.

Jöreskog, K. G., & Sörbom, D. (1988). *PRELIS: A program for multivariate data screening and data summarization. A preprocessor for LISREL* (2nd ed.). Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

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. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*, 48-65.

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

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model. Theoretical integration and user's guide*. Mooresville, IN: Scientific Software.

Muthén, B. (1989a). Dichotomous factor analysis of symptom data. In M. Eaton & G. W. Bohrnstedt (Eds.), Latent variable models for dichotomous outcomes: Analysis of data from the Epidemiological Catchment Area Program [Special issue]. *Sociological Methods & Research, 18*, 19-65.

Muthén, B. (1989b). Latent variable modeling in heterogeneous populations: Presidential address to the Psychometric Society, July 1989. *Psychometrika, 54,* 557-585.

Muthén, B. (1989c). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42,* 55-62.

Muthén, B. (1989d). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology, 42,* 241-250.

Muthén, B. (1992, April). *A new inference technique for factor analyzing binary items using tetrachoric correlations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika, 53,* 563-578.

Muthén, B., Huba, G., & Short, L. (1985). *Applications of LISCOMP structural equation modeling for ordered categorical variables.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Muthén, B., & 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, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19-30.

Muthén, B., Wisnicki, K. S., & Hasin, D. (1991). *Factor analysis of ICD-10 symptom items in the 1988 National Health Interview Survey on Alcohol Dependence.* Unpublished manuscript.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54,* 131-151.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. In P. Marsden (Ed.), *Sociological methodology 1992.* Washington, DC: American Sociological Association.

Satorra, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis, 10,* 235-249.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309-316.

10

Some New Covariance Structure Model Improvement Statistics

P. M. BENTLER
CHIH-PING CHOU

A standard methodology for improving an ill-fitting structural equation or covariance structure model is to remove parameter restrictions (e.g., to free up previously fixed parameters). The Lagrange Multiplier (LM) or score test was developed to evaluate hypotheses on whether a restriction is statistically inconsistent with the data (Lee & Bentler, 1980; Satorra, 1989; Sörbom, 1989). Another method suggests evaluating the estimated value, or estimated parameter change (EPC), that a specific fixed parameter may take if it is freed (Saris, Satorra, & Sörbom, 1987). In this chapter, we propose extending the ideas behind these model improvement methods and develop three other criteria that concentrate on the impacts to the current model of reducing constraints. As each restriction is removed, or a fixed parameter is freed, from the current model, it may cause changes in parameter estimates as well as an increment in the estimated sampling variabilities of the free parameters (Bentler & Mooijaart, 1989). Without reevaluating the model, statistics reflecting these features can be computed based on the existing model.

AUTHORS' NOTE: This research was supported by grants DA01070, DA00017, and DA03976 from the National Institute on Drug Abuse, and was reported at the Joint Statistical Meetings of the American Statistical Association and Biometric Society, Anaheim, CA, August 1990. The helpful comments of several anonymous reviewers and the editors are gratefully acknowledged.