Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

A Simulation Study on the Estimation of Some Simple Structural Equations Models when the Original Variables are Categorized

Emmanuel Aris[1]

Abstract

The present study is focused on the behavior of Structural Equation Model (SEM) parameter, standard-error and goodness-of-fit estimates when some or all of the variables on which the model is fitted have been categorized. The behavior of the estimates is studied for different models, underlying and observed distributions, parameter values, sample sizes, number of categorical variables, and SEM programs. The purpose of this study is to determine which are the factors that may disturb the most the estimation of parameter, standard-error or goodness-of-fit values. If the estimation of parameters seems to be satisfactory for polychoric and polyserial procedures, the standard-error estimates are shown to be often biased, especially for models with polyserial correlations. The use of bootstrap is shown here to be a possible solution to this problem.

1 Introduction

Structural Equation Models were originally developed for continuous variables. They are based on the modeling of the covariance structure of the multivariate distribution (Bollen, 1989). In order to include categorical variable in these models, polychoric and polyserial procedures were developed. In the polychoric estimation procedure, for example, the correlation between each pair of categorical variables is estimated as the correlation that would be obtained between two corresponding underlying continuous variables. Details about these procedures can be found in Olsson (1979), Olsson, Drasgow and Dorans (1982), Lee, Poon, and Bentler (1990, 1992, 1995), Jöreskog (1994), or Muthén and Satorra (1995).

[1] CTO/Biometrics, Organon, Oss, The Netherlands: emmanuel.aris@organon.com.

The estimation procedures considered in this article are implemented in EQS 5.7 (Bentler and Wu, 1993), PRELIS 2.30 / LISREL 8.30 (Jöreskog and Sörbom, 1996a, 1996b), and Mplus 1.0 (Muthén and Muthén, 1999).

This article is organized as follows. Section 2 details the criteria chosen to evaluate the bias and precision of parameter, standard-error, and goodness-of-fit estimates. Section 3 presents results obtained in a preliminary study where polychoric estimation procedures are compared to product-moment procedures with or without optimal scaling. Sections 4 to 6 present the results obtained from the simulation studies.

2 Evaluation criteria

The simulations performed here are Monte Carlo simulations. For a description of the process of the simulations see Appendix A. Two different types of models are studied in this article: the bivariate regression models and the 4-indicator, 1-factor analysis models. For each model, in order to evaluate the behavior of the estimation procedures, the following criteria are introduced.

The *relative bias* and *absolute relative bias* are the criteria chosen to evaluate the bias and precision of the parameter and standard-error estimates. The *relative bias* ($B(\hat{\gamma}_{ir})$) and the *absolute relative bias* ($AB(\hat{\gamma}_{ir})$) of a parameter estimate $\hat{\gamma}_{ir}$ corresponding to observation number r, are:

$$B(\hat{\gamma}_{ir}) = \frac{\hat{\gamma}_{ir} - \gamma_i}{\gamma_i} \qquad \text{and} \qquad AB(\hat{\gamma}_{ir}) = \left| \frac{\hat{\gamma}_{ir} - \gamma_i}{\gamma_i} \right|,$$

with γ_i being the original parameter value chosen in the original model in order to simulate the samples. Note that the denomination *absolute relative bias* is used for $AB(.)$ as in Boomsma and Hoogland (2001), although *absolute relative deviation* could be a more appropriate denomination, $AB(.)$ being informative not only on the bias but also on the precision of the estimates.

Similarly, the *relative bias* and the *absolute relative bias* of the estimate $\hat{se}_{\hat{\gamma}_{ir}}$ of the standard error of γ_i corresponding to observation r are defined as:

$$B(\hat{se}_{\gamma_{ir}}) = \frac{\hat{se}_{\hat{\gamma}_{ir}} - se_{\gamma_i}}{se\gamma_i} \qquad \text{and} \qquad AB(\hat{se}_{\gamma_{ir}}) = \left| \frac{\hat{se}_{\hat{\gamma}_{ir}} - se_{\gamma_i}}{se\gamma_i} \right|,$$

with se_{γ_i} being the empirical standard deviation of the estimates of γ_i corresponding to all observations being in the same design cell as observation r. For a discussion about the choice of the empirical standard error as the reference standard-error value, see Appendix B.

Finally, the values of the goodness-of-fit indices are compared to their expected value, corresponding to the number of degree of freedom of the model. The distribution of the estimates is also studied by means of Q-Q plots.

3 Preliminary study: Polychoric vs. product - moment

In this section, results yielded by LISREL, EQS, and Mplus polychoric estimation procedures are compared to the ones yielded by the product-moment (PM) procedure and the product-moment-with-optimal-scaling (PM+OS) procedure. In both PM and PM+OS procedures the categorical variables are treated as if they were continuous. With the PM procedure, the correlation is calculated using the original values of the categories. With the PM+OS procedure, the value of the categories are optimized in order to obtain the highest nontrivial (product-moment) correlation coefficient possible. Both PM and PM+OS methods have been performed using the Maximum Likelihood estimation procedure from EQS 5.7.

The procedures are compared on the basis of the estimates yielded if bivariate regression models, with an original standardized effect coefficient equal to 0.2, 0.4, 0.6, or 0.8, are fitted when both variables are trichotomized (the underlying and observed distributions considered are presented in Appendix A). For each combination of observed and underlying distribution, 50 replications of samples of size 1000 are drawn.

Every estimation method was fitted on all the 12600 samples, LISREL, EQS and Mplus, yielding estimates on 12600, 12600 and 5405 of them, respectively. Mplus was thus rather unstable as it did not yield estimates for more 50% of the replications. This seems to come from a division by zero during the estimation of the robust chi-square statistic (L. Muthén, personal communication). This problem should be solved in version 1.03 of Mplus.

Table 1: Grand average $B(.)s$ and $AB(.)s$ over all experimental conditions (bivariate regression models).

	$\overline{B}(\hat{\gamma})$	$\overline{AB}(\hat{\gamma})$	$\overline{B}(\hat{se}_\gamma)$	$\overline{AB}(\hat{se}_\gamma)$
Estimation method				
Product Moment	-0.28	0.28	-0.73	0.73
PM + Optimal Scaling	-0.25	0.25	-0.70	0.70
Polychoric EQS 5.7	0.01	0.10	-0.51	0.61
Polychoric LISREL 8.20	0.01	0.09	0.10	0.21
Polychoric Mplus 1.0	0.01	0.09	-0.13	0.17

The regression parameter estimates yielded by the polychoric procedures were rather similar and on average much less biased (0.01 against -0.28 and -0.25) and more precise (average absolute biases of 0.09 or 0.10 against 0.28 and 0.25) than the

ones yielded by PM and PM+OS (see Table 1). This pattern was also found for each specific observed distribution, unobserved distribution, and parameter size, which corresponds with findings from several earlier simulation studies (see, e.g., Coenders, Satorra and Saris, 1997). The PM+OS procedure always performed slightly better than the PM one. A possible explanation of this last result may be that, as the PM procedure often underestimates the true correlation, by optimizing the categories values, the PM+OS procedure corrects the estimated correlation slightly upwards, yielding a smaller relative bias.

The precision of the polychoric estimates ($AB(\hat{\gamma})$) was found to decrease with the value of the original regression parameter. For example, with regression parameters 0.2, 0.4, 0.6 and 0.8, the average $AB(\hat{\gamma})$s found for EQS estimates were 0.19, 0.09, 0.06 and 0.03, respectively. The bias and precision of the PM or PM+OS estimates also depended on the type of observed distributions: strongly biased estimates were obtained for models with variables having highly skewed or leptokurtic observed distributions. As a result, for a low original regression parameter and a not too skewed or leptokurtic observed distribution, the difference *in precision* between the estimates of the five procedures is small. For example, for an original regression parameter of 0.2 with an equiprobable distribution (C2), the average $AB(\hat{\gamma})$ was equal to 0.19 for PM, 0.16 for PM+OS and 0.14, 0.14 and 0.17 for the three polychoric procedures. However, if the observed distribution chosen was highly leptokurtic (C8), the average $AB(\hat{\gamma})$ was equal to 0.49 for PM, 0.44 for PM+OS and 0.21, 0.21 and 0.23 for the three polychoric procedures. Note that these results are similar to those of O'Brien and Homer (1987).

The standard-error estimates were less precise than the parameter ones. In particular, the estimates yielded by PM or PM+OS underestimated strongly their empirical value. Further, the estimates yielded by the polychoric procedures were much different from each other. EQS procedure strongly underestimated the standard errors, Mplus underestimated them and LISREL overestimated them. Mplus obtained the best results regarding the precision with an average $AB(\hat{se}_{\gamma})$ of 0.17.

In the study presented above, better parameter estimates were obtained when using polychoric procedures than when using PM or PM+OS procedures. It seems then often advantageous to use them instead of PM or PM+OS procedures. Even if Mplus yielded quite satisfactory estimates compared to LISREL of EQS, due to the large number of replications where no results were provided, only LISREL and EQS estimates are considered in the following.

4 The bivariate regression models

The continuous data is simulated here from a bivariate regression model where variable X_1 has an effect on variable X_2, which is represented by the standardized regression coefficient γ ($\gamma = 0.2$, 0.4, 0.6, or 0.8). Either both variables are catego-

rized, or only X_2 is categorized. The variable "type of treatment" is defined in order to indicate whether both variables are categorized (polychoric models) or only X_2 is categorized (polyserial models). The various conditions are thus: 2 types of treatment, 4 different regression parameter values, 9 observed distributions, 7 underlying distributions, 2 sample sizes (300 and 1000). For each condition, 100 replications are drawn. In total, each of the two estimation programs are performed 100800 times.

4.1 Previous research

Polychoric correlations are unbiased for underlying normal distribution and observed distributions with zero skewness and kurtosis (Olsson, 1979). However, polychoric estimation procedures were shown to yield poor estimates whenever: the observed distributions are skewed (Faber, 1988), the underlying skewness and kurtosis are high (O'Brien and Homer, 1987), or the sample size is small (Lee and Lam, 1988). Furthermore, the higher the original correlation, the lower the average relative bias (O'Brien and Homer, 1987; Lee and Lam, 1988). Note that standard-error estimates were not considered in most previous researches. Olsson et al. (1982) compared the behavior of the Full ML and the Two-Step ML polyserial estimations procedure with the point-polyserial estimation procedure. The first two procedures yielded estimates with $B(\hat{\gamma})$s lower than 0.05, while the point-polyserial procedure yielded estimates with $B(\hat{\gamma})$s higher than 0.15. A recent study by Coenders et al. (1997) found polychoric and polyserial estimates rather robust against nonnormality of the underlying continuous variables, and polyserial estimates fairly sensitive to nonnormality of the (observed) continuous variable.

4.2 Simulations results

4.2.1 Behavior of parameter estimates

An analysis of variance was performed on the relative bias of the parameter estimates ($B(\hat{\gamma})$) yielded by LISREL and on the ones yielded by EQS. Several remarks may be drawn from the results obtained:

- Both ANOVA-models containing all possible interaction effects did not explain more than 11 % of the total variance. Hence, when a parameter estimate of a certain sample varies from its original value, this is, in general, more due to sampling errors than to a specific bias of the estimation procedure due to particular design conditions.
- The behavior of the LISREL and EQS estimates procedures are most of the time similar. The average difference between the two estimates was of 0.003 and the mean absolute difference was of 0.02.
- The total average of all relative biases is rather low (< 0.01) for both estimation procedures. The factors found to affect the most the relative bias are the type

of observed distribution and the type of underlying distribution. However, the effects were found to be low as all average $B(\hat{\gamma})$s given one level of each design factors are always between -0.05 and +0.05.

Table 2: Average $B(\hat{\gamma})$s (bivariate regression models).

- *Polychoric models*

		EQS			LISREL		
		Unobserved distr.			*Unobserved distr.*		
		D1	D2	D3	D1	D2	D3
	C1	0.01	-0.02	-0.11	0.01	-0.03	-0.12
	C3	-0.01	-0.02	-0.08	-0.01	-0.02	-0.08
Observed	C5	0.01	0.06	0.11	0.01	0.06	0.12
distr.	C9	0.00	0.06	0.12	-0.01	0.04	0.10
	C7	0.00	-0.06	-0.14	0.00	-0.06	-0.14
	C8	0.00	-0.10	-0.16	0.01	-0.09	-0.16

- *Polyserial models*

		EQS			LISREL		
		Unobserved distr.			*Unobserved distr.*		
		D1	D2	D3	D1	D2	D3
Observed							
distr.	C8	-0.01	-0.04	-0.10	-0.01	-0.06	-0.11

The combinations of type of treatment, observed distributions, and underlying distributions resulting in average $B(\hat{\gamma})$s higher than 0.10 are presented in Table 2. Although none of these average biases are significantly different from zero which support the claim that polychoric estimates are quite robust against nonnormality (Olsson, 1979; Coenders et al., 1997), several conclusions may be drawn from this table:

- Higher parameter relative biases appear whenever both underlying and observed distributions deviate from multivariate normality: the highest average $|B(\hat{\gamma})|$ with D1 is 0.01, while it is 0.16 with D3. Note that this was also found by O'Brien and Homer (1987).
- Underlying distributions and observed distributions do not affect the value of the relative bias independently. Considering LISREL estimates for example,

Table 3: Average $B(\hat{se}_\gamma)$s (bivariate regression).

		EQS		LISREL	
		Polychoric	Polyserial	Polychoric	Polyserial
	(0.2)	-0.77*	-0.63*	0.01	0.37
Value	*(0.4)*	-0.72*	-0.54*	0.01	1.01
of γ	*(0.6)*	-0.58*	-0.28	0.07	1.53
	(0.8)	-0.02	1.15	0.36	2.16

NB: The average $B(\hat{se}_\gamma)$s being significantly different from zero are indicated by a star.

the interaction underlying/observed distributions accounts for 31% of the variance explained, whereas underlying and observed distributions account independently for 10% and 17% respectively. The influence of skewness or kurtosis in the observed distributions is much more important if the underlying variable is highly skewed than if it is normally distributed.

- Estimates from polyserial models are slightly less affected by nonnormality than estimates from polychoric models. For example, with LISREL, the average $B(\hat{\gamma})$ is -0.01 for polyserial models against 0.02 for polychoric models. Furthermore, it can be seen from Table 2. that polyserial estimation procedures seem to be at least as robust as polychoric ones in these cases. Although all original continuous variables are supposed to have the same distribution here and all variables are observed, this last result is different from the one found by Coenders et al. (1997).

4.2.2 Behavior of standard-error estimates

The grand average of $B(\hat{se}_\gamma)$ over all experimental conditions was -0.29 for EQS and 0.74 for LISREL, indicating a much less accurate estimation than for the effect parameters. Similarly to what was done previously, two analyses of variance were performed on the estimated-standard-error relative biases obtained.

Note first that very few sampling variation was found between the standard-errors estimates of the 100 replications per design cell. Indeed, the total variance of $B(\hat{se}_\gamma)$ is explained for 94% and for 99% by the variation of the different design conditions for EQS and LISREL, respectively.

The factors found to affect the value of $B(\hat{se}_\gamma)$ most are the type of treatment (polychoric or polyserial) and the value of the original regression parameter. The type of observed distribution did also, to a lesser extent, affect LISREL estimates. In order to see more precisely, which effect do type of treatment and parameter value have, mean relative biases are calculated for all cells of the crosstable formed by these two factors. The results are presented in Table 3. Several remarks may be done:

- For low values of γ, EQS $B(\hat{se}_\gamma)$ are significantly different from zero. The standard-error estimates are then underestimated.
- LISREL standard-error estimates are very close to their empirical value for polychoric models, but are overestimating quite strongly their empirical values for polyserial models. Indeed, given an original value of γ and one of the observed distributions C1, C2, C4 or C6, the average $B(\hat{se}_\gamma)$ obtained for polyserial models is significantly overestimating the empirical standard deviations.
- The average relative bias value of the standard-error estimates almost always increases (i.e., estimates become more overestimated for LISREL or less underestimated for EQS) as the original value of the regression parameter increases, for polychoric or polyserial models.
- Dependence on the original regression parameter value is more pronounced for polyserial models than for polychoric ones. For example, EQS mean relative bias increases from -0.63 ($\gamma = 0.2$) to 1.15 ($\gamma = 0.8$) for polyserial models, while it increases from -0.77 ($\gamma = 0.2$) to 0.02 ($\gamma = 0.8$) for polychoric models.

5 The Confirmatory Factor Analysis models

In this section, a four-indicator, one-factor model is studied. Variables X_1, X_2, X_3 and, X_4 are supposed to be the indicators of a latent variable ξ. The equations implied by this model are:

$$\begin{cases} X_1 &= \lambda_{11}\xi + \epsilon_1, \\ X_2 &= \lambda_{21}\xi + \epsilon_2, \\ X_3 &= \lambda_{31}\xi + \epsilon_3, \\ X_4 &= \lambda_{41}\xi + \epsilon_4, \end{cases}$$

with λ_{i1} and ϵ_i being the factor loading and the error of measurement, respectively, associated with latent variable ξ and indicator X_i (for $i = 1, ..., 4$). Variables X_1 to X_4 are observed. In order to identify the model, the variance of ξ is set to 1. If all variables are categorical, the models are called polychoric models. If only X_3 and X_4 are categorical while X_1 and X_2 are continuous, models are called polyserial models. The bias is calculated here as the average bias obtained between the last two factor loadings (the factor loadings of X_3 and X_4, the variables that are always categorized). All four factor loadings are supposed to be equal. Selected values of the (standardized) factor loadings are 0.55, 0.71, and 0.84. The analyses of variance are performed with the usual design factors. The nonconvergent cases (134 for LISREL and 2228 for EQS) were deleted from the study. Note that, although the model here is much simpler than the ones considered in Boomsma and Hoogland (2001), the number of nonconvergent cases obtained from these estimation procedures is smaller to the ones obtained by Generalized-Least-Squares, or Asymptotic-Distribution-Free estimation procedures for similar sample sizes.

Table 4: Grand average estimated values over all experimental conditions (CFA models).

	$\overline{B}(\hat{\lambda}_{-})$	$\overline{B}(\hat{se}_{\lambda_{-}})$	$\overline{AB}(\hat{\lambda}_{-})$	$\overline{AB}(\hat{se}_{\lambda_{-}})$	Chi-square
EQS	0.00	-0.51	0.07	0.56	7.8
LISREL	0.01	-0.04	0.07	0.23	3.3

5.1 Previous research

For models with categorical variables only and factor loadings around 0.71, Liscomp estimated factor loadings and standard errors were found to be close to their true value (Muthén and Kaplan, 1985; Potthast, 1993). The bias of factor-loading estimates (with original value around 0.87), in models with categorical data having underlying nonnormal variables, was found to be low by Coenders et al. (1997). If the model size increased (more latent variables having four indicators each), standard errors were found to be underestimated whereas the bias of the factor-loading estimates remained fairly small (Potthast, 1993). The parameter estimate bias was found to be smaller for large sample sizes than for small ones (Parry and McArdle, 1991), although for sample sizes higher than 500 this difference was reported to be small (Potthast, 1993). Parry and McArdle (1991) also found that the quality of parameter estimation increased as the sample size and/or as the true factor loading values increased. LISREL and Liscomp estimation procedures, were found to yield similar results for categorical models (Dolan, 1994). For models with both categorical and continuous variables, Lee et al. (1992, 1995) found that parameter and goodness-of-fit estimates were satisfactory with EQS and Liscomp for samples of 200 observations or more.

5.2 Simulations results

5.2.1 Behavior of factor-loading estimates

The values of the total average of the factor-loading relative biases and average relative biases were close to zero for both EQS and LISREL: lower than 0.01 for $B(\hat{\lambda}_{-})$ and lower than 0.08 for $AB(\hat{\lambda}_{-})$ (see Table 4). For both methods, less than 11% of the variance of $B(\hat{\lambda}_{-})$ could be explained by the design factors. Similarly to what happened with the bivariate regression models, most of the variation of the estimates are due to sampling fluctuations. Although the most important factors are the observed and underlying distributions, all average $B(\hat{\lambda}_{-}^{(1)})$s, given an underlying and an observed distribution, are lower than 10 %. Hence, these estimates are rather

robust against observed and unobserved nonnormality. Note that this is in line with what was found by Coenders et al. (1997).

For both procedures, the absolute relative biases of the factor loadings ($AB(\hat{\lambda}_{-})$) were found to depend mainly on the factor-loading original value. The higher the original factor-loading, the lower the average $AB(\hat{\lambda}_{-})$, thus the higher the relative precision of the estimates. For example, with LISREL, the average $AB(\hat{\lambda}_{-})$ was equal to 0.12, 0.06 and 0.04 for $\lambda = 0.55$, $\lambda = 0.71$ and $\lambda = 0.84$, respectively. Note that this is similar to what was found for Liscomp by Parry and McArdle (1991).

5.2.2 Behavior of standard-error estimates

The grand average of the standard-error relative biases deviates strongly from zero for EQS (-0.51) but less for LISREL (-0.04) (see Table 4). Furthermore, the average relative biases are also high for both procedures: higher than 20% for LISREL and higher than 50% for EQS.

Similarly to the results for regression models, a large part of the variance of $B(\hat{se}_{\lambda_{-}})$ is explained by the saturated ANOVA-model (90% for LISREL and 96% for EQS) and little sampling variation is found across the 100 replications per design cell. LISREL estimates are found to be mainly affected by the observed distribution and by the interaction between observed distribution and type of treatment. EQS estimates are mainly affected by the value of the original parameter and by the observed distribution. Average $B(\hat{se}_{\lambda_{-}})$ obtained for each of these levels are presented in Table 5. From examination of this table, several points are worth remark:

- LISREL standard-error estimates are relatively close to their empirical value for polychoric models but relatively further away for polyserial models, especially when the observed distributions are skewed or leptokurtic. For example, with C3 (no skewness or kurtosis) the average $B(\hat{se}_{\lambda_{-}})$ are -0.07 and 0.11 for polychoric models and polyserial models, respectively, whereas with C9 (high skewness and kurtosis) the average are -0.09 and -0.57, respectively.
- EQS estimates are often further from their empirical values than are LISREL ones. EQS estimates seem to be closer to their empirical values for high original factor-loadings than for low ones: average $B(\hat{se}_{\lambda_{-}})$s of -0.76 and 0.69 are found for polychoric and polyserial models with all λs equal to 0.55, against -0.32 and -0.01 for polychoric and polyserial models with all λs equal to 0.84. They are also slightly less biased for polyserial models than for polychoric models: the average $B(\hat{se}_{\lambda_{-}})$ is of -0.58 for polychoric models while it is of -0.44 for polyserial models.

LISREL estimates are thus close to their empirical values for models with categorical variables only. Given the fact that LISREL and Liscomp procedures were found to yield similar results (Dolan, 1994), these results are in line with the results

Table 5: Average $B(\hat{se}_{\lambda\_})$ given some combinations of different conditions.

• *For LISREL*									
	Observed distributions								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Polychoric	-0.07	-0.01	-0.07	-0.08	-0.08	-0.05	-0.09	-0.13	-0.09
Polyserial	0.31	0.53*	0.11	-0.05	-0.32*	0.49*	-0.23	-0.36*	-0.57*
	Original factor-loading values								
	0.55	*0.71*	*0.84*						
Polychoric	-0.08	-0.06	-0.08						
Polyserial	-0.01	-0.01	-0.01						
• *For EQS*									
	Observed distributions								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Polychoric	-0.51*	-0.34	-0.55*	-0.45	-0.58*	-0.54*	-0.69*	-0.78*	-0.74*
Polyserial	-0.36	-0.21	-0.38	-0.30	-0.45	-0.40	-0.58*	-0.69*	-0.64*
	Original factor-loading values								
	0.55	*0.71*	*0.84*						
Polychoric	-0.76*	-0.65*	-0.32						
Polyserial	-0.69*	-0.55*	-0.01						

NB: The average $B(\hat{se}_{\lambda_})$s being significantly different from zero are indicated by a star.

from Potthast (1993). Furthermore, LISREL polychoric estimates have been shown here to be rather robust against observed or underlying nonnormality.

5.2.3 Behavior of goodness-of-fit values

The 4-indicator, 1-factor model, has 2 degrees of freedom. The expected χ^2-goodness-of-fit value is 2, and has a 95 % probability of being contained in the interval from 0.05 to 7.38. With a grand mean of 7.8, EQS goodness-of-fit values were very often too high to be acceptable. LISREL goodness-of-fit values were much less inflated (grand mean around 3.3). Analyses of variance were performed on the estimated goodness-of-fit values. The most notable results, shown in Table 6, are the following ones:

- The most important factors explaining the variation of LISREL estimates are the observed distribution, the type of treatment, and their interaction. The average estimated goodness-of-fit values of the polychoric models are close to the expected value of 2 for all observed distributions. They are not too inflated for the polyserial models either, provided that not too skewed or leptokurtic observed distributions are used. For example, with C1 to C4 and also C6 the average goodness-of-fit values are lower than 4.0.
- The most important factors explaining the variation of EQS estimates are the observed distribution and the interaction between the type of treatment and

Table 6: Average χ^2-goodness-of-fit estimates.

• *For LISREL*									
	Observed distributions								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Polychoric	2.1	2.0	2.0	1.9	2.2	2.0	2.0	2.2	1.9
Polyserial	2.9	2.0	3.6	3.1	4.8	1.7	6.4	9.3	8.0
	Original factor-loading values								
	0.55	*0.71*	*0.84*						
Polychoric	2.1	2.0	2.0						
Polyserial	4.4	4.7	4.8						
• *For EQS*									
	Observed distributions								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Polychoric	4.1	3.0	5.3	4.5	7.8	6.8	12.7	25.7	26.0
Polyserial	3.1	2.0	2.5	2.8	4.2	3.1	4.8	11.9	9.3
	Original factor-loading values								
	0.55	*0.71*	*0.84*						
Polychoric	22.5	14.1	1.7						
Polyserial	7.3	3.6	2.8						

the original factor-loading value. Estimation procedures produce somewhat better estimates for polyserial models than for polychoric ones. For both types of treatment however, the goodness-of-fit estimates are too high for low factor-loading values: for example with $\lambda = 0.55$, the average χ^2-goodness-of-fit value was 22.5 for polychoric models and 7.3 for polyserial models.

6 Bootstrapping used with polychoric estimation procedure

As the average LISREL goodness-of-fit estimates are close to 2 for polychoric models, the shape of the distribution of the estimates is studied. Two Q-Q plots are presented here in Figure 1 for both observed and underlying distributions with no skewness or kurtosis (D1C3), and for highly skewed and leptokurtic observed and underlying distributions (D3C9). For both combinations, the shape of the distributions is approximately chi-squared distributed (the relation between expected and observed quantiles is approximately linear). However, the tails of the distributions can be rather different: for example, the distribution of the estimates for D1C3 has thicker tails than the chi-squared one, whereas the one of D3C9 has thinner tails than the chi-squared one. Indeed, higher values of the goodness-of-fit estimates are

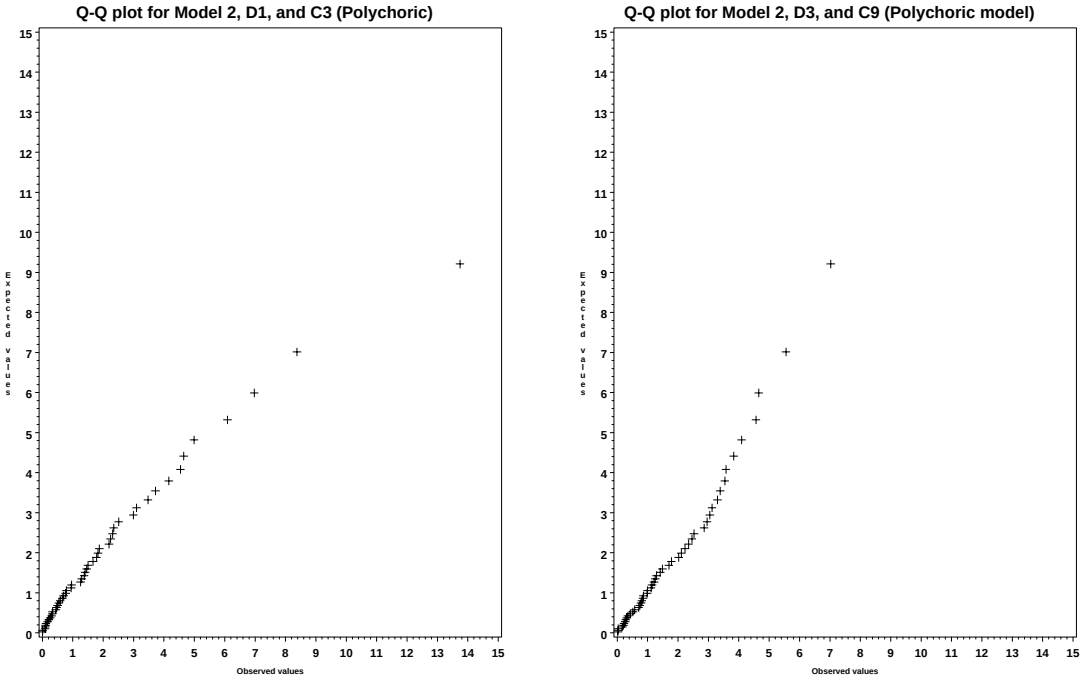

Figure 1: Q-Q Plots for polychoric models (CFA model with $\lambda = 0.71$, LISREL, N=1000).

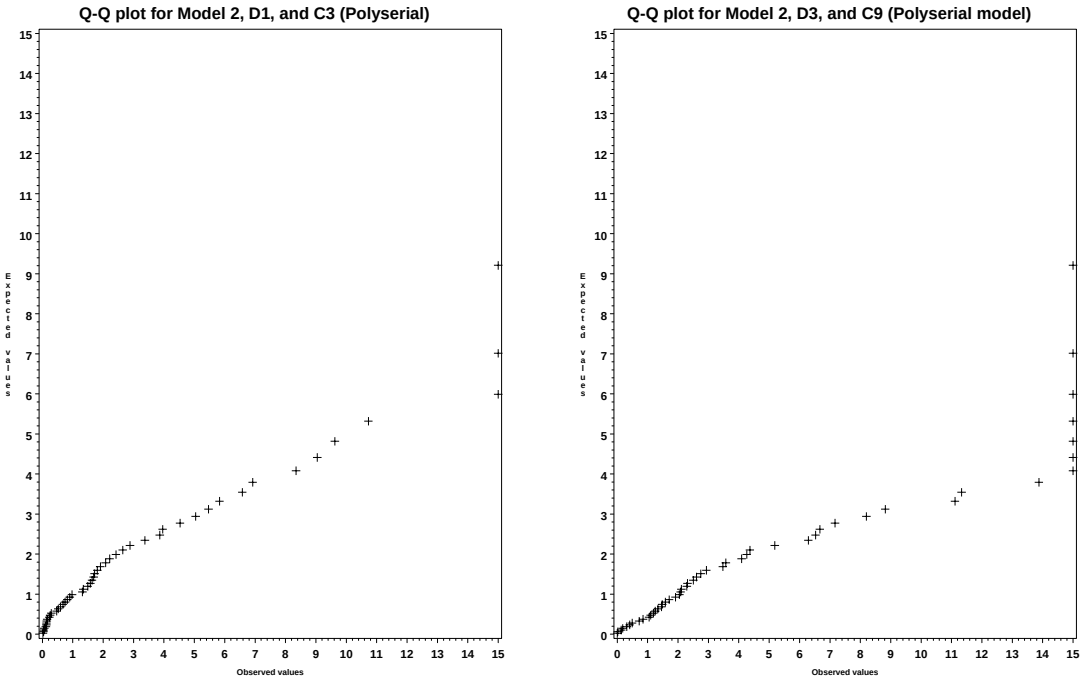

Figure 2: Q-Q Plots for polyserial models (CFA model with $\lambda = 0.71$, LISREL, N=1000).

more likely to happen for D1C3 than for D3C9. The same kind of Q-Q plots have been created for polyserial models, and indicate that the distributions deviate somewhat more from the chi-squared distributions, and that the right tail is most of the time much thicker than for the theoretical distribution (see, Figure 2).

As Bollen and Stine (1990) have shown, bootstrap methods can be used to obtain standard-error estimates. The purpose of this section is to evaluate whether the use of bootstrapping may solve the problem of inaccuracy in the estimation of the standard errors from EQS. A nonparametric bootstrap is used here associated with the estimation procedure from EQS 5.7.

The results of the simulation studies in this section are obtained with the 4-indicator, 1-factor models, whenever all indicators are categorical. The underlying distributions are normally distributed; 50 replications of samples of size 1000 are drawn. For each replication, 200 (bootstrap) samples of size 1000 are drawn with replacement. The standard error of the factor-loading estimates is calculated for each replication as the empirical standard deviation of the 200 bootstrap samples.

The factor-loading estimates produced by the bootstrap procedure are very close to their original values and they are not presented here. Results of the standard-error estimates from the bootstrap procedures for given specific trichotomous distributions are displayed in Table 7. The estimates provided by EQS in combination with the bootstrap procedure are very close to the empirical standard errors. Furthermore, this result holds for all original values of the factor loadings and all observed distributions, even the most skewed and leptokurtic: the average relative bias is always lower than 10% for all observed distributions.

Table 7: $B(\hat{se}_{\lambda_})$s for EQS with and without using bootstrap.

		EQS (direct) estimates			*EQS bootstrap estimates*		
		"True" factor loadings			*"True" factor loadings*		
Observed distribution		*0.55*	*0.71*	*0.84*	*0.55*	*0.71*	*0.84*
Skewness	*Kurtosis*						
-0.8	*-0.5*	-0.71	-0.54	-0.13	-0.01	0.04	-0.01
0.0	*-1.5*	-0.66	-0.47	0.05	-0.00	0.02	-0.02
0.0	*0.0*	-0.69	-0.64	-0.26	-0.05	-0.05	-0.01
1.5	*1.1*	-0.77	-0.66	-0.32	-0.00	-0.00	-0.04
2.5	*5.4*	-0.88	-0.81	-0.54	-0.04	-0.07	0.01

7 Conclusion

The results presented here are based on a Monte-Carlo study. The conclusions drawn in this section are thus restricted to the models and distributions used here.

In particular, variables are supposed here to have the same observed and underlying distributions, and the underlying distributions are supposed to be approximately unimodal. Other simulation studies have shown that polychoric and polyserial estimation procedures may be much less accurate when dealing with underlying multimodal distributions or variables having different degree of skewness and kurtosis (see, e.g., Aris 2001). The main results obtained from the present study are summarized below.

EQS and LISREL parameter estimates were found to be very often similar and close to their original value. High relative parameter estimate biases were obtained for simultaneous observed and underlying skewness and/or kurtosis. The effects of observed and underlying skewness/kurtosis have been shown to interact: the more the underlying distribution was skewed, the stronger the effect of observed skewness on the over- or underestimation of the parameters.

EQS standard-error estimates were very often found to significantly underestimate their empirical values, and this especially for models with low parameter values. LISREL standard-error estimates were found to be close to their empirical values for polychoric models but overestimated significantly their empirical values for polyserial models with certain skewed, leptokurtic or platykurtic observed distributions. For both estimation procedures, the values of the relative bias of the standard-error estimates were found to depend much more on the type of treatment or on the value of the original parameters than on the underlying and/or observed distributions. As the parameter estimates are rather close for LISREL and EQS, the large differences between the standard-error estimates found may very probably come from the estimation of the asymptotic variance-covariance matrix of the correlation estimates. Different estimates for this matrix are used in EQS 5.7 and LISREL 8.3 (Jöreskog, 1994: equation 35, Lee et al., 1995: equation 31). If these estimates are still far away to their asymptotic value, the Generalized Least Squares procedure used to estimate the parameters may not be efficient, and yield standard-error estimates far from their correct value.

A bootstrap procedure was performed in combination with EQS, and did help to correct the estimation of the standard errors. The relative bias of the (bootstrap) standard-error estimates was much lower than the original ones.

LISREL and EQS yielded also very different goodness-of-fit estimates. LISREL goodness-of-fit values were almost always acceptable for polychoric models, while for polyserial models they were too large for variables with high observed skewness and/or kurtosis. The distribution of LISREL goodness-of-fit estimates was often close to the theoretical one although for polyserial models, the upper tail of the observed distribution was almost always too thick. EQS goodness-of-fit values were also affected by high observed skewness and kurtosis, yielding goodness-of-fit values too inflated for polychoric and polyserial models. EQS goodness-of-fit values did also fairly depend on the original factor-loading values. In particular, too high

goodness-of-fit estimates were obtained for models with low original factor-loading values.

Finally, it is interesting to note that skewness and kurtosis in the observed distributions seem to affect the results much more that skewness and kurtosis in the underlying distributions, although for polychoric and polyserial procedures, only assumptions about the underlying distributions are made. Even though samples considered here are not that small, in the light of results from Boomsma and Hoogland (2001), the effect of observed distributions could possibly be due not to mere skewness of kurtosis, but to their consequences such as zero cells, or unprecise estimation of the asymptotic variance-covariance matrix.

Acknowledgments

This research was conducted during the PhD thesis of the author at Tilburg University and was supported by grants from Tilburg University and from The Netherlands Organization for Scientific Research. The author is grateful to Jacques Hagenaars and Marcel Croon for their comments and advices, to Germa Coenders for a discussion on simulation studies, and to two anonymous reviewers for their comments on a previous draft of this article.

References

[1] Aris, E.M.D. (2001): *Statistical Causal Models for Categorical Variables.* Oisterwijk: Tilburg University Press.

[2] Bentler, P.M. and Wu, E.J.C. (1993): *EQS/Windows User's Guide, version 4.* Los Angeles: BMDP Statistical Software, Inc.

[3] Bollen, K.A. (1989): *Structural Equations with Latent Variables.* New York: Wiley.

[4] Bollen, K.A. and Stine, R. (1990): Direct and indirect effects: classical and bootstrap estimates of variability. In C. C. Clogg (Ed.): *Sociological Methodology 1991.* Oxford: Basil Blackwell, 235-262.

[5] Boomsma, A. and Hoogland, J.J. (2001): The robustness of LISREL modeling revisited. In R. Cudeck, S. Du Toit, and D. Sörbom (Eds.): *Structural Equation Modeling: Present and Future.* Lincolnwood: SSI Scientific Software International, 139-168.

[6] Coenders, G., Satorra, A., and Saris, W.E. (1997): Alternative approaches to structural modeling of ordinal data: a Monte Carlo study. *Structural Equation Modeling*, **4**, 261-282.

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

[8] Faber, J. (1988): Consistent estimation of correlations between observed interval variables with skewed distributions. *Quality and Quantity*, **22**, 381-392.

[9] Fleishman, A.I. (1978): A method for simulating nonnormal distributions. *Psychometrika*, **43**, 521-532.

[10] Jöreskog, K.G. and Sörbom, D. (1996): *LISREL 8: User's Reference Guide.* Chicago: Scientific Software, Inc.

[11] Jöreskog, K.G. and Sörbom, D. (1996): *PRELIS 2: User's Reference Guide.* Chicago: Scientific Software, Inc.

[12] Jöreskog, K.G. (1994): On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, **59**, 381-389.

[13] Lee, S.Y. and Lam, M.L. (1988): Estimation of polychoric correlation with elliptical latent variables. *Journal of Statistical Computation and Simulation*, **30**, 173-188.

[14] Lee, S.Y., Poon, W.Y., and Bentler, P.M (1990): A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, **55**, 45-51.

[15] Lee, S.Y., Poon, W.Y., and Bentler, P.M. (1992): Structural equation models with continuous and polytomous variables. *Psychometrika*, **57**, 89-105.

[16] Lee, S.Y., Poon, W.Y., and Bentler, P.M. (1995): A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, **48**, 339-358.

[17] Muthén, B.O. and Kaplan, D. (1985): A comparison of some methodologies for the factor analysis of nonnormal Likert variables. *British Journal of Mathematical and Statistical Psychology*, **38**, 171-189.

[18] Muthén, B.O. and Muthén, L. (1999): *Mplus 1.0.* Los Angeles: Muthén and Muthén.

[19] Muthén, B.O. and Satorra, A. (1995): Technical aspects of Muthén's Liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, **60**, 489-503.

[20] O'Brien, R.M. and Homer, P. (1987): Correction for coarsely categorized measures: LISREL's polyserial and polychoric correlations. *Quality and Quantity*, **21**, 349-360.

[21] Olsson, U. (1979): Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, **44**: 443-460.

[22] Olsson, U., Drasgow, F., and Dorans, N.J. (1982): The polyserial correlation coefficient. *Psychometrika*, **47**, 337-347.

[23] Parry, C.D.H. and McArdle, J.J. (1991): An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, **15**, 35-46.

[24] Potthast, M.J. (1993): Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, **46**, 273-286.

[25] Vale, C.D. and Maurelli, V.A. (1983): Simulating multivariate nonnormal distributions. *Psychometrika*, **48**, 465-471.

Appendix A

Organization of the simulations

The simulation studies are conducted in the following way. A model $\mathcal{M}$, with a set of parameter values (θ) is chosen to represent the relationships among several continuous variables. The variance-covariance matrix for the entire population can be calculated and is denoted by Σ. A certain degree of skewness and kurtosis is chosen for the continuous variables and a sample of N observations of these continuous variables having this degree of skewness and kurtosis given Σ is generated. Then, using certain fixed threshold values, some variables are categorized in order to have approximately a chosen categorical distribution. The model $\mathcal{M}$ is finally fitted on these transformed samples. The estimates of $\mathcal{M}$'s parameters, standard deviations and goodness-of-fit indices are then compared to the references values.

In the simulations, 7 different possible underlying distributions (denoted D1, D2, ..., D7) and 9 different observed ones (C1, C2, ..., C9) will be considered. A description of these distributions can be found in the following.

Underlying distributions

The 7 continuous distributions called D1, D2, ..., D7, are generated using results from Fleishman (1978) and Vale and Maurelli (1983). The nonnormal variables simulated here are obtained from a fourth-degree-polynomial transformation of a standardized normal variable.

The first four moments of the 7 underlying distributions are shown below.

Name	Distribution type	Values of Mean	Variance	Skewness	Kurtosis
D 1	Normal variable	0.0	1.0	0.0	0.0
D 2	Mildly skewed variable	0.0	1.0	0.8	0.0
D 3	Highly skewed variable	0.0	1.0	1.5	2.5
D 4	Platykurtic variable	0.0	1.0	0.0	-1.0
D 5	Mildly leptokurtic variable	0.0	1.0	0.0	2.5
D 6	Highly leptokurtic variable	0.0	1.0	0.0	4.0
D 7	Highly leptokurtic + skewed	0.0	1.0	1.5	4.0

Observed distributions

The number of possible categories per variable chosen is three. In a previous study (Dolan, 1994), the use of polychoric correlation and of specific estimation procedures was advised until a maximum of five categories per variables.

The observed three-category distributions were chosen to have certain values of skewness and kurtosis. These values and the observed distributions chosen are shown below.

Name	Distribution type	Observed frequency categories (in %) Cat1	Cat2	Cat3	Skewness	Kurtosis
C 1	Negatively skewed	10	35	55	-0.8	-0.5
C 2	Equal categories	33.3	33.3	33.3	0.0	-1.5
C 3	Normally distributed	16.7	66.7	16.7	0.0	0.0
C 4	Positively skewed	55	35	10	0.8	-0.5
C 5	Highly posit. skewed	71.4	21.4	7.4	1.5	1.1
C 6	Platykurtic	46	8	46	0.0	-1.9
C 7	Mildly leptokurtic	10	80	10	0.0	2.0
C 8	Highly leptokurtic	7.1	85.7	7.1	0.0	4.0
C 9	Extremely posit. skewed and leptokurtic	85	10	5	2.5	5.4

Appendix B

Reference value for the standard-error estimates

For the standard-error estimates, two reference values can be chosen: the value of the standard deviation calculated from the original continuous model, or the empirical value calculated from the variations of the parameters across the different replications.

It may be interesting to use the original value of the standard error as reference value since this value is the same for all estimation procedures and all types of categorization of one (underlying) continuous variable. Furthermore, it is the original

Table 8: $se_{\lambda_}$s for LISREL (polychoric model, N=1000, D1, 200 replic).

			Standard errors		
			Continuous model	*LISREL*	
			True values	*Empirical*	*Expected*
Value of	*Observed dist.:*				
λ	*Skew.*	*Kurt.*			
	0.0	*-1.9*	0.039	0.060	0.052
0.55	*0.0*	*0.0*	0.039	0.047	0.047
	0.0	*2.0*	0.039	0.062	0.058
	0.0	*-1.9*	0.031	0.039	0.035
0.71	*0.0*	*0.0*	0.031	0.033	0.032
	0.0	*2.0*	0.031	0.044	0.040
	0.0	*-1.9*	0.027	0.024	0.022
0.84	*0.0*	*0.0*	0.027	0.023	0.021
	0.0	*2.0*	0.027	0.027	0.026

NB : all standard deviations of the expected std are between 0.0012 and 0.0070.

standard deviation value of the parameter from the model that should "ideally" be estimated.

Given *NR* replications of a data set from a cell of the multivariate crosstable formed by crossing all design factors, the empirical standard deviation value can be calculated from:

$$se_{\gamma_i}^{emp} = \sqrt{\frac{1}{NR-1}\sum_{r=1}^{NR}\left(\hat{\gamma}_{ir} - \gamma_i\right)^2},$$

with γ_i being the parameter from which the standard deviation is estimated. This value mirrors exactly the variation of the estimates obtained by the estimation procedures. Hence, it may also be interesting to use this value as reference value for the standard-error estimates.

In Table 8, values of the original standard error, of the empirical standard deviation, and of the (expected) estimated standard error yielded by LISREL polychoric estimation procedure are presented for several CFA models. The original and the empirical values differ somewhat from each other. As a result, the standard-error estimates can be relatively close to their empirical value, even though they are quite far away from their original true value, and vice versa.

Because the stress will be put here on the correctness of the modeling of the parameters variation and because the empirical standard deviations of the estimates from the two procedures are often very close, the empirical standard deviations are considered to be the reference standard errors here.