

# A Threshold-Free Approach to the Study of the Structure of Binary Data

Karl Schweizer<sup>1</sup>

<sup>1</sup> Department of Psychology, Goethe University Frankfurt, Frankfurt a. M., Germany

Correspondence: Karl Schweizer, Department of Psychology, Goethe University Frankfurt, Frankfurt a. M. 60323, Germany. Tel: 49-69-7983-5355. E-mail: K.Schweizer@psych.uni-frankfurt.de

Received: February 24, 2013 Accepted: April 3, 2013 Online Published: April 16, 2013

doi:10.5539/ijsp.v2n2p67

URL: <http://dx.doi.org/10.5539/ijsp.v2n2p67>

## Abstract

A major characteristic of the threshold-free approach to the investigation of the structure of binary data is the step from binary to continuous by computing probabilities instead of estimating thresholds. Another characteristic is the consideration of the shift from the distribution of the binary data to the normal distribution of the latent variables at the level of variances and covariances. Two ways of relating the distributions are considered: standardization and modifying the assumed model of measurement accordingly. Furthermore, there is the consideration of the change in the proportion of true variance. A method for estimating the effect on the completely standardized factor loadings is proposed. In an example the various steps of this threshold-free approach of investigating the structure of binary data are demonstrated. A major advantage of this approach is the avoidance of estimating thresholds that requires especially large samples.

**Keywords:** binary data, structural equation modeling, generalized linear model, link function, probability-based covariance

## 1. Introduction

The investigation of binary data by means of factor-analytic methods means relating binary variables following a binomial distribution to continuous variables showing a normal distribution. The preferred way of bridging the gap between binary and continuous variables requires the estimation of latent thresholds (Muthen, 1984; Raykov & Mels, 2009). It is denoted *threshold approach* in this paper. An investigation of the relationship between binary and continuous variables, as it is established as part of factor-analytic methods, revealed the equivalence of this relationship and the relationship established as part of methods based on item response theory (Takane & de Leeuwe, 1987). An unfortunate characteristic of methods requiring the estimation of latent parameter serving as thresholds is the need for very large samples. The *threshold-free approach* is proposed in order to overcome this need. Instead of estimating thresholds probabilities are computed that are achievable on the basis of a smaller sample.

The threshold-free approach includes three steps for bridging the gap between binary and continuous variables in confirmatory factor analysis. The first step is concerning the scale level. The switch from binary to continuous is achieved by computing probabilities that are subsequently transformed into probability-based covariances and probability-based correlations. Probabilities also play a major role in switching the levels in item response theory (Hambleton, Swaminathan, & Rogers, 1991) although they contribute in a different way. The second step is regarding the difference between the distributions since binary variables show a binomial distribution and the latent continuous variables included in factor-analytic methods follow a normal distribution. Variables following different distributions need to be related to each other by means of a link function (McCullagh & Nelder, 1985; Nelder & Wedderburn, 1972). In the third step the focus is on the proportion of true variance. Binary variables can be thought of as the result of dichotomizing continuous variables which means a loss of information respectively a diminution of the proportion of true variance. Such variables can be assumed to show a smaller proportion of true variance than corresponding continuous variables.

In the following sections the threshold-free approach is described by considering the three steps. Furthermore, it is applied to an example in order to demonstrate its usefulness and to provide instruction for users.

## 2. The Step from Ordered Categories to the Continuous Scale

This section presents the method of transforming binary data into covariances that is in agreement with three requirements: first, the transformation does not include assumptions concerning the number of underlying dimensions. Second, the mathematical operations are in agreement with the characteristics of the data. Third, probabilities provide the outset.

The definition of the covariance considered in this paper can be found in old textbooks on statistics. It was proposed in order to have an economic way of computing covariances during a time when computer-capacities were rather limited. Presenting the definition requires the assumption of the real-valued random variables  $R$  and  $S$ . Given these random variables and also the expected values  $E()$  of  $R$  and  $S$  and of their product the definition of the covariance  $\text{cov}(R, S)$  is provided by

$$\text{cov}(R, S) = E(R \cdot S) - E(R)E(S) \quad (1)$$

By adding weights this covariance is transformed into a covariance corresponding to the covariance based on cross-products. This definition was developed for continuous variables and can easily be transformed into a definition for binary variables.

The definition of the probability-based covariance omits weights. Furthermore, the continuous variables are restricted to binary variables where the two possible events are coded as 0 and 1. In this case the expected values correspond to the probabilities of the selected values. Let  $R$  be a binary random variable with 0 and 1 as values,  $\Pr(R = 0)$  and  $\Pr(R = 1)$  the corresponding probabilities and 1 the selected value. In this case the expected value is defined as the sum of the values assigned to the events that are weighted by the probabilities:

$$E(R) = 1 \cdot \Pr(R = 1) + 0 \cdot \Pr(R = 0) = 1 \cdot \Pr(R = 1) + 0 \cdot [1 - \Pr(R = 1)] = \Pr(R = 1) \quad (2)$$

Consequently, for the binary random variables  $R$  and  $S$  Equation (1) can be rewritten as

$$\text{cov}(R, S) = \Pr(R = 1 \wedge S = 1) - \Pr(R = 1)\Pr(S = 1) \quad (3)$$

where the computation of probabilities precedes subtraction and multiplication. This definition of the probability-based covariance also applies to the variance of  $R$  (and also of  $S$ ) since

$$\text{cov}(R, R) = \Pr(R = 1 \wedge R = 1) - \Pr(R = 1)\Pr(R = 1) = \Pr(R = 1) - \Pr(R = 1)^2 \quad (4)$$

The rearrangement of the components of this equation leads to the well-known formula for the computation of the variances of binary variables  $\text{var}(R)$ :

$$\text{var}(R) = \text{cov}(R, R) = \Pr(R = 1)[1 - \Pr(R = 1)] \quad (5)$$

Equation (3) can be used for computing the elements of the empirical  $q \times q$  covariance matrix  $\mathbf{S}$ . In the  $q$  binary random variables  $X_1, \dots, X_q$  with 0 and 1 as values it is given by

$$\mathbf{S} = [\text{cov}(X_i, X_j)]_{q \times q} \quad (6)$$

## 3. The Step from Binomial to Normal

The generalized linear model (McCullagh & Nelder, 1985; Nelder & Wedderburn, 1972) provides the framework for establishing a relationship between random variables following different distributions. This model assumes two random latent variables  $\eta$  and  $\mu$ . They serve as linear predictor showing a normal distribution and as criterion that is to be perceived as expected value following a distribution of the exponential family in corresponding order. The relationship between these variables is established by the link function  $g()$  such that

$$\eta = g(\mu) \quad (7)$$

(McCullagh & Nelder, 1985, pp. 19-20). Various link functions have been considered for this purpose, as for example the logit, the complementary log-log function, the inverse normal function and the inverse Cauchy.

Structural equation modeling is at its core a method for the investigation of covariance matrices (Jöreskog, 1970). Therefore this paper concentrates not on relating predictor and criterion variables but on relating the corresponding variances and covariances. In the following the focus is on variances. So in this paper the possibilities of establishing a relationship between variances by means of a link function are considered. Starting from Equation (7) a

relationship between the variance of the linear predictor and the variance of the criterion has to be established. The question is whether it is possible to specify  $g()$  in such a way that the following equation holds:

$$\text{var}(\eta) = g[\text{var}(\mu)] \quad (8)$$

The variance of  $\mu$  originates from binary data and the variance of  $\eta$  from continuous data with a normal distribution.

There are two major characteristics of distributions that need to be considered: shape and size. However, shape is only of importance as far as it influences the size of the variance. As illustrated by Equation (5), the size of the variance of the criterion variable depends on the probability of the selected event. Furthermore, it varies between 0 and 0.25. In the case of a probability of 0.5 the shape is symmetric and the variance is at its maximum. In contrast, the variance of the linear predictor is a property of the assumed model. It is mostly assumed to be one or one multiplied by a parameter. Consequently, the variances differ in two ways as a result of the difference between the distributions: there is a specific difference because of the deviation from symmetry and there is a general difference in size.

Two ways of bridging the specific difference are considered. The first means the standardization of the variances and covariances of the various criterion variables. They can be thought of as being associated with the binary variables of a dataset. Since these variances and covariances are based on probabilities serving as expected values, standardization is to be conducted with respect to these variances and covariances. The variance of the  $j$ th binary random variable  $X_j$   $\text{var}(X_j)$  computed according to Equation (5) can provide the outset for the computation of the weight  $w_j$ :

$$w_j = \left\{ \frac{1}{\text{Pr}(X_j)[1 - \text{Pr}(X_j)]} \right\}^{1/2} \quad (9)$$

The weight serves as a multiplier in the standardization of the covariances of  $X_j$  with the other random variables. The variances are standardized in a similar way. The standardization of the empirical  $q \times q$  covariance matrix  $\mathbf{S}$  of Equation (6) by means of such weights leads to a correlation matrix: the empirical  $q \times q$  matrix of probability-based correlations  $\mathbf{R}$ . This way is denoted the *criterion-based way*.

A possible disadvantage of the criterion-based way is that it applies equally to true and error components of the model of measurement. However, according to McCullagh and Nelder (1985, pp. 19-20)  $\mu$  as expected value does not include an error component. Therefore, a second way that concentrates on the linear predictor is considered. This change of perspective requires that the inverse function  $f()$  of  $g()$ , which in this case is a weight, is considered and Equation (8) is changed accordingly:

$$\text{var}(\mu) = f[\text{var}(\eta)] \quad (10)$$

In Equation (10) the variance of the linear predictor is adapted to the variance of the criterion.

This change of perspective enables the consideration of the model of the covariance matrix that is closely associated with the model of measurement. According to the model of the covariance matrix (e.g., Bollen, 1989, p. 18) the variance of the  $j$ th criterion variable  $\sigma_j$  is composed of a true component which is the product of the  $j$ th factor loading  $\lambda_j$  and the variance of the latent variable  $\phi$  and of the  $j$ th error variance  $\theta_j$ :

$$\sigma_j = \lambda_j \phi \lambda_j + \theta_j \quad (11)$$

The first summand of the right-hand part represents the true component of variance. The second way requires that this component is adapted adequately. Such an adaptation can be achieved by constraining the factor loadings to specific numbers, for example to the number 1, and adding a weight that reflects the probability. The weighted version of the tau-equivalent model (Schweizer, 2012a) includes such a weight:

$$\lambda_{\tau|j} = \left\{ \frac{\text{Pr}(X_j)[1 - \text{Pr}(X_j)]}{0.25} \right\}^{1/2} \lambda_{\tau} \quad (12)$$

The factor loading  $\lambda_{\tau}$  refers to the original tau-equivalent model and is usually a constant and  $\lambda_{\tau|j}$  the factor loading of the weighted tau-equivalent model that is expected to reflect the effect of the deviation from symmetry on the variance of the  $j$ th manifest variable  $X_j$ . This way is denoted the *predictor-based way*. It is in line with methods considered in previous research on the effect of the item position (Schweizer, 2012b; Schweizer, Schreiner, & Gold, 2009).

Furthermore, there is the general difference in the sizes of the variances associated with the criterion and predictor variables. Fortunately, there is no need for an additional link function because as part of the estimation process

of confirmatory factor analysis the model of the covariance matrix is adjusted to the empirical covariance matrix. The difference between the variances and covariances of the empirical and theoretical matrices is minimized by estimating either the factor loadings (e.g.  $\lambda_j$ ) or the variance of the latent variable ( $\phi$ ) as well as the error variance accordingly. Consequently, the estimates can be expected to reflect the general difference in the size of variances. Therefore the model fit which serves the evaluation of the appropriateness of the model is not influenced by this general difference, and no further modification is necessary.

#### 4. The Disattenuation of the Completely Standardized Factor Loadings

Another important issue is the effect of the switch from binary to continuous on the proportion of true variance. The concept of reliability of classical test theory (Lord & Novick, 1968) suggests that increasing test length changes the relationship of true and error variance. It was proposed with tests composed of binary items in mind. According to this concept the switch from binary to continuous can be assumed to be associated with the disproportional increase of the types of variance. The true variance should increase faster than the error variance. This disproportional increase is reflected by the completely standardized factor loadings since the square of a completely standardized factor loading is interpreted as the proportion of true variance (Brown, 2006, p. 133).

The formula developed for the variance of the sum of two random variables provides the basis for the reasoning. Let  $R$  and  $S$  be two random variables. Then the variance of the sum  $\text{var}(R + S)$  is given by

$$\text{var}(R + S) = \text{var}(R) + \text{var}(S) + r_{RS}\text{var}(R)^{1/2}\text{var}(S)^{1/2} \quad (13)$$

The correlation of  $R$  and  $S$  is represented by  $r_{RS}$ . For reasons of simplicity it is assumed that the variances of the two random variables correspond. It is obvious that in the case of uncorrelated variables there is a doubling of the original variance characterizing both variables whereas in the case of perfectly correlated variables the variance of the sum is four times the original variance. So, in the case of true variance that can be assumed to be due to perfectly correlated random variables the multiplier must be four while it must be only two in the case of error variance.

The change in the proportion of true variance becomes especially apparent in the completely standardized factor loading that is obtained in relating the true variance, for example as it is defined in Equation (11), to the complete variance, as it is also given by Equation (11). In order to highlight the change resulting from doubling the length, multipliers are added to the various parts of the ratio with respect to the  $j$ th random variable  $X_j$  in the left-hand part of the following Equation:

$$\frac{4 \cdot \lambda_j \phi \lambda_j}{4 \cdot \lambda_j \phi \lambda_j + 2 \cdot \theta_j} = \frac{(2^{1/2} \cdot \lambda_j) \phi (2^{1/2} \cdot \lambda_j)}{2 \cdot \lambda_j \phi \lambda_j + \theta_j} \quad (14)$$

The multipliers added to the true and error components are 4 and 2 in corresponding order. The right-hand part of Equation (14) is achieved by simple transformations and reordering.

We assume that the shift from the variance of the binary variable showing a symmetric distribution that is 0.25 to the variance of the predictor variable of 1.0 is achieved by doubling the measurement. In this case Equation (14) provides the opportunity to estimate the effect on the completely standardized factor loading. Since  $\phi$  is usually set equal to one, Equation (14) suggests that the disattenuated and completely standardized factor loading  $\lambda_{\text{disattenuated-completely standardized}|j}$  can be estimated from the original completely standardized factor loading  $\lambda_{\text{completely standardized}|j} (= \lambda_j)$  in the following way:

$$\lambda_{\text{disattenuated-completely standardized}|j} = 2^{1/2} \lambda_{\text{completely standardized}|j} \quad (15)$$

However, since the assumption suggesting the doubling of the measure can only be true if there is no error variance, this way of estimating disattenuated and completely standardized factor loadings may be regarded as a lower limit in disattenuation.

#### 5. Demonstration

Random data were generated according to a specific population pattern. This pattern was a  $9 \times 9$  correlation matrix. All the correlations showed the same size: .25. The upper half of Table 1 includes the lower triangle of this population pattern. The next step served the generation of a  $400 \times 9$  matrix of random data. The numbers of the columns of this matrix were distributed normally with a mean of zero and a variance of one. In order to generate an internal structure according to the population pattern, this matrix was re-computed in using weights achieved by means of a procedure proposed by Jöreskog and Sörbom (2001). The result was a  $400 \times 9$  matrix of simulated data.

In the following step a covariance matrix was computed in order to illustrate the deviations from the population pattern resulting from the use of random data. It is provided in the lower half of Table 1. This covariance matrix computed from simulated data provided the outset for the investigation of structure. It also served as a comparison level since it was based on continuous data that could be assumed to be normally distributed and structural equation modeling of such data was undisputed.

Table 1. Population pattern used for data generation (upper half) and covariances computed from simulated data before dichotomization (lower half)

Population pattern								
1.000								
.250	1.000							
.250	.250	1.000						
.250	.250	.250	1.000					
.250	.250	.250	.250	1.000				
.250	.250	.250	.250	.250	1.000			
.250	.250	.250	.250	.250	.250	1.000		
.250	.250	.250	.250	.250	.250	.250	1.000	
.250	.250	.250	.250	.250	.250	.250	.250	1.000
Covariances based on simulated data								
1.071								
0.361	1.086							
0.304	0.351	0.979						
0.291	0.367	0.217	0.998					
0.286	0.302	0.203	0.201	0.967				
0.293	0.218	0.253	0.171	0.211	0.960			
0.271	0.373	0.292	0.251	0.237	0.282	1.021		
0.310	0.399	0.261	0.287	0.235	0.219	0.276	1.077	
0.283	0.248	0.291	0.249	0.234	0.252	0.300	0.285	0.927

In the next step the numbers of the columns of the re-computed matrix were dichotomized by transforming them into zeros and ones. Nine proportions were selected for the splits: .10, .20, .30, .40, .50, .60, .70, .80 and .90. In the first column the 40 smallest numbers (10 percent) were transformed into zeros and the remaining numbers into ones. In the second column the 80 smallest numbers (20 percent) were replaced by zeros and the remaining numbers by ones. Columns three to nine were processed accordingly. These data were used for computing probabilities as expected values and probability-based covariances. The upper half of Table 2 provides the probabilities serving as expected values for the pairs of columns of the matrix of simulated data. The main diagonal includes the probabilities for individual columns. It can be seen that in two cases there is no exact correspondence between the expected value and the split. These deviations were due to the fact that in some cases the generation of random data yielded the same number repeatedly so that it was not possible to automatically split them exactly according to the given proportion. The probabilities of the combination of columns varied between .035 and .735.

The transformation of the expected values according to Equation (3) led to the probability-based covariances that are presented in the lower part of Table 2. The main diagonal of this lower triangle matrix includes the variances. They varied between 0.09 and 0.25. Furthermore, the covariances varied between 0.005 and 0.042.

### 5.1 The Confirmatory Factor Models

All confirmatory factor models for investigating the data included one latent variable and nine manifest variables since they represented the assumption that there should be one underlying source of systematic variation. These models were designed according to the proposed ways of investigating binary data. They were used to investigate the population pattern and the covariance matrix based on continuous data in order to conduct comparisons. Confirmatory factor analysis was originally proposed as a method for the investigation of covariances (Jöreskog, 1970). Therefore, preference was given to the investigation of covariance matrices. However, there were also cases demanding the investigation of a correlation matrix.

The first model was constructed according to the criterion-based way. This included a simple congeneric model (Jöreskog, 1971) comprising one latent and nine manifest variables. It was to be applied to probability-based

correlations. The other model served the realization of the predictor-based way. It was constructed according to the weighted version of the tau-equivalent model (Schweizer, 2012a) that also including one latent and nine manifest variables. The factor loadings were set to one and multiplied by a weight, as shown in Equation (12). This model had to be applied to probability-based covariances.

Table 2. Probabilities as expected values obtained for the dichotomized data (upper half) and probability-based covariances (lower half)

Probabilities serving as expected values									
.900									
.735	.797								
.642	.590	.697							
.560	.520	.445	.600						
.470	.422	.385	.332	.500					
.385	.347	.305	.260	.235	.400				
.285	.270	.235	.222	.170	.155	.300			
.192	.187	.172	.160	.125	.107	.090	.200		
.095	.085	.082	.077	.060	.065	.047	.035	.100	
Probability-based covariances									
0.090									
0.017	0.161								
0.014	0.033	0.211							
0.020	0.041	0.026	0.240						
0.020	0.023	0.036	0.032	0.250					
0.025	0.028	0.026	0.020	0.035	0.240				
0.015	0.030	0.025	0.042	0.020	0.035	0.210			
0.012	0.028	0.033	0.040	0.025	0.027	0.030	0.160		
0.005	0.005	0.012	0.017	0.010	0.025	0.017	0.015	0.090	

The models for investigating the population pattern and the covariance matrix based on continuous data were congeneric models. Since the congeneric model most often characterized confirmatory factor analysis, it was considered as the *standard* model and is addressed accordingly in this paper. Each one of these models included one latent and nine manifest variables. In order to investigate the population pattern it was necessary to specify the sample size although there was no sample. It was set to the same number as in all the other models.

Some of the columns of the matrix of dichotomized data showed considerable deviations from symmetry, i. e. equal numbers of zeros and ones. This lack of symmetry in a way meant nonnormality and was considered as a major problem for the investigation of model fit (Curran, West, & Finch, 1996; Fan & Hancock, 2012). Therefore a specific estimation method had to be applied in combination with the criterion-based and predictor-based ways of investigating binary data. It was the robust estimation method proposed by Satorra and Bentler (1994; Bryant & Satorra, 2012).

The investigations were conducted by means of LISREL (Jöreskog & Sörbom, 2006). Fit results and completely standardized factor loading were considered in the evaluation. The report of the results of investigating model fit includes the following statistics: chi-squares, degrees of freedom, normed chi-squares, *RMSEA*, *SRMR*, *CFI*, *TLI* and *AIC*. Cut-offs provided by Hu and Bentler (1999) served the evaluation of the results (*RMSEA* .06, *SRMR* .08, *CFI* .95, *TLI* .95). Furthermore, following Bollen (1989) normed chi-squares below 2 were considered as an indication of a good model fit while values of below 3 were considered acceptable. The remaining statistics were not associated with a cut-off.

## 5.2 The Results

The fit results observed in investigating the covariances and correlations are presented in Table 3. The first and second row were based on continuous data and constituted the comparison level for the results presented in the other rows. As is obvious from the first row, the investigation of the population pattern yielded incomplete results probably because of the lack of variability. All the statistics indicated a good model fit. *RMSEA* and *CFI* even signified a perfect degree of model fit for all investigations of models. The other fit statistics showed minor variations that did not exceed the selected cut-offs. The difference in chi-squares between the standard model

applied to continuous data and the models according to the threshold-free approach was surprisingly large. It could be explained by the difference between the estimation methods: normal ML estimation versus robust estimation.

Since the models were not nested, it was not possible to apply the chi-square difference test. The comparison by means of *AIC* revealed that the predictor-based way led to the best model fit and that the criterion-based way ranked second in the absence of a result for the population pattern. The good model fit for the predictor-based way was partly due to the high number of degrees of freedom resulting from the constraint of the factor loadings.

Most interesting was the observation that the various ways did not differ according to the *CFI* results. This was a crucial observation since according to the work of Cheung and Rensvold (2002) *CFI* differences larger than 0.01 would have indicated a substantial difference between the models. Apparently, the various ways of investigating the structure of binary data and corresponding continuous data did not differ substantially according to model fit. Since all models showed the same structure and the binary data were derived from the continuous data, this outcome suggested that the two ways of the threshold-free approach of investigating binary data did reasonably well.

Table 3. Results of investigating model fit by means of the standard model and models according to the threshold-free approach

Way	Data type	$\chi^2$	df	Normed $\chi^2$	RMSEA	SRMR	CFI	TLI	AIC
Continuous data as basis									
Standard	Pattern	0.00	27	0.00	-	-	-	-	-
Standard	Covariances	25.04	27	0.93	0.000	0.027	1.00	1.00	61.0
Binary data as basis									
Criterion-based <sup>1</sup>	PCor <sup>2</sup>	4.15	27	0.15	0.000	0.027	1.00	1.08	40.1
Predictor-based <sup>1</sup>	PCov <sup>3</sup>	0.12	35	0.01	0.000	0.036	1.00	1.08	20.1

<sup>1</sup> Parameter estimation with robust method.

<sup>2</sup> Probability-based correlations.

<sup>3</sup> Probability-based covariances.

Next, the completely standardized factor loadings were considered. Table 4 provides these factor loadings. Again the results obtained for continuous data are presented first and for binary data subsequently. As expected, investigating the population pattern by the standard model led to a uniform loading pattern. All factor loadings were .50 and reached the level of significance. The investigation of the covariance matrix computed from continuous data by means of the standard model revealed factor loadings that varied between .45 and .62. The mean of these factor loadings was .52 which slightly surmounted the mean for the population pattern. As expected, there was no need for disattenuation of the factor loadings achieved in continuous data.

Table 4. Completely standardized factor loadings obtained by means of the standard model and models according to the threshold-free approach

Way	Data type	Number of manifest variables								
		1	2	3	4	5	6	7	8	9
Continuous data as basis										
Standard	Pattern	0.50*	0.50*	0.50*	0.50*	0.50*	0.50*	0.50*	0.50*	0.50*
Standard	Covariances	0.56*	0.62*	0.53*	0.49*	0.46*	0.45*	0.54*	0.53*	0.52*
Binary data as basis										
Criterion-based <sup>1</sup>	PCor <sup>2</sup>	0.32*	0.42*	0.37*	0.42*	0.32*	0.36*	0.39*	0.43*	0.26
	disattenuated	0.45	0.59	0.52	0.59	0.45	0.51	0.55	0.61	0.37
Predictor-based <sup>1</sup>	PCov <sup>3</sup>	0.36*	0.37*	0.37*	0.37*	0.36*	0.37*	0.37*	0.37*	0.36*
	disattenuated	0.51	0.52	0.52	0.52	0.51	0.52	0.52	0.52	0.51

<sup>1</sup> Parameter estimation with robust method.

<sup>2</sup> Probability-based correlations.

<sup>3</sup> Probability-based covariances.

\*  $p < .05$  (concerning factor loading).

\*  $p < .05$  (concerning variance of the corresponding latent variable).

In contrast, two factor loadings were computed for each manifest variable by applying the criterion-based and predictor-based ways of the threshold-free approach: one without disattenuation and one with disattenuation. In the criterion-based way the range of factor loadings was between .26 and .42 with a mean factor loading of .36. All factor loadings with the exception of the last one reached the level of significance. After disattenuation these loadings varied between 0.37 and .59. The mean of disattenuated factor loadings was .51. The completely standardized factor loadings of the predictor-based way that were modified constraints varied between .36 and .37 without disattenuation and between .51 and .52 with disattenuation. The means were .37 and .52 in corresponding order. It was also important to inspect these constraints since the weights according to Equation (12) could have caused a systematic deviation from equal sizes after standardization.

Obviously, disattenuation was necessary in order to achieve means in investigating binary data, which were close to the means observed in investigating the population pattern and continuous data. It was interesting to find that the disattenuated factor loadings of the predictor-based way showed the lowest overall deviations from the factor loadings obtained for the population pattern.

Finally, the binary data were investigated according to the threshold approach. Tetrachoric correlations were computed and investigated by the congeneric model in considering the robust estimation method. The fit results are presented in the first row of Table 5. For comparison the results observed by the criterion-based and predictor-based ways of the threshold-free approach are given in the other rows of this table. Despite the robust estimation method only the *SRMR* results indicated a good model fit for the threshold approach. Comparing the upper and lower parts of this Table revealed considerable differences. Apparently, the model fit achieved in the investigation according to the threshold approach was not acceptable.

Table 5. Fit results achieved in investigations according to the threshold and threshold-free approaches

Way	Data type	$\chi^2$	df	Normed $\chi^2$	RMSEA	SRMR	CFI	TLI	AIC
Threshold approach									
Standard	TCor <sup>2</sup>	98.65	27	3.65	0.082	0.053	0.88	0.84	134.6
Threshold-free approach									
Criterion-based <sup>1</sup>	PCor <sup>3</sup>	4.15	27	0.15	0.000	0.027	1.00	1.08	40.1
Predictor-based <sup>1</sup>	PCov <sup>4</sup>	0.12	35	0.01	0.000	0.036	1.00	1.08	20.1

<sup>1</sup> Parameter estimation with robust method.

<sup>2</sup> Tetrachoric correlations.

<sup>3</sup> Probability-based correlations.

<sup>4</sup> Probability-based covariances.

## 6. Discussion

The introduction of a threshold-free approach for the investigation of the structure of binary data by means of confirmatory factor analysis is described as the major aim in the introductory section of this paper since the consideration of thresholds implies a considerable demand to the estimation process in the investigation of structure. Meeting this demand usually means providing a very large dataset. Although only one dataset of agreeable size was investigated for a demonstration of the threshold approach, this investigation provided a very convincing corroboration of the demand. Furthermore, there is another point which may additionally influence the outcome and needs to be mentioned. Data analysis according to the threshold approach implies two separate estimation processes: the estimation of the thresholds as part of the computation of tetrachoric correlations and the estimation of the parameters of the structural model. These are estimation processes according to different models. In datasets showing a high quality the combination of these estimation processes can be expected to do very well. However, if the quality of the data is less than optimal, the first estimation process can cause an exaggeration of the deviations from the expected structure.

In the threshold-free approach the repeated estimation of parameters is avoided by concentrating on the parameters of the structural model. The estimation of thresholds is replaced by the computation and adaptation of sample statistics. This approach initially requires the computation of probabilities. This includes the transformation of the probabilities into probability-based covariances or probability-based correlations. Adjustments for bridging the difference in variance between the binomial and normal distributions add up to a second step. These adjustments follow the logic of the generalized linear model by McCullagh and Nelder (1985, p. 21) at the level of variances

and covariances. Finally, there is the disattenuation of completely standardized factor loadings in order to achieve values close to the completely standardized factor loadings for the population pattern. However, the disattenuation step is only an auxiliary component of the threshold-free approach which can be helpful in checking the appropriateness of the outcome. Corresponding results achieved for binary and continuous data can not really be expected since the original step from continuous to binary means a loss of information.

Two ways of dealing with the distributional differences are considered. The criterion-based way requires the application of the congeneric model of measurement (Jöreskog, 1971), which is the standard model of confirmatory factor analysis, to probability-based correlations. The predictor-based way is more demanding because it requires the investigation of probability-based covariances by means of the weighted version of the tau-equivalent model of measurement (Schweizer, 2012a). In the demonstration both ways did equally well according to the fit statistics. In contrast, the disattenuated and completely standardized factor loadings obtained in the predictor-based way were much closer to the population statistics than the disattenuated and completely standardized factor loadings for the other way. The advantage of the predictor-based way is presumably due to the fact that both the tau-equivalent and data-generation models assumed one latent source that equally contributes to all the manifest variables. If there were a discrepancy between the models, the predictor-based way could be expected to yield the less favourable results.

Although the simulated data did not mean a severe challenge to the methods according to the two ways of the threshold-free approach, the demonstration made characteristic similarities and differences obvious. Furthermore, the demonstration substantiated the claim that the threshold and threshold-free approaches are differently demanding to the quality of the data.

## References

- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143. <http://dx.doi.org/10.1007/s11336-008-9100-1>
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.
- Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling*, *19*, 373-398. <http://dx.doi.org/10.1080/10705511.2012.687671>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. <http://dx.doi.org/10.1207/S15328007SEM0902-5>
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification errors in confirmatory factor analysis. *Psychological Methods*, *1*, 16-29. <http://dx.doi.org/10.1037/1082-9>
- Fan, W., & Hancock, G. R. (2012). Robust means modeling: an alternative for hypothesis testing of independent means under variance heterogeneity and nonnormality. *Journal of Educational and Behavioral Statistics*, *37*, 137-156. <http://dx.doi.org/10.3102/1076998610396897>
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, *66*, 930-944. <http://dx.doi.org/10.1177/0013164406288165>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. <http://dx.doi.org/10.1080/10705519909540118>
- Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika*, *57*, 239-257.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133. <http://dx.doi.org/10.1007/BF02291393>
- Jöreskog, K. G., & Sörbom, D. (2001). *Interactive LISREL: Users Guide*. Lincolnwood, IL: Scientific Software International Inc.
- Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80*. Lincolnwood, IL: Scientific Software International Inc.
- McCullagh, P., & Nelder, J. A. (1985). *Generalized linear models*. London: Chapman and Hall.

- Muthen, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variables indicators. *Psychometrika*, *32*, 1-13. <http://dx.doi.org/10.1007/BF02294210>
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, *135*, 370-384.
- Raykov, T., & Mels, G. (2009). Interval estimation of inter-item and item-total correlations for ordinal items of multiple-component measuring instruments. *Structural Equation Modeling*, *16*, 99-108. <http://dx.doi.org/10.1080/10705510802561337>
- Satorra, A., & Bentler, P. M. (1994). Corrections to the test statistics and standard errors on covariance structure analysis. In A. von Eye, & C. C. Glogg (Eds.), *Latent variable analysis* (pp. 399-419). Thousand Oaks, CA: Sage.
- Schweizer, K. (2012a). A weighted version of the tau-equivalent model of measurement for items with ordered response categories. *International Journal of Statistics and Probability*, *1*, 151-163. <http://dx.doi.org/10.5539/ijsp.v1n2p151>
- Schweizer, K. (2012b). The position effect in reasoning items considered from the CFA perspective. *International Journal of Educational and Psychological Assessment*, *11*, 44-58.
- Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: a two-dimensional model of APM. *Psychology Science Quarterly*, *51*, 47-64.
- Takane, Y., & de Leeuwe, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408. <http://dx.doi.org/10.1007/BF02294363>