Lund University

Department of Statistics

Factor Analysis of Ordinal Variables

Polychoric and Pearson correlations in the exploratory approach

Author

Theodor Emanuelsson

Bachelor's thesis in Statistics.

Course code: STAH11 (15HP)

Supervisor: Björn Holmquist

February 5, 2021

Abstract

This thesis considers the problem of applying exploratory factor analysis to data obtained though the Likert scale. It is often the case that this sort of data is treated as an interval level of measurement and used in analyses that require continuous variables, although given the categorization and nature of the scale, it should rather be treated as of ordinal level. Exploratory factor analysis is a prevailing technique for studying the construct validity of data, a method that relies on the correlation matrix of the data to obtain factor solutions. Previous research recommend the use of polychoric correlations as opposed to the common Pearson correlation method when attempting to apply factor analysis to ordinal data. Preceding research is complemented with further consideration of asymmetry in the data. The results, obtained through simulation studies, show that the polychoric correlations provide a more accurate reproduction of the theoretical model used to generate the data in three cases of item skewness.

Keywords ·Exploratory factor analysis ·Polychoric correlation ·Pearson correlation

Introduction

Likert scale is a commonly used psychometric tool in numerous scientific fields of study, specifically within psychology, education and social science. It is considered in this study because of the prevalent usage of factor analysis on Likert scale items for dimensionality reduction and construct validity. Researchers aiming to use this type of model on a set of Likert scale items generally face a difficulty with the statistical assumptions that are prevalent for estimation of correlation coefficients. The aim of this thesis is therefore to illustrate the possible advantages and disadvantages of utilizing two types of correlation matrices, the Pearson and the polychoric, for exploratory factor analysis.

The general goal of a factor analysis is to take a number of observable interrelated *manifest variables* and infer one or many latent variables, also called *factors*. There may be an underlying explanation when a group of variables has considerable interrelatedness, that they in fact are measurements of a latent factor. How much these variables are related could be quantified through correlation. If one, for example, measures the length of each finger from a group of people and correlate the observations, one could expect to find an association between measurements. This association constitutes a factor, underlying the observed measurements. In this example the underlying factor is most likely related to the size of the hand.

The Likert scale is in some ways the mean to a difficult end, aiming to measure and quantify an attitude, a perception or opinion. These attitudes are by nature qualitative and researchers need to be careful if aiming to transform them into something quantitative. The Likert scale was introduced in 1932 as a possible solution to the issue of quantifying the subjective preferential attitude in a scientifically accepted and validated manner (Joshi et al. 2015).

A long lasting debate ensued from differing aspects of the scale and how to properly analyze responses. A common issue is if the scale should be symmetrical, in addition if and where a "neutral" or "don't know" option could be positioned within the continuum. Another issue is the length of the scale and how the differences between using, for example, a 5 point scale performs compared to a 7 or even a 10 point scale. When analyzing the scale, the most frequent and in many regards fundamental issue a researcher faces is which level of measurement the Likert scale is. There are differing schools of thoughts - some consider the Likert scale as measured at ordinal level, while others as measured on interval level, the latter school treating the measurement as continuous. It logically becomes a question about whether the points on an item are equivalent or equidistant.

They generally cannot be considered equidistant, although applied researchers sometimes makes this assumption (Joshi et al. 2015; Lantz 2013).

Variables of ordinal level are in essence categorical variables whose values can be ordered. In terms of a Likert scale, for example, we know someone who responds "4. Agree" to a given statement has rated their attitude towards the statement lower than someone who responds "5. Strongly agree". It is apparent that there exists an underlying continuum of opinion and researchers ask respondents to rate where on this continuum they position themselves, given a number of alternatives. It is however uncertain what the distance between "4. Agree" and "5. Strongly Agree" is in relation of the underlying opinion. Moreover, it is unclear if that distance is equal to the distance between for instance "1. Strongly Disagree" and "2. Disagree". The consequence of this ordering of categorical variables without specifically defined distances between each point of the discrete scale makes it so that these variables cannot be added or multiplied. It is therefore not particularly interesting to consider for example the mean or the covariance of a number of observations from a Likert scale. For the mean to actually make sense, the distances of measurement need to be continuous. Instead statisticians usually view pairs of ordinal variables in contingency tables where the rows and columns represents each ordinal variable. The corresponding frequencies of observation are presented in the cells. Consider the *Inheritance of Eye-colour in Man* data presented in Table 1.1 as an example. Pearson used this type of data in his development of what is now called the polychoric correlation coefficient in his article *On the Correlation of Characters not Quantitatively Measurable* (Pearson 1900).

Table 1.1: Karl Pearson's table of eye-color between maternal grandmothers and their granddaughters

Tint	Granddaughter	Maternal grandmother		Totals
		Gray or lighter	Dark gray or darker	
	Gray or lighter	254	136	390
	Dark gray or darker	156	193	349
Totals		410	329	739

Source: See page 39 in K. Pearson. "Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable." In: *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 195. (1900)

There are comparative studies of the two correlation coefficients, studied in this thesis, when dealing with ordinal data (Choi, Peters, and Mueller 2010). In addition, the coefficients has been compared in terms of method of estimation when used in confirmatory factor analysis (Angeles Morata-Ramirez and F. P. Holgado-Tello 2013). This thesis complement the research of F. Holgado-Tello et al. (2010), by comparing the accuracy of

Pearson and polychoric correlation in the case of exploratory factor analysis with a more extensive consideration of asymmetry in the data.

1.1 Two types of factor analysis

There are two basic types of factor analysis available to researchers depending on the purpose of constructing the models. DeCoster (1998) outlines the differences and criteria in a proper way. *Exploratory* factor analysis attempts to discover latent constructs that influence the set of responses, while *confirmatory* factor analysis is used to test whether a specified set of constructs is influencing responses in a predicted manner.

1.1.1 Confirmatory factor analysis

The primary objective of confirmatory factor analysis is to determine the ability of a predefined factor model to fit an observed set of data. It is commonly used to:

- Establish the validity of a single factor model.
- Compare the ability of two different model to account for the same set of data.
- Test the significance of a specific factor loading.
- Test the relationship between two or more factor loadings.
- Test whether a set of factors are correlated or uncorrelated.
- Assess the convergent and discriminant validity of a set of measures.

DeCoster (ibid.) points out six important steps to a confirmatory model:

1. **Define the factor model**. This sort of model builds on an already proposed theoretical basis, making it necessary to already have a defined model that one wishes to test. This includes the number of factors selected and the nature of the loadings between factors and measures. The factor loadings can be allowed to vary freely, to be fixed at a specific constant or to be able to vary under specified constraints. For models analyzing multiple factors simultaneously, cross-loadings are typically fixed at zero, meaning that observations where there are no theoretical grounds for a relationship with a specific factor have their loadings fixed.
2. **Collect measurements**. Variables should be measured on the same (or matched) experimental units.
3. **Obtain the correlation matrix**. Obtain the correlations (or covariances) between each of the variables.
4. **Fit the model to the data**. A method to obtain estimates of factor loadings if they were free to vary must be considered. The most common method is maximum likelihood estimation. It has been shown to be fairly robust against departures

from multivariate normality (Finney and DiStefano 2014). If the observed variables, however, seriously lack multivariate normality, asymptotically distribution free estimation is an option.

5. **Evaluate model adequacy**. When the factor model is fit to the data, the factor loadings are chosen to minimize the discrepency between the correlation matrix implied by the model and the actual observed matrix. The amount of discrepency after the best parameters are chosen can be used as a measure of how consistent the model is with the data. The model adequacy can be tested through χ^2 *goodness-of-fit* test.

6. **Compare with other models**. To compare a reduced model with the initial model one can perform a likelihood ratio test utilizing the difference of χ^2 statistics of the two models. This difference also follows a χ^2 distribution. If comparing two models where one is not a reduced version of the other, the *Root mean square error of approximation* (RMSEA) statistic can be used.

1.1.2 Exploratory factor analysis

The primary objective of exploratory factor analysis is to determine the number of common factors influencing a set of measures. In addition, researchers also need to determine the strength of the relationship between each factor and each observed measure.

- Identify the nature of the constructs underlying responses in a specific content area.
- Determine which sets of items that are related to each other in for example a questionnaire.
- Demonstrate the dimensionality of a measurement scale.
- Determine what features are most important when classifying a group of items.
- Generate "factor scores" representing values of the underlying constructs for use in other analyses.

DeCoster (1998) points out seven important steps to an exploratory model:

1. **Collect measurements**. Variables should be measured on the same (or matched) experimental units.

2. **Obtain the correlation matrix**. Obtain the correlations (or covariances) between each of the variables.

3. **Select the number of factors for inclusion**. This step needs some consideration. It could be the case that a researcher has a specific hypothesis that will determine the number of factors to extract. While at other times it is a question of choosing the number of factors that account for as much of the covariance in the data, with as

few factors as possible. The *Kaiser criterion* states that one should use the number of factors that is equal to the number of eigenvalues of the correlation matrix that are greater than one. Another approach is to plot the eigenvalues of the correlation matrix in descending order and then use the number of factors equal to the number of eigenvalues that occurs prior to the last major drop in the magnitude of the eigenvalues.

4. **Extract initial set of factors**. Submit correlation or covariance matrices to a computer program to extract factors. In this thesis I use R 4.0.3 (R Core Team 2020). There are a few options for extraction methods here, such as maximum likelihood, principal component and principal axis extraction.

5. **Rotate factors to a final solution**. For any given set of correlations and number of factors there is an infinite number of ways that the factors can be defined and still account for the same account of covariance in the measurements. Some of these definitions are however easier to interpret theoretically than others. It is therefore important to rotate factors in an attempt to find a more interpretable factor solution and it is still equal to that obtained in the initial extraction. For a comprehensive overview of rotations see Abdi (2004).

6. **Interpret factor structure**. In these analyses each measure has a linear relationship with each of the factors. The strength of this relationship is contained in the factor loading, produced by the rotation. The loading is similar to that of a standardized regression coefficient, regressing the factor on the measures.

7. **Construct factor scores for further analysis**. If the obtained model is to be used for further analysis using the factors as variables in, for example, a multiple regression analysis, one would need to estimate factor scores. There are estimates of the factor given the model. They are a linear combination of all of the measures, weighted by the corresponding factor loading.

It is important to note that exploratory factor analysis is more than just a dimensionality reduction tool, such as principal component analysis. Principal component analysis is used for data reduction, where researchers does not want to use all of the original measures, but instead wants to work with information they contain. Exploratory factor analysis on the other hand is to be used when researchers are interested in making statements about the factors that are responsible for a set of observed responses (DeCoster 1998). The confirmatory and exploratory approach is mathematically not very different. It mainly comes down to whether factor loadings are fixed at some value. In the exploratory case, loadings are always allowed to vary freely while cross-loadings, without theoretical grounding, are typically set at zero in the confirmatory case (Fabrigar and Wegener 2014).

Definitions and literature review

2.1 Ordinal data for factor analysis

As mentioned in the introduction, Likert scale items should be considered to be of ordinal level. For these sort of variables, we assume that an item is designed in order to measure a theoretical concept and the observed item responses are realizations of a small number of categories. Again, the distances between categories are unknown and in most cases unmeasureable. If we, however, use Muthén's view on the connection between ordinal and continuous variables for this type of data, then it is possible to estimate threshold parameters that can be used to estimate probabilities of two observed values on two ordinal variables (B. Muthén 1983).

There is assumed to exist a continuous variable, ${x_i}^*$, that underlies the ordinal variable x_i, $i = 1, 2, ..., p$. The continuous variable is assumed to be the true measure for a given attitude, underlying the ordered responses of x_i and it has a domain from $-\infty$ to ∞. For an ordinal variable, x_i with m_i categories, the relationship between the ordinal variable x_i and the underlying continuous variable ${x_i}^*$ is

$$x_i = c \iff \tau_{c-1}^{(i)} < {x_i}^* < {\tau_c}^{(i)}, \quad c = 1, 2, \ldots, m_i \,, \tag{2.1}$$

where

$$\tau_0^{(i)} = -\infty, \; \tau_1^{(i)} < \tau_2^{(i)} < ... < {\tau_{m_i-1}}^{(i)}, \; {\tau_{m_i}}^{(i)} = \infty, \tag{2.2}$$

For the observed ordinal variable x_i, there are $m_i - 1$ strictly increasing threshold parameters ${\tau_1}^{(i)}, {\tau_2}^{(i)}, ..., {\tau_{m_i-1}}^{(i)}$. Since only ordinal information about x_i is measured, the distribution of ${x_i}^*$ is determined only by a monotonic transformation. If one assumes a standard normal distribution for ${x_i}^*$, with density function $\phi(\bullet)$ and distribution function $\Phi(\bullet)$, the probability ${\pi_c}^{(i)}$ of a response in category c on variable x_i, is

$$\pi_c^{(i)} = Pr[x_i = c] = Pr[{\tau_{c-1}}^{(i)} < {x_i}^* < {\tau_c}^{(i)}] = \int_{{\tau_{c-1}}^{(i)}}^{{\tau_c}^{(i)}} \phi(u)du = \Phi({\tau_c}^{(i)}) - \Phi({\tau_{c-1}}^{(i)}) \tag{2.3}$$

for $c = 1, 2, \ldots, m_i - 1$ so that

$${\tau_c}^{(i)} = \Phi^{-1}({\pi_1}^{(i)} + \pi_2^{(i)} + \cdots + {\pi_c}^{(i)}) \tag{2.4}$$

Φ^{-1} is the inverse standard Gaussian distribution function and the quantity $({\pi_1}^{(i)} + \pi_2^{(i)} + ... + {\pi_c}^{(i)})$ is the probability of a response in category c or lower. The probabilities ${\pi_c}^{(i)}$

are unknown population quantities but can be estimated consistently by a corresponding percentage ${p_c}^{(i)}$ of responses in category c on variable x_i. Estimates of the thresholds can therefore be obtained as

$$\hat{\tau}_c{}^{(i)} = \Phi^{-1}({p_1}^{(i)} + {p_2}^{(i)} + ... + {p_c}^{(i)}), \quad c = 1, \ldots, m_i - 1 \tag{2.5}$$

The quantity $({p_1}^{(i)} + {p_2}^{(i)} + ... + {p_c}^{(i)})$ is the proportion of cases in the sample responding in category c or lower on variable x_i

2.2 Pearson correlation

Pearson correlation, also called the Pearson product-moment correlation coefficient, measures a linear correlation between two variables X and Y and is defined over the interval $I \in [-1, 1]$. It is simply the covariance of the random variables X and Y divided by their standard deviations.

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{2.6}$$

It is based upon a number of assumptions and has been shown to have nonrobust properties with respect to outliers (Devlin, Gnanadesikan, and Kettenring 1975). The assumptions are (I) the two correlated variables are continuous, (II) the relationship between the two variables is rectilinear, (III) the joint distribution of the scores is a bivariate normal distribution and (IV) the scores have been obtained in independent pairs, where each pair is not connected to other pairs (Havlicek and Peterson 1976). In the case where data is of ordinal level of measurement, the first methodological issue with using Pearson correlations for factor analysis is the first assumption of continuous variables.

2.3 Polychoric correlation

Polychoric correlation is a coefficient that measures the association for ordinal variables, and was proposed by Karl Pearson in the early 1900. Suppose we have two ordinal variables x_i and x_j with m_i and m_j categories, respectively. Suppose also that underlying x_i and x_j there exists some latent variable ξ and η, which are bivariate normally distributed with $E[\xi] = E[\eta] = 0$ and unit variances. The marginal distribution in the sample is represented by a contingency table

$$\begin{pmatrix} n_{11}^{(ij)} & n_{12}^{(ij)} & \cdots & n_{1m_j}^{(ij)} \\ n_{21}^{(ij)} & n_{22}^{(ij)} & \cdots & n_{2m_j}^{(ij)} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ n_{m_i1}^{(ij)} & n_{m_i2}^{(ij)} & \cdots & n_{m_im_j}^{(ij)} \end{pmatrix} \tag{2.7}$$

where n_{ab}^{ij} is the number of cases in the sample in category a on variable x_i and in category b on variable x_j. The correlation between ξ and η, ρ_{ij}, is the polychoric correlation. We can estimate this correlation by maximizing the log-likelihood of the multinomial distribution (Olsson 1979).

We do this in a two-step procedure. First thresholds are estimated from the univariate marginal distributions in Equation (2.5). Second the polychoric correlations are estimated from the bivariate marginal distributions by maximizing the log-likelihood for given thresholds.

If C is a constant then the likelihood function of the sample is:

$$L = C \cdot \prod_{i}^{m_i} \prod_{j}^{m_j} \pi_{ij}^{n_{ab}^{(ij)}}$$

Taking the natural logarithms,

$$l = \ln\ L = \ln C + \sum_{a=1}^{m_i} \sum_{b=1}^{m_j} n_{ab}^{(ij)} \ln \pi_{ab}^{(ij)} \tag{2.8}$$

where

$$\pi_{ab}^{(ij)} = Pr[x_i = a, x_j = b] = \int_{{\tau_{a-1}}^{(i)}}^{{\tau_a}^{(i)}} \int_{{\tau_{b-1}}^{(j)}}^{{\tau_b}^{(j)}} \phi_2(u, v) du dv$$

and

$$\phi_2(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho_{ij}^2)}(u^2 - 2\rho uv + v^2)}$$

is the standard bivariate normal density with correlation ρ_{ij}. Maximizing $\ln L$ gives the sample polychoric correlation, r_{ij}. In theory it is necessary to test the assumption of bivariate normality before calculating the polychoric correlation. To test the model the likelihood ratio (LR) test statistic can be used (Jöreskog 2005). Let $p_{ab}^{(ij)} = n_{ab}^{(ij)}/N$ be the sample proportions, then the LR-test statistic can be expressed as:

$$\chi_{LR}^2 = 2 \sum_{a=1}^{m_i} \sum_{b=1}^{m_j} n_{ab}^{(ij)} \ln \left(\frac{p_{ab}^{(ij)}}{\hat{\pi}_{ab}^{(ij)}} \right) \tag{2.9}$$

However, the polychoric correlation coefficient has been shown to be fairly robust to violations of the bivariate normality assumption (Coenders, Satorra, and Saris 1997). Although the normality assumption will not be evaluated by hypothesis testing in this thesis, the robustness of violations will be explored in comparison to the robustness of the bivariate normality assumption also present when performing Pearson correlations.

2.4 The mathematics of factor analysis

Confirmatory factor analysis is a common method within the larger research field of structural equation modeling. The mathematics of this type of model is briefly outlined here because of the mathematical similarity to the exploratory model. The model assumes a linear relationship between manifest variables and the underlying factor. Given the logic of manifest and factor variables outlined in the introduction, consider $x_1, x_2, \ldots, x_p$ to be observed manifest variables and $\xi_1, \xi_2, \ldots, \xi_m$ to be underlying factors, where $m < p$. The underlying factors account for the inner correlation of the observed manifest variables.

$$\boldsymbol{x} = \boldsymbol{\Lambda\xi} + \boldsymbol{\delta} \tag{2.10}$$

where $\boldsymbol{x}$ is a $p\times1$ vector and $\boldsymbol{\Lambda_{p\times m}}$ is the factor loadings matrix. $\boldsymbol{\xi}$ is the $m\times1$ vector of latent factors and $\boldsymbol{\delta}$ is a $p\times1$ vector of measurement errors. The measurement errors are assumed to be uncorrelated and corresponds to each of the p observed manifest variables. Based on Equation 2.10, we can formulate covariance matrices. Let $\boldsymbol{\Phi_{m\times m}}$ be the covariance matrix of $\boldsymbol{\xi}$ and $\boldsymbol{\Theta_{p\times p}}$ be the covariance matrix of $\boldsymbol{\delta}$. Since we are assuming measurement error to be uncorrelated, the $\boldsymbol{\Theta_{p\times p}}$ covariance matrix is diagonal. The covariance matrix for $\boldsymbol{x}$ is subsequently

$$\boldsymbol{\Sigma(\Lambda, \Theta) = \Lambda\Phi\Lambda' + \Theta} \tag{2.11}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the manifest variables and is a function of the free parameters in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$. The basic idea in estimating confirmatory models is to minimize the differences between the sample covariance matrix and the model implied covariance matrix. Regard a typical confirmatory model with five manifest variables and two factors in matrix form. The model can be understood as shown in the path diagram in Figure 2.1.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{pmatrix} \tag{2.12}$$

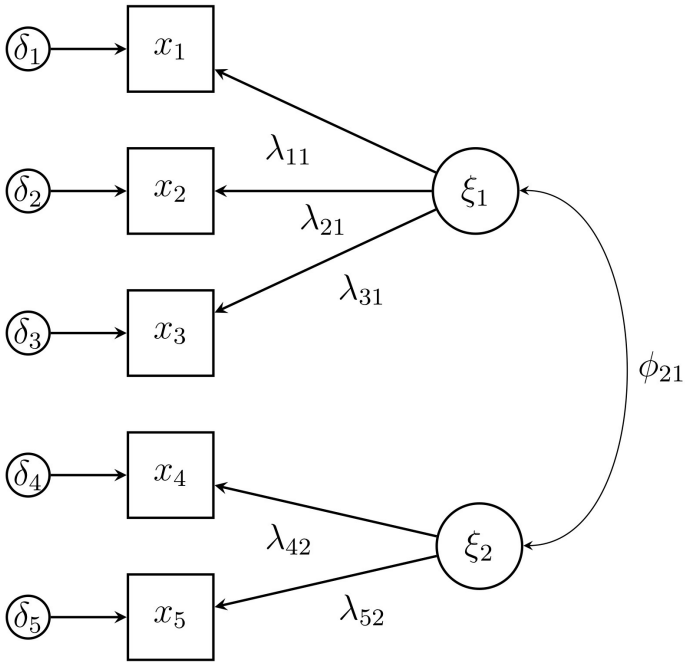

Figure 2.1: Path Diagram of a confirmatory factor analysis model with five manifest variables and two factors

If we are to use the polychoric correlation instead of Pearson correlation for the model we need to adapt Equation (2.11). Instead consider that we are modeling the underlying continuous variable x^*, as described in Section 2.1. The model function would be

$$\boldsymbol{x}^* = \boldsymbol{\Lambda\xi} + \boldsymbol{\delta} \tag{2.13}$$

As by assumptions stated in Section 2.3 the underlying variables now have unit variance and $\boldsymbol{\Theta}$ can be expressed as

$$\boldsymbol{\Theta} = \mathbf{I} - \boldsymbol{diag}(\boldsymbol{\Lambda\Phi\Lambda'}) \tag{2.14}$$

where $\boldsymbol{\Phi}$ now is a correlation matrix. If we substitute in Equation (2.13) into Equation (2.11) we get the appropriate model using polychoric correlations for the underlying continuous data.

$$\boldsymbol{\Sigma^*(\Lambda, \Theta)} = \boldsymbol{\Lambda\Phi\Lambda'} + \mathbf{I} - \boldsymbol{diag}(\boldsymbol{\Lambda\Phi\Lambda'}) \tag{2.15}$$

We are aiming to find a model where parameters can make $\boldsymbol{\Sigma^*(\Lambda, \Theta)}$ as close as possible to the sample-implied polychoric correlation matrix. To find the maximum likelihood fit function, let $\boldsymbol{R}$ be the polychoric correlation matrix and $\boldsymbol{\Sigma^*}$ be the function defined in Equation (2.15). The fit function can then be defined as

$$F_{ML}(\boldsymbol{R, \Lambda, \Phi}) = \ln(|\boldsymbol{\Sigma^*}|) + tr(\boldsymbol{R\Sigma^{*^{-1}}}) - \ln(|\boldsymbol{R}|) - p \tag{2.16}$$

which is then minimized to the free elements in $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ (Yang-Wallentin, Jöreskog, and Luo 2010). The principle of ML estimation in this case is to find the model parameter estimates that maximize the probability of observing the available data if the data are collected from the same population again. Namely, to maximize the likelihood of the parameters, given the data. This is an iterative process, where the used computer program[1] begins with an initial set of parameter estimates and repeatedly refines these estimates in an attempt to reduce the value of F_{ML}. In the polychoric case this implies finding model parameter estimates that minimize the difference between $\mathbf{\Sigma}^*$ and $\boldsymbol{R}$. Convergence is reached when the iteration cannot reduce the difference further (Brown 2006).

If the theoretically implied basis for the zero relation between for example x_1 and ξ_2 is doubtful, there may be an issue with the confirmatory approach, apparent by Figure 2.1. Then an exploratory approach may be more accurate. In other words, the theory and prior analysis that the confirmatory model is based on must be so that it can fix a number of factor loadings at zero to reflect the hypothesis that only certain factors influence certain manifest variables. If apriori substantive knowledge is lacking, fixing cross-loadings at zero may force a researcher to specify a more parsimonious model than is suitable for the data. Also, the misspecification of zero loadings tend to give distorted factors, and overestimated factor correlations, subsequently leading to distorted structural relations (Asparouhov and Bengt Muthén 2009). The mathematical difference in the exploratory approach is that no loadings are fixed at zero, rather all measures can potentially load on all factors, making the approach better for identifying latent structures where the researcher lacks theoretical basis (Fabrigar and Wegener 2014). Now consider instead how the two factor model and path diagram in Figure 2.1 would look if the model instead was an exploratory model. As can be seen, there are more defined λ_{ij} coefficients and the factors ξ_1 and ξ_2 have a potential relationship with all observed manifest variables. These new λ_{ij}, if the confirmatory model is correct, would likely be estimated close to zero.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{pmatrix} \tag{2.17}$$

[1] In this thesis R 4.0.3 is used. See R Core Team (2020).

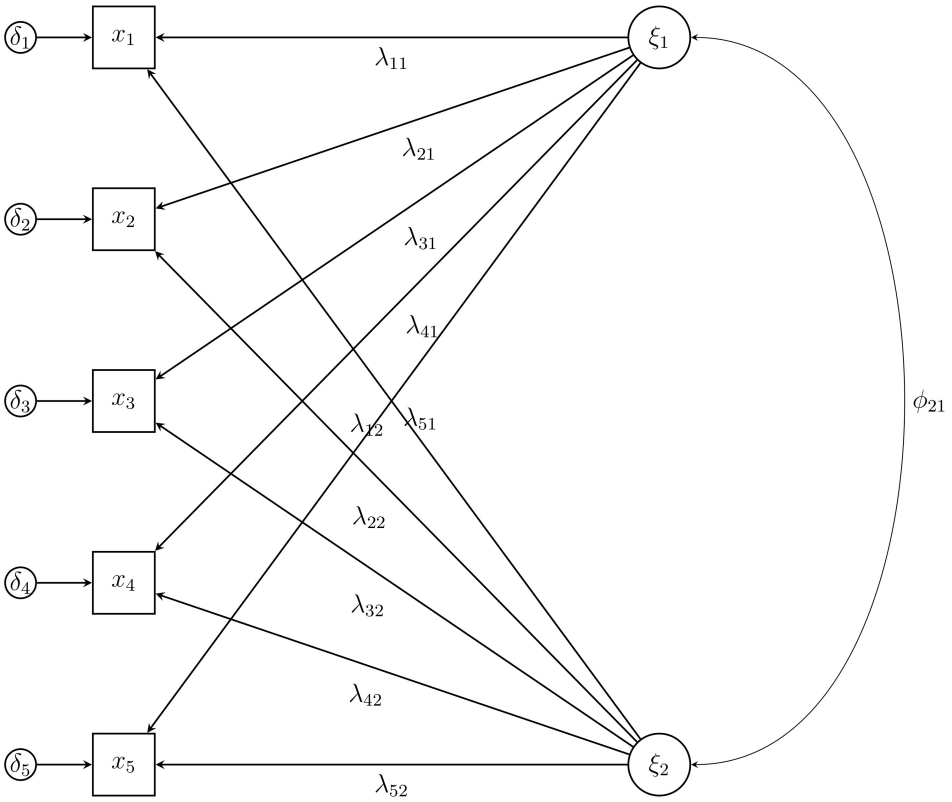

Figure 2.2: Path Diagram of an exploratory factor analysis model with five manifest variables and two factors

Simulation study

3.1 Data generation and categorization

First assume a normal distribution $N(0, 1)$ and generate data for a three, four and five factors model on 1000 subjects to 12 items following a theoretical model.[1] Second the observations are categorized on a five-point scale so that (a) the distribution of the answers to all items, except one, are symmetric, (b) moderate skewness is introduced to the items and (c) severe skewness is introduced to the items. See figure 3.1 for a graphical representation of the symmetrical categorization. Figures over the distributions following categorization can be found in Appendix A: Figure A.1.

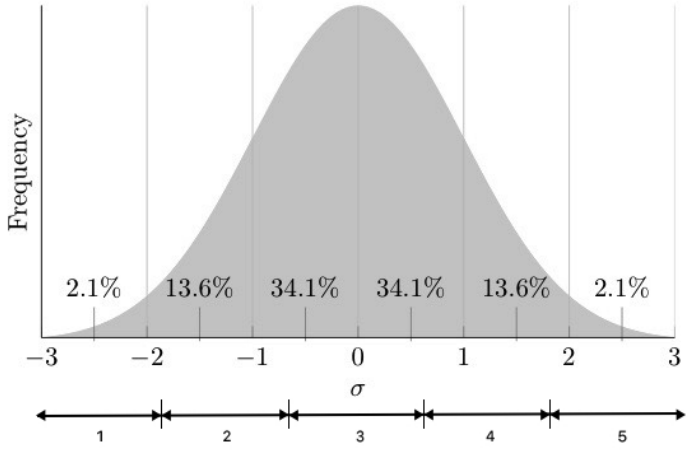

Figure 3.1: Graphical representation of the symmetric categorization of items based on a standard normal distribution.

Let z be a simulated observation from the standard normal distribution.

(1) In the symmetric categorization:

$$\begin{aligned} \text{If } z \leq -1.8 \quad & z \text{ is codified as 1} \\ \text{If } -1.8 < z \leq -0.6 \quad & z \text{ is codified as 2} \\ \text{If } -0.6 < z \leq 0.6 \quad & z \text{ is codified as 3} \\ \text{If } 0.6 < z \leq 1.8 \quad & z \text{ is codified as 4} \\ \text{If } 1.8 < z \quad & z \text{ is codified as 5} \end{aligned}$$

[1] For the data generating algorithm see Schneider (2020) and corresponding paper Schneider (2013).

(2) In the moderate negative asymmetric categorization:

$$\begin{aligned} \text{If } z \leq 0 \quad & z \text{ is codified as 1} \\ \text{If } 0 < z \leq 0.6 \quad & z \text{ is codified as 2} \\ \text{If } 0.6 < z \leq 1.2 \quad & z \text{ is codified as 3} \\ \text{If } 1.2 < z \leq 1.8 \quad & z \text{ is codified as 4} \\ \text{If } 1.8 < z \quad & z \text{ is codified as 5} \end{aligned}$$

(3) In the moderate positive asymmetric categorization:

$$\begin{aligned} \text{If } z \leq -1.8 \quad & z \text{ is codified as 1} \\ \text{If } -1.8 < z \leq -1.2 \quad & z \text{ is codified as 2} \\ \text{If } -1.2 < z \leq -0.6 \quad & z \text{ is codified as 3} \\ \text{If } -0.6 < z \leq 0 \quad & z \text{ is codified as 4} \\ \text{If } 0 < z \quad & z \text{ is codified as 5} \end{aligned}$$

(4) In the severe negative asymmetric categorization:

$$\begin{aligned} \text{If } z \leq 1 \quad & z \text{ is codified as 1} \\ \text{If } 1 < z \leq 1.5 \quad & z \text{ is codified as 2} \\ \text{If } 1.5 < z \leq 2 \quad & z \text{ is codified as 3} \\ \text{If } 2 < z \leq 2.5 \quad & z \text{ is codified as 4} \\ \text{If } 2.5 < z \quad & z \text{ is codified as 5} \end{aligned}$$

(5) In the severe positive asymmetric categorization:

$$\begin{aligned} \text{If } z \leq -2.5 \quad & z \text{ is codified as 1} \\ \text{If } -2.5 < z \leq -2 \quad & z \text{ is codified as 2} \\ \text{If } -2 < z \leq -1.5 \quad & z \text{ is codified as 3} \\ \text{If } -1.5 < z \leq 1 \quad & z \text{ is codified as 4} \\ \text{If } 1 < z \quad & z \text{ is codified as 5} \end{aligned}$$

Third, an exploratory factor analysis is carried out from the matrices of both Pearson and polychoric correlations. This is done in order to compare the estimated coefficients in the model against the theoretical model used to generate the data. Parameters are estimated through maximum likelihood method and an oblique rotation, specifically oblimin, was performed on the loadings. This procedure is replicated 1000 times. For each of the models, the assumption is made that the number of factors for inclusion is already established and this corresponds to the theoretical number of factors. Theoretical models are inspired by F. Holgado-Tello et al. (2010) but some adjustments have been made in terms of theoretical factor saturation. Particularly the saturation for items saturated on multiple factors have been gradually decreased when number of factors increase, in order to consider a more difficult model to estimate accurately.

3.2 Three factors model

The resulting matrix Λ defines the factor saturation for the generation of data following a three factor theoretical model, as well as the Φ matrix representing the correlation between latent factors:

$$\Lambda = \begin{pmatrix} .7 & 0 & 0 \\ .6 & 0 & 0 \\ .5 & 0 & 0 \\ .4 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & .6 & 0 \\ 0 & .5 & 0 \\ 0 & .4 & 0 \\ 0 & 0 & .7 \\ 0 & 0 & .6 \\ 0 & 0 & .5 \\ .4 & .4 & .4 \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 1 & & \\ .456 & 1 & \\ .673 & .309 & 1 \end{pmatrix}$$

The presented results are (a) a model with symmetric categorization for all items except one and (b) a model with moderate asymmetric categorization for all items and (c) a model with severe asymmetric categorization for all items. As can be seen in the saturation matrix above the 12th item is saturated on all three factors. This is also the item that has asymmetry in the models based on (a). For the three factor model this 12th item has moderate positive asymmetry in (a) and (b) while severe positive asymmetry in (c). In the models, where all items are asymmetric, the distributions are alternating between positive and negative asymmetry on items 1 to 11 while the 12th item keeps the positive asymmetry.

In a comparison of estimated correlation matrices and the theoretical correlation matrix (See Table A.1 in Appendix A), both methods perform well in the case where all items are symmetrical except one, see Table 3.1. The polychoric coefficients (above the diagonal dash) perform better than the Pearson coefficients (below the diagonal dash) but the differences are not large.

Table 3.1: Matrix Pearson (below) and polychoric (above) correlations for the three factors model with one item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.441	.355	.311	.259	.198	.150	.117	.331	.289	.224	.588
2	.393	-	.308	.218	.237	.188	.170	.106	.288	.263	.210	.507
3	.315	.276	-	.224	.173	.122	.143	.030	.304	.200	.204	.469
4	.277	.194	.199	-	.128	.096	.161	.066	.212	.154	.139	.361
5	.230	.210	.155	.114	-	.440	.355	.268	.216	.257	.162	.545
6	.176	.168	.110	.086	.391	-	.273	.200	.153	.243	.069	.429
7	.132	.151	.127	.144	.316	.243	-	.202	.182	.160	.035	.398
8	.104	.094	.027	.059	.238	.178	.179	-	.108	.184	.098	.273
9	.293	.257	.270	.188	.193	.137	.161	.096	-	.404	.320	.595
10	.256	.233	.178	.138	.229	.217	.142	.163	.359	-	.286	.552
11	.198	.187	.182	.123	.145	.062	.031	.087	.283	.255	-	.383
12	.490	.429	.392	.298	.454	.355	.332	.227	.493	.462	.325	-

In the case where items are moderate asymmetrical, see Table 3.2, the polychoric coefficients produces similar results as in the symmetrical case, while the Pearson coefficients show a clear drop in accuracy.

Table 3.2: Matrix of Pearson (below) and polychoric (above) correlations for the three factors model with all moderate item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.402	.337	.305	.288	.235	.144	.152	.325	.269	.220	.598
2	.288	-	.320	.198	.207	.194	.144	.137	.283	.263	.264	.492
3	.274	.237	-	.215	.190	.154	.132	.045	.285	.208	.180	.468
4	.224	.163	.155	-	.117	.116	.137	.038	.188	.181	.200	.361
5	.245	.152	.149	.098	-	.444	.351	.233	.220	.262	.196	.547
6	.180	.157	.117	.098	.319	-	.285	.233	.199	.255	.047	.451
7	.117	.113	.125	.114	.309	.215	-	.205	.164	.138	.037	.383
8	.118	.099	.039	.031	.176	.188	.151	-	.131	.137	.102	.26
9	.277	.216	.235	.141	.174	.146	.140	.102	-	.421	.313	.587
10	.198	.217	.153	.154	.198	.195	.114	.105	.311	-	.348	.543
11	.174	.190	.138	.154	.166	.025	.026	.083	.268	.258	-	.396
12	.514	.355	.394	.270	.469	.326	.328	.200	.512	.388	.340	-

In Table 3.3, results are shown when asymmetry gets more extreme. The Pearson coefficients continue to produce even less accurate results. The polychoric correlation does show a slight drop in accuracy but still manages to estimate most of the correlation from the theoretical model.

Table 3.3: Matrix of Pearson (below) and polychoric (above) correlations for the three factors model with all severe item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.326	.332	.327	.243	.244	.121	.161	.319	.272	.241	.582
2	.161	-	.285	.253	.244	.165	.217	.105	.302	.239	.190	.498
3	.214	.136	-	.210	.149	.164	.206	.064	.213	.221	.209	.419
4	.158	.154	.115	-	.082	.090	.060	-.022	.177	.199	.213	.370
5	.127	.129	.078	.036	-	.497	.408	.333	.160	.171	.137	.498
6	.117	.091	.073	.037	.247	-	.269	.244	.255	.109	.021	.428
7	.054	.105	.103	.028	.244	.132	-	.138	.208	.126	.035	.411
8	.073	.057	.028	-.030	.157	.119	.067	-	.256	.100	.177	.330
9	.193	.152	.139	.095	.089	.125	.120	.114	-	.394	.277	.606
10	.126	.141	.111	.106	.078	.049	.069	.063	.189	-	.311	.487
11	.169	.092	.115	.111	.091	.014	.023	.080	.160	.146	-	.396
12	.357	.248	.263	.189	.294	.206	.251	.152	.385	.240	.243	-

In regards to loadings from the fitted exploratory model, the Pearson based models produce adequate results when only one item has skewness, see Table 3.4. Over replications, the Pearson based models have issue closely estimating the loading that theoretically is saturated on multiple factors. When moderate asymmetry is introduced it produces low loadings on the 12th item, that was saturated on multiple factors. Both models do however produce fairly accurate results when moderate asymmetry is introduced, with the polychoric based models generally being closer to the theoretical model.

Table 3.4: Lambda matrix, (Λ), for the three factors model with one item skewness using exploratory factor analysis (and all moderate item skewness in brackets).

Item	Pearson correlations			Polychoric correlations		
	F1	F2	F3	F1	F2	F3
1	**.692(.660)**	-.003(.001)	-.031(-.050)	**.739(.707)**	-.009(.025)	-.037(-.060)
2	**.533(.352)**	.035(-.015)	.013(.117)	**.566(.513)**	.034(-.024)	.002(.070)
3	**.455(.505)**	-.060(-.054)	.103(.009)	**.484(.556)**	-.069(-.049)	.120(.012)
4	**.401(.304)**	-.012(-.033)	.002(.063)	**.429(.437)**	-.019(-.060)	.010(.022)
5	.005(-.011)	**.679(.649)**	-.003(.014)	.005(.004)	**.721(.688)**	-.010(.017)
6	-.012(.011)	**.579(.499)**	-.026(-.037)	-.016(-.022)	**.621(.633)**	-.035(-.017)
7	.041(.018)	**.458(.506)**	-.021(-.072)	.025(.038)	**.490(.534)**	.000(-.078)
8	-.105(-.068)	**.351(.309)**	.095(.058)	-.112(-.043)	**.375(.354)**	.093(.040)
9	.053(.139)	-.069(-.029)	**.633(.527)**	.032(.143)	-.081(-.031)	**.707(.576)**
10	-.079(-.075)	.098(.095)	**.579(.515)**	-.081(-.093)	.107(.064)	**.606(.693)**
11	.019(-.024)	-.093(-.078)	**.479(.543)**	.029(.078)	-.095(-.131)	**.486(.517)**
12	**.313(.451)**	.298(.294)	**.412(.303)**	**.342(.475)**	**.341(.329)**	**.469(.340)**

In bold, factor loadings higher than 0.3.

See Table 3.5 for the factor loading when data is subject to severe asymmetry. The Pearson based model has issues accurately estimating the higher loadings (0.7 and 0.6) and generally produces less accurate recreation of the theoretical model.

Table 3.5: Lambda matrix, (Λ), for the three factors model with all severe item skewness using exploratory factor analysis.

Item	Pearson correlations			Polychoric correlations		
	F1	F2	F3	F1	F2	F3
1	**.472**	.010	.035	**.597**	.051	.036
2	**.303**	.073	.030	**.448**	.087	.066
3	**.385**	-.003	.004	**.495**	.015	-.022
4	**.451**	-.081	-.107	**.618**	-.114	-.116
5	.001	**.687**	-.055	.018	**.858**	-.086
6	-.023	**.346**	.092	-.013	**.561**	.098
7	-.016	**.339**	.111	.072	**.440**	.069
8	-.120	.207	.209	.176	**.347**	**.321**
9	.006	-.047	**.610**	.030	-.048	**.765**
10	.185	-.023	.202	.267	-.050	**.358**
11	.256	-.029	.124	.296	-.068	.228
12	**.371**	.205	**.353**	**.531**	.277	**.376**

In bold, factor loadings higher than 0.3.

Table 3.6: Phi matrix, (Φ), for three factors model one item skewness using exploratory factor analysis (all moderate item skewness in brackets).

Dimension	F1	F2	F3
F1	-	.514(.554)	.710(.680)
F2	.500(.585)	-	.527(.520)
F3	.702(.692)	.520(.505)	-

Pearson (below diagonal) and Polychoric (above).

Table 3.7: Phi matrix, (Φ), for three factors model all severe item skewness using exploratory factor analysis.

Dimension	1	2	3
1	-	.415	.594
2	.400	-	.394
3	.629	.372	-

Pearson (below diagonal) and Polychoric (above).

3.3 Four factors model

The resulting matrix Λ defines the factor saturation for the generation of data following a four factors theoretical model, as well as the Φ matrix representing the correlation between latent factors:

$$\Lambda = \begin{pmatrix} .7 & 0 & 0 & 0 \\ .6 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .6 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .6 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .6 \\ .34 & .34 & .34 & .34 \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 1 & & & \\ .353 & 1 & & \\ .510 & .177 & 1 & \\ .628 & .223 & .321 & 1 \end{pmatrix}$$

In terms of symmetry and asymmetry the presented results are similar to that of the three factors model. As can be seen in the saturation matrix above the 12th item is saturated on all four factors. This is also the item that has asymmetry in the models based on (a). For this model the 12th item has moderate negative asymmetry in (a) and (b), while severe negative asymmetry in (c). In the model, where all items are asymmetric, the distributions are alternating between positive and negative asymmetry and the 12th item keeps the negative asymmetry.

Table 3.8: Matrix Pearson (below) and polychoric (above) correlations for the four factors model with one item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.426	.330	.153	.190	.121	.203	.167	.143	.342	.288	.610
2	.380	-	.279	.171	.193	.131	.189	.166	.146	.223	.254	.527
3	.294	.247	-	.090	.054	.102	.095	.155	.129	.178	.203	.407
4	.137	.152	.078	-	.453	.336	.113	.069	.102	.056	.085	.408
5	.170	.170	.046	.401	-	.342	.079	.128	.120	.088	.049	.404
6	.108	.115	.091	.298	.302	-	.035	.098	.031	.110	.087	.346
7	.181	.169	.083	.100	.069	.031	-	.404	.326	.084	.112	.408
8	.149	.148	.138	.062	.113	.086	.357	-	.326	.089	.139	.388
9	.128	.130	.115	.090	.107	.027	.288	.291	-	.109	.105	.393
10	.306	.197	.158	.048	.078	.098	.076	.079	.097	-	.431	.476
11	.256	.225	.181	.076	.043	.077	.099	.123	.093	.383	-	.455
12	.516	.444	.338	.334	.329	.281	.335	.330	.327	.398	0.382	-

Table 3.9: Matrix of Pearson (below) and polychoric (above) correlations for the four factors model with all moderate item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.428	.330	.155	.190	.102	.203	.156	.143	.324	.288	.610
2	.361	-	.264	.141	.185	.098	.208	.153	.139	.250	.244	.514
3	.294	.223	-	.106	.054	.084	.095	.130	.129	.208	.203	.407
4	.137	.124	.084	-	.466	.314	.088	.065	.101	.030	.073	.388
5	.170	.160	.046	.390	-	.312	.079	.117	.120	.072	.049	.404
6	.088	.072	.065	.267	.264	-	.038	.089	.044	.105	.069	.322
7	.181	.173	.083	.074	.069	.025	-	.433	.326	.096	.112	.408
8	.140	.131	.118	.054	.104	.076	.358	-	.325	.092	.114	.394
9	.128	.116	.115	.088	.107	.027	.288	.277	-	.130	.105	.393
10	.274	.195	.173	.018	.051	.098	.079	.072	.104	-	.401	.488
11	.256	.211	.181	.059	.043	.058	.099	.097	.093	.340	-	.455
12	.516	.448	.338	.325	.329	.258	.335	.340	.327	.412	.382	-

Table 3.8 shows the correlation matrices when data has symmetrical distributions in all items except one. Table 3.9 when moderate asymmetry is introduced to all items. The results show the same trend as in the three factors models, where Pearson correlation generally underestimates coefficients and the polychoric correlation are able to approximately recreate the theoretical matrix.

Table 3.10: Matrix of Pearson (below) and polychoric (above) correlations for the four factors model with all severe item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.474	.354	.246	.144	.084	.172	.240	.052	.312	.300	.657
2	.234	-	.302	.177	.085	-.010	.202	.189	.134	.201	.264	.543
3	.208	.165	-	.061	.099	.035	.136	.278	.136	.199	.234	.403
4	.122	.095	.048	-	.451	.392	.114	.102	.081	-.003	.097	.369
5	.075	.053	.073	.220	-	.325	.078	.102	.061	-.006	.098	.312
6	.062	.010	.020	.271	.159	-	-.012	.123	.006	.231	.058	.274
7	.105	.108	.070	.046	.046	.004	-	.500	.354	.032	.121	.358
8	.133	.090	.145	.038	.063	.066	.243	-	.449	.020	.191	.430
9	.022	.068	.074	.041	.040	-.007	.207	.212	-	.014	.103	.277
10	.144	.108	.094	.003	.002	.112	.016	.020	.009	-	.464	.494
11	.157	.135	.134	.058	.055	.034	.100	.088	.073	.222	-	.512
12	.338	.356	.198	.243	.162	.206	.167	.256	.136	.321	.246	-

In Table 3.10 correlation estimates are presented, where items have severe asymmetry. The trend continues with the Pearson method showing poorer results as asymmetry gets more extreme and the polychoric method still managing to capture most of the theoretical correlation.

See Table 3.11 for the factor loadings for the exploratory model in the case where data is symmetrical (all items except one) and moderately asymmetrical. The Pearson and Polychoric methods produce acceptable results, with the Pearson based models generally estimating the loading of the 12th item low.

In Table 3.12, loading when items are subject to severe asymmetry is presented. Similarly to the three factors model, the Pearson based model has difficulties with estimating the higher factor loadings (0.7 and 0.6). Both models lack in their accuracy for estimating the 12th item loading, that are saturated on all factors. The polychoric based model still produces more accurate results.

Table 3.11: Lambda matrix, (Λ), for the four factors model with one item skewness using exploratory factor analysis (and all moderate item skewness in brackets).

Item	Pearson correlations				Polychoric correlations			
	F1	F2	F3	F4	F1	F2	F3	F4
1	**.672(.657)**	-.027(-.010)	-.040(-.033)	.042(.030)	**.717(.768)**	.035(-.021)	-.046(-.039)	.040(.001)
2	**.595(.578)**	.025(.008)	.001(.008)	-.058(-.039)	**.623(.553)**	.026(.020)	-.002(..037)	-.059(.010)
3	**.482(.411)**	-.068(-.047)	.005(.001)	-.019(.051)	**.528(.381)**	-.070(-.023)	.003(.012)	-.023(.114)
4	-.007(.014)	**.651(.652)**	-.014(-.031)	-.042(-.064)	-.015(-.002)	**.699(.709)**	-.017(-.034)	-.049(-.059)
5	-.001(.030)	**.636(.608)**	-.005(.001)	-.038(-.060)	.001(035)	**.671(.676)**	-.012(-.010)	-.050(-.072)
6	-.042(-.137)	**.497(.466)**	-.045(-.019)	.079(.153)	-.047(-132)	**.541(.500)**	-.052(-.021)	.083(.153)
7	.025(.053)	-.046(-.063)	**.608(.605)**	-.054(-.073)	.023(.070)	-.054(-.065)	**.659(.674)**	-.064(-.090)
8	-.037(-.059)	-.013(-.008)	**.605(.628)**	-.006(-.010)	-.038(-.063)	-.022(-.015)	**.624(.675)**	-.012(-.010)
9	-.046(-.054)	.013(.027)	**.521(.493)**	.030(.052)	-.059(-.086)	.014(.040)	**.573(.521)**	.037(.101)
10	-.026(-.009)	-.016(-.044)	-.035(-.032)	**.711(.678)**	-.026(.020)	-.024(-.062)	-.044(-.049)	**.747(.680)**
11	.062(.129)	-.037(-.047)	.030(.001)	**.524(.445)**	.047(.038)	-.043(-.050)	.026(-.011)	**.578(.577)**
12	**.387(.353)**	.256(.269)	.272(.279)	.237(**.309**)	**.415(.319)**	**.302(.313)**	**.307(.303)**	.272(**.390**)

In bold, factor loadings higher than 0.3.

Table 3.12: Lambda matrix, (Λ), for the four factors model with all severe item skewness using exploratory factor analysis.

Item	Pearson correlations				Polychoric correlations			
	F1	F2	F3	F4	F1	F2	F3	F4
1	**.487**	.007	-.009	.007	**.727**	.028	-.039	.052
2	**.591**	-.078	-.043	-.052	**.690**	-.072	-.005	-.041
3	**.303**	-.053	.087	.006	**.365**	-.071	.157	.068
4	.087	**.553**	-.038	-.102	.187	**.678**	-.047	-.166
5	.042	**.338**	.038	-.066	.097	**.595**	-.002	-.120
6	-.102	**.529**	.007	.085	-.227	**.639**	.052	.245
7	.019	-.045	**.453**	-.029	.069	-.043	**.601**	-.050
8	.024	.000	**.538**	-.034	-.003	-.003	**.808**	-.043
9	-.068	-.026	**.452**	-.010	-.060	-.024	**.600**	-.015
10	.005	-.012	-.013	**.871**	.053	-.027	-.049	**.904**
11	.215	-.020	.089	.179	.283	-.024	.094	**.391**
12	**.503**	.199	.148	.162	**.622**	.215	.235	.249

In bold, factor loadings higher than 0.3.

Table 3.13: Phi matrix, (Φ), for four factors model one item skewness using exploratory factor analysis (all moderate item skewness in brackets).

Dimension	F1	F2	F3	F4
F1	-	.453(.410)	.515(.460)	.655(.663)
F2	.438(.419)	-	.338(.308)	.279(.288)
F3	.503(.492)	.318(.303)	-	.337(.364)
F4	.639(.627)	.253(.244)	.314(.329)	-

Pearson (below diagonal) and Polychoric (above).

Table 3.14: Phi matrix, (Φ), for four factors model all severe item skewness using exploratory factor analysis.

Dimension	F1	F2	F3	F4
F1	-	.290	.418	.368
F2	.361	-	.207	.161
F3	.456	.218	-	.125
F4	.332	.160	.114	-

Pearson (below diagonal) and Polychoric (above).

3.4 Five factors model

The resulting matrix Λ defines the factor saturation for the generation of data following a five factors theoretical model, as well as the Φ matrix representing the correlation between latent factors:

$$\Lambda = \begin{pmatrix} .7 & 0 & 0 & 0 & 0 \\ .6 & 0 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 & .25 \\ 0 & .7 & 0 & 0 & 0 \\ 0 & .6 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 & 0 \\ 0 & 0 & .7 & 0 & 0 \\ 0 & 0 & .6 & 0 & 0 \\ 0 & 0 & 0 & .7 & 0 \\ 0 & 0 & 0 & .6 & 0 \\ 0 & 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & 0 & .6 \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 1 & & & & \\ .367 & 1 & & & \\ .501 & .171 & 1 & & \\ .627 & .231 & .324 & 1 & \\ .662 & .254 & .319 & .405 & 1 \end{pmatrix}$$

The same approach in terms of symmetry and asymmetry, is used here as in the three factors model. This time however the 3rd item is saturated fairly low on all factors, making it a particularly difficult model to estimate correctly. This is also the item that has moderate positive asymmetry in the models based on (a) and (b) while severe positive asymmetry in (c). In the model, where all items are asymmetric, the distributions are alternating between positive and negative asymmetry, as it did in the three factors model, and the 3rd item keeps the positive asymmetry.

Table 3.15: Matrix Pearson (below) and polychoric (above) correlations for the five factors model with one item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.368	.549	.171	.135	.159	.207	.207	.263	.293	.348	.276
2	.327	-	.446	.107	.126	.142	.262	.199	.216	.247	.305	.240
3	.461	.374	-	.357	.262	.222	.382	.315	.435	.364	.516	.395
4	.150	.095	.295	-	.443	.388	.016	.039	.140	.074	.156	.102
5	.120	.114	.222	.395	-	.315	-.016	.046	.082	.124	.103	.070
6	.139	.127	.181	.343	.281	-	.034	.007	.057	.069	.110	.093
7	.182	.233	.322	.012	-.015	.030	-	.440	.152	.052	.203	.157
8	.184	.176	.267	.033	.040	.006	.389	-	.093	.103	.161	.123
9	.234	.190	.362	.123	.072	.052	.137	.083	-	.366	.197	.175
10	.258	.218	.307	.065	.109	.061	.047	.091	.323	-	.217	.171
11	.310	.271	.428	.137	.091	.098	.179	.143	.175	.192	-	.417
12	.245	.215	.335	.092	.063	.083	.140	.109	.157	.152	.369	-

Table 3.16: Matrix of Pearson (below) and polychoric (above) correlations for the five factors model with all moderate item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.412	.540	.204	.101	.157	.216	.193	.292	.253	.330	.256
2	.302	-	.431	.127	.114	.140	.230	.180	.191	.194	.319	.223
3	.466	.314	-	.345	.245	.186	.390	.290	.430	.346	.511	.391
4	.150	.092	.255	-	.434	.368	-.023	.015	.170	.097	.126	.127
5	.076	.094	.203	.319	-	.286	-.025	.021	.047	.099	.054	.075
6	.117	.131	.137	.316	.210	-	.038	.015	.062	.095	.135	.080
7	.193	.169	.332	-.016	-.019	.022	-	.482	.100	.074	.188	.159
8	.145	.141	.223	.005	.022	.009	.355	-	.092	.077	.123	.118
9	.251	.147	.364	.134	.040	.052	.081	.062	-	.403	.202	.125
10	.187	.163	.254	.072	.068	.071	.055	.074	.296	-	.221	.145
11	.289	.237	.442	.108	.046	.106	.157	.106	.171	.179	-	.436
12	.187	.176	.289	.095	.054	.072	.114	.099	.098	.120	.316	-

The polychoric correlation coefficients are again more accurate compared to the Pearson coefficients in recreating the theoretical matrix. As can be seen in Table 3.15 and 3.16, the Pearson correlation gets less accurate as moderate asymmetry is introduced and the polychoric coefficients stays similar in most estimates.

Table 3.17: Matrix of Pearson (below) and polychoric (above) correlations for the five factors model with all severe item skewness.

	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.411	.493	.176	.188	.200	.200	.247	.284	.258	.302	.239
2	.201	-	.370	.039	.081	.146	.249	.178	.212	.188	.267	.203
3	.307	.194	-	.367	.249	.229	.410	.337	.459	.348	.481	.334
4	.089	.007	.183	-	.406	.464	.023	-.006	.071	.062	.156	.105
5	.115	.042	.152	.208	-	.321	.047	.037	.095	.074	.111	.083
6	.115	.067	.122	.296	.192	-	.010	.060	.099	.039	.055	.145
7	.131	.120	.228	-.002	.037	.005	-	.447	.253	.021	.159	.127
8	.132	.088	.180	-.003	.026	.044	.229	-	.143	.052	.158	.194
9	.172	.111	.295	.042	.063	.049	.127	.077	-	.389	.191	.205
10	.128	.120	.163	.035	.038	.032	-.018	.044	.189	-	.209	.130
11	.184	.135	.324	.069	.077	.038	.091	.090	.100	.103	-	.380
12	.108	.095	.150	.062	.044	.065	.053	.081	.119	.058	.182	-

The trend of the Pearson coefficients becoming less accurate as even more asymmetry is introduced is prevalent in the five factors model as well. The polychoric estimate is still accurate to a certain extent with severe asymmetry, as can be seen in Table 3.17.

As more factors are introduced both methods of correlation become worse at recreating the theoretical model. Both models tend to overestimate the theoretically higher loading of 0.7 in some cases. Still the polychoric models is better at estimating the item that is saturated on all factors (item 3 in Table 3.11) compared to the Pearson based models.

Table 3.18: Lambda matrix, (Λ), for the five factors model with one item skewness using exploratory factor analysis (and all moderate item skewness in brackets).

Item	Pearson correlations					Polychoric correlations				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
1	**.562**(.143)	.013(.067)	.007(.117)	-.008(.187)	.080(.241)	**.622(.811)**	.009(-.014)	.005(-.022)	-.013(.000)	.068(-.026)
2	**.430**(-.047)	-.002(.100)	.106(.169)	-.012(.118)	.072(**.307**)	**.455(.425)**	-.010(.017)	.114(.081)	-.008(-.025)	.074(.140)
3	**.410(.966)**	.180(.011)	.149(.015)	.091(.014)	**.230**(.022)	**.441(.248)**	**.205(.227)**	.169(**.233**)	.119(**.204**)	**.261(.320)**
4	-.064(.047)	**.750(.644)**	.000(-.031)	.023(.030)	.029(-.026)	-.064(-.011)	**.817(.757)**	-.003(-.033)	.022(.032)	.023(-.016)
5	.134(.133)	**.533(.463)**	-.045(-.024)	-.032(-.065)	-.089(-.073)	.145(-.027)	**.549(.625)**	-.052(-.002)	-.036(-.059)	-.090(-.027)
6	.063(-.133)	**.461(.496)**	.004(.056)	-.040(-.001)	-.019(.120)	.082(.030)	**.479(.464)**	.002(.011)	-.049(-.066)	-.030(.027)
7	-.017(.069)	-.008(-.034)	**.869(.642)**	.011(-.025)	-.002(-.014)	-.021(-.029)	-.011(-.021)	**.906(.880)**	.012(-.011)	-.003(-.004)
8	**.212**(-.054)	-.016(.028)	**.403(.553)**	-.048(.023)	-.042(-.002)	**.217**(.077)	-.015(.003)	**.442(.547)**	-.054(.000)	-.045(-.050)
9	.001(.043)	.000(-.005)	.004(-.011)	**.997(.663)**	-.003(-.042)	-.001(-.015)	.000(-.010)	.004(-.007)	**.999(.873)**	-.005(-.021)
10	**.440**(-.066)	-.048(.018)	-.109(.009)	.183(**.420**)	-.019(.142)	**.452**(.098)	-.057(.004)	-.121(-.037)	**.216(.400)**	-.014(.107)
11	.018(.110)	-.004(-.025)	-.010(-.028)	-.011(.003)	**.680(.582)**	.005(-.014)	-.007(-.030)	-.014(-.026)	-.014(.000)	**.750(.810)**
12	.014(.010)	-.021(.023)	-.013(.013)	.014(-.032)	**.534(.466)**	.030(.017)	-.027(.030)	-.012(.019)	.014(-.041)	**.547(.548)**

In bold, factor loadings higher than 0.2.

Table 3.19: Lambda matrix, (Λ), for the five factors model with all severe item skewness using exploratory factor analysis.

Item	Pearson correlations					Polychoric correlations				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
1	**.516**	.028	.004	-.003	.004	**.665**	.026	-.008	.071	.027
2	**.409**	-.072	.021	-.025	.002	**.520**	-.090	.087	.007	.069
3	**.305**	.147	.134	.156	**.209**	.195	**.254**	**.272**	**.355**	**.202**
4	-.052	**.647**	-.004	-.001	.013	-.055	**.907**	-.005	-.001	.018
5	.084	**.328**	.011	.001	.005	.134	**.436**	-.014	.000	-.017
6	.109	**.454**	-.024	-.024	-.068	**.262**	**.493**	-.054	-.066	.104
7	-.006	-.006	**.749**	.002	-.007	-.024	-.009	**.841**	.009	-.009
8	.193	-.030	**.249**	-.029	.003	**.203**	-.060	**.487**	-.063	.015
9	-.006	-.005	.000	**.824**	-.008	-.005	-.011	.111	**.664**	-.046
10	**.259**	-.035	-.143	.157	.022	.055	-.042	-.176	**.620**	.031
11	-.004	-.008	-.008	-.011	**.784**	-.002	-.005	-.012	-.014	**.955**
12	.097	.041	.000	.068	.168	.177	.018	.034	.052	**.301**

In bold, factor loadings higher than 0.2.

As more severe asymmetry is introduced, results are consistent with previous identified trends in loss of accuracy for the Pearson model, but here the polychoric based model tend to overestimate the higher loading of 0.7 more frequently. The polychoric based model shows remarkable accuracy in identifying the factor saturated on all factors (item 3), given the low theoretical value and severe item skewness.

Table 3.20: Phi matrix, (Φ), for five factors model one item skewness using exploratory factor analysis (all moderate item skewness in brackets).

Dimension	1	2	3	4	5
1	-	.341(.373)	.354(.389)	.412(.470)	.661(.591)
2	.337(.338)	-	.058(.055)	.171(.254)	.287(.278)
3	.341(.453)	.046(.003)	-	.165(.185)	.336(.336)
4	.389(.591)	.156(.220)	.152(.185)	-	.288(.346)
5	.663(.523)	.282(.232)	.327(.341)	.269(.373)	-

Pearson (below diagonal) and Polychoric (above).

Table 3.21: Phi matrix, (Φ), for five factors model all severe item skewness using exploratory factor analysis.

Dimension	1	2	3	4	5
1	-	.280	.366	.513	.410
2	.302	-	.068	.182	.195
3	.348	.037	-	.301	.234
4	.414	.125	.213	-	.361
5	.457	.172	.182	.189	-

Pearson (below diagonal) and Polychoric (above).

3.5 Results from replications

As mentioned previously, the simulation study was replicated 1000 times in order to assess the accuracy of drawn conclusions. In this section, summarizing results of replications are presented after a criterion is put on each of the loadings for a given factor model. For the three factors model the criterion is set at 0.3, meaning that a model is deemed to be correctly identified if factor loadings are higher than 0.3 for items with theoretical $\lambda > 0$ and if loadings are lower than 0.3 for items with theoretical $\lambda = 0$. For the four and five factors models the criterion is set at 0.25 and 0.2 respectively. It should be noted that other assessments have been taken into consideration for the conclusions, such as the deviance of loadings from theoretical Λ and cases where the estimated models only falsely estimate one loading higher or lower than criterion.

Table 3.22: Proportion of replicated three factors models correctly identifying the theoretical model in three cases of item skewness

	F1	F2	F3
Symmetrical* Pearson	0.701	0.832	0.783
Symmetrical* polychoric	0.843	0.963	0.931
Moderate asymmetrical Pearson	0.495	0.668	0.708
Moderate asymmetrical polychoric	0.802	0.954	0.901
Severe asymmetrical Pearson	0	0	0
Severe asymmetrical polychoric	0	0	0

* Symmetrical in all items except one.

Table 3.23: Proportion of replicated four factors models correctly identifying the theoretical model in three cases of item skewness.

	F1	F2	F3	F4
Symmetrical* Pearson	0.745	0.823	0.832	0.756
Symmetrical* polychoric	0.869	0.969	0.961	0.891
Moderate asymmetrical Pearson	0.731	0.856	0.811	0.768
Moderate asymmetrical polychoric	0.876	0.973	0.969	0.872
Severe asymmetrical Pearson	0	0	0	0
Severe asymmetrical polychoric	0	0	0	0

* Symmetrical in all items except one.

Table 3.24: Proportion of replicated five factors models correctly identifying the theoretical model in three cases of item skewness.

	F1	F2	F3	F4	F5
Symmetrical* Pearson	0.259	0.662	0.641	0.546	0.473
Symmetrical* polychoric	0.302	0.807	0.752	0.658	0.540
Moderate asymmetrical Pearson	0.035	0.162	0.214	0.171	0.133
Moderate asymmetrical polychoric	0.188	0.777	0.699	0.554	0.474
Severe asymmetrical Pearson	0	0	0	0	0
Severe asymmetrical polychoric	0	0	0	0	0

* Symmetrical in all items except one.

As can be seen in Tables 3.22 - 3.24, the factor analysis models that were estimated by polychoric correlations have a higher proportion of correctly identifying the theoretical models compared with the models estimated by Pearson correlations. As items become asymmetrical in distribution, the robustness feature of the polychoric coefficient is especially apparent. Also as factors increase, the polychoric based models show superiority in identifying the theoretical model when asymmetry is present in the data. Although none of the models are able to identify the theoretical model based on the criterion when items are severely asymmetrical in distribution, the polychoric based models show a smaller mean deviance from the theoretical Λ compared with the Pearson models.

Discussion and conclusions

This thesis aimed at examining which correlation matrix is most suitable to use when attempting to create factor analytical solutions and in analyzing the results given an ordinal level of measurement on observed variables. In social science and psychology it is common to see Likert scale data, collected though surveys, to be analyzed though interval-based measures. In relation to this, methodologists need to ensure that inferences made from the obtained results are as rigorous as possible. Furthermore, in social science, measurement often implies certain degrees of both random and systematic error. This possible error may bias the estimates of the relation between variables measured. In turn, this could lead to bias in substantive conclusions.

The replicated simulation study show that when ordinal data, obtained from Likert scales, is analyzed the results show a better fit to the theoretical model when factorization is carried out using the polychoric in comparison to the Pearson correlation matrix. Three levels of asymmetry in the observed variable is also considered and the results show that the polychoric method of estimating correlation is more robust to the violation of normality assumption and could be preferably used, compared to the Pearson correlation, when data is not approximately normal and of ordinal level. In relation to number of extracted factors, the Pearson correlation shows significant inaccuracy in the factor solution as they increase, while the polychoric stays fairly consistent in reproducing the measurement model. Results are comparable to results presented by F. Holgado-Tello et al. (2010) in the case of symmetrical (in all items except one) and severe asymmetrical items. This research was complemented by consideration of a less extreme item skewness, in this thesis called moderate asymmetrical items, where the polychoric correlations also outperform the Pearson correlation when used for exploratory factor analysis.

When analyzing construct validity, it is therefore clear that a polychoric correlation matrix could be advantageously used to analyze factors of asymmetrical ordinal data. No emphasis has been put on the power and effectiveness of such solutions and consequently possible drawn conclusions. It could still be deemed important, in terms of correct substantive conclusions, that factor solutions are more in keeping with the original measurement model.

There are many aspects of factor analytical solutions to ordinal levels of measurement that can be subject to further study, for example identifying at which degree of skewness factor solutions fail to reproduce the theoretical model or which correlation matrix is preferable with higher order factors, more or fewer items and a smaller sample size.

Bibliography

Abdi, Hervé (2004). "Factor Rotations in Factor Analyses." In: *The SAGE Encyclopedia of Social Science Research Methods*. Ed. by Michael Lewis-Beck, Alan Bryman, and Tim Liao. Thousand Oaks, California: SAGE Publications, Inc. (page 6).

Angeles Morata-Ramirez, Maria de los and Francisco Pablo Holgado-Tello (2013). "Construct Validity of Likert Scales through Confirmatory Factor Analysis: A Simulation Study Comparing Different Methods of Estimation Based on Pearson and Polychoric Correlations". In: *International Journal of Social Science Studies* 1.1, pp. 54–61 (page 3).

Asparouhov, Tihomir and Bengt Muthén (2009). "Exploratory Structural Equation Modeling". In: *Structural Equation Modeling: A Multidisciplinary Journal* 16.3, pp. 397–438 (page 12).

Brown, Timothy A. (2006). "Confirmatory factor analysis for applied research". In: 2nd ed. New York: Guliford Press (page 12).

Choi, Jaehwa, Michelle Peters, and Ralph O. Mueller (2010). "Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation". In: *Asia Pacific Education Review* 11, pp. 459–466 (page 3).

Coenders, Germà, Albert Satorra, and Willem E. Saris (1997). "Alternative approaches to structural modeling of ordinal data: A Monte Carlo study". In: *Structural Equation Modeling: A Multidisciplinary Journal* 4.4, pp. 261–282 (page 9).

DeCoster, J. (1998). *Overview of Factor Analysis*. URL: http://www.stat-help.com/notes.html (pages 4–6).

Devlin, Susan J., R. Gnanadesikan, and J. R. Kettenring (Dec. 1975). "Robust estimation and outlier detection with correlation coefficients". In: *Biometrika* 62.3, pp. 531–545 (page 8).

Fabrigar, Leandre R. and Duane T. Wegener (2014). "Exploring Causal and Noncausal Hypotheses in Nonexperimental Data". In: *Handbook of Research Methods in Social and Personality Psychology*. Ed. by Harry T. Reis and Charles M. Judd. 2nd ed. Chap. 19 (pages 6, 12).

Finney, Sara J. and Christine DiStefano (2014). "Nonnormal and Categorical Data in Structural Equation Modeling". In: *Structural Equation Modeling: A Second Course*. Ed. by Gregory R. Hancock and Ralph O. Mueller. 2nd ed. Chap. 9 (page 5).

Havlicek, Larry L. and Nancy L. Peterson (1976). "Robustness of the Pearson correlation against violations of assumptions". In: *Perceptual and Motor Skills* 43.3, pp. 1319–1334 (page 8).

Holgado-Tello, F.P. et al. (2010). "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables". In: *Qual Quant* 44, pp. 153–166 (pages 3, 15, 32).

Jöreskog, Karl G. (2005). *Structural equation modeling with ordinal variables using lisrel.* Technical report, Scientific Software International, Inc., Lincolnwood, IL (page 9).

Joshi, Ankur et al. (Jan. 2015). "Likert Scale: Explored and Explained". In: *British Journal of Applied Science Technology* 7, pp. 396–403 (pages 2, 3).

Lantz, B. (Jan. 2013). "Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations". In: *Electronic Journal of Business Research Methods* 11, pp. 16–28 (page 3).

Muthén, B. (1983). "Latent variable structural equation modeling with cateogrial data". In: *Journal of Econometrics* 22.1-2, pp. 43–65 (page 7).

Olsson, U. (1979). "Maximum likelihood estimation of the polychoric correlation coefficient". In: *Psychometrika* 44.4, pp. 443–460 (page 9).

Pearson, K. (1900). "Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable." In: *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 195, pp. 1–47 (page 3).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/ (pages 6, 12).

Schneider, W. Joel (2013). "What If We Took Our Models Seriously? Estimating Latent Scores in Individuals". In: *Journal of Psychoeducational Assessment* 31.2, pp. 186–201 (page 14).

– (2020). *simstandard: Generate Standardized Data*. R package version 0.6.0. URL: https://CRAN.R-project.org/package=simstandard (page 14).

Yang-Wallentin, Fan, Karl G. Jöreskog, and Hao Luo (2010). "Confirmatory Factor Analysis of Ordinal Variables With Misspecified Models". In: *Structural Equation Modeling: A Multidisciplinary Journal* 17, pp. 392–432 (page 12).

Tables

Table A.1: Theoretical correlation matrix three factors model

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	0.420	0.350	0.280	0.596	0.223	0.192	0.160	0.128	0.330	0.283	0.236
2	0.420	1.000	0.300	0.240	0.511	0.192	0.164	0.137	0.109	0.283	0.242	0.202
3	0.350	0.300	1.000	0.200	0.426	0.160	0.137	0.114	0.091	0.236	0.202	0.168
4	0.280	0.240	0.200	1.000	0.341	0.128	0.109	0.091	0.073	0.188	0.162	0.135
12	0.596	0.511	0.426	0.341	1.000	0.517	0.443	0.369	0.295	0.578	0.495	0.413
5	0.223	0.192	0.160	0.128	0.517	1.000	0.420	0.350	0.280	0.191	0.164	0.136
6	0.192	0.164	0.137	0.109	0.443	0.420	1.000	0.300	0.240	0.164	0.140	0.117
7	0.160	0.137	0.114	0.091	0.369	0.350	0.300	1.000	0.200	0.136	0.117	0.098
8	0.128	0.109	0.091	0.073	0.295	0.280	0.240	0.200	1.000	0.109	0.094	0.078
9	0.330	0.283	0.236	0.188	0.578	0.191	0.164	0.136	0.109	1.000	0.420	0.350
10	0.283	0.242	0.202	0.162	0.495	0.164	0.140	0.117	0.094	0.420	1.000	0.300
11	0.236	0.202	0.168	0.135	0.413	0.136	0.117	0.098	0.078	0.350	0.300	1.000

Table A.2: Theoretical correlation matrix four factors model

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	0.420	0.350	0.593	0.173	0.148	0.124	0.250	0.214	0.178	0.308	0.264
2	0.420	1.000	0.300	0.508	0.148	0.127	0.106	0.214	0.184	0.153	0.264	0.226
3	0.350	0.300	1.000	0.423	0.124	0.106	0.088	0.178	0.153	0.128	0.220	0.188
12	0.593	0.508	0.423	1.000	0.417	0.358	0.298	0.478	0.410	0.341	0.517	0.443
4	0.173	0.148	0.124	0.417	1.000	0.420	0.350	0.087	0.074	0.062	0.109	0.094
5	0.148	0.127	0.106	0.358	0.420	1.000	0.300	0.074	0.064	0.053	0.094	0.080
6	0.124	0.106	0.088	0.298	0.350	0.300	1.000	0.062	0.053	0.044	0.078	0.067
7	0.250	0.214	0.178	0.478	0.087	0.074	0.062	1.000	0.420	0.350	0.157	0.135
8	0.214	0.184	0.153	0.410	0.074	0.064	0.053	0.420	1.000	0.300	0.135	0.116
9	0.178	0.153	0.128	0.341	0.062	0.053	0.044	0.350	0.300	1.000	0.112	0.096
10	0.308	0.264	0.220	0.517	0.109	0.094	0.078	0.157	0.135	0.112	1.000	0.420
11	0.264	0.226	0.188	0.443	0.094	0.080	0.067	0.135	0.116	0.096	0.420	1.000

Table A.3: Theoretical correlation matrix five factors model

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	0.420	0.554	0.180	0.154	0.128	0.250	0.214	0.307	0.263	0.324	0.278
2	0.420	1.000	0.475	0.154	0.132	0.110	0.214	0.184	0.263	0.226	0.278	0.238
3	0.554	0.475	1.000	0.354	0.303	0.253	0.407	0.349	0.453	0.388	0.462	0.396
4	0.180	0.154	0.354	1.000	0.420	0.350	0.084	0.072	0.113	0.097	0.124	0.107
5	0.154	0.132	0.303	0.420	1.000	0.300	0.072	0.062	0.097	0.083	0.107	0.091
6	0.128	0.110	0.253	0.350	0.300	1.000	0.060	0.051	0.081	0.069	0.089	0.076
7	0.250	0.214	0.407	0.084	0.072	0.060	1.000	0.420	0.159	0.136	0.156	0.134
8	0.214	0.184	0.349	0.072	0.062	0.051	0.420	1.000	0.136	0.117	0.134	0.115
9	0.307	0.263	0.453	0.113	0.097	0.081	0.159	0.136	1.000	0.420	0.198	0.170
10	0.263	0.226	0.388	0.097	0.083	0.069	0.136	0.117	0.420	1.000	0.170	0.146
11	0.324	0.278	0.462	0.124	0.107	0.089	0.156	0.134	0.198	0.170	1.000	0.420
12	0.278	0.238	0.396	0.107	0.091	0.076	0.134	0.115	0.170	0.146	0.420	1.000

Figures

A.1 Categorizations

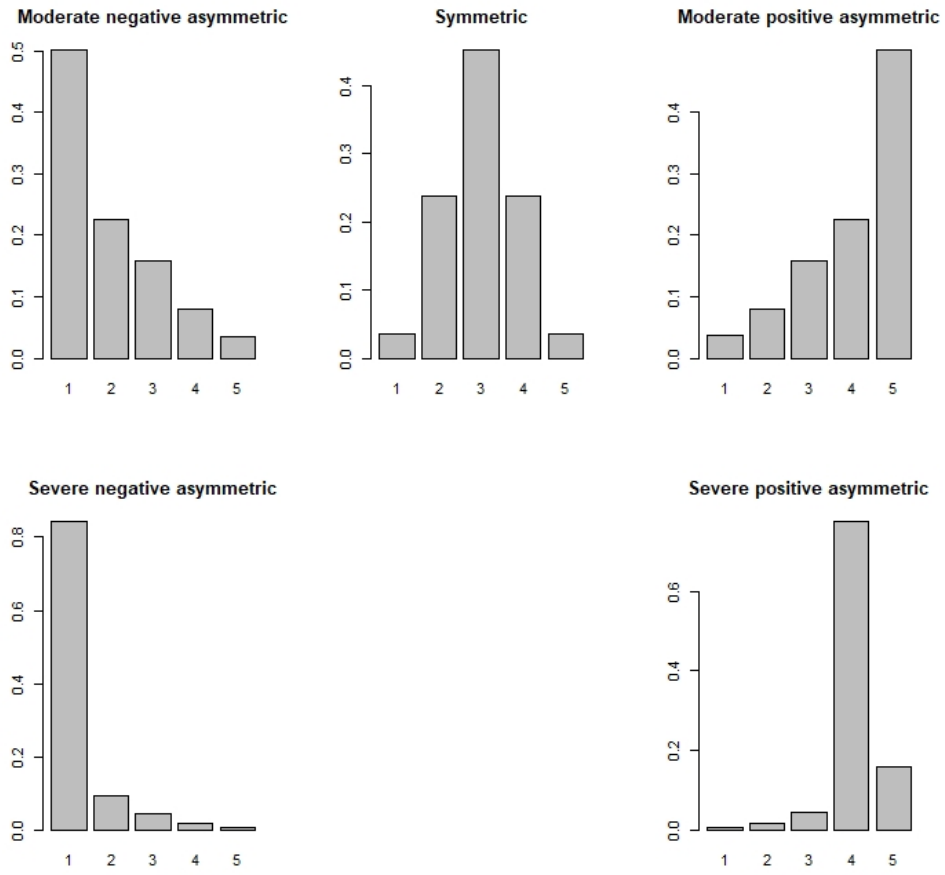

Figure A.1: Bar plots over categorizations.