Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Statistics and Machine Learning
2024 | LIU-IDA/STAT-A--24/022--SE

Effects of symmetry conditions, correlation matrix type and unmatched factor numbers on factor loading in Exploratory Factor Analysis with ordinal data

Jin Yan

Supervisor : Annika Tillander
Examiner : Frank Miller

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

Exploratory Factor Analysis (EFA) is a statistical technique used to uncover unobserved (latent) structures that explain patterns of correlations among observed variables. When combined with Likert scales, it becomes a powerful tool in social science research. Given the ordinal nature of data derived from such scales, evaluating the performance of EFA under various experimental conditions is a topic of significant interest. This thesis investigates the performance of EFA with ordinal data across three key aspects. The first aspect examines the impact of symmetry conditions on EFA results. The second focuses on the differences between Pearson correlation matrix and the polychoric correlation matrix. The third explores the effect of using unmatched factor numbers in the EFA process. To address these aspects, simulation methods were used to generate normally distributed data, which were subsequently transformed into ordinal data to mimic real-world conditions. The EFA was then applied to the simulated datasets, and the root mean square error (RMSE) was employed as a metric to evaluate the accuracy of the factor loadings. Additionally, analysis of variance (ANOVA) was implemented to provide statistical insights into the effects of the experimental conditions. Overall, the results indicate that using the polychoric correlation matrix in combination with symmetric conditions yields the most reliable EFA performance for ordinal data. However, under other conditions, the performance of EFA was found to be less stable.

Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Annika Tillander, whose professionalism, patience, and unwavering sense of responsibility deeply inspired me during the past five months. Her guidance during our regular meetings not only steered my research in the right direction but also provided me with the confidence and encouragement to tackle the challenges encountered along the way.

I would also like to extend my sincere thanks to my examiner, Professor Frank Miller, and my opponent, Alan Cacique Tamariz, for their invaluable feedback and constructive suggestions, which have significantly improved the quality of this thesis.

Lastly, I am profoundly grateful to my parents and my girlfriend for their steadfast emotional support and understanding throughout this journey. Their encouragement has been a source of strength that allowed me to persevere and successfully complete this thesis.

Contents

List of Figures

List of Tables

List of Notations and Abbreviations

Symbol/Abbreviation	Definition
$\mathbf{X}$	Continuous data matrix with n observations and p variables
n	Number of observations (sample size)
p	Number of variables
x_{ij}	Value of the j-th variable for the i-th observation
$\mathbf{X}^*$	Ordinal data matrix with n observations and p variables
x_{ij*}	Value of the j-th variable for the i-th observation
$\mathbf{R}$	Correlation matrix
r_{ij}	Pearson correlation coefficient between variable i and variable j
x_{ki}	k-th element in the i-th column of $\mathbf{X}$
x_{kj}	k-th element in the j-th column of $\mathbf{X}$
$\bar{x}_i$	Mean of the i-th column of $\mathbf{X}$
$\bar{x}_j$	Mean of the j-th column of $\mathbf{X}$
X_1^*, X_2^*	Observed ordinal variables
ρ_{12}	Correlation between the first and second latent variables
X_1, X_2	Simulated latent continuous variables associated with ordinal variables X_1^* and X_2^*
n_{ij}	Number of cases where X_1^* is in category i and X_2^* in category j
P_{ij}	Joint probability of observing category i for X_1^* and category j for X_2^*
a_{i-1}, a_i	Thresholds for variable X_1 in category i

b_{j-1}, b_j	Thresholds for variable X_2 in category j
x_1 and x_2	The corresponding latent continuous variables
P^*_{ij}	The observed proportion in the cell (i, j) of a contingency table
$P^*_{i.}$	The observed cumulative marginal proportions of the contingency table
$P^*_{.j}$	The observed cumulative marginal proportions of the contingency table
$P^*_{i-1.}$	The observed cumulative marginal proportions of the contingency table
$P^*_{.j-1}$	The observed cumulative marginal proportions of the contingency table
s	The number of rows or columns in the contingency table
$\Phi^{-1}(\cdot)$	The Probit function
$\ln L$	Log-likelihood function
μ_i	Expected value of variable X_i
λ_{ik}	Loading of factor k on variable i
f_{jk}	Common factor k for individual j
ϵ_{ij}	Specific factor for variable i for individual j
$\mathbf{x}$	Vector of observed variables for an individual
$\mathbf{f}$	Vector of common factors
$\boldsymbol{\epsilon}$	Vector of specific factors
$\mathbf{\Lambda}$	Factor loading matrix
λ_{ij}	Loading of factor j on variable i
$\mathbf{\Phi}$	Correlation matrix among common factors
$\mathbf{D}_\psi$	Covariance matrix of unique factors
$\mathbf{x_i}$	Vector of observed variables for the i-th individual
$\lvert\mathbf{R}\rvert$	Determinant of the Pearson correlation matrix
$\mathbf{R}^{-1}$	Inverse of the Pearson correlation matrix
Y	A continuous variable
N	Sample Size
k	The number of subgroups
N_i	The number of elements in the ith subgroup
W	The Levene test statistic
$\bar{Y}_{i.}$	The mean of the i-th subgroup

Z_{ij}	$\lvert Y_{ij} - \bar{Y}_{i.} \rvert$
$\bar{Z}_{i.}$	The group mean of the Z_{ij}
$\bar{Z}_{..}$	The overall mean of the Z_{ij}
$F_{\alpha,k-1,N-K}$	The upper critical value of the F distribution
α	The significance level
ANOVA	Analysis of Variance
H_0 and H_1	Null hypothesis and Alternative hypothesis
SS_T	The sum of squares total
SS_A or SS_B	The sum of squares between groups
SS_{AB}	Sum of squares for interactions
SS_E	Error sum of squares
$RMSE_{ijk}$	Observed value for the k-th replicate in cell (i, j)
$\overline{RMSE}_{..}$	Mean of all observed values across all levels of Factors A and B
$\overline{RMSE}_{i..}$	Mean of observations for the i-th level of Factor A
m	Number of levels of Factor B
r	Number of replicates per cell
$\overline{RMSE}_{.j.}$	Mean of observations for the j-th level of Factor B
n	Number of levels of Factor A (Only for Two-Way ANOVA)
$\overline{RMSE}_{ij.}$	Mean of observations for the cell (i, j), averaged over replicates
$RMSE_{ijk}$	Observed value for the k-th replicate in cell (i, j).
$\overline{RMSE}_{ij.}$	Mean of observations for the cell (i, j), averaged over replicates
MS_A, MS_B	Mean Square for A or B
MS_{AB}	Mean Square for the interaction between A and B
MS_E	Mean Square for Error
df_A, df_B	Degrees of Freedom for Factor A or B
df_{AB}	Degrees of Freedom for the Interaction Between Factors A and B
df_E	Degrees of Freedom for the Error (Residuals)
μ_{tjk}	The trimmed mean for the group
C	A $k \times p$ contrast matrix of rank k, representing the hypothesis being tested
j'_J	A $1 \times J$ vector of ones

C_J	The matrix used in constructing C
$\otimes$	The (right) Kronecker product
v_{jj}	Yuen's estimate of the squared standard error of the sample trimmed mean for the jth group
Q	Test statistics for the robust two-way ANOVA
RMSE	Root Mean Square Error
y_i	i-th element in the factor loading matrix used to simulate $\mathbf{X}$
$\hat{y}_i$	The i-th element in the factor loading matrix produced by EFA
n	The number of elements in the loading matrix (Only for RMSE)
Q_1	The first quartile
Q_3	The third quartile
IQR	Interquartile range
n	The number of elements in the group (Only for Trimmed Mean, SE and confidence interval)
$\bar{x}_{tk}$	k-times trimmed mean
$\bar{x}_{wk}$	Winsorized mean
s^2_{wk}	Winsorized sum of squared deviations
$SE(\bar{x}_{tk})$	Standard Error of the trimmed mean
γ	The trim rate and is equal to k/n

1 Introduction

1.1 Motivation

Exploratory Factor Analysis (EFA) is a widely used statistical technique for identifying underlying relationships among observed variables. It enables researchers to uncover latent structures that explain patterns of correlations within a dataset.

In social science research, EFA is often employed alongside questionnaires that collect participants' self-reported characteristics, frequently utilizing the Likert response scale. The Likert scale is a psychometric tool that aggregates responses across several Likert items. Each item consists of a statement, and respondents are asked to indicate their level of agreement or disagreement using a symmetric agree-disagree scale.[2]

An example of this application is demonstrated in the work of [8], who used EFA to identify factors influencing residential location choices in South Africa. The study utilized a Likert scale consisting of 90 variables. A total of 266 families participated in the research and provided self-reported data for the scale. Following the analysis, the researchers identified eight factors influencing residential location choices, including demographic characteristics, government-provided services, and green building features.

While the combination of EFA and the Likert scale is a robust approach for addressing such research questions, two types of errors may arise when analyzing Likert-scaled data: categorization errors and transformation errors. Categorization errors occur when continuous data is transformed into categorical data, while transformation errors emerge when Likert scales are poorly designed, resulting in unequal category widths.[3]

Furthermore, [7] investigated the impact of transformation errors through simulation studies. They first simulated 12 items with continuous values and then categorized these values asymmetrically. Their results demonstrated that with three factors used in the simulation and EFA process, and with asymmetric ordinal data analyzed using a polychoric correlation matrix, the loading matrix obtained was consistent with that in the simulation. Additionally, [17] noted that when the number of ordered categories exceeds five, the Pearson correlation matrix may replace the polychoric correlation matrix and still be effectively used in the EFA

process.

These findings suggest that some traditional assumptions about the prerequisites for optimal EFA performance may not always hold. This motivates further investigation to evaluate the validity and extent of these claims.

1.2 Aim

The primary aim of this thesis is to evaluate the performance of Exploratory Factor Analysis (EFA) under three experimental conditions involving ordinal data.

1.3 Research Questions

To achieve the stated aim, this study seeks to address the following research questions:

1. Does asymmetric ordinal data influence the factor loading matrix derived through EFA?
2. Is there a difference in EFA performance between ordinal data analyzed with Pearson correlation matrix and that analyzed with the polychoric correlation matrix?
3. Does the use of incorrect factor numbers, compared to those specified in the simulation design, impact the performance of EFA on ordinal data?

1.4 Ethical Considerations

This thesis employs simulation studies to evaluate the performance of Exploratory Factor Analysis (EFA). As the research relies solely on simulated data, there are no ethical concerns related to data collection or data privacy. Consequently, no ethical issues are associated with this study.

2 Theory

This chapter presents the theoretical framework that supports the Method chapter and is organized into five sections. The first section introduces the data utilized in this study, laying the groundwork for the subsequent analyses. It describes two types of data matrices: one tailored for continuous data and another designed for ordinal data. The second section examines the Pearson correlation matrix and the polychoric correlation matrix, detailing their structures and the computational methods used to derive their elements. The third section is divided into two parts. The first part delves into Exploratory Factor Analysis (EFA) and comprises three subsections: the first subsection outlines the common factor model, the second explains how the model is derived from the data, and the third discusses approaches to obtaining a loading matrix that aids in interpretation. The second part of this section introduces the two-way ANOVA, detailing its purpose and its application within this study. The fourth section addresses the experimental metric, Root Mean Square Error (RMSE), which serves as a measure of the accuracy of the factor loadings. Additionally, it defines the criteria for identifying and handling outliers in the dataset.

2.1 Introduction (Data)

Let $\mathbf{X}$ be the **continuous data matrix**, where:

- $\mathbf{X}$ is an $n \times p$ matrix and shows there are n individuals and p variables.
- Each row of $\mathbf{X}$ represents an individual.
- Each column represents a latent continuous variable in this thesis.

Mathematically, the data matrix $\mathbf{X}$ can be written as:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \tag{2.1}$$

Where x_{ij} represents the value of the j-th variable for the i-th individual.

Let $\mathbf{X}^*$ be the **ordinal data matrix**, where:

- $\mathbf{X}^*$ is an $n \times p$ matrix, and shows there are n individuals and p variables.
- Each row of $\mathbf{X}^*$ represents an individual.
- Each column represents a observable ordinal variable in this thesis.

Mathematically, the data matrix $\mathbf{X}^*$ can be written as:

$$\mathbf{X}^* = \begin{pmatrix} x_{11}* & x_{12}* & \dots & x_{1p}* \\ x_{21}* & x_{22}* & \dots & x_{2p}* \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}* & x_{n2}* & \dots & x_{np}* \end{pmatrix} \tag{2.2}$$

Where $x_{ij}*$ represents the value of the j-th variable for the i-th individual.

2.2 Correlation Matrix

Pearson Correlation Matrix

The Pearson correlation coefficient is a value ranging from -1 to 1 that measures the linear relationship between two variables, indicating both the strength and direction of the relationship.[15] A Pearson correlation matrix is a symmetric matrix in which each element represents a Pearson correlation coefficient, describing the pairwise linear correlations between the variables in the dataset. The matrix is represented as follows:

$$\mathbf{R} = \begin{pmatrix} 1 & \dots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \dots & 1 \end{pmatrix} \tag{2.3}$$

The Pearson correlation coefficient between the i-th and j-th columns (r_{ij}) is calculated as follows:

$$r_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i\right) \left(x_{kj} - \bar{x}_j\right)}{\sqrt{\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i\right)^2} \sqrt{\sum_{k=1}^{n} \left(x_{kj} - \bar{x}_j\right)^2}}$$

Where:

- x_{ki} is the k-th individual element of the i-th column in $\mathbf{X}$.
- x_{kj} is the k-th individual element of the j-th column in $\mathbf{X}$.
- $\bar{x}_i$ is the mean of the i-th column in $\mathbf{X}$.
- $\bar{x}_j$ is the mean of the j-th column in $\mathbf{X}$.

Polychoric Correlation Matrix

The polychoric correlation coefficient estimates the linear relationship between two continuous latent variables that are represented by observed ordinal variables.[9] A polychoric correlation matrix is a symmetric matrix in which each element represents a polychoric correlation coefficient. It is expressed as follows:

$$\mathbf{R} = \begin{pmatrix} 1 & \dots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \dots & 1 \end{pmatrix} \tag{2.4}$$

Where ρ_{ij} represents the polychoric correlation coefficient between the i-th ordinal variable and the j-th ordinal variable from 2.2

In the following part of this subsection, the two ordinal variables X_1^* and X_2^*, as introduced in 2.2, correspond to the first two columns of the ordinal data matrix. These variables are employed to illustrate the process of deriving the linear relationship (ρ_{12}) between the corresponding latent variables X_1 and X_2, which are the first two columns in the continuous data matrix described in 2.1. The ordinal variables X_1^* and X_2^* are assumed to have m_1 and m_2 categories, respectively. To capture the joint distribution of these two ordinal variables, a contingency table can be constructed. For example, if $m_1 = m_2 = 5$, the contingency table can be presented as follows:

$$\begin{array}{ccccc} n_{11} & n_{12} & n_{13} & n_{14} & n_{15} \\ n_{21} & n_{22} & n_{23} & n_{24} & n_{25} \\ n_{31} & n_{32} & n_{33} & n_{34} & n_{35} \\ n_{41} & n_{42} & n_{43} & n_{44} & n_{45} \\ n_{51} & n_{52} & n_{53} & n_{54} & n_{55} \end{array} \tag{2.5}$$

Where n_{ij} represents the count of observations whose variable X_1^* falls into the *i-th* category and the variable X_2^* falls into the *j-th* category.

The joint probability (P_{ij}) of observing the the *i-th* category in X_1^* and the *j-th* category in X_2^* is:

$$P_{ij} = \int_{a_{i-1}}^{a_i} \int_{b_{j-1}}^{b_j} \frac{1}{2\pi\sqrt{1-\rho_{12}^2}} \exp\left(-\frac{1}{2(1-\rho_{12}^2)} \left(x_1^2 - 2\rho_{12} x_1 x_2 + x_2^2 \right) \right) dx_1 dx_2 \tag{2.6}$$

Where:

- a_{i-1} is the lower threshold of the variable X_1^* for the ith category.
- a_i is the upper threshold of the variable X_1^* for the ith category.
- b_{j-1} is the lower threshold of the variable X_2^* for the jth category.
- b_j is the upper threshold of the variable X_2^* for the jth category.
- ρ_{12} is the linear relationship between the two latent variables (X_1 and X_2).
- x_1 and x_2 are the corresponding latent continuous variables.

Before estimating the correlation parameter ρ_{12} using maximum likelihood estimation, the remaining parameters in the likelihood function, aside from ρ_{12}, can be determined in advance. In particular, the thresholds for the categories, such as a_{i-1}, a_i, b_{j-1}, and b_j, can be estimated. As noted by [10], these thresholds can be derived based on the proportions of responses in each category.

Let P_{ij}^* denote the observed proportion in the cell (i, j) of a contingency table, as illustrated in 2.5. Similarly, $P_{i.}^*$, $P_{.j}^*$, $P_{i-1.}^*$, and $P_{.j-1}^*$ represent the observed cumulative marginal proportions of the contingency table. Assuming the contingency table consists of s rows and s columns, these proportions are used for estimating the thresholds.

$$P_{i.}^* = \sum_{k=1}^{i} \sum_{j=1}^{s} P_{kj}^* \quad P_{.j}^* = \sum_{i=1}^{s} \sum_{k=1}^{j} P_{ik}^* \quad P_{i-1.}^* = \sum_{k=1}^{i-1} \sum_{j=1}^{s} P_{kj}^* \quad P_{.j-1}^* = \sum_{i=1}^{s} \sum_{k=1}^{j-1} P_{ik}^*$$

a_i, b_j, a_{i-1} and b_{j-1} can be calculated as follows.

$$a_i = \Phi^{-1}(P_{i.}^*) \quad b_j = \Phi^{-1}(P_{.j}^*) \quad a_{i-1} = \Phi^{-1}(P_{i-1.}^*) \quad b_{j-1} = \Phi^{-1}(P_{.j-1}^*)$$

- $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the standard normal distribution.

After that, one can estimate ρ_{12} by maximizing the following likelihood function [7]:

$$\ln L = \sum_{i=1}^{m1} \sum_{j=1}^{m2} n_{ij} \log P_{ij} \tag{2.7}$$

2.3 Statistic Methods

Exploratory Factor Analysis

Common Factors Model

[9] mentioned that Exploratory Factor Analysis (EFA) aims to use **m** common factors to explain a set of p continuous variables as:

$$\begin{aligned} x_{j1} &= \mu_1 + \lambda_{11} f_{j1} + \lambda_{12} f_{j2} + \cdots + \lambda_{1k} f_{jk} + \epsilon_{j1} \\ x_{j2} &= \mu_2 + \lambda_{21} f_{j1} + \lambda_{22} f_{j2} + \cdots + \lambda_{2k} f_{jk} + \epsilon_{j2} \\ &\vdots \\ x_{jp} &= \mu_p + \lambda_{p1} f_{j1} + \lambda_{p2} f_{j2} + \cdots + \lambda_{pk} f_{jk} + \epsilon_{jp} \end{aligned} \tag{2.8}$$

Where:

- μ_i means the expected value of variable X_i.
- λ_{ik} represents the loading of factor $k (k \in \{1, \ldots, m\})$ on variable i.
- f_{jk} is the common factor k for individual $j (j \in \{1, \ldots, n\})$.
- ϵ_{ij} represent the specific factor associated with variable $i (i \in \{1, \ldots, p\})$ on individual j.

It is important to highlight that for continuous datasets, standardization is a standard pre-processing step prior to analysis [9]. This process transforms all variables to have a mean of zero and a standard deviation of one, thereby ensuring comparability by placing them on a uniform scale. Similarly, in the context of ordinal datasets, it is typically assumed that the corresponding latent continuous variables are also standardized.

Following the standardization of the dataset, the model described in Equation 2.8 can be expressed in matrix form without the inclusion of the μ_p term, as it is effectively eliminated through the standardization process.

$$\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{\epsilon} \tag{2.9}$$

Where

- $\mathbf{x}$ is a vector of observed variables for an individual.
- $\mathbf{f}$ refers to a vector of common factors shared across the variables.
- $\boldsymbol{\epsilon}$ is a vector that contains a specific factor and error of measurement.
- $\mathbf{\Lambda}$ is referred to as the factor loading matrix, representing the strength and direction of the linear influence that the common factors exert on the observed variables.[4] It is presented as follows:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pm} \end{pmatrix} \tag{2.10}$$

Extraction Methods

In Exploratory Factor Analysis (EFA), deriving a correlation matrix from the data matrix is a crucial step, as demonstrated in Equations 2.3 and 2.4. The choice of the correlation matrix depends on the dataset's characteristics, with Pearson correlation matrix being commonly employed for continuous data. Conversely, for ordinal datasets, the polychoric correlation matrix is typically utilized to better capture the relationships between variables.

After deriving the correlation matrix, the next step in Exploratory Factor Analysis (EFA) involves extracting the underlying factors. Following the approach outlined by [4], the correlation matrix is typically represented using matrix algebra as follows:

$$\mathbf{R} = \mathbf{\Lambda \Phi \Lambda}^{\mathrm{T}} + \mathbf{D}_{\psi} \tag{2.11}$$

In the above equation, $\mathbf{\Phi}$ refers to a correlation matrix among the common factors. It is presented as follows:

$$\mathbf{\Phi} = \begin{pmatrix} 1 & \phi_{12} & \dots & \phi_{1m} \\ \phi_{21} & 1 & \dots & \phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m1} & \phi_{m2} & \dots & 1 \end{pmatrix} \tag{2.12}$$

$\mathbf{D}_\psi$ represents the covariance matrix of the unique factors. The variances of these unique factors are captured by the diagonal elements of the matrix, while the off-diagonal elements represent the covariances between them. In this case, the unique factors are assumed to be orthogonal, which means all off-diagonal elements are zero.[4] The matrix is expressed as follows:

$$\mathbf{D}_\psi = \begin{pmatrix} D_{\psi 1.1} & 0 & \dots & 0 \\ 0 & D_{\psi 2.2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D_{\psi p.p} \end{pmatrix} \tag{2.13}$$

In this thesis, **Maximum Likelihood (ML)** is employed as the extraction method, as the data satisfy two key requirements of this approach. First, the original data must be randomly sampled from the population. Second, the data must follow a multivariate normal distribution.[17] The likelihood function is expressed as:

$$L(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{D}_\psi \mid \mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{x_i} \mid \mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{D}_\psi) \tag{2.14}$$

Assuming that the observed data follow a multivariate normal distribution and have already been standardized, the probability density function $f(\mathbf{x}_i \mid \mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{D}_\psi)$ can be expressed as:

$$f(\mathbf{x_i} \mid \mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{D}_\psi) = \frac{1}{(2\pi)^{p/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x_i}^T\mathbf{R}^{-1}\mathbf{x_i}\right) \tag{2.15}$$

Where

- $\mathbf{x_i}$ refers to the i-th of n individuals.
- $|\mathbf{R}|$ is the determinant of the correlation matrix $\mathbf{R}$.
- $\mathbf{R}^{-1}$ is the inverse of the correlation matrix $\mathbf{R}$.
- p is the number of variables.

Rotation

After obtaining the optimal $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{D}_\psi$, the next step is to perform a rotation, as the loading matrix is often difficult to interpret in its initial form. A common issue is that factors may exhibit varying degrees of relationships with multiple variables, making it challenging to discern their meaning. The **rotation** step addresses this issue by simplifying the structure

of the loading matrix.

During rotation, the factor axes in the factor space are adjusted around their origin. This adjustment facilitates a clearer interpretation of the factor loadings without altering the underlying structure of the data.[12]

The principles underlying the use of rotation in factor analysis are well-documented in [5], which highlights that the spatial representation of measurable variables facilitates the interpretation of factor analysis models. As depicted in Figure 2.1, the axes in this spatial framework represent the common factors, while the coordinates of each variable correspond to the factor loadings. Moreover, the distances between variables in this representation indicate the strength and direction of the correlations among the measured variables. Importantly, as long as the rotation preserves these distances, the correlations between variables remain unchanged.

Regarding the determination of optimal axes, [13] emphasizes that a loading matrix exhibiting a **simple structure** is ideal. A key characteristic of such a matrix is that each measured variable loads significantly on only a subset of the common factors, rather than on all factors.

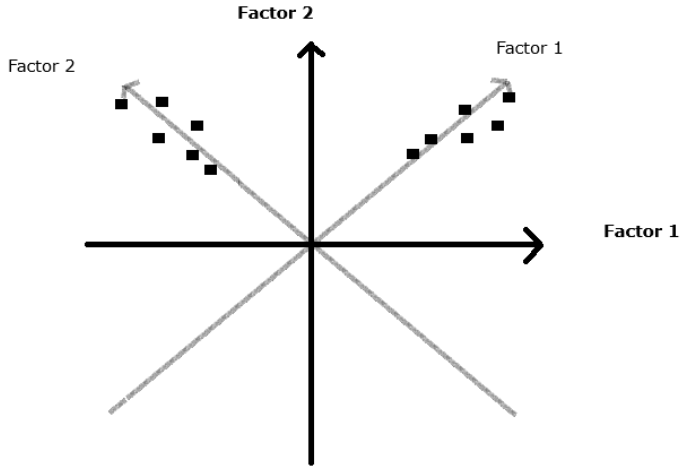

Figure 2.1: Two-Factor Model Geometrical Representation (Orthogonal Rotation of Axes). The black lines and points represent the factor structure before rotation, while the gray lines and points show the factor structure after rotation.

To achieve a loading matrix with the simple structure, two main types of rotation are commonly employed:

- **Orthogonal rotation**: In this type, the factor axes are constrained to remain at right angles, ensuring that the factors remain uncorrelated.
- **Oblique rotation**: In this type, there is no restriction on the angles between the factor axes, allowing the factors to be correlated.[17]

Generally speaking, oblique rotation is a more flexible rotation method. If there is truly no correlation between the common factors, oblique rotation yields the same results as orthogonal rotation. Furthermore, in scenarios where the angle between variable clusters deviates

from 90 degrees and orthogonal rotation is less effective, oblique rotation can still produce a factor matrix with a simple structure. This is illustrated in Figure 2.2.

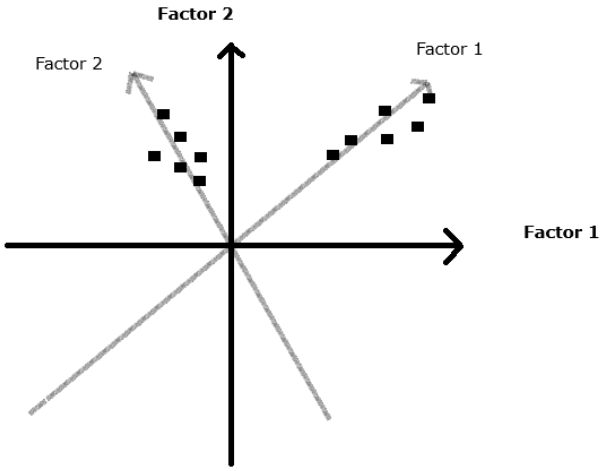

Figure 2.2: Two-Factor Model Geometrical Representation (Oblique Rotation of Axes). The black lines and points represent the factor structure before rotation, while the gray lines and points show the factor structure after rotation.

Levene Test

The Levene test is commonly employed to assess the assumption of homogeneity of variances across multiple groups in analysis of variance (ANOVA). Specifically, it tests whether the variances of k groups are equal, which is a critical assumption for performing ANOVA [1].

The null and alternative hypotheses for the Levene test are as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_a : \sigma_i^2 \neq \sigma_j^2 \text{at least for one specific pair} (i, j).$$

Let Y be a continuous variable with a sample size of N, consisting of k subgroups, where the i-th group contains N_i elements. The Levene test statistic is defined as:

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2},$$

where

- $\bar{Y}_{i.}$ is the mean of the i-th subgroup.
- $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$.
- $\bar{Z}_{i.}$ is the group mean of the Z_{ij}.
- $\bar{Z}_{..}$ is the overall mean of the Z_{ij}.

The Levene test would reject the null hypothesis, which asserts that the variances across different groups are equal, when:

$$W > F_{\alpha,k-1,N-K} \tag{2.16}$$

Where

- $F_{\alpha,k-1,N-K}$ is the upper critical value of the F distribution.
- $k-1$ is one degree of freedom.
- $N-K$ is one degree of freedom.
- α is the significance level.

Two-Way ANOVA

Analysis of Variance (ANOVA) is a widely used statistical method for testing differences between group means. Two-way ANOVA is a specific case of this method, aiming to determine the influence of two categorical variables and the effect of their interaction on a quantitative continuous variable.[16]

Before conducting the analysis, we formulate hypotheses corresponding to the ANOVA results.[16]

- **Main Effect of Categorical Variable A:**
  - H_0: The mean values of the quantitative continuous variable are equal across the different levels of categorical variable A.
  - H_1: The mean values of the quantitative continuous variable differ across the different levels of categorical variable A.
- **Main Effect of Categorical Variable B:**
  - H_0: The mean values of the quantitative continuous variable are equal across the different levels of categorical variable B.
  - H_1: The mean values of the quantitative continuous variable differ across the different levels of categorical variable B.
- **Interaction Effect Between Categorical Variables A and B:**
  - H_0: The effect of categorical variable A on the quantitative continuous variable is the same across all levels of categorical variable B (i.e., there is no interaction effect).
  - H_1: The effect of categorical variable A on the quantitative continuous variable varies across levels of categorical variable B (i.e., there is an interaction effect).

When using this method, several key terms must be considered. The first is SS_T, the total sum of squares, which reflects the overall variability in the data. The second includes SS_A and SS_B, the sum of squares due to factors A and B, respectively, representing the variability between the group means of each factor. The third is SS_{AB}, the sum of squares due to the interaction between factors A and B. In this thesis, the two factors can be **Number of Categories**, **Symmetry Condition**, **Matrix and Data** and **Number of Factors Used in EFA**. The last

is SS_E, the error sum of squares, reflecting the variability within groups. Their relationship is shown as follows:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E \tag{2.17}$$

Below is the explanation and formulas for each term:

1. **Total Sum of Squares (SS_T)**:

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(RMSE_{ijk} - \overline{RMSE}_{..})^2$$

Where:

- $RMSE_{ijk}$: Observed value for the k-th replicate in cell (i, j).
- $\overline{RMSE}_{..}$: Mean of all observed values across all levels of Factors A and B.

2. **Sum of Squares for Factor A (SS_A)**:

$$SS_A = m \cdot r \sum_{i=1}^{n}(\overline{RMSE}_{i..} - \overline{RMSE}_{..})^2$$

Where:

- $\overline{RMSE}_{i..}$: Mean of observations for the i-th level of Factor A, averaged over levels of Factor B and replicates.
- m: Number of levels of Factor B.
- r: Number of replicates per cell.

3. **Sum of Squares for Factor B (SS_B)**:

$$SS_B = n \cdot r \sum_{j=1}^{m}(\overline{RMSE}_{.j.} - \overline{RMSE}_{..})^2$$

Where:

- $\overline{RMSE}_{.j.}$: Mean of observations for the j-th level of Factor B, averaged over levels of Factor A and replicates.
- n: Number of levels of Factor A.
- r: Number of replicates per cell.

4. **Sum of Squares for Interaction (SS_{AB})**:

$$SS_{AB} = r \sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{RMSE}_{ij.} - \overline{RMSE}_{i..} - \overline{RMSE}_{.j.} + \overline{RMSE}_{..})^2$$

Where:

- $\overline{RMSE}_{ij.}$: Mean of observations for the cell (i, j), averaged over replicates.
- r: Number of replicates per cell.

5. **Sum of Squares for Error (Residuals) (SS_E)**:

$$SS_E = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} (RMSE_{ijk} - \overline{RMSE}_{ij.})^2$$

Where:

- $RMSE_{ijk}$: Observed value for the k-th replicate in cell (i, j).
- $\overline{RMSE}_{ij.}$: Mean of observations for the cell (i, j), averaged over replicates.

Key Notes:

- n: Number of levels of Factor A.
- m: Number of levels of Factor B.
- r: Number of replicates per cell (combination of Factor A and B levels).
- The total number of observations is $n \cdot m \cdot r$, which is the product of the levels of A, B, and the number of replicates r.

By dividing each sum of squares by its corresponding degrees of freedom, we obtain the mean squares MS_A, MS_B, MS_{AB}, and MS_E.

1. **Mean Square for Factor A (MS_A)**:

$$MS_A = \frac{SS_A}{df_A}$$

Where df_A is Degrees of Freedom for Factor A.

2. **Mean Square for Factor B (MS_B)**:

$$MS_B = \frac{SS_B}{df_B}$$

Where df_B is Degrees of Freedom for Factor B.

3. **Mean Square for the Interaction of Factors A and B (MS_{AB})**:

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$$

Where df_{AB} is Degrees of Freedom for the Interaction Between Factors A and B.

4. **Mean Square for the Error Term (MS_E)**:

$$MS_E = \frac{SS_E}{df_E}$$

Where df_E is Degrees of Freedom for the Error (Residuals).

To determine whether we should reject the null hypothesis regarding the **main effect of categorical variable A**, we perform an F-test comparing MS_A to the mean square error MS_E. Similarly, we perform an F-test comparing MS_{AB} to MS_E to assess whether to reject the null

hypothesis concerning the **interaction effect between categorical variables A and B**. In all cases, if the resulting p-value is less than 0.05, we reject the null hypothesis.[6]

Before implementing ANOVA, it is important to ensure that several assumptions are satisfied.[16] These assumptions are:

- **Variable Type**: The dependent variable must be quantitative and continuous, while variables A and B must be categorical.
- **Independence**: Observations must be independent; each experimental unit is measured only once.
- **Normality**: The dependent variable should be approximately normally distributed within each group. This assumption is crucial for small sample sizes but less critical for larger groups (typically when the number of observations per group exceeds 30).
- **Homogeneity of Variances**: The variances across different groups should be equal.
- **Outliers**: The data should be free of significant outliers that could affect the results.

Ensuring these assumptions are met is essential for the validity of the ANOVA results.

Robust ANOVA

The robust two-way ANOVA implemented in the 't2way' function in R utilizes trimmed means for increased robustness.[18] For a two-way factorial design, the trimmed means are summarized in Table 2.1.

	Factor B				
Factor A	μ_{t11}	μ_{t12}	$\cdots$	μ_{t1K}	$\mu_{t1\cdot}$
	μ_{t21}	μ_{t22}	$\cdots$	μ_{t2K}	$\mu_{t2\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	μ_{tJ1}	μ_{tJ2}	$\cdots$	μ_{tJK}	$\mu_{tJ\cdot}$
	$\mu_{t\cdot 1}$	$\mu_{t\cdot 2}$	$\cdots$	$\mu_{t\cdot K}$	

Table 2.1: Trimmed Means Corresponding to a J-by-K Design.

In Table 2.1, μ_{tjk} represents the trimmed mean for the group corresponding to the jth level of Factor A and the kth level of Factor B.

To evaluate the main effects and interaction effects between the two factors, it is essential to verify whether the following assumption holds.

$$H_0 : C\mu_t = 0$$

where

- $\mu_t = (\mu_{t1}, \ldots, \mu_{tJK})'$.
- C is a $k \times p$ contrast matrix of rank k, representing the hypothesis being tested.

Table 2.2 illustrates how the contrast matrix C is constructed for testing various effects. Where

- j_J' denotes a $1 \times J$ vector of ones. For instance, $j_4' = (1, 1, 1, 1)$.

Effect	C
A	$C_J \otimes j'_K$
B	$j'_J \otimes C_K$
A × B	$C_J \otimes C_K$

Table 2.2: The construction of the contrast matrix, C, for a two-way design.

- C_J refers to the matrix used in constructing C, as shown in Table 2.3.
- $\otimes$ represents the (right) Kronecker product.

The following is how C_J looks like:

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}$$

Where $C_{ii} = 1$ and $C_{i,i+1} = -1, i = 1, ..., J-1$.

Assume that V is a $p \times p$ diagonal matrix, where the diagonal elements are defined as follows:

$$v_{jj} = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)},$$

Where

- $j = 1, ..., p$.
- $p = JK$.
- v_{jj} represents Yuen's estimate of the squared standard error of the sample trimmed mean for the jth group.

The test statistics is defined as below:

$$Q = \mu_t' C' (CVC')^{-1} C\mu_t$$

If Q exceeds the $(1-\alpha)$-quantile of a chi-square distribution with k degrees of freedom, the null hypothesis (H_0) should be rejected.

2.4 Measure of Deviation

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is an index used to measure the differences between the actual values and the predicted values. It is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.18}$$

Where

- y_i is the i-th element in the factor loading matrix used to simulate $\mathbf{X}$.
- $\hat{y_i}$ is the i-th element in the factor loading matrix produced by Exploratory Factor Analysis (EFA).
- n means the number of elements in the loading matrix.

A smaller value of RMSE indicates better EFA performance.

Definition of Outliers

Outliers refer to values that are significantly distant from others in a dataset. When analyzing a group of data values, it is essential to identify these outliers to ensure accurate analysis. The process begins by calculating the first quartile (Q_1) and the third quartile (Q_3). The distance between these two values is known as the interquartile range (IQR), which is defined as:

$$\text{IQR} = Q_3 - Q_1.$$

Next, the data values are evaluated against the following thresholds:

- Data values less than $Q_1 - 1.5 \times \text{IQR}$ are considered lower outliers.
- Data values greater than $Q_3 + 1.5 \times \text{IQR}$ are considered upper outliers.

By applying these criteria, values that deviate significantly from the central tendency of the dataset are identified as outliers. This method is simple yet effective for detecting anomalies in the data.

Trimmed Mean, SE and confidence interval

Trimmed Mean

The k-times trimmed mean is calculated after removing the largest k values and the smallest k values from a group containing n elements, as shown below [11].

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_i$$

SE

To compute the Standard Error of the trimmed mean, it is necessary to first determine the Winsorized mean and the Winsorized sum of squared deviations.

Winsorized mean

$$\bar{x}_{wk} = \frac{1}{n} \left((k+1)x_{k+1} + \sum_{i=k+2}^{n-k-1} x_i + (k+1)x_{n-k} \right)$$

Winsorized sum of squared deviations

$$s^2_{wk} = (k+1)(x_{k+1} - \bar{x}_{wk})^2 + \sum_{i=k+2}^{n-k-1} (x_i - \bar{x}_{wk})^2 + (k+1)(x_{n-k} - \bar{x}_{wk})^2$$

Based on the information provided in [14], the formula for calculating the Standard Error of the trimmed mean is presented below.

$$SE(\bar{x}_{tk}) = \sqrt{\frac{s^2_{\text{wk}}}{n(1-2\gamma)^2}}$$

- n is the size of the group before the trim.
- γ is the trim rate and is equal to k/n

Confidence Interval

The confidence interval for the trimmed mean can be determined using the following formula:

$$\bar{x}_{tk} \pm t_{(1-\frac{\alpha}{2},n-2k-1)} SE(\bar{x}_{tk})$$

3 Method

This chapter outlines the experimental process and is divided into two main parts. The first part describes the method for calculating the RMSE, which serves as the metric for evaluating EFA performance under various experimental conditions. The second part explains how the RMSE values are analyzed using ANOVA to identify which experimental conditions significantly influence EFA performance and which do not. The overall process is illustrated in the flow diagram shown in Figure 3.4.

3.1 Part One: Continuous Data Simulation, Data Transformation, Exploratory Factor Analysis, and Root Mean Square Error

Continuous Data Simulation

In this study, the research objectives focus on examining the effects of varying symmetry conditions and evaluating the performance of Pearson correlation matrix versus the polychoric correlation matrix on factor loadings for ordinal data. Additionally, this study investigates the impact of mismatched factor numbers on the performance of Exploratory Factor Analysis (EFA). Through data simulation, we can establish the underlying truth, allowing for a comparison with experimental results under different conditions to assess performance differences.

In this study, the **sim.normal** function from the R package **MonteCarloSEM** is used to simulate data. This function was chosen for two reasons. First, it generates multivariate normally distributed data, which can be used to represent the latent traits underlying respondents' Likert scale answers. Second, it provides parameters to specify the factor loading matrix and the factor correlation matrix. This capability is crucial, as it allows us to establish a benchmark for comparing EFA results with the true simulated factor structure.

The parameters **ss** and **nd** in the function represent the sample size and the number of variables in the generated dataset, respectively. From a practical perspective, **ss** corresponds to the number of respondents, while **nd** represents the number of statements in a Likert scale. For this study, the sample size is set to $ss = 2000$, and the number of variables is set to $nd = 12$. A sample size of 2000 is sufficiently large to ensure that the values for each variable

exhibit a normal distribution. A total of 12 variables were selected, as this number exceeds the maximum number of factors considered in this thesis (5). This ensures sufficient scope for establishing relationships between variables and factors while accounting for computational constraints. It is important to note that the choice of 12 is arbitrary; any number greater than 5 would fulfill this purpose.

The parameters **loading** and **fcors** are used to define the factor loading matrix and the factor correlation matrix, denoted as $\mathbf{\Lambda}$ and $\mathbf{\Phi}$, respectively, in this thesis. Since the study explores EFA performance under varying numbers of factors (3, 4, and 5), three sets of $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ matrices are specified. These are presented below.

The matrices $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$, and $\mathbf{\Phi_3}$ are not identity matrices, indicating relationships between factors in each case. This setting reflects realistic scenarios, as it is common for factors to be correlated in a latent structure. Meanwhile, the matrices $\mathbf{\Lambda_1}$, $\mathbf{\Lambda_2}$, and $\mathbf{\Lambda_3}$ show that each factor is related to only a subset of the variables. This clear distinction makes it easier to monitor EFA results during the simulation process. If the differences between the obtained EFA results and the true simulated loading matrix become noticeable, the process can be stopped, and adjustments can be made accordingly.

Based on 2.11, it is important to note that $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ are independent during the simulation process. In other words, the values in these matrices can be designed relatively arbitrarily, subject to certain constraints. For instance, $\mathbf{\Phi}$ must be positive-definite to ensure validity.

$$\mathbf{\Lambda_1} = \begin{pmatrix} 0.8 & 0 & 0 \\ 0.7 & 0 & 0 \\ 0.6 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.65 & 0 \\ 0 & 0.55 & 0 \\ 0 & 0.45 & 0 \\ 0 & 0 & 0.4 \\ 0 & 0 & 0.3 \\ 0 & 0 & 0.2 \\ 0 & 0 & 0.1 \end{pmatrix} \quad \mathbf{\Lambda_2} = \begin{pmatrix} 0.8 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0.6 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0.3 \end{pmatrix} \quad \mathbf{\Lambda_3} = \begin{pmatrix} 0.8 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0.3 \end{pmatrix}$$

$$\mathbf{\Phi_1} = \begin{pmatrix} 1 & & \\ 0.556 & 1 & \\ 0.773 & 0.409 & 1 \end{pmatrix} \quad \mathbf{\Phi_2} = \begin{pmatrix} 1 & & & \\ 0.453 & 1 & & \\ 0.610 & 0.277 & 1 & \\ 0.728 & 0.323 & 0.421 & 1 \end{pmatrix} \quad \mathbf{\Phi_3} = \begin{pmatrix} 1 & & & & \\ 0.467 & 1 & & & \\ 0.601 & 0.271 & 1 & & \\ 0.727 & 0.331 & 0.424 & 1 & \\ 0.762 & 0.354 & 0.419 & 0.505 & 1 \end{pmatrix}$$

Data Transformation

In this step, the mean (μ) and standard deviation (σ) of each variable are utilized to categorize the continuous data into distinct groups, assigning a unique label to each group. It is important to highlight that, for certain items, the transformed ordinal data may not exhibit symmetry. The specific details of this process are provided below, particularly for the case where the transformed data in a column is divided into five categories:

Data with Symmetry:

If $x < \mu - 3\sigma \quad x^*$ is codified as 1

If $\mu - 3\sigma \leqslant x < \mu - \sigma \quad x^*$ is codified as 2

If $\mu - \sigma \leqslant x < \mu + \sigma \quad x^*$ is codified as 3

If $\mu + \sigma \leqslant x < \mu + 3\sigma \quad x^*$ is codified as 4

If $\mu + 3\sigma \leqslant x \quad x^*$ is codified as 5

Data with Positive Asymmetry:

If $x < \mu - 3\sigma \quad x^*$ is codified as 1

If $\mu - 3\sigma \leqslant x < \mu - 2\sigma \quad x^*$ is codified as 2

If $\mu - 2\sigma \leqslant x < \mu - \sigma \quad x^*$ is codified as 3

If $\mu - \sigma \leqslant x < \mu \quad x^*$ is codified as 4

If $\mu \leqslant x \quad x^*$ is codified as 5

Data with Negative Asymmetry:

If $x < \mu \quad x^*$ is codified as 1

If $\mu \leqslant x < \mu + \sigma \quad x^*$ is codified as 2

If $\mu + \sigma \leqslant x < \mu + 2\sigma \quad x^*$ is codified as 3

If $\mu + 2\sigma \leqslant x < \mu + 3\sigma \quad x^*$ is codified as 4

If $\mu + 3\sigma \leqslant x \quad x^*$ is codified as 5

In the symmetric case, values falling outside the range of $\mu \pm 3\sigma$ are categorized as types 1 and 5, respectively. The remaining range is evenly divided into three sections, with the values in these sections assigned the labels 2, 3, and 4. This process is illustrated in Figure 3.1.

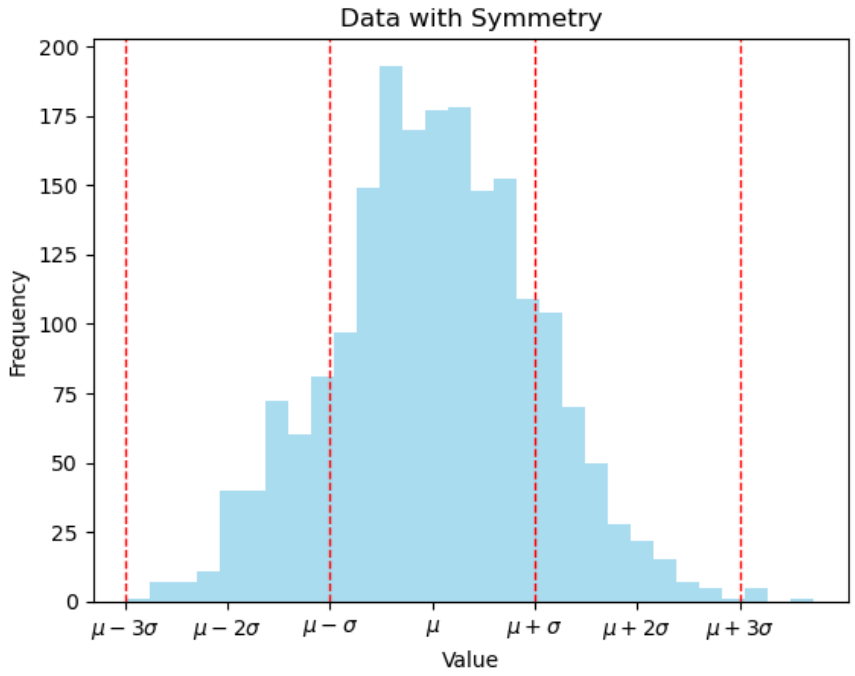

Figure 3.1: Data with Symmetry

In the positively skewed case, the 5th category encompasses the majority of the population, as values greater than μ are assigned the label 5. Similarly, values less than $\mu - 3\sigma$ are categorized as type 1. The remaining range is evenly divided into three sections, with the labels 2, 3, and 4 assigned to the values in these intervals, respectively. This distribution is illustrated in Figure 3.2.

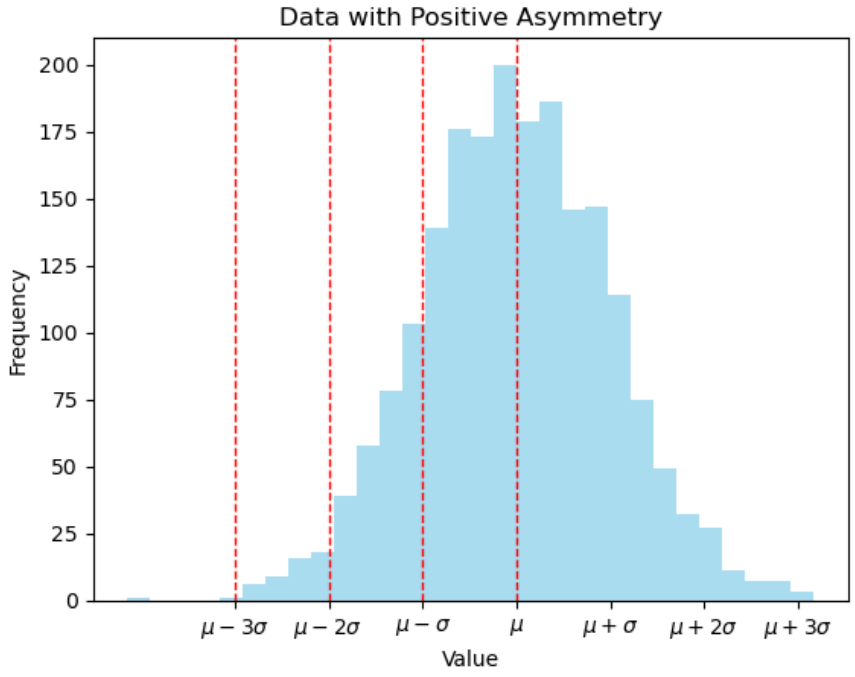

Figure 3.2: Data with Positive Asymmetry

In the negatively skewed case, the opposite pattern is observed. The 1st category encompasses the majority of the population, as values less than μ are assigned the label 1. Similarly, values greater than $\mu + 3\sigma$ are categorized as type 5. The remaining range is evenly divided into three sections, with the labels 2, 3, and 4 assigned to the values in these intervals, respectively. This distribution is effectively illustrated in Figure 3.3.

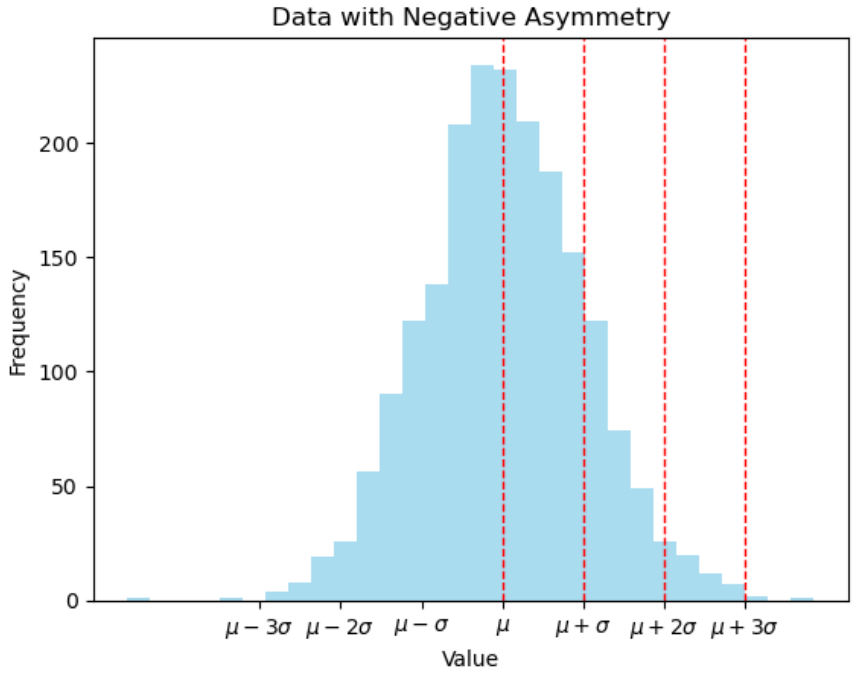

Figure 3.3: Data with Negative Asymmetry

Symmetry Condition In the experimental design, three settings are defined for the **Symmetry Condition**: **Symmetry**, **Positive Asymmetry**, and **Positive and Negative Asymmetry**. For the first two conditions (**Symmetry** and **Positive Asymmetry**), all variables in the simulated data are transformed to ordinal data based on the specific condition. For example, under the **Positive Asymmetry** condition, all 12 variables, after transformation, exhibit a positive asymmetry distribution. However, for the third condition, **Positive and Negative Asymmetry**, the transformation follows a different pattern. In this case, positive and negative asymmetry alternates across the 12 variables, resulting in an alternating sequence of asymmetry types after the transformation.

Exploratory Factor Analysis

Obtaining the Polychoric Correlation Matrix

In this step, the **hetcor** function from the **polycor** R package is used to compute the polychoric correlation matrix. The input for this function is the ordinal dataset obtained from the previous step. After executing the function, the polychoric correlation matrix is extracted using the **correlations** attribute of the resulting object.

Obtaining the Pearson Correlation Matrix

For this step, the **cor** function is applied to compute the Pearson correlation matrix. The input for this function is the ordinal dataset from the previous step, converted into numeric format

before processing. The resulting Pearson correlation matrix is then used in subsequent EFA steps.

EFA Implementation

In this step, the **fa** function from the **psych** R package is used to perform Exploratory Factor Analysis (EFA). The input for this function is the correlation matrix obtained in the previous step. Two key parameters are specified for this function. The first is **rotate**, which is set to "Promax," indicating that oblique rotation is applied. The second is **fm="mle"**, specifying that the extraction method used is maximum likelihood estimation. After executing the function, the factor **loading matrix $\mathbf{\Lambda}$** is extracted using the **loadings** attribute of the output. This matrix is subsequently used in the evaluation phase.

Root Mean Square Error (RMSE)

The final step involves comparing the factor loading matrix obtained from 3.1 with the original **loading matrix $\mathbf{\Lambda}$** used to generate the continuous data. The Root Mean Square Error (RMSE) metric is used for this comparison. Several critical details must be considered when performing this comparison.

First, the loading matrix obtained from EFA does not guarantee that the sequence of columns matches that of the matrix used for simulation. This discrepancy is inconsequential if the goal is to identify the latent structure only once, as the column sequence does not affect the identification of relationships between factors and variables. However, for RMSE calculations, this misalignment impacts the results, as RMSE measures the differences between corresponding elements in the matrices. Thus, the column sequence must be adjusted to ensure meaningful RMSE results. Without this adjustment, the RMSE may incorrectly indicate large differences between the matrices, even if the true latent structure is accurately identified.

Second, when examining the impact of unmatched factor numbers in EFA, only the first three columns of the obtained matrix are used for comparison.

3.2 Part Two: Analysis of Variance (ANOVA)

After completing the first part, sufficient data are available to conduct ANOVA and use the results to address the research questions. In this thesis, for each research question, the relationship between RMSE values and two categorical variables is examined. Consequently, a two-way ANOVA is applied. This section provides details on the hypothesis tests for each experimental condition, explains how to test the assumptions of a two-way ANOVA, and illustrates the implementation of the method.

Hypothesis Tests

Effects of Symmetric Conditions

For the first research question, two aspects are investigated. The first examines the relationship between RMSE values and category numbers under different symmetric conditions. The second determines whether the mean RMSE values differ across symmetric conditions. The corresponding **hypothesis tests** are as follows:

- **Main effect of symmetric conditions on RMSE values:**
  - H_0: The mean RMSE values are equal across different symmetric conditions.

  - H_1: The mean RMSE values differ across different symmetric conditions.

- **Interaction between category numbers and symmetric conditions:**
  - H_0: There is no interaction between category numbers and symmetric conditions. In other words, the relationship between RMSE values and category numbers is consistent across different symmetric conditions.
  - H_1: There is an interaction between category numbers and symmetric conditions. In other words, the relationship between RMSE values and category numbers varies across different symmetric conditions.

Effects of Correlation Matrix Type

For the second research question, two aspects are explored. The first concerns the relationship between RMSE values and category numbers under different correlation matrix types. The second evaluates whether the mean RMSE values differ between the Pearson and polychoric correlation matrices. The corresponding **hypothesis tests** are as follows:

- **Main effect of correlation matrix type on RMSE values:**
  - H_0: The mean RMSE values are equal between the Pearson correlation matrix and the polychoric correlation matrix.
  - H_1: The mean RMSE values differ between the Pearson correlation matrix and the polychoric correlation matrix.
- **Interaction between category numbers and correlation matrix type:**
  - H_0: There is no interaction between category numbers and correlation matrix type. In other words, the relationship between RMSE values and category numbers is consistent across different correlation matrix types.
  - H_1: There is an interaction between category numbers and correlation matrix type. In other words, the relationship between RMSE values and category numbers varies across different correlation matrix types.

Effects of Unmatched Factor Numbers in EFA

For the third research question, two aspects are examined. The first focuses on the relationship between RMSE values and category numbers under different unmatched factor numbers in EFA. The second investigates whether the mean RMSE values differ across unmatched factor numbers in EFA. The corresponding **hypothesis tests** are as follows:

- **Main effect of unmatched factor numbers in EFA:**
  - H_0: The mean RMSE values are equal across unmatched factor numbers in EFA.
  - H_1: The mean RMSE values differ across unmatched factor numbers in EFA.
- **Interaction between category numbers and unmatched factor numbers in EFA:**
  - H_0: There is no interaction between category numbers and unmatched factor numbers in EFA. In other words, the relationship between RMSE values and category numbers is consistent across different unmatched factor numbers in EFA.
  - H_1: There is an interaction between category numbers and unmatched factor numbers in EFA. In other words, the relationship between RMSE values and category numbers varies across different unmatched factor numbers in EFA.

Assumptions of a Two-Way ANOVA

To apply two-way ANOVA, several assumptions must be met. This subsection addresses **Variable type**, **Independence**, **Normality**, **Equality of variances**, and **Outliers**.

Variable type The data used are Root Mean Square Error (RMSE) values, which are quantitative and continuous, thereby satisfying this assumption.

Independence As discussed in Section 3.1, RMSE values are generated independently for each combination of factor levels. Each experimental unit is measured only once, ensuring the independence assumption is satisfied.

Outliers To address the presence of outliers, **ggplot** is utilized to visualize RMSE values across different groups. Following the definition of outliers provided in Section 2.4, the **maximum proportion of outliers across all combinations of factor levels** is determined. This proportion is subsequently applied as the trimming rate in the **t2way** function in R. By adopting this approach, the final results are adjusted to minimize the influence of outliers, ensuring greater robustness and reliability.

Normality The assumption of normality is satisfied in all tests, as the sample size for each group exceeds 30 after trimming under all experimental conditions. Prior to trimming, the group sizes are all 100, and the trimming rate never exceeds 0.2, as shown in Tables 4.2, 4.5, and 4.8. Consequently, the Central Limit Theorem is applicable in this context.

Equality of variances In this thesis, Levene's test is used to evaluate the equality of variances. Although the raw data occasionally produces significant results in this test, as presented in Tables 4.1, 4.4, and 4.7, ANOVA can still be conducted. This is because the robust **t2way** function is utilized in this thesis to minimize the effects of violations of the equal variances assumption.

Implementation of a Two-Way ANOVA

In this thesis, the **aov** function is used to implement the two-way ANOVA. The parameter **formula** in this function is utilized to specify the dependent and independent variables within the dataset. The categorical variables differ across the research questions. For the first research question, the variables are **Number of Categories** and **Symmetric Condition**. For the second, they are **Number of Categories** and **Matrix and Data Type**. For the third, the variables are **Number of Categories** and **Number of Factors Used in EFA**.

For example, in the first research question, the results generated by the function, such as those in Table 4.3, can indicate whether the main effect of Symmetric Condition is significant by comparing the p-value to 0.05. If the p-value is significantly less than 0.05, the null hypothesis for the main effect of Symmetric Condition (in 3.2) is rejected. Similarly, the p-value for the interaction effect (e.g., **Number of Categories : Symmetric Condition**) determines whether the interaction is significant. If the p-value is less than 0.05, the null hypothesis for the interaction effect is rejected. To conclude that symmetric conditions do not significantly affect RMSE values, both p-values must exceed 0.05.

It is important to note that while the **Number of Categories** is consistently included as one of the two independent variables across all experimental situations, it is not the primary focus of this research. Instead, the analysis concentrates on the main effects of the other independent variables: **Symmetric Condition**, **Matrix and Data Type**, and **Number of Factors Used in EFA**. Additionally, the interaction effects between the **Number of Categories** and

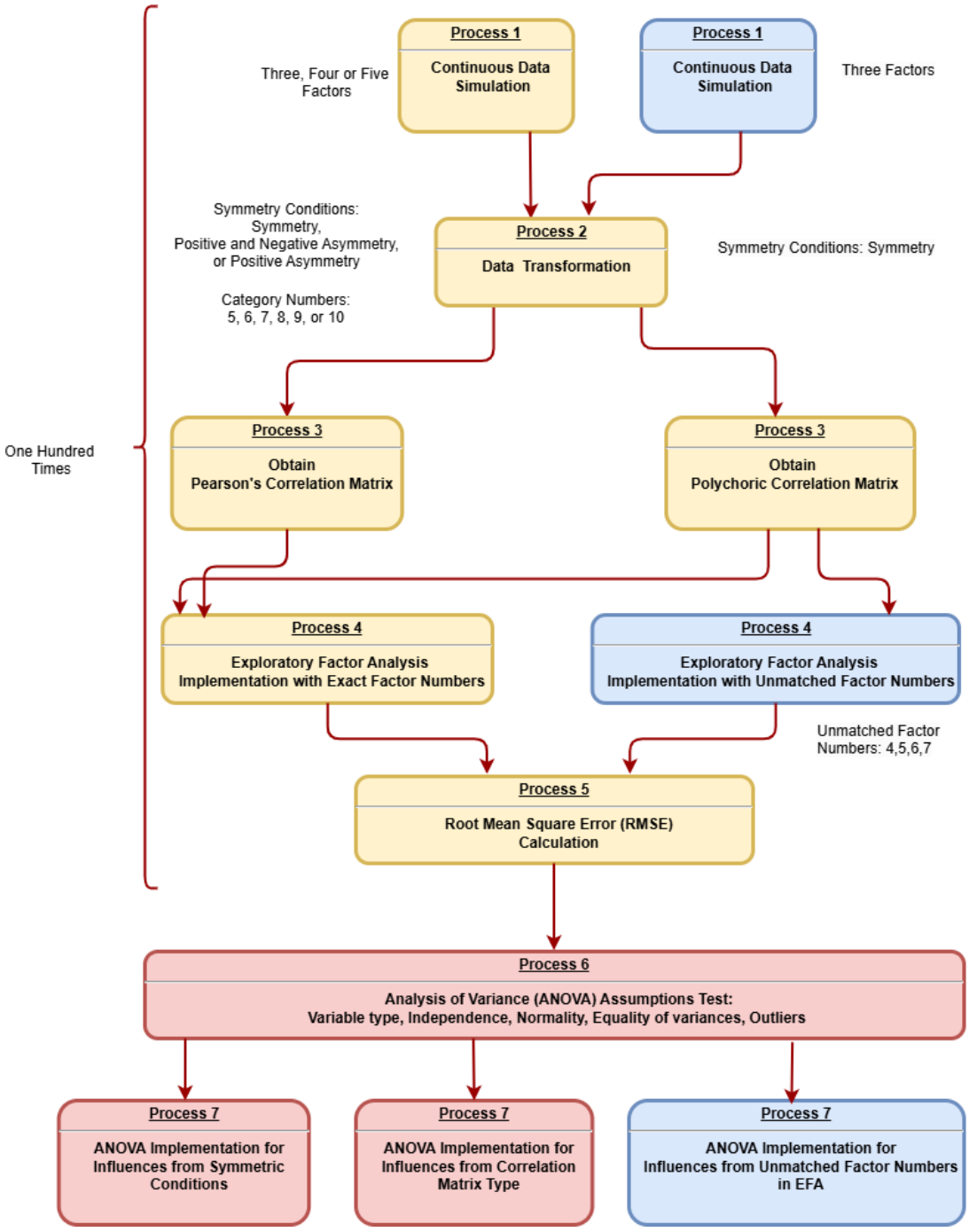

Figure 3.4: Flow Plot for Experiments

these other variables are of particular interest. By adopting this approach, the results do not explicitly display the main effect of the **Number of Categories**, as it is not relevant to the central research questions of this thesis.

4 Results

This chapter shows the Two-way ANOVA results for different experimental conditions, aiming to answer the three research questions. 4.1 contains three sets. In each set, we can find two outlier plots and one main plot. From the main plot, we can see the changes of EFA performances caused by the different symmetry situations roughly. Combined with the corresponding ANOVA summary table, the conclusion can be drawed for the first research question. As for the second research question, we can get the information from 4.2. This section contains 9 sets, among which the frist three ones are all with ordinal data with symmetry situation, the middle three ones are all with ordinal data with positive and negative asymmetry situation and the last three ones are all with ordinal data with positive asymmetry situation. In each set of the nine ones, we can also see two outlier plots and a main plot, but in this kind of main plot, the comparison is mainly between the matrix and data settings. For matrix and data, we have two ones to choose: polychoric matrix with ordinal data and pearson matrix with ordinal data. 4.3 only contains one set, and it shows the changes in the EFA performance as gap between the real factor number and the factor number used in EFA in reality increases.

4.1 EFA performance when ordinal data with polychoric matrix is with different symmetry conditions

4.1.1

Levene's test for homogeneity of variances, summarized in Table 4.1, resulted in a p-value of 0.5247, which exceeds the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups can be considered homogeneous.

Figure 4.1 highlights the presence of outliers in the RMSE values, with Table 4.2 showing that the maximum proportion of outliers among all combinations of factor levels reached 6%. To address the impact of these outliers, a trimming rate of 6% was employed during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.3, indicate that the interaction term "Number of Categories:Symmetry Condition" is not statistically significant, with a p-value of 0.101. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not meaningfully influenced by the "Symmetry Condition."

On the other hand, the main effect of "Symmetry Condition" is highly significant, with a p-value of 0.001 (denoted by ***), demonstrating that the mean RMSE varies significantly across different symmetry conditions.

In conclusion, when performing EFA with three factors using a polychoric correlation matrix on ordinal data, the symmetry or asymmetry of the ordinal variables has a significant impact on the accuracy of the EFA results.

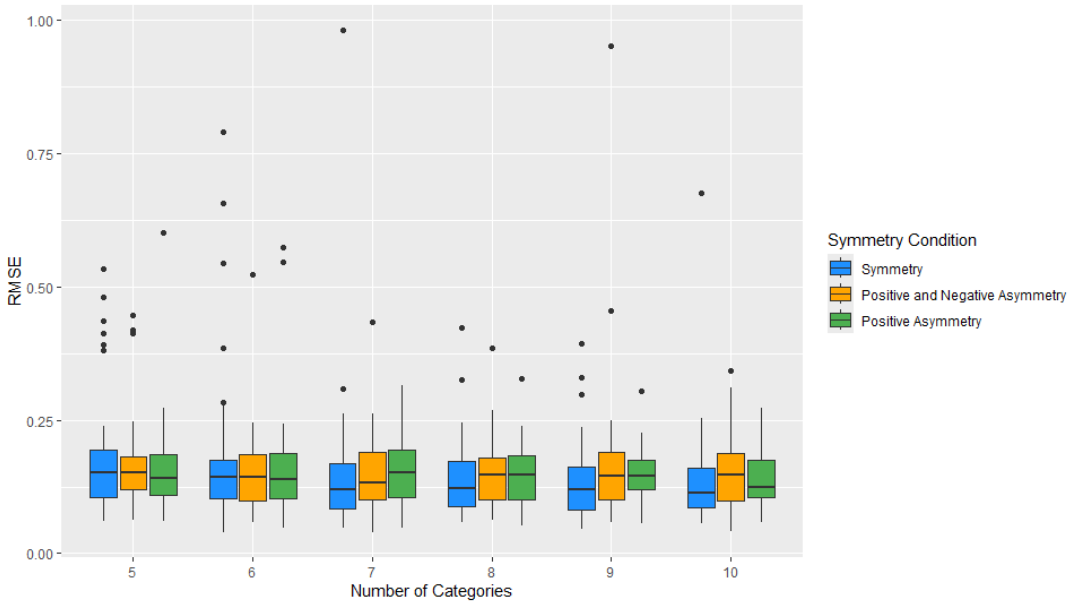

Figure 4.1: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis with Three Factors in Simulation and EFA Process (Raw Data / Three Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

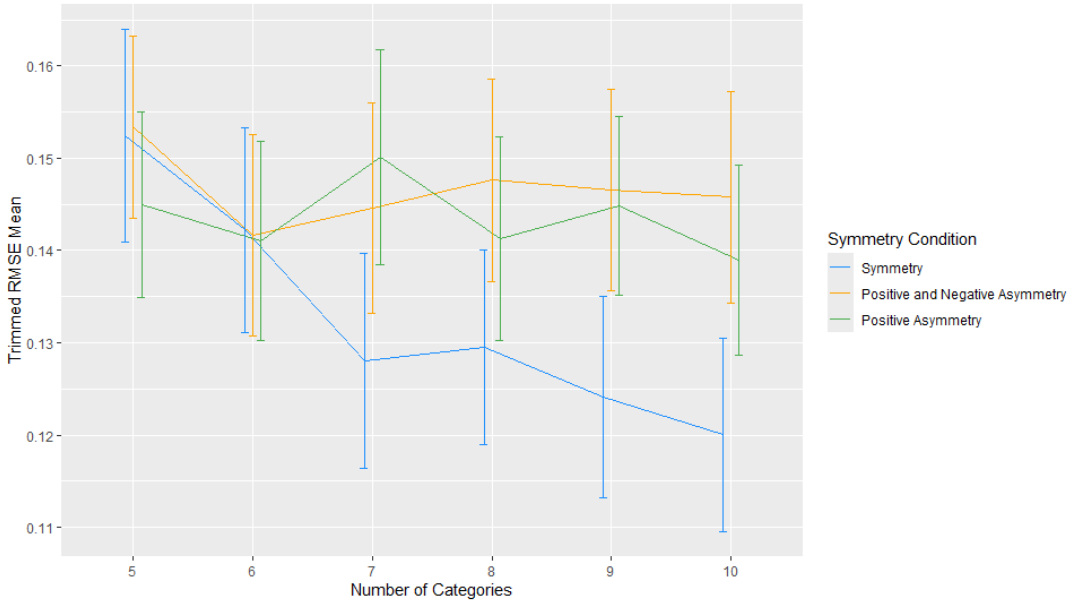

Figure 4.2: RMSE trimmed means with 95% confidence interval (Three Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

4.1.2

Levene's test for homogeneity of variances, summarized in Table 4.1, resulted in a p-value of 0.0141 (denoted by *), which falls below the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are not homogeneous.

Figure 4.3 demonstrates the presence of outliers in the RMSE values, with Table 4.2 showing that the maximum proportion of outliers among all combinations of factor levels reached 9%. To address the influence of these outliers, a trimming rate of 9% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.3, reveal that the interaction term "Number of Categories:Symmetry Condition" is statistically significant, with a p-value of 0.001 (indicated by ***). This suggests that the relationship between "Number of Categories" and RMSE in EFA is significantly influenced by the "Symmetry Condition."

Additionally, the main effect of "Symmetry Condition" is highly significant, with a p-value of 0.001 (denoted by ***), indicating substantial differences in the mean RMSE across the various symmetry conditions.

In conclusion, when performing EFA with three factors using a polychoric correlation matrix on ordinal data, the symmetry or asymmetry of the ordinal variables has a significant impact on the accuracy of the EFA results.

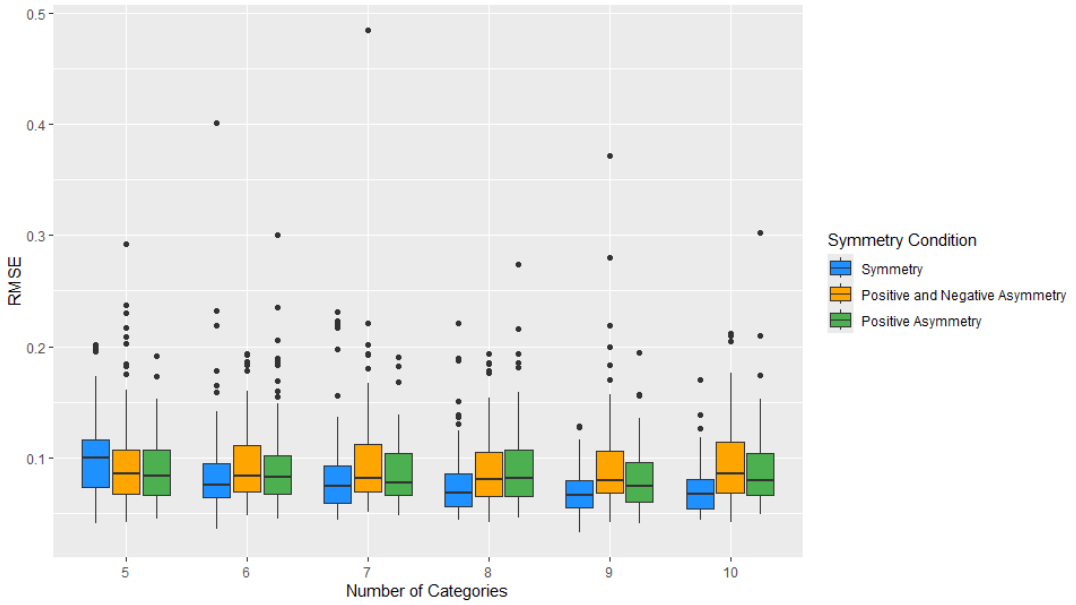

Figure 4.3: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis with Four Factors in Simulation and EFA Process (Raw Data / Four Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

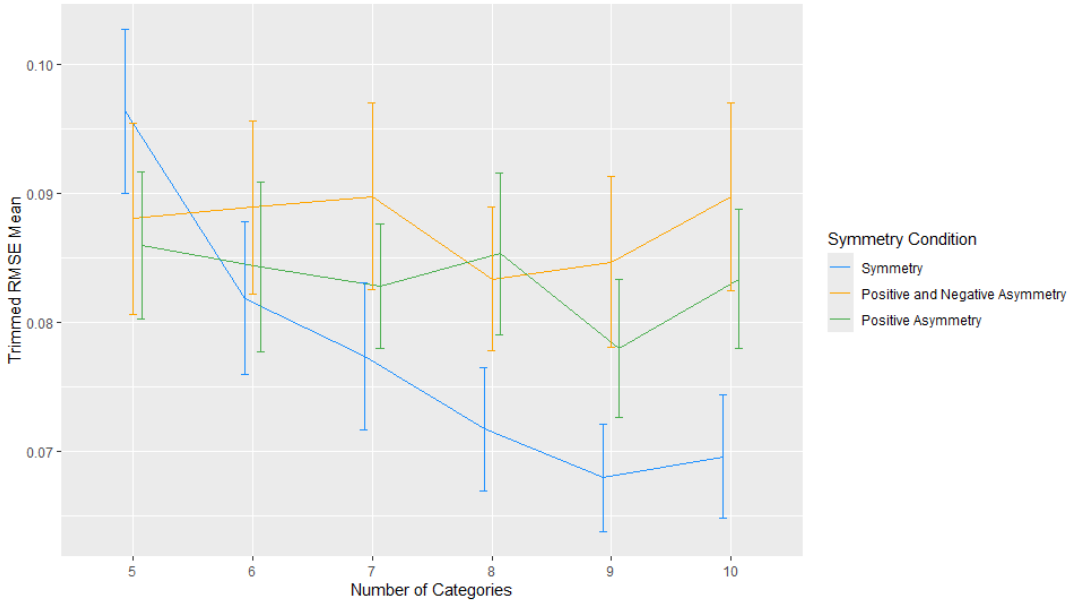

Figure 4.4: RMSE trimmed means with 95% confidence interval (Four Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

4.1.3

Levene's test for homogeneity of variances, summarized in Table 4.1, yielded a p-value of 0.3304, which exceeds the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are homogeneous.

Figure 4.5 highlights the presence of outliers in the RMSE values, with Table 4.2 showing that the maximum proportion of outliers among all combinations of factor levels reached 11%. To address the influence of these outliers, a trimming rate of 11% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.3, indicate that the interaction term "Number of Categories:Symmetry Condition" is statistically significant, with a p-value of 0.022 (denoted by *). This finding suggests that the relationship between "Number of Categories" and RMSE in EFA is significantly influenced by the "Symmetry Condition."

Additionally, the main effect of "Symmetry Condition" is statistically significant, with a p-value of 0.018 (denoted by *), demonstrating that the mean RMSE varies significantly across different symmetry conditions.

In conclusion, when performing EFA with three factors using a polychoric correlation matrix on ordinal data, the symmetry or asymmetry of the ordinal variables has a notable impact on the accuracy of the EFA results.

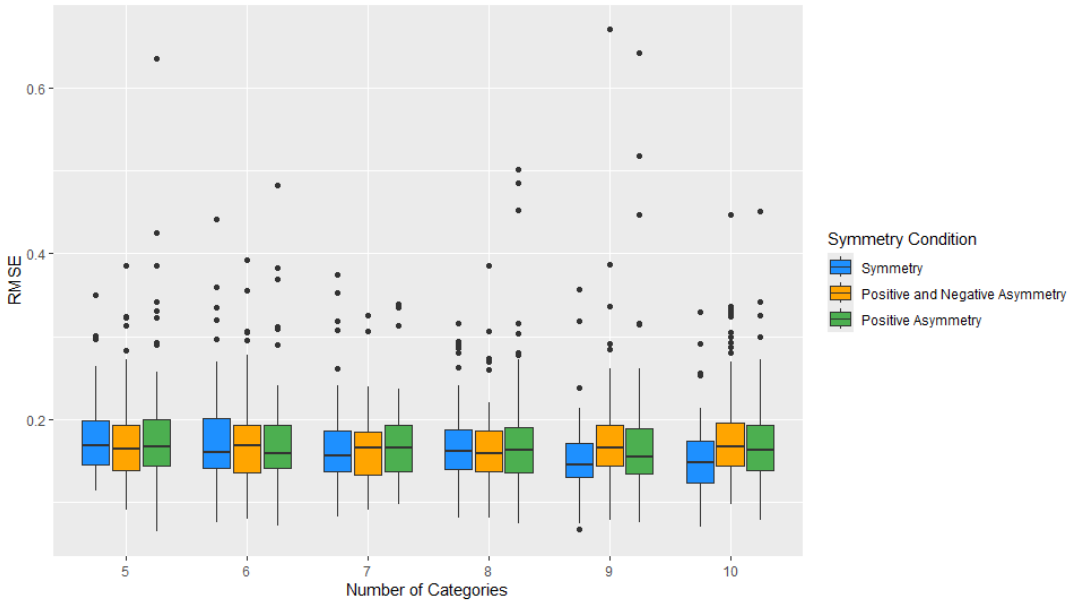

Figure 4.5: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis with Five Factors in Simulation and EFA Process (Raw Data / Five Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

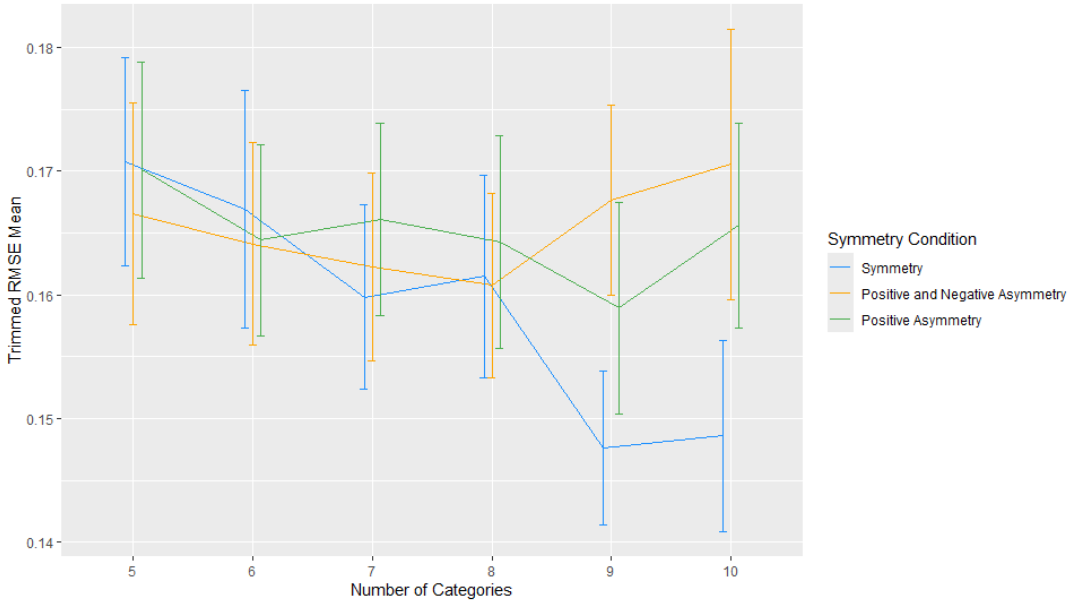

Figure 4.6: RMSE trimmed means with 95% confidence interval (Five Factors in Simulation and EFA, Two Independent Categorical Variables - Number of Categories and Symmetry Conditions).

Factor Numbers	Levene's Test (F-value / p-value)
3	0.9407 / 0.5247
4	1.9049 / 0.0141 *
5	1.1169 / 0.3304

Table 4.1: Levene's Test for Homogeneity of Variance Summary Table (Two Independent Categorical Variables - Number of Categories and Symmetry Conditions)

Factor Numbers	Highest Proportion of Outliers
3	6%
4	9%
5	11%

Table 4.2: Highest Proportion of Outliers Summary Table (Two Independent Categorical Variables - Number of Categories and Symmetry Conditions)

Factor Numbers	Symmetry Condition Main Effect (F-value / p-value)
3	20.3758 / 0.001 ***
4	32.3292 / 0.001 ***
5	8.1736 / 0.018 *
Factor Numbers	**Interaction Effect (F-value / p-value)**
3	16.2270 / 0.101
4	33.6730 / 0.001 ***
5	21.3667 / 0.022 *

Table 4.3: ANOVA Summary Table (Two Independent Categorical Variables - Number of Categories and Symmetry Conditions)

4.2 EFA performance comparison between polychoric matrix with ordinal data and Pearson matrix with ordinal data

4.2.1

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.5933, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.7 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels was 6%. To

mitigate the impact of these outliers, a trimming rate of 6% was applied during the robust ANOVA analysis.

The results of the ANOVA, presented in Table 4.6, show that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.945. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not substantially affected by "Matrix and Data."

Additionally, the main effect of "Matrix and Data" does not reach statistical significance, with a p-value of 0.088, indicating that the mean RMSE does not differ significantly across the various matrix and data conditions.

In summary, under symmetric conditions and using three factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—does not significantly impact the accuracy of the EFA results.

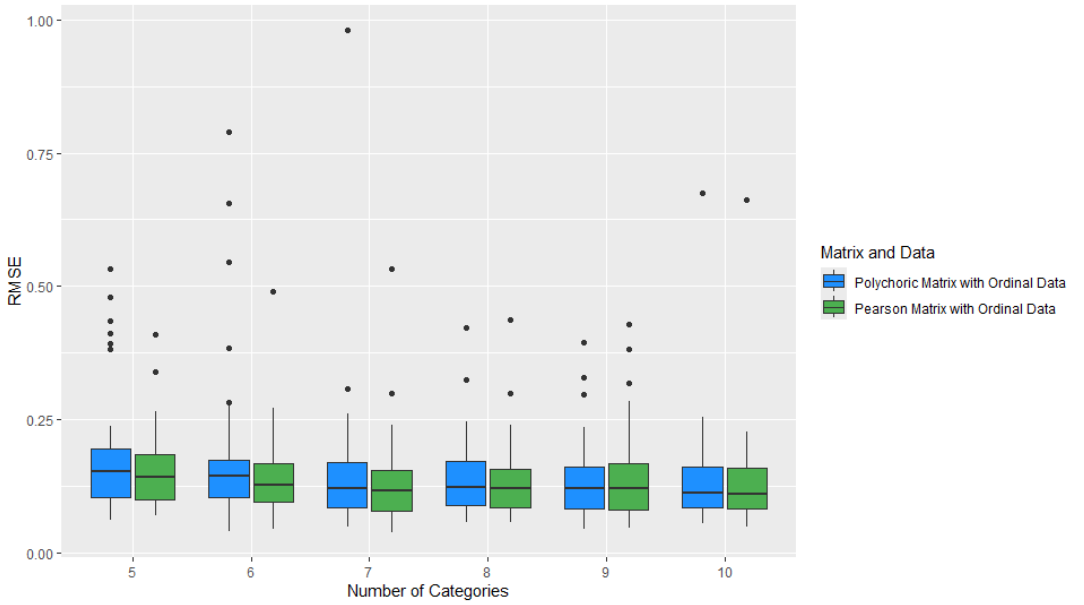

Figure 4.7: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data /Symmetry Condition: Symmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

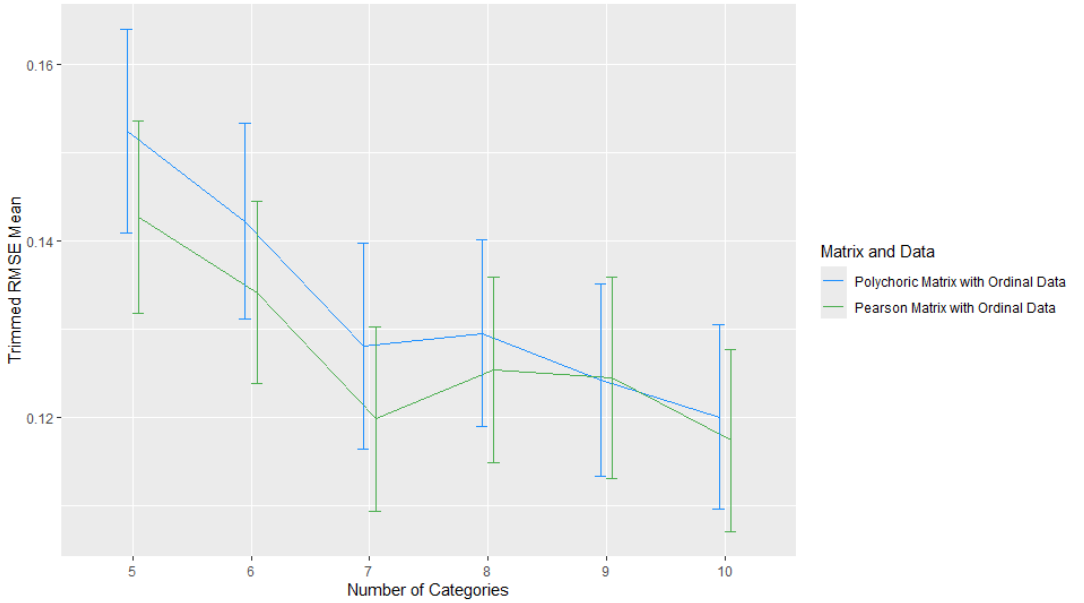

Figure 4.8: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Symmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.2

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.0009616 (denoted by ***), which is below the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are inconsistent and not homogeneous.

Figure 4.9 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 10%. To mitigate the influence of these outliers, a trimming rate of 10% was applied during the robust ANOVA analysis.

The results of the ANOVA, presented in Table 4.6, reveal that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.999. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not significantly affected by "Matrix and Data."

Additionally, the main effect of "Matrix and Data" does not reach statistical significance, with a p-value of 0.554, indicating that the mean RMSE does not differ significantly across the various matrix and data conditions.

In conclusion, under symmetric conditions and using four factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—does not significantly impact the accuracy of the EFA results.

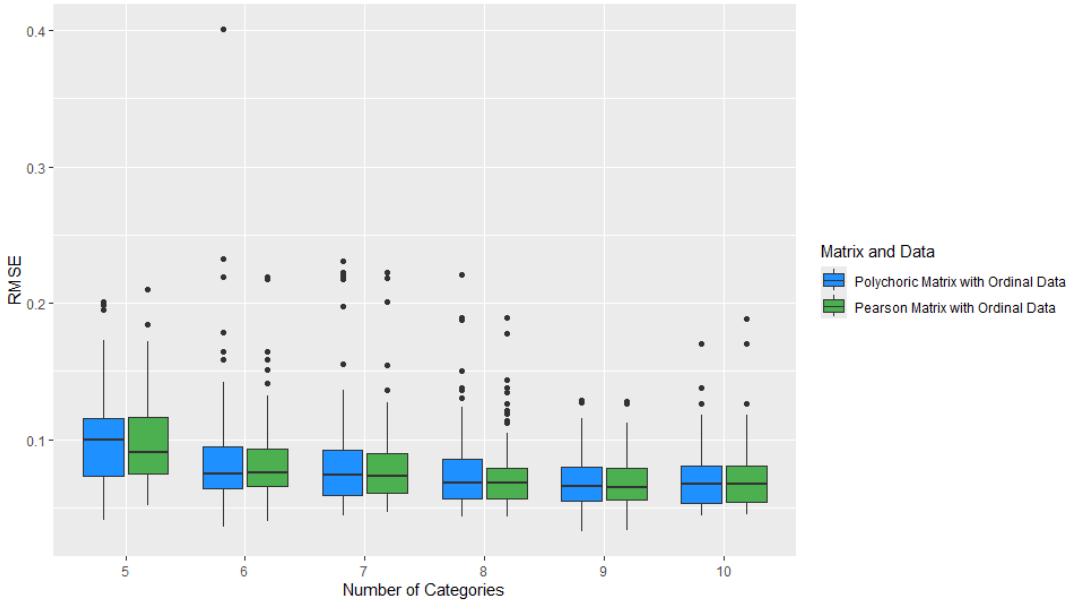

Figure 4.9: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Symmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

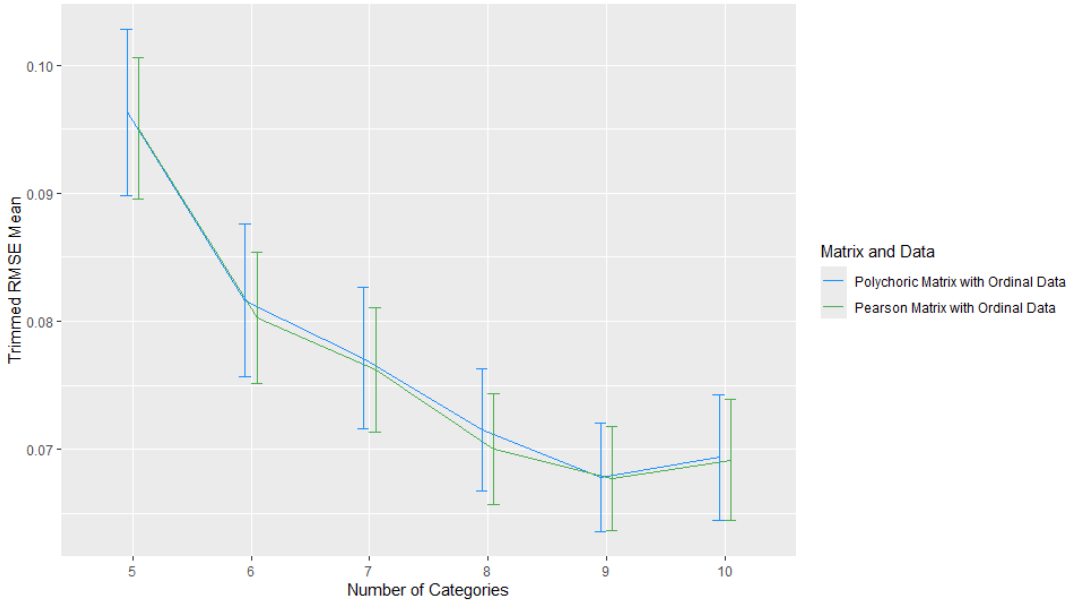

Figure 4.10: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Symmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.3

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.02992 (denoted by *), which is below the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are inconsistent and not homogeneous.

Figure 4.11 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 8%. To

mitigate the influence of these outliers, a trimming rate of 8% was applied during the robust ANOVA analysis.

The results of the ANOVA, presented in Table 4.6, reveal that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.689. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not substantially affected by "Matrix and Data."

However, the main effect of "Matrix and Data" is statistically significant, with a p-value of 0.003 (denoted by **), indicating that the mean RMSE varies significantly across the different matrix and data conditions.

In conclusion, under symmetric conditions and using five factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—significantly impacts the accuracy of the EFA results.

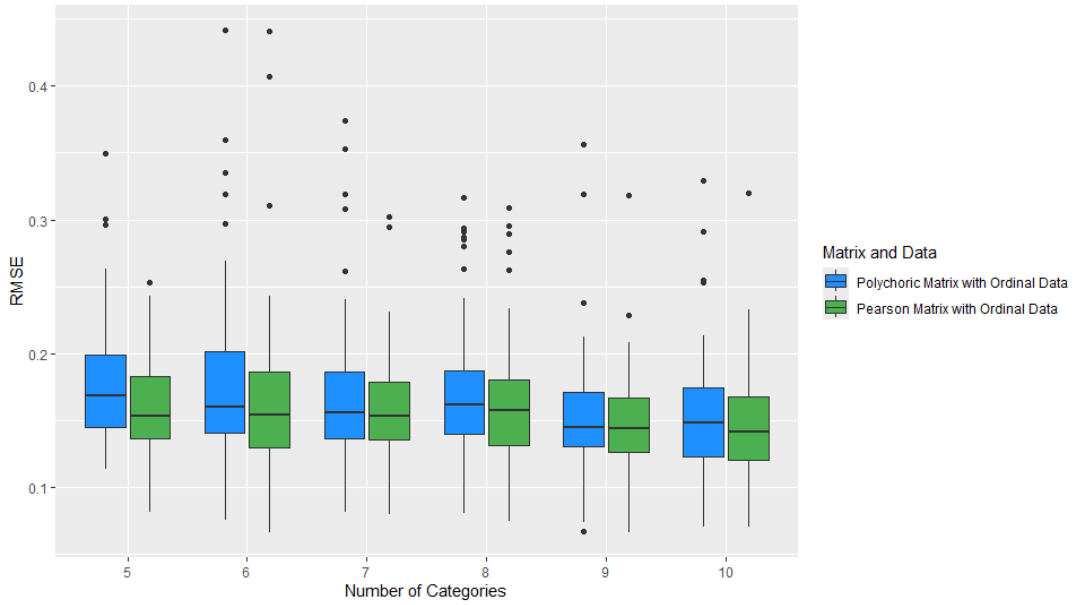

Figure 4.11: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Symmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

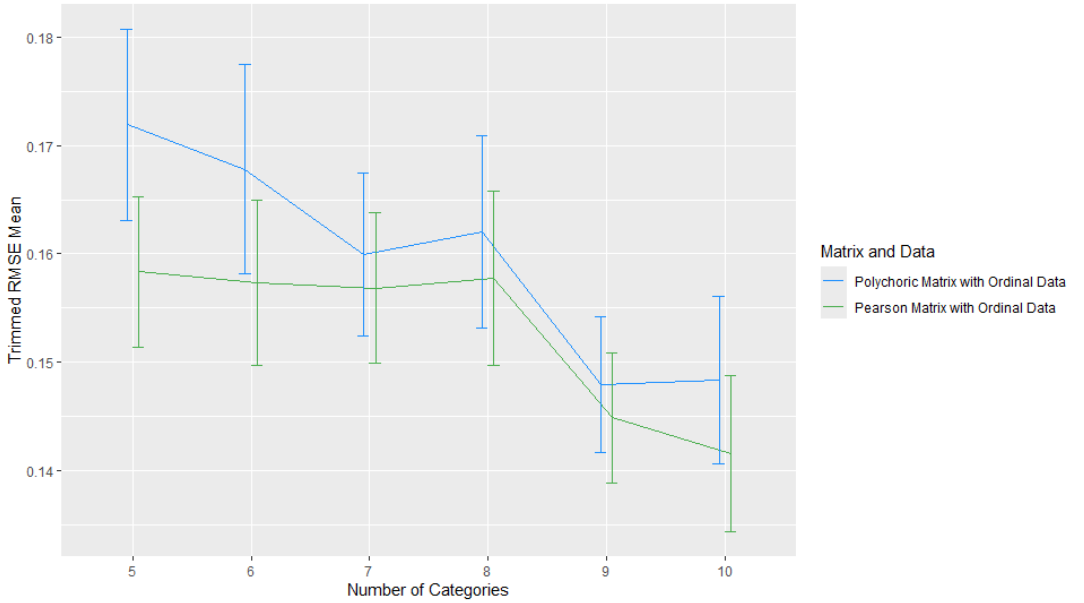

Figure 4.12: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Symmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.4

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 4.463e-15 (denoted by ***), which is well below the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are inconsistent and not homogeneous.

Figure 4.13 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels was 5%. To mitigate the influence of these outliers, a trimming rate of 5% was applied during the robust ANOVA analysis.

The results of the ANOVA, presented in Table 4.6, reveal that the interaction term "Number of Categories:Matrix and Data" is statistically significant, with a p-value of 0.013 (denoted by *). This indicates that the relationship between "Number of Categories" and RMSE in EFA is significantly influenced by "Matrix and Data."

Additionally, the main effect of "Matrix and Data" is highly significant, with a p-value of 0.001 (denoted by ***), demonstrating that the mean RMSE varies significantly across the different matrix and data settings.

In summary, under conditions of positive and negative asymmetry and with three factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—has a significant impact on the accuracy of the EFA results.

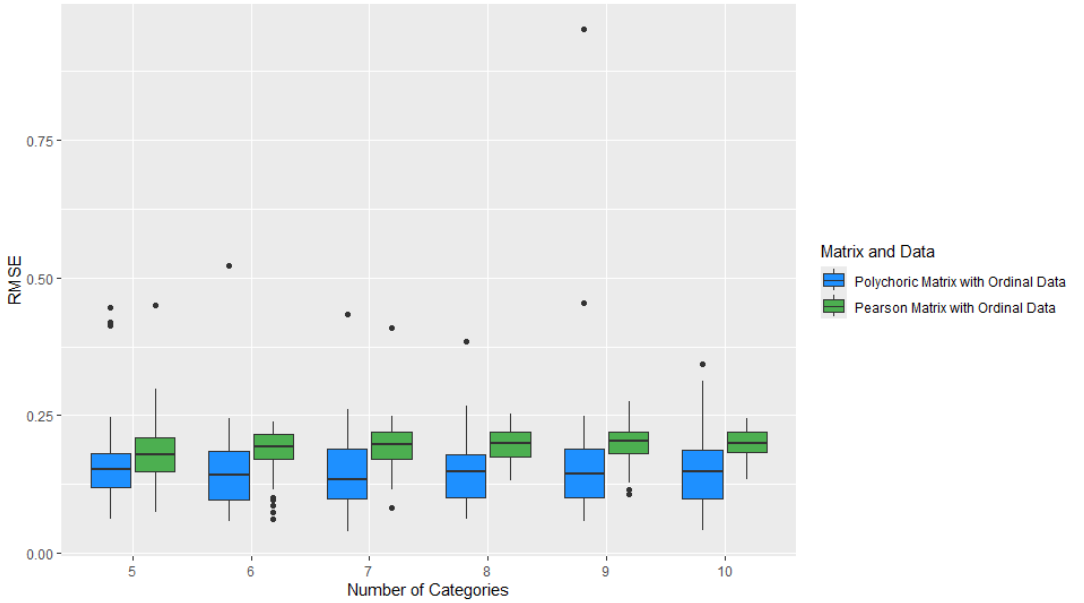

Figure 4.13: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

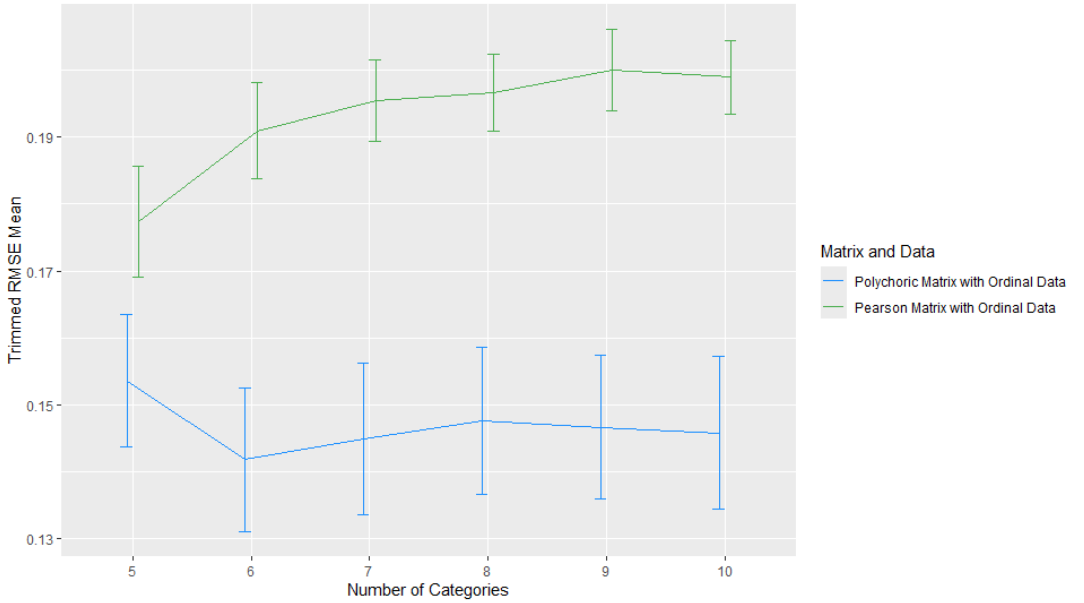

Figure 4.14: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.5

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.8819, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.15 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 12%. To mitigate the impact of these outliers, a trimming rate of 12% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.6, indicate that the interaction term "Number of Categories:Matrix and Data" is statistically significant, with a p-value of 0.047 (denoted by *). This finding suggests that the relationship between "Number of Categories" and RMSE in EFA is significantly influenced by "Matrix and Data."

Furthermore, the main effect of "Matrix and Data" is highly significant, with a p-value of 0.001 (denoted by ***), demonstrating that the mean RMSE differs significantly across the various matrix and data conditions.

In conclusion, under conditions of positive and negative asymmetry and using four factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—significantly affects the accuracy of the EFA results.

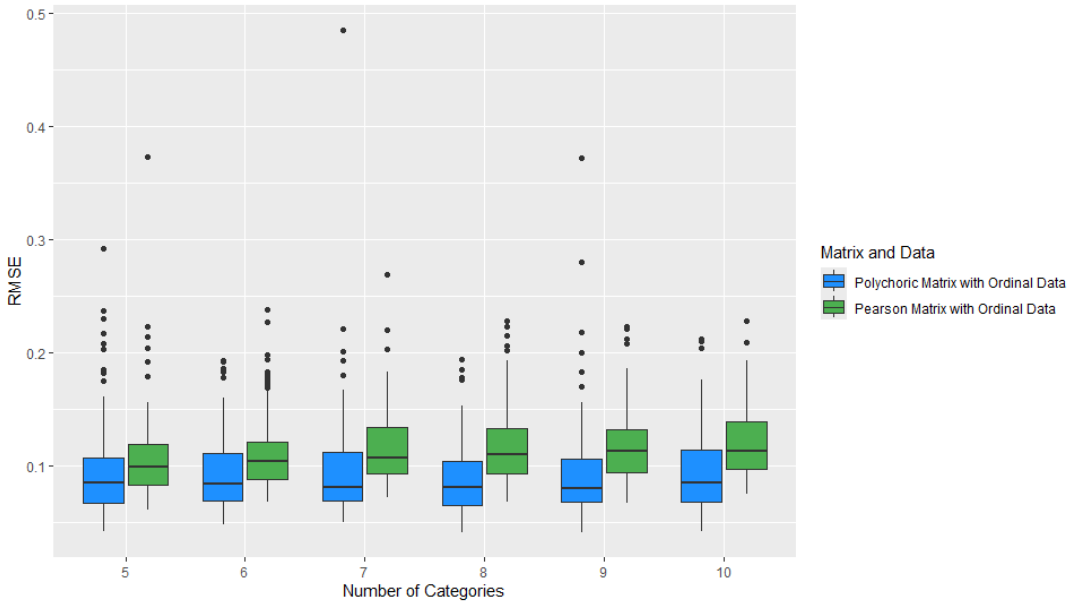

Figure 4.15: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

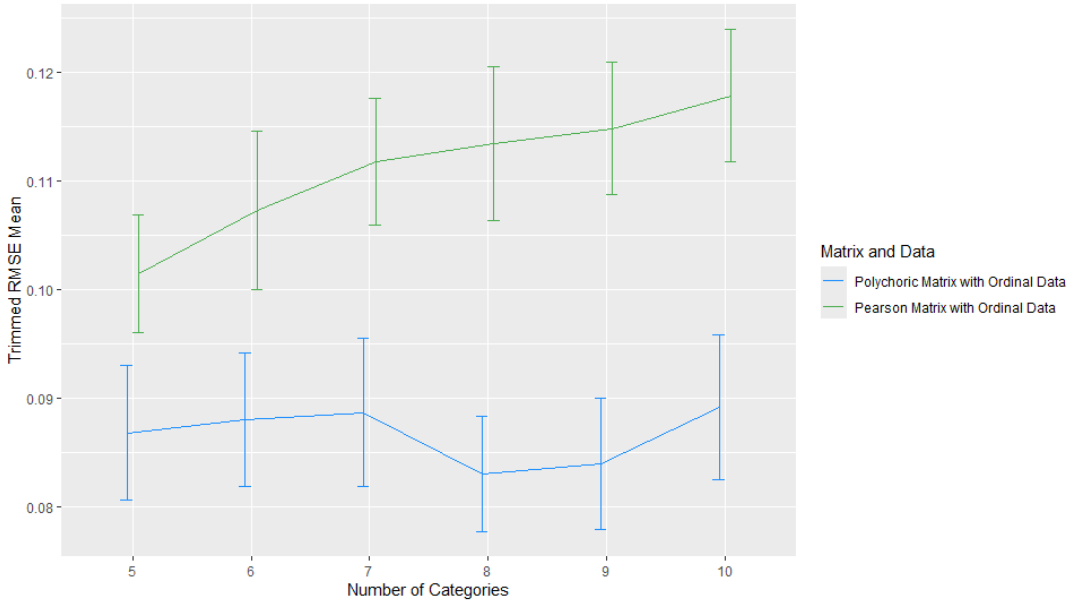

Figure 4.16: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.6

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.3187, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.17 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 11%. To mitigate the impact of these outliers, a trimming rate of 11% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.6, show that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.926. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not substantially influenced by "Matrix and Data."

Additionally, the main effect of "Matrix and Data" also fails to reach statistical significance, with a p-value of 0.927, indicating that the mean RMSE does not differ significantly across the various matrix and data conditions.

In conclusion, under conditions of positive and negative asymmetry and using five factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—does not significantly affect the accuracy of the EFA results.

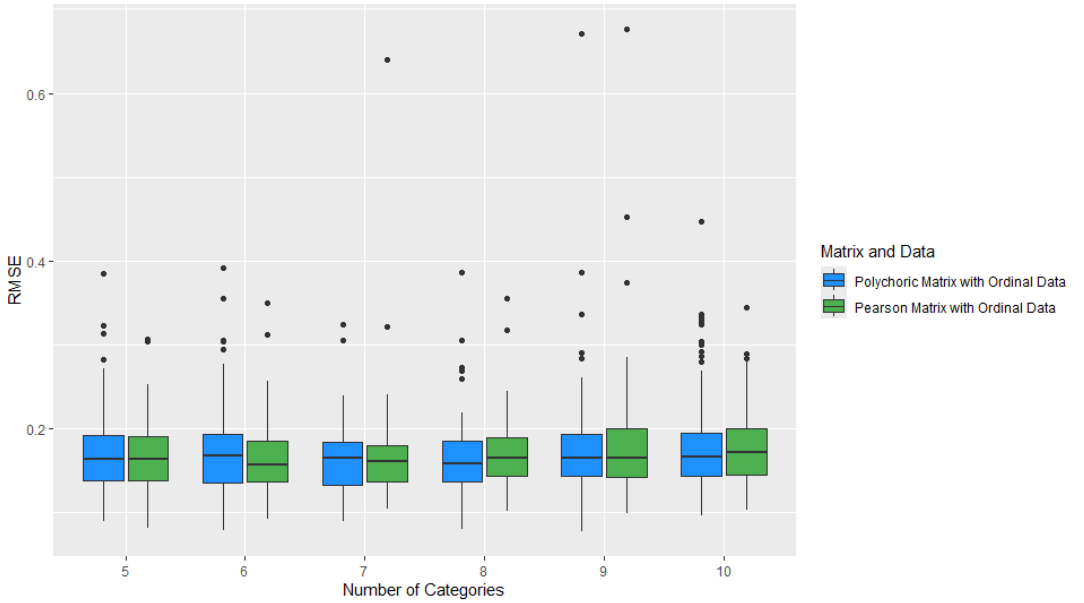

Figure 4.17: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

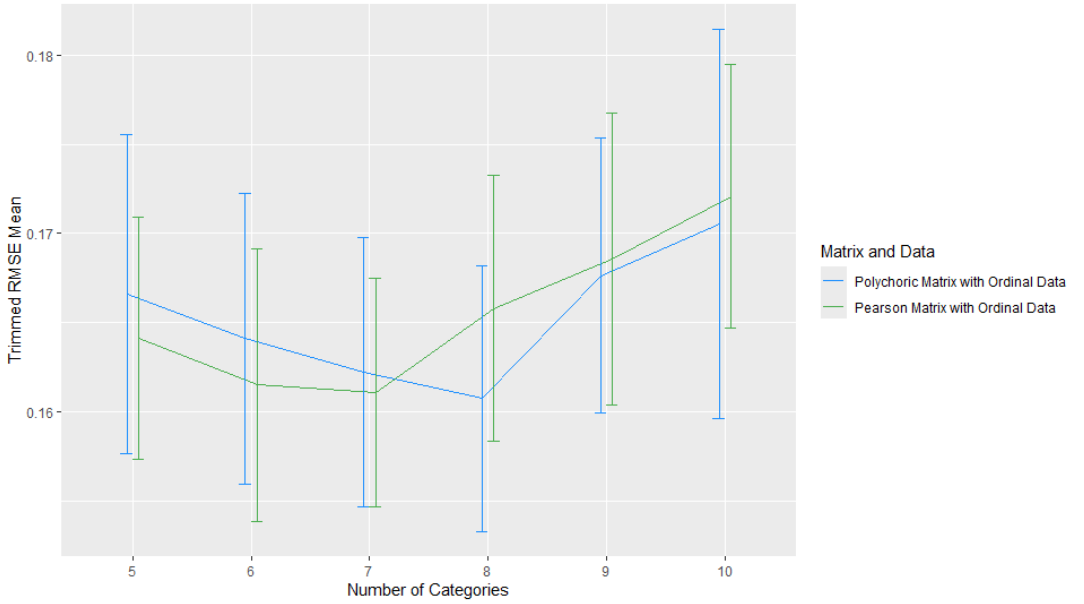

Figure 4.18: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Positive and Negative Asymmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.7

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.4386, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.19 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels was 2%. To

minimize the impact of these outliers, a trimming rate of 2% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.6, show that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.994. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not substantially influenced by "Matrix and Data."

However, the main effect of "Matrix and Data" is statistically significant, with a p-value of 0.037 (denoted by *), indicating that the mean RMSE varies significantly across the different matrix and data conditions.

In conclusion, under conditions of positive asymmetry and with three factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—significantly affects the accuracy of the EFA results.

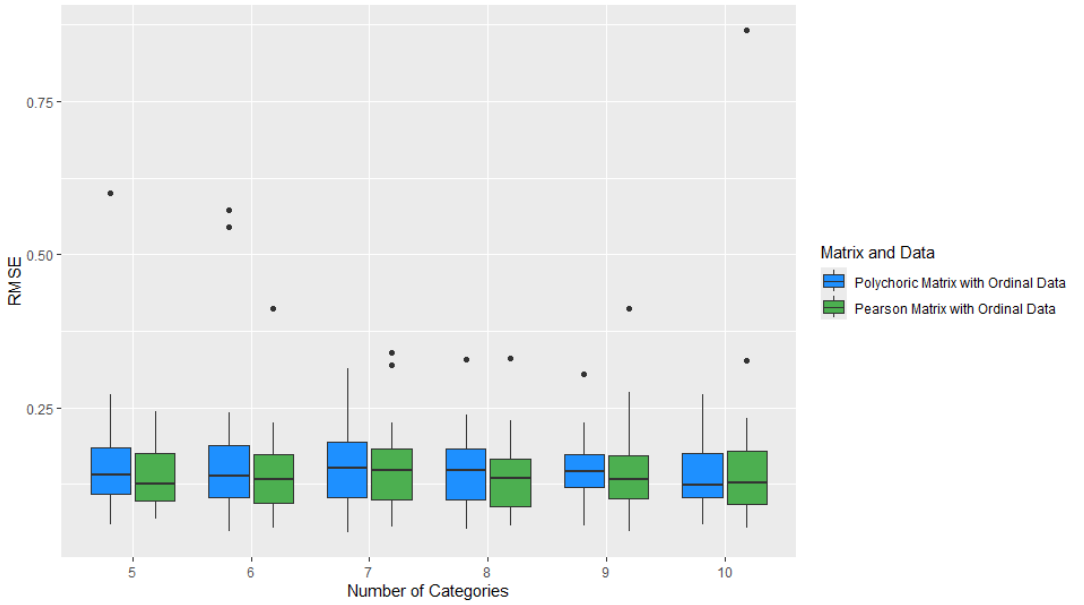

Figure 4.19: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive Asymmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

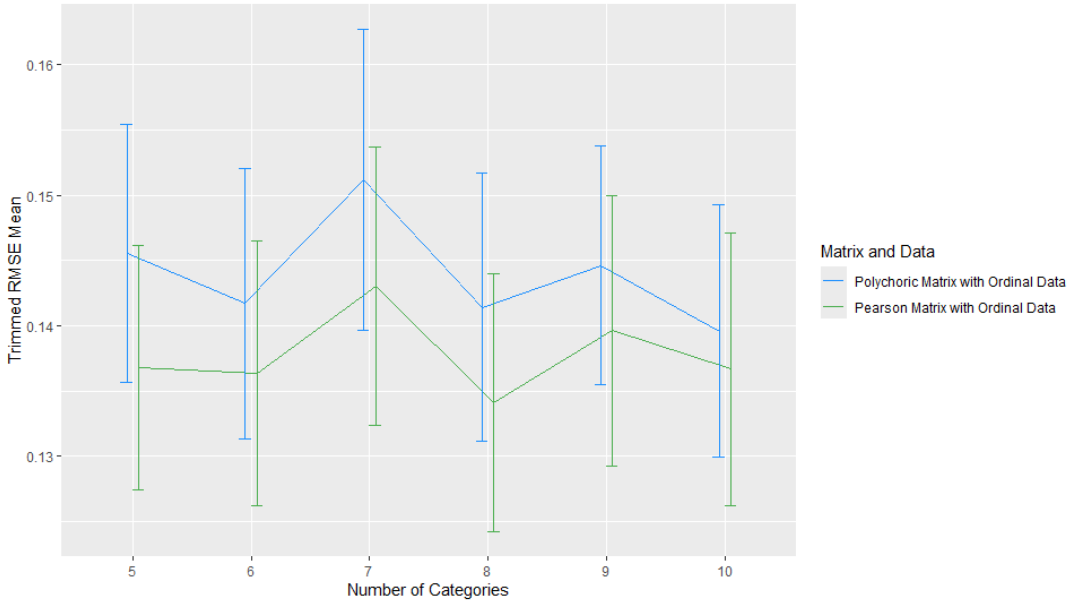

Figure 4.20: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Positive Asymmetry / Factor Number: Three / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.8

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.2906, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.21 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 9%. To mitigate the influence of these outliers, a trimming rate of 9% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.6, show that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.978. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not significantly influenced by "Matrix and Data."

Additionally, the main effect of "Matrix and Data" does not achieve statistical significance, with a p-value of 0.256, indicating that the mean RMSE does not differ significantly across the various matrix and data conditions.

In conclusion, under conditions of positive asymmetry and using four factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—does not significantly affect the accuracy of the EFA results.

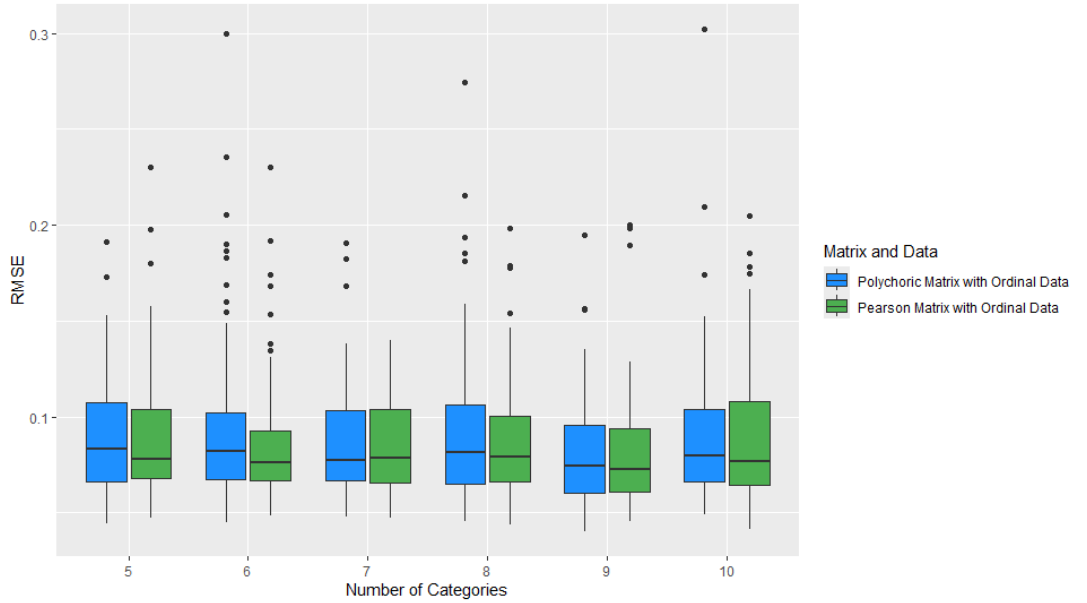

Figure 4.21: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive Asymmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

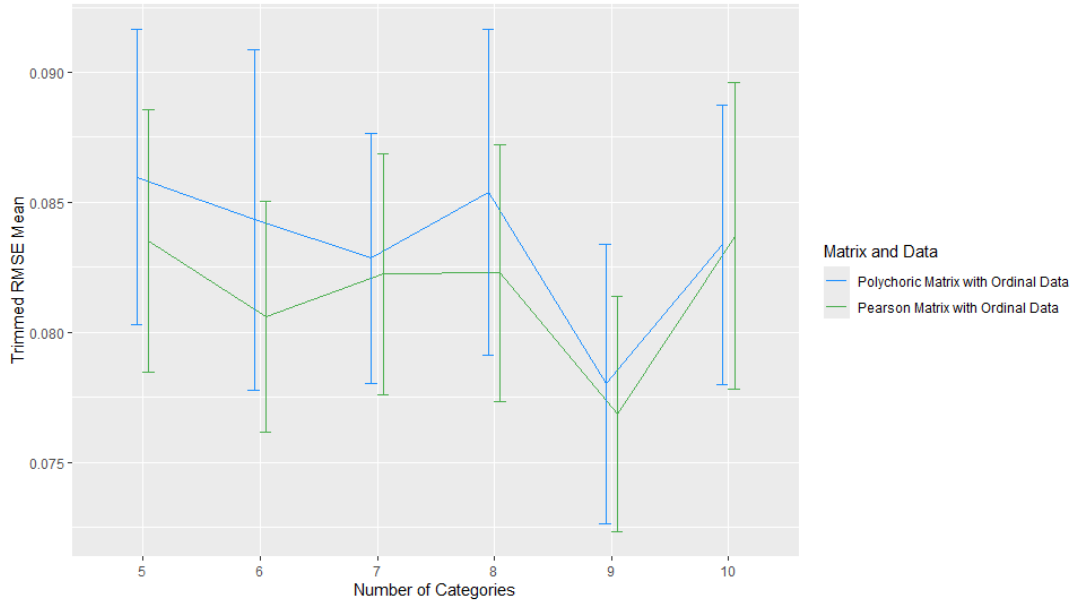

Figure 4.22: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Positive Asymmetry / Factor Number: Four / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

4.2.9

Levene's test for homogeneity of variances, summarized in Table 4.4, yielded a p-value of 0.4804, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.23 illustrates the presence of outliers in the RMSE values, with Table 4.5 showing that the highest proportion of outliers among all combinations of factor levels reached 8%. To

minimize the impact of these outliers, a trimming rate of 8% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.6, show that the interaction term "Number of Categories:Matrix and Data" is not statistically significant, with a p-value of 0.984. This suggests that the relationship between "Number of Categories" and RMSE in EFA is not significantly influenced by "Matrix and Data."

However, the main effect of "Matrix and Data" is statistically significant, with a p-value of 0.002 (denoted by **), indicating that the mean RMSE varies significantly across the different matrix and data conditions.

In conclusion, under conditions of positive asymmetry and using four factors in both the simulation and EFA, the choice of matrix type—whether a polychoric correlation matrix or a Pearson correlation matrix with ordinal data—has a significant impact on the accuracy of the EFA results.

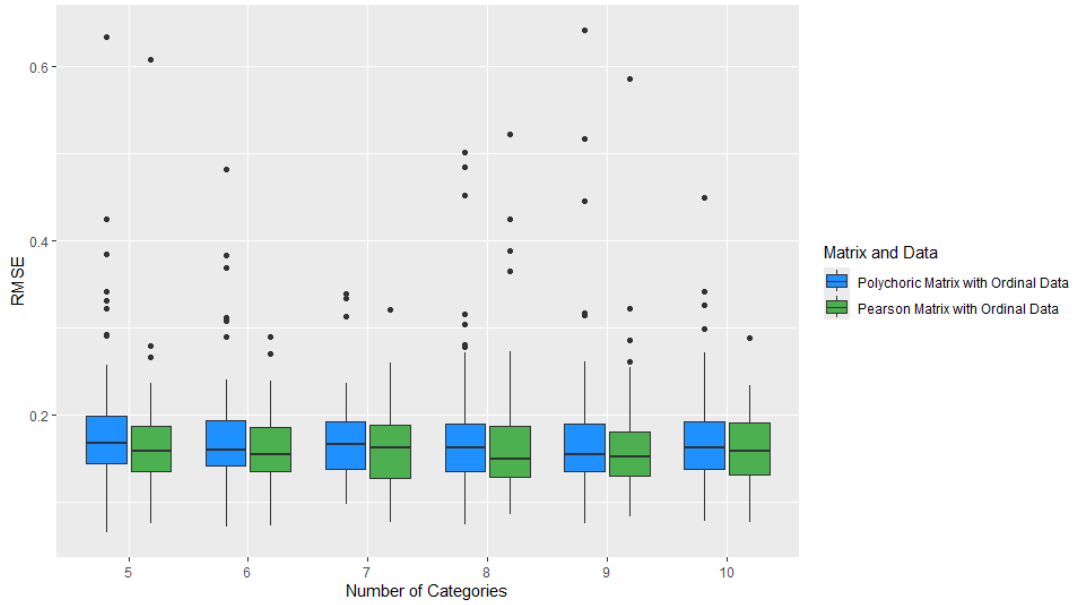

Figure 4.23: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Positive Asymmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

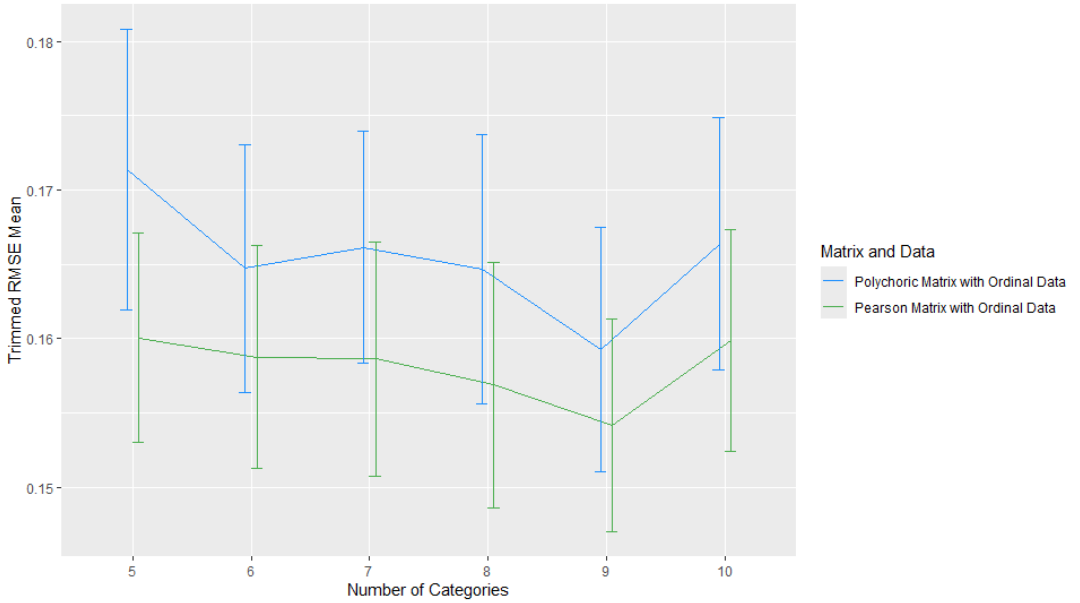

Figure 4.24: RMSE trimmed means with 95% confidence interval (Trimmed Data / Symmetry Condition: Positive Asymmetry / Factor Number: Five / Two Independent Categorical Variables - Number of Categories and Matrix and Data).

Symmetry Condition / Factor Numbers	**Levene's Test (F-value / p-value)**
Symmetry / 3	0.8465 / 0.5933
Symmetry / 4	2.8789 / 0.0009616 ***
Symmetry / 5	1.951 / 0.02992 *
Positive and Negative Asymmetry / 3	8.7518 / 4.463e-15 ***
Positive and Negative Asymmetry / 4	0.5329 / 0.8819
Positive and Negative Asymmetry / 5	1.1493 / 0.3187
Positive Asymmetry / 3	1.006 / 0.4386
Positive Asymmetry / 4	1.1876 / 0.2906
Positive Asymmetry / 5	0.9611 / 0.4804

Table 4.4: Levene's Test for Homogeneity of Variance Summary Table (Two Independent Categorical Variables - Number of Categories and Matrix and Data)

Symmetry Condition / Factor Numbers	Highest Proportion of Outliers
Symmetry / 3	6%
Symmetry / 4	10%
Symmetry / 5	8%
Positive and Negative Asymmetry / 3	5%
Positive and Negative Asymmetry / 4	12%
Positive and Negative Asymmetry / 5	11%
Positive Asymmetry / 3	2%
Positive Asymmetry / 4	9%
Positive Asymmetry / 5	8%

Table 4.5: Highest Proportion of Outliers Summary Table (Two Independent Categorical Variables - Number of Categories and Matrix and Data)

Symmetry Condition / Factor Numbers	Matrix and Data (F-value / p-value)
Symmetry / 3	2.9231 / 0.088
Symmetry / 4	0.3516 / 0.554
Symmetry / 5	9.2255 / 0.003 **
Positive and Negative Asymmetry / 3	309.8543 / 0.001 ***
Positive and Negative Asymmetry / 4	176.4870 / 0.001 ***
Positive and Negative Asymmetry / 5	0.0085 / 0.927
Positive Asymmetry / 3	4.3614 / 0.037 *
Positive Asymmetry / 4	1.2955 / 0.256
Positive Asymmetry / 5	9.6340 / 0.002 **
Symmetry Condition / Factor Numbers	**Interaction Effect (F-value / p-value)**
Symmetry / 3	1.2036 / 0.945
Symmetry / 4	0.1672 / 0.999
Symmetry / 5	3.0960 / 0.689
Positive and Negative Asymmetry / 3	14.9367 / 0.013 *
Positive and Negative Asymmetry / 4	11.4595 / 0.047 *
Positive and Negative Asymmetry / 5	1.3973 / 0.926
Positive Asymmetry / 3	0.4688 / 0.994
Positive Asymmetry / 4	0.7884 / 0.978
Positive Asymmetry / 5	0.6921 / 0.984

Table 4.6: ANOVA Summary Table (Two Independent Categorical Variables - Number of Categories and Matrix and Data)

4.3 EFA performance comparison between different factor numbers in EFA, while the factor number in simulation is 3

4.3.1

Levene's test for homogeneity of variances, summarized in Table 4.7, yielded a p-value of 0.3886, which is greater than the commonly accepted significance threshold of 5%. This indicates that the variances across the subgroups are consistent and homogeneous.

Figure 4.25 illustrates the presence of outliers in the RMSE values, with Table 4.8 showing that the highest proportion of outliers among all combinations of factor levels reached 12%. To mitigate the influence of these outliers, a trimming rate of 12% was applied during the robust ANOVA analysis.

The ANOVA results, presented in Table 4.9, show that the interaction term "Number of Categories:Number of Factors Used in EFA" is statistically significant, with a p-value of 0.001 (denoted by ***). This indicates that the relationship between "Number of Categories" and RMSE in EFA is significantly influenced by the "Number of Factors Used in EFA."

Additionally, the main effect of "Number of Factors Used in EFA" is also statistically significant, with a p-value of 0.001 (denoted by ***), demonstrating that the mean RMSE varies significantly across different settings for the number of factors used in the EFA.

In conclusion, when three factors are used in the simulation and a polychoric correlation matrix is applied to symmetric ordinal data, the alignment between the number of factors used in EFA and the number of factors in the simulation is critical for the accuracy of the EFA results. Specifically, as the discrepancy between these two numbers increases, the error in EFA performance also becomes larger.

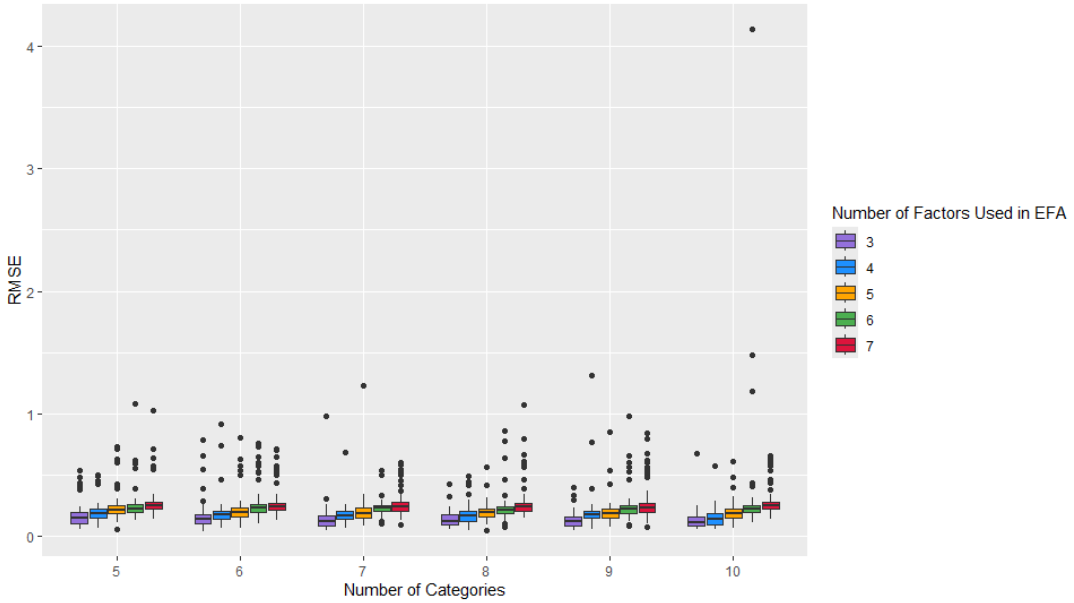

Figure 4.25: Boxplot for visualization of distribution of RMSE after Exploratory Factor Analysis (Raw Data / Symmetry Condition: Symmetry / Factor Number Used in Simulation: Three / Two Independent Categorical Variables - Number of Categories and Number of Factors Used in EFA).

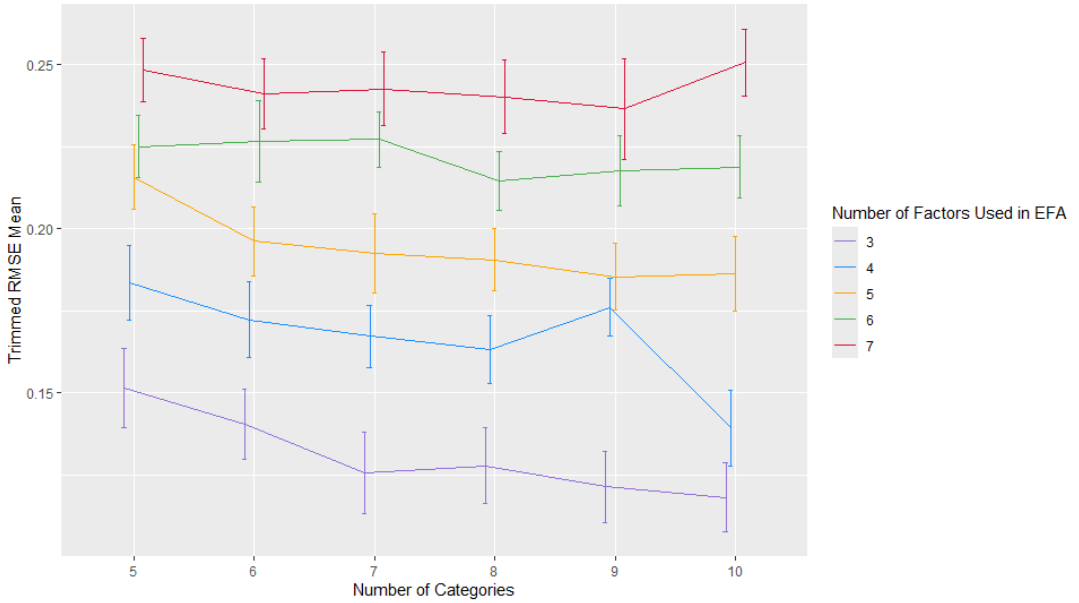

Figure 4.26: RMSE trimmed means with 95% confidence interval (Symmetry Condition: Symmetry / Factor Number Used in Simulation: Three / Two Independent Categorical Variables - Number of Categories and Number of Factors Used in EFA).

Symmetry Condition / Real Factor Numbers	Levene's Test (F-value / p-value)
Symmetry / 3	1.0529 / 0.3886

Table 4.7: Levene's Test for Homogeneity of Variance Summary Table (Two Independent Categorical Variables - Number of Categories and Number of Factors Used in EFA)

Symmetry Condition / Real Factor Numbers	Highest Proportion of Outliers
Symmetry / 3	12%

Table 4.8: Highest Proportion of Outliers Summary Table (Two Independent Categorical Variables - Number of Categories and Number of Factors Used in EFA)

Number of Factors Used in EFA (F-value / p-value)
1462.4580 / 0.001 ***
Number of Categories : Number of Factors Used in EFA (F-value / p-value)
48.2721 / 0.001 ***

Table 4.9: ANOVA Summary Table (Two Independent Categorical Variables - Number of Categories and Number of Factors Used in EFA)

5 Discussion

This chapter consists of three sections. In the results section, we provide possible explanations for the findings and compare them with those from previous studies. The limitations section discusses the constraints related to the methods and data used in the experiments. Finally, the last section offers recommendations for applying EFA in academic research.

5.1 Results

First Research Question

From Section 4.1, we observe consistent differences in EFA results (RMSE) between symmetric and asymmetric ordinal data, regardless of the number of factors used in simulation and EFA. In general, symmetric data consistently outperforms asymmetric data. Notably, all cases involved ordinal data analyzed with the polychoric correlation matrix. This difference arises from the mismatch between the distribution shape of the ordinal data and the original continuous data. Specifically, asymmetric ordinal data fails to accurately reflect the characteristics of the continuous data, introducing bias and leading to poorer performance.

As noted in [7], when three factors are used in simulation and EFA, with five categories and ordinal data analyzed using the polychoric correlation matrix, the same three-factor structures are consistently obtained, irrespective of symmetry. However, our findings suggest that this conclusion cannot be safely generalized to other category numbers, as evident from the changes in RMSE observed in our analysis. While differences in RMSE do not necessarily mean that the final factor structures will differ, they do increase the risk of errors.

Second Research Question

In Section 4.2, the results indicate that most EFA outcomes differ between ordinal data analyzed with Pearson correlation matrix and polychoric correlation matrix, regardless of the symmetry condition. As [7] notes, assigning the same score to values within intervals reduces data variability, introducing bias. On the other hand, [17] states that Pearson correlation matrix can be used for categorical data with at least five ordered categories. Although

our findings differ, this does not invalidate our results. In our study, we examined the effect of the correlation matrix type on loading matrix values. While it is possible to achieve correct structures using Pearson correlation matrix, the higher RMSE values indicate a higher likelihood of generating incorrect structures, requiring repeated attempts to obtain accurate results.

Third Research Question

Finally, Section 4.3 demonstrates that as the gap between the number of factors used in EFA and the actual number of factors in the simulation increases, the accuracy of the loading matrix from EFA decreases. This occurs because a greater number of available factors provides more room for errors during the EFA process.

5.2 Limitations

Several limitations in the study should be acknowledged due to time and resource constraints. First, the rotation method used in EFA was consistently "promax", a widely used oblique rotation method [17]. Although minimal differences are reported among oblique rotation methods, a comparison would be beneficial. Second, the study only considered six category numbers (5, 6, 7, 8, 9, 10) and three factor numbers (3, 4, 5). Third, for asymmetric conditions, only two types were tested: alternating positive and negative asymmetry, and consistently positive asymmetry. It would be valuable to transform specific variables into ordinal data with asymmetry and compare the results. Fourth, for the third research question, only symmetric conditions were analyzed with three factors, leaving other asymmetry conditions unexplored. Lastly, the factor-loading matrix used in the simulations remained fixed, and alternative matrices were not considered.

The study demonstrates replicability and reliability, as the use of simulated multivariate normally distributed data and a series of R functions ensures that other researchers can achieve similar results by following the same methodology and settings. However, certain limitations in validity should be acknowledged.

First, a small number of NA values in the loading matrix generated during the EFA process were replaced with zeros to facilitate comparisons. This adjustment may have introduced minor modifications to the results. Second, RMSE (Root Mean Square Error) was employed to assess the differences between the simulated and obtained loading matrices. While RMSE provides a measure of similarity, it does not directly verify structural equivalence. Similar RMSE values indicate comparable structures, but differing RMSE values do not necessarily suggest fundamentally distinct structures. Finally, to address the issues of outliers and the violation of the assumption of homogeneity of variances, a robust ANOVA method was applied in this thesis. However, this approach correspondingly reduced the statistical power of the analyses.

5.3 The Work in a Wider Context

EFA is a widely used statistical technique for analyzing ordinal data in various research fields. The findings of this study highlight important considerations for questionnaire design, the choice of correlation matrices, and assumptions regarding factor numbers before conducting EFA. By addressing these aspects, researchers are more likely to achieve accurate and reliable experimental results.

5.4 Use of Generative AI Tools

In this thesis, generative AI tools were primarily utilized to correct grammatical errors, paraphrase text, and enhance the academic tone of the writing. Additionally, these tools were employed to generate certain formulas and tables, significantly improving efficiency and saving time during the writing process.

6 Conclusion

In this study, the influences of three key factors on the factor loading matrix derived from EFA were examined. Based on the results, it is evident that the symmetry condition plays a critical role in EFA. Asymmetric ordinal data can introduce bias into the factor loading matrix, highlighting the importance of addressing symmetry in the data. Additionally, the choice of correlation matrix significantly impacts the results. Pearson correlation matrix cannot substitute for the polychoric correlation matrix, as the latter consistently performs better, even when the number of categories exceeds five. Finally, mismatched factor numbers used in EFA can significantly affect the resulting loading matrix, with the magnitude of the impact increasing as the gap between the specified and true factor numbers widens. For researchers aiming to identify latent structures from questionnaires, careful consideration of these three aspects is essential when conducting EFA.

Future studies could explore the following areas. First, the selection of loading matrices and factor correlation matrices should be expanded to include scenarios where no relationships exist between factors. Second, additional rotation methods in EFA could be investigated to evaluate their influence on results. By addressing these points in future research, novel and valuable insights may be uncovered.

Bibliography

[1] *1.3.5.10. Levene Test for Equality of Variances*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm (visited on 01/17/2025).

[2] Alvin C. Burns. *Basic marketing research : using Microsoft Excel data analysis*. eng. Upper Saddle River, N.J. : Pearson Prentice Hall, 2008. ISBN: 978-0-13-205958-9 978-0-13-135421-0. URL: http://archive.org/details/basicmarketingre0000burn (visited on 08/28/2024).

[3] Christine DiStefano. "The Impact of Categorization With Confirmatory Factor Analysis". In: *Structural Equation Modeling: A Multidisciplinary Journal* 9.3 (July 2002). Publisher: Routledge _eprint: https://doi.org/10.1207/S15328007SEM0903_2, pp. 327–346. ISSN: 1070-5511. DOI: 10.1207/S15328007SEM0903_2. URL: https://doi.org/10.1207/S15328007SEM0903_2 (visited on 08/30/2024).

[4] Leandre R. Fabrigar and Duane T. Wegener. "Introductory Factor Analysis Concepts". In: *Exploratory Factor Analysis*. Ed. by Leandre R. Fabrigar and Duane T. Wegener. Oxford University Press, Dec. 2011, p. 0. ISBN: 978-0-19-973417-7. DOI: 10.1093/acprof:osobl/9780199734177.003.0001. URL: https://doi.org/10.1093/acprof:osobl/9780199734177.003.0001 (visited on 09/17/2024).

[5] Leandre R. Fabrigar and Duane T. Wegener. "Requirements and Decisions for Implementing Exploratory Common Factor Analysis". In: *Exploratory Factor Analysis*. Ed. by Leandre R. Fabrigar and Duane T. Wegener. Oxford University Press, Dec. 2011, p. 0. ISBN: 978-0-19-973417-7. DOI: 10.1093/acprof:osobl/9780199734177.003.0003. URL: https://doi.org/10.1093/acprof:osobl/9780199734177.003.0003 (visited on 12/25/2024).

[6] Mark C. Greenwood. *Chapter 4 Two-Way ANOVA | Intermediate Statistics with R*. URL: https://greenwood-stat.github.io/GreenwoodBookHTML/chapter4.html (visited on 11/19/2024).

[7] Francisco Pablo Holgado–Tello, Salvador Chacón–Moscoso, Isabel Barbero–García, and Enrique Vila–Abad. "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables". en. In: *Quality & Quantity* 44.1 (Jan. 2010), pp. 153–166. ISSN: 1573-7845. DOI: 10.1007/s11135-008-9190-y. URL: https://doi.org/10.1007/s11135-008-9190-y (visited on 08/30/2024).

[8] Mosala Phillip Lesia, Clinton O. Aigbavboa, and Wellington D. Thwala. "Factors influencing residential location choice in South Africa: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA)". en. In: *Journal of Housing and the Built Environment* 39.1 (Mar. 2024), pp. 133–160. ISSN: 1573-7772. DOI: 10.1007/s10901-023-10070-w. URL: https://doi.org/10.1007/s10901-023-10070-w (visited on 08/30/2024).

[9] João Marôco. "Factor Analysis of Ordinal Items: Old Questions, Modern Solutions?" en. In: *Stats* 7.3 (Sept. 2024). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, pp. 984–1001. ISSN: 2571-905X. DOI: 10.3390/stats7030060. URL: https://www.mdpi.com/2571-905X/7/3/60 (visited on 09/25/2024).

[10] Ulf Olsson. "Maximum likelihood estimation of the polychoric correlation coefficient". en. In: *Psychometrika* 44.4 (Dec. 1979), pp. 443–460. ISSN: 1860-0980. DOI: 10.1007/BF02296207. URL: https://doi.org/10.1007/BF02296207 (visited on 12/22/2024).

[11] Frank Schoonjans. *Calculation of Trimmed Mean, SE and confidence interval*. en. URL: https://www.medcalc.org/manual/note-trimmedmean.php (visited on 01/18/2025).

[12] Barbara G. Tabachnick and Linda S. Fidell. *Using multivariate statistics*. en. Seventh edition. Always learning. New York, NY: Pearson, 2019. ISBN: 978-0-13-479054-1.

[13] Thurstone L. L. *Multiple -factor Analysis A Development and Expansion Of The Vectors Of Mind*. eng. The University Of Chicago Press, Chicago, Illinois, 1947. URL: http://archive.org/details/dli.ernet.18325 (visited on 12/25/2024).

[14] *Trimmed Mean Standard Error*. URL: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/trimmse.htm?utm_source=chatgpt.com (visited on 01/18/2025).

[15] Shaun Turney. *Pearson Correlation Coefficient (r) | Guide & Examples*. en-US. May 2022. URL: https://www.scribbr.com/statistics/pearson-correlation-coefficient/ (visited on 10/08/2024).

[16] *Two-way ANOVA in R*. en. URL: https://statsandr.com/blog/two-way-anova-in-r/ (visited on 11/19/2024).

[17] Marley W. Watkins. *A step-by-step guide to exploratory factor analysis with R and Rstudio*. eng. 1st ed. New York, NY: Routledge, 2021. ISBN: 978-1-00-312000-1.

[18] Rand R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Chantilly, UNITED STATES: Elsevier Science & Technology, 2012. ISBN: 978-0-12-387015-5. URL: http://ebookcentral.proquest.com/lib/linkoping-ebooks/detail.action?docID=858695 (visited on 01/18/2025).

7 Appendix

Number of Categories	Symmetry Condition	Winsorized Var	Trimmed RMSE Mean
5	Symmetry	0.00268	0.152
5	Positive and Negative Asymmetry	0.00196	0.153
5	Positive Asymmetry	0.00204	0.145
6	Symmetry	0.00249	0.142
6	Positive and Negative Asymmetry	0.00240	0.142
6	Positive Asymmetry	0.00237	0.141
7	Symmetry	0.00273	0.128
7	Positive and Negative Asymmetry	0.00263	0.145
7	Positive Asymmetry	0.00272	0.150
8	Symmetry	0.00224	0.130
8	Positive and Negative Asymmetry	0.00240	0.148
8	Positive Asymmetry	0.00245	0.141
9	Symmetry	0.00240	0.124
9	Positive and Negative Asymmetry	0.00241	0.147
9	Positive Asymmetry	0.00189	0.145
10	Symmetry	0.00221	0.120
10	Positive and Negative Asymmetry	0.00266	0.146
10	Positive Asymmetry	0.00213	0.139

Table 7.1: Summary of Number of Categories, Symmetry Condition, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Symmetry Condition	Winsorized Var	Trimmed RMSE Mean
5	Symmetry	0.000703	0.0964
5	Positive and Negative Asymmetry	0.000965	0.0881
5	Positive Asymmetry	0.000567	0.0860
6	Symmetry	0.000618	0.0819
6	Positive and Negative Asymmetry	0.000784	0.0889
6	Positive Asymmetry	0.000753	0.0843
7	Symmetry	0.000567	0.0774
7	Positive and Negative Asymmetry	0.000909	0.0898
7	Positive Asymmetry	0.000405	0.0828
8	Symmetry	0.000399	0.0718
8	Positive and Negative Asymmetry	0.000546	0.0834
8	Positive Asymmetry	0.000687	0.0854
9	Symmetry	0.000302	0.0680
9	Positive and Negative Asymmetry	0.000764	0.0847
9	Positive Asymmetry	0.000506	0.0780
10	Symmetry	0.000398	0.0696
10	Positive and Negative Asymmetry	0.000925	0.0898
10	Positive Asymmetry	0.000507	0.0834

Table 7.2: Summary of Number of Categories, Symmetry Condition, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Symmetry Condition	Winsorized Var	Trimmed RMSE Mean
5	Symmetry	0.00111	0.171
5	Positive and Negative Asymmetry	0.00127	0.167
5	Positive Asymmetry	0.00121	0.170
6	Symmetry	0.00145	0.167
6	Positive and Negative Asymmetry	0.00105	0.164
6	Positive Asymmetry	0.00094	0.164
7	Symmetry	0.00088	0.160
7	Positive and Negative Asymmetry	0.00091	0.162
7	Positive Asymmetry	0.00095	0.166
8	Symmetry	0.00106	0.161
8	Positive and Negative Asymmetry	0.00088	0.161
8	Positive Asymmetry	0.00117	0.164
9	Symmetry	0.00062	0.148
9	Positive and Negative Asymmetry	0.00094	0.168
9	Positive Asymmetry	0.00115	0.159
10	Symmetry	0.00094	0.149
10	Positive and Negative Asymmetry	0.00189	0.171
10	Positive Asymmetry	0.00109	0.166

Table 7.3: Summary of Number of Categories, Symmetry Condition, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00268	0.152
5	Pearson Matrix with Ordinal Data	0.00239	0.143
6	Polychoric Matrix with Ordinal Data	0.00249	0.142
6	Pearson Matrix with Ordinal Data	0.00214	0.134
7	Polychoric Matrix with Ordinal Data	0.00273	0.128
7	Pearson Matrix with Ordinal Data	0.00220	0.120
8	Polychoric Matrix with Ordinal Data	0.00224	0.130
8	Pearson Matrix with Ordinal Data	0.00223	0.125
9	Polychoric Matrix with Ordinal Data	0.00240	0.124
9	Pearson Matrix with Ordinal Data	0.00264	0.124
10	Polychoric Matrix with Ordinal Data	0.00221	0.120
10	Pearson Matrix with Ordinal Data	0.00215	0.117

Table 7.4: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.000698	0.0963
5	Pearson Matrix with Ordinal Data	0.000502	0.0951
6	Polychoric Matrix with Ordinal Data	0.000593	0.0816
6	Pearson Matrix with Ordinal Data	0.000439	0.0803
7	Polychoric Matrix with Ordinal Data	0.000508	0.0771
7	Pearson Matrix with Ordinal Data	0.000395	0.0762
8	Polychoric Matrix with Ordinal Data	0.000378	0.0715
8	Pearson Matrix with Ordinal Data	0.000310	0.0700
9	Polychoric Matrix with Ordinal Data	0.000296	0.0678
9	Pearson Matrix with Ordinal Data	0.000277	0.0677
10	Polychoric Matrix with Ordinal Data	0.000394	0.0694
10	Pearson Matrix with Ordinal Data	0.000375	0.0692

Table 7.5: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00143	0.172
5	Pearson Matrix with Ordinal Data	0.000895	0.158
6	Polychoric Matrix with Ordinal Data	0.00172	0.168
6	Pearson Matrix with Ordinal Data	0.00107	0.157
7	Polychoric Matrix with Ordinal Data	0.00104	0.160
7	Pearson Matrix with Ordinal Data	0.000878	0.157
8	Polychoric Matrix with Ordinal Data	0.00145	0.162
8	Pearson Matrix with Ordinal Data	0.00119	0.158
9	Polychoric Matrix with Ordinal Data	0.000714	0.148
9	Pearson Matrix with Ordinal Data	0.000669	0.145
10	Polychoric Matrix with Ordinal Data	0.00110	0.148
10	Pearson Matrix with Ordinal Data	0.000952	0.142

Table 7.6: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00207	0.154
5	Pearson Matrix with Ordinal Data	0.00143	0.177
6	Polychoric Matrix with Ordinal Data	0.00242	0.142
6	Pearson Matrix with Ordinal Data	0.00109	0.191
7	Polychoric Matrix with Ordinal Data	0.00273	0.145
7	Pearson Matrix with Ordinal Data	0.000779	0.196
8	Polychoric Matrix with Ordinal Data	0.00252	0.148
8	Pearson Matrix with Ordinal Data	0.000707	0.197
9	Polychoric Matrix with Ordinal Data	0.00243	0.147
9	Pearson Matrix with Ordinal Data	0.000785	0.200
10	Polychoric Matrix with Ordinal Data	0.00275	0.146
10	Pearson Matrix with Ordinal Data	0.000644	0.199

Table 7.7: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.000586	0.0868
5	Pearson Matrix with Ordinal Data	0.000442	0.101
6	Polychoric Matrix with Ordinal Data	0.000569	0.0879
6	Pearson Matrix with Ordinal Data	0.000799	0.107
7	Polychoric Matrix with Ordinal Data	0.000699	0.0887
7	Pearson Matrix with Ordinal Data	0.000511	0.112
8	Polychoric Matrix with Ordinal Data	0.000423	0.0830
8	Pearson Matrix with Ordinal Data	0.000754	0.113
9	Polychoric Matrix with Ordinal Data	0.000552	0.0840
9	Pearson Matrix with Ordinal Data	0.000557	0.115
10	Polychoric Matrix with Ordinal Data	0.000671	0.0891
10	Pearson Matrix with Ordinal Data	0.000560	0.118

Table 7.8: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00127	0.167
5	Pearson Matrix with Ordinal Data	0.000733	0.164
6	Polychoric Matrix with Ordinal Data	0.00105	0.164
6	Pearson Matrix with Ordinal Data	0.000926	0.162
7	Polychoric Matrix with Ordinal Data	0.000909	0.162
7	Pearson Matrix with Ordinal Data	0.000646	0.161
8	Polychoric Matrix with Ordinal Data	0.000881	0.161
8	Pearson Matrix with Ordinal Data	0.000880	0.166
9	Polychoric Matrix with Ordinal Data	0.000939	0.168
9	Pearson Matrix with Ordinal Data	0.00106	0.169
10	Polychoric Matrix with Ordinal Data	0.00189	0.171
10	Pearson Matrix with Ordinal Data	0.000859	0.172

Table 7.9: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00234	0.146
5	Pearson Matrix with Ordinal Data	0.00210	0.137
6	Polychoric Matrix with Ordinal Data	0.00259	0.142
6	Pearson Matrix with Ordinal Data	0.00246	0.136
7	Polychoric Matrix with Ordinal Data	0.00318	0.151
7	Pearson Matrix with Ordinal Data	0.00272	0.143
8	Polychoric Matrix with Ordinal Data	0.00252	0.141
8	Pearson Matrix with Ordinal Data	0.00234	0.134
9	Polychoric Matrix with Ordinal Data	0.00201	0.145
9	Pearson Matrix with Ordinal Data	0.00259	0.140
10	Polychoric Matrix with Ordinal Data	0.00224	0.140
10	Pearson Matrix with Ordinal Data	0.00261	0.137

Table 7.10: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.000567	0.0860
5	Pearson Matrix with Ordinal Data	0.000445	0.0835
6	Polychoric Matrix with Ordinal Data	0.000753	0.0843
6	Pearson Matrix with Ordinal Data	0.000346	0.0806
7	Polychoric Matrix with Ordinal Data	0.000405	0.0828
7	Pearson Matrix with Ordinal Data	0.000377	0.0822
8	Polychoric Matrix with Ordinal Data	0.000687	0.0854
8	Pearson Matrix with Ordinal Data	0.000428	0.0823
9	Polychoric Matrix with Ordinal Data	0.000506	0.0780
9	Pearson Matrix with Ordinal Data	0.000359	0.0769
10	Polychoric Matrix with Ordinal Data	0.000507	0.0834
10	Pearson Matrix with Ordinal Data	0.000607	0.0837

Table 7.11: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Matrix and Data	Winsorized Var	Trimmed RMSE Mean
5	Polychoric Matrix with Ordinal Data	0.00163	0.171
5	Pearson Matrix with Ordinal Data	0.000912	0.160
6	Polychoric Matrix with Ordinal Data	0.00128	0.165
6	Pearson Matrix with Ordinal Data	0.00103	0.159
7	Polychoric Matrix with Ordinal Data	0.00112	0.166
7	Pearson Matrix with Ordinal Data	0.00114	0.159
8	Polychoric Matrix with Ordinal Data	0.00151	0.165
8	Pearson Matrix with Ordinal Data	0.00125	0.157
9	Polychoric Matrix with Ordinal Data	0.00125	0.159
9	Pearson Matrix with Ordinal Data	0.000934	0.154
10	Polychoric Matrix with Ordinal Data	0.00131	0.166
10	Pearson Matrix with Ordinal Data	0.00102	0.160

Table 7.12: Summary of Number of Categories, Matrix and Data, Winsorized Var, and Trimmed RMSE Mean

Number of Categories	Number of Factors Used in EFA	Winsorized Var	Trimmed RMSE Mean
5	3	0.00220	0.152
5	4	0.00194	0.184
5	5	0.00145	0.216
5	6	0.00138	0.225
5	7	0.00144	0.248
6	3	0.00171	0.140
6	4	0.00202	0.172
6	5	0.00169	0.196
6	6	0.00232	0.227
6	7	0.00169	0.241

Table 7.13: Summary of Number of Categories, Number of Factors Used in EFA, Winsorized Var, and Trimmed RMSE Mean