NEW ERA AGRICULTURE MAGAZINE

# FACTOR ANALYSIS
## A Comprehensive Methodological Guide with Application

Pooja Aditya[1*], Simadri Rajasri[2] and Sanasam Angousana[3]

**Abstract: -**

*Factor Analysis (FA) is a powerful multivariate statistical technique used to discover the underlying latent structure in a set of observed variables by reducing dimensionality while preserving the essential information. This article offers a systematic and comprehensive treatment of Factor Analysis covering its historical development, theoretical foundations, assumptions, extraction and rotation methods, and model adequacy measures. Both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are discussed in depth. Special emphasis is placed on agricultural applications: crop yield factor identification, soil fertility parameter reduction, farmer adoption behaviour studies, socio-economic index construction, and climate variable grouping. Software implementation in R, SPSS, SAS, and Python is outlined. Forty peer-reviewed APA citations are provided.*

**Keywords:** Factor Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis, Dimension Reduction, Latent Variables, KMO, Bartlett's Test, Varimax Rotation.

## 1. Introduction:

### 1.1 Concept of Factor Analysis

Factor analysis has its origins in the early 1900s with Charles Spearman's interest in human ability and his development of the Two-Factor Theory; this eventually led to a burgeoning of work on the theories and mathematical principles of factor analysis (Harman, 1976) [9]. The method involved using simulated data where the answers were already known to test factor analysis (Child, 2006). The technological improvements of computers have led to the utilisation of factor

***Pooja Aditya[1*], Simadri Rajasri[2] and Sanasam Angousana[3]***
*[1*]PhD Research Scholar, Department of Agricultural Statistics,*
*[2]PhD Research Scholar, Department of Agricultural Extention,*
*[3]PhD Research Scholar, Department of Pomology and Post Harvest Technology,*
***Uttar Banga Krishi Vishwavidyalaya, Cooch Behar, West Bengal, India***

analysis in a variety of sectors, including the behavioural and social sciences, medicine, economics, and geography, among others.

The two main factor analysis techniques are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). CFA attempts to confirm hypotheses and uses path analysis diagrams to represent variables and factors, whereas EFA tries to uncover complex patterns by exploring the dataset and testing predictions (Child, 2006). Factor analysis is a key technique in multivariate statistics that reduces a large number of correlated variables into a smaller set of meaningful latent factors, thereby simplifying complex datasets and improving interpretation. The method was first introduced by Charles Spearman (1904) [3] through the concept of the general intelligence factor and was later refined by Louis Leon Thurstone (1931) [17] with the theory of multiple factors. It is widely used for dimension reduction in research to eliminate redundancy, reduce multicollinearity, and enhance model efficiency.

Mathematically, FA decomposes the covariance or correlation matrix Σ of p observed variables into components attributable to m < p common factors and unique factors. For an observation vector $X = (X_1, X_2, \ldots, X_p)^T$, the common factor model states:

$$X_i = \lambda_{i1}F_1 + \lambda_{i2}F_2 + \ldots + \lambda_{im}F_m + \varepsilon_i \quad (Eq.\ 1)$$

where $\lambda_{ij}$ are factor loadings quantifying the relationship between variable i and factor j, $F_1 \ldots F_m$ are the common factors, and $\varepsilon_i$ is the unique factor or specific error [5,6].

**1.2 Importance in Multivariate Statistics**

In high-dimensional datasets, traditional univariate methods fail to capture the complex inter-relationships among variables [7]. Factor Analysis addresses this by reducing the dimensionality of data without major information loss, revealing latent constructs that are not directly measurable, improving interpretability and enabling parsimonious modelling, and providing composite variables (factor scores) for subsequent analyses [8,9].

**1.3 Need for Dimension Reduction in Research**

High-dimensional data introduces the curse of dimensionality as dimensions increase, data become sparse, parameter estimates become unreliable, and overfitting occurs [10]. In survey research, experimental designs, and observational studies, researchers routinely collect dozens of correlated variables. Dimension reduction through FA enables concise modelling, reduces multicollinearity, and facilitates clearer scientific communication [11,12].

**1.4 Applications Across Disciplines**

Factor Analysis finds application across many domains [13-16]:

- Psychology & Education: Intelligence testing, personality assessment, scale development
- Social Sciences: Socio-economic status indices, quality of life measures
- Marketing: Brand perception, customer satisfaction studies
- Finance: Portfolio risk factors, credit scoring models
- Health Sciences: Quality of life instruments, symptom clustering
- Agriculture: Crop yield factor identification, soil fertility parameter reduction, farmer adoption behaviour studies, socio-economic index construction, and climate variable grouping

## 2. THEORETICAL BACKGROUND

### 2.1 History and Development

Factor Analysis traces its origins to Charles Spearman's 1904 seminal work on intelligence, where he proposed a general factor (g) underlying performance on diverse cognitive tests [3]. Spearman's two-factor theory distinguished between general (g) and specific (s) factors. His use of tetrachoric correlations and rank-order methods laid the groundwork for modern FA.

Louis Leon Thurstone (1931, 1947) [17,18] revolutionised the field by introducing multiple-factor analysis and the concept of simple structure---the criterion that each variable should load highly on as few factors as possible. Thurstone's Primary Mental Abilities (PMAs) theory challenged the unidimensional view and demonstrated that intelligence comprises multiple independent factors.

The advent of electronic computing in the 1960s-70s made FA computationally feasible for large datasets. Lawley and Maxwell (1963) [19] introduced maximum likelihood estimation, and Joreskog (1969) [20] developed LISREL, paving the way for Confirmatory Factor Analysis (CFA) and Structural Equation Modelling (SEM).

### 2.2 Concept of Latent Variables

A latent variable is a theoretical construct that cannot be directly measured but is inferred from a set of observed (manifest) variables [21]. Examples include: intelligence (inferred from test scores), anxiety (inferred from questionnaire items), and socio-economic status (inferred from income, education, occupation). Latent variable models provide a principled framework for operationalising abstract constructs [22].

### 2.3 Correlation and Covariance Structure

FA models the correlation/covariance matrix Σ of observed variables. Under the common factor model with m factors:

$$\Sigma = \Lambda\Phi\Lambda^T + \Psi \quad (Eq.\ 2)$$

where Λ (p×m) is the factor loading matrix, Φ is the factor correlation matrix (identity under orthogonal rotation), and Ψ is the diagonal matrix of unique variances [5]. The communality h² of variable i is the proportion of its variance explained by the common factors:

$$h^2_i = \lambda^2_{i1} + \lambda^2_{i2} + \ldots + \lambda^2_{im} \quad (Eq.\ 3)$$

## 3. TYPES OF FACTOR ANALYSIS

### 3.1 Exploratory Factor Analysis (EFA)

EFA is used when the researcher has no prior hypothesis about the number or structure of factors. It explores the data empirically to uncover latent structure [23]. EFA is appropriate in early stages of scale development or when theoretical frameworks are incomplete.

### 3.2 Confirmatory Factor Analysis (CFA)

CFA is a hypothesis-driven approach used to test whether a specified factor structure derived from theory or prior EFA fits the observed data [24]. CFA requires a priori specification of which variables load on which factors. Model fit is evaluated using indices such as RMSEA, CFI, TLI, and SRMR.

### 3.3 Principal Component Analysis (PCA) vs. Factor Analysis

PCA and FA are often confused. PCA is a data reduction technique that transforms variables into orthogonal components explaining maximum variance. FA is a structural model positing latent factors as causal influences on observed variables [25]. In PCA, all variance (including unique variance) is analysed; in FA, only common (shared) variance is modelled.

## 4. ASSUMPTIONS OF FACTOR ANALYSIS

Valid FA application requires satisfaction of the following assumptions [26,27]:

### 4.1 Linearity

FA assumes linear relationships among variables. Non-linear associations may cause factors to reflect artefacts rather than true structure. Linearity is assessed via scatter plot matrices or Box-Cox transformations [26].

### 4.2 Adequate Sample Size

FA is sensitive to sample size. Common rules: minimum 100-200

**Table 1. Comparison of Exploratory and Confirmatory Factor Analysis.**

| Criterion | Exploratory FA | Confirmatory FA |
|---|---|---|
| **Prior hypothesis** | Not required | Required (theory-driven) |
| **Factor-variable links** | Data-determined | Specified a priori |
| **Number of factors** | Determined empirically | Specified in advance |
| **Model fit indices** | Not applicable | RMSEA, CFI, TLI, SRMR |
| **Software** | R (psych), SPSS | R (lavaan), AMOS, Mplus |
| **Use case** | Scale development | Theory testing |

observations, subject-to-variable ratio ≥ 5:1 (preferably 10:1), or absolute minimum n = 50 for well-defined factors [28]. Comrey and Lee (1992) [29] proposed: 50 = very poor, 100 = poor, 200 = fair, 300 = good, 500 = very good, ≥1000 = excellent.

### 4.3 Multivariate Normality

Maximum Likelihood extraction requires multivariate normality, though PCA and PAF are more robust. Normality is assessed via Mardia's test or Henze-Zirkler test [5]. Skewed variables should be transformed prior to FA [30].

### 4.4 Absence of Extreme Multicollinearity

While FA requires sufficient inter-correlation, extreme multicollinearity (determinant ≈ 0) causes computational instability. The correlation matrix determinant should exceed 0.00001. Variables with VIF > 10 should be reviewed [27].

### 4.5 Factorability of the Correlation Matrix

The correlation matrix must contain sufficient correlations to justify FA. This is formally tested using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity. KMO ≥ 0.60 and significant Bartlett's test ($p < 0.05$) are required [31].

## 5. STEPS IN CONDUCTING FACTOR ANALYSIS

A rigorous FA investigation follows an eight-step protocol [23,32]:

**Step 1: Formulation of Research Problem**

Clearly define objectives and identify observable variables. The choice of variables must be guided by subject-matter knowledge, theory, or literature [16].

**Step 2: Data Collection and Variable Selection**

Collect data systematically ensuring representativeness. Handle missing data via listwise deletion, pairwise deletion, or multiple imputation. Variables should be continuous or ordinal; binary variables require tetrachoric correlations [33].

**Step 3: Testing Sampling Adequacy - KMO Test**

Compute the Kaiser-Meyer-Olkin (KMO) statistic to evaluate whether partial correlations are small relative to ordinary correlations:

$$KMO = \Sigma\Sigma(r_{ij})^2 / [\Sigma\Sigma(r_{ij})^2 + \Sigma\Sigma(a_{ij})^2] \quad (Eq.\ 4)$$

where $r_{ij}$ are simple correlations and $a_{ij}$ are partial correlations. Kaiser (1974) [31] ratings: 0.90+ = marvellous, 0.80-0.89 = meritorious, 0.70-0.79 = middling, 0.60-0.69 = mediocre, 0.50-0.59 = miserable, <0.50 = unacceptable.

**Step 4: Bartlett's Test of Sphericity**

Bartlett's (1950) [34] test evaluates $H_0$: the correlation matrix is an identity matrix. A significant result ($p < 0.05$) indicates sufficient correlations:

$$\chi^2 = -[n - 1 - (2p + 5)/6] \times \ln|R| \quad (Eq.\ 5)$$

where n = sample size, p = number of variables, |R| = determinant of correlation matrix. Degrees of freedom = p(p−1)/2.

**Step 5: Extraction of Initial Factors**

Apply a chosen extraction method (PCA, PAF, or ML) to obtain initial unrotated factor loadings.

**Step 6: Determining Number of Factors**

Three criteria are commonly applied [23,35]: retain factors with eigenvalue $\lambda \geq 1.0$ (Kaiser criterion); plot eigenvalues and retain factors above the 'elbow' (Cattell, 1966 [36]); and retain factors cumulatively explaining ≥60-70% of variance [23].

**Step 7: Factor Rotation**

Rotate to achieve simple structure. See Section 7 for Varimax, Quartimax, Promax, and Oblimin rotations [37].

**Step 8: Interpretation and Naming**

Examine rotated loadings. Variables with |loading| ≥ 0.40-0.50 are substantially related to a factor [23]. Assign conceptually meaningful names based on high-loading variables.

## 6. METHODS OF FACTOR EXTRACTION

### 6.1 Principal Component Method (PCM)

PCM decomposes the full correlation matrix R to minimise unexplained variance. Eigenvalues and eigenvectors of R provide the solution [5,25]:

$$R = P\Lambda P^T \quad (Eq.\ 6)$$

where P is the eigenvector matrix (factor loadings) and Λ is the diagonal eigenvalue matrix. PCM is preferred when the goal is pure data reduction.

### 6.2 Principal Axis Factoring (PAF)

PAF operates on a reduced correlation matrix with communality estimates on the diagonal (replacing 1's). Initial communalities are obtained from squared multiple correlations (SMC). Communalities are iteratively updated until convergence [38]. PAF is the most recommended extraction method for EFA because it properly estimates common factor variance [23].

### 6.3 Maximum Likelihood Estimation (MLE)

MLE assumes multivariate normality and finds the loading matrix Λ that maximises the log-likelihood of observing the sample correlation matrix [19]. MLE produces a goodness-of-fit chi-square statistic, enabling

*Table 2. Comparison of factor extraction methods.*

| Method | Variance Analysed | Normality Required | Fit Test | Best For |
|---|---|---|---|---|
| **Principal Component** | Total (common + unique) | No | No | Data reduction |
| **Principal Axis Factoring** | Common only | No | No | EFA (general purpose) |
| **Maximum Likelihood** | Common only | Yes | Yes | EFA + model fit testing |

formal evaluation of model fit [24]. MLE also enables confidence intervals for loadings.

## 7. FACTOR ROTATION TECHNIQUES

The initial (unrotated) factor solution is indeterminate infinitely many rotations yield equally good fit. Rotation transforms the loading matrix to achieve simple structure [18]. Rotations are classified as orthogonal (factors uncorrelated) or oblique (factors allowed to correlate).

### 7.1 Varimax Rotation (Orthogonal)

Varimax (Kaiser, 1958 [39]) maximises the variance of squared loadings within each factor column, driving loadings toward 0 or ±1. This simplifies factor structure by ensuring each variable loads highly on as few factors as possible. Varimax is the default in SPSS and most widely used [23].

$$V_j = (1/p)\ \Sigma_i[\lambda^4_{ij}] - [(1/p)\ \Sigma_i\ \lambda^2_{ij}]^2 \rightarrow \text{maximise } \Sigma_j V_j \quad (Eq.\ 7)$$

### 7.2 Quartimax Rotation (Orthogonal)

Quartimax maximises variance of squared loadings within each variable row, simplifying variable profiles rather than factor profiles [37]. It often yields a dominant general factor, making it less popular than Varimax.

### 7.3 Promax Rotation (Oblique)

Promax (Hendrickson & White, 1964 [40]) first applies Varimax, then raises loadings to a power (typically $\kappa = 3$ or 4) to sharpen simple structure while allowing factor correlations. It is computationally efficient and widely used when factors are expected to correlate [13].

### 7.4 Direct Oblimin (Oblique)

Direct Oblimin minimises covariance of squared loadings across factor columns, providing flexibility via the delta ($\delta$) parameter [37]. $\delta = 0$ yields maximum obliqueness; negative $\delta$ restricts correlations.

## 8. MEASURES OF MODEL ADEQUACY

### 8.1 Kaiser-Meyer-Olkin (KMO) Measure

Described in Step 3 (Eq. 4), the KMO statistic quantifies the proportion of variance in variables that might be caused by underlying factors [31]. KMO $\geq 0.70$ is generally acceptable. Individual MSA values identify variables that may weaken the structure.

*Table 3. Summary of rotation methods.*

| Rotation | Type | Factors Correlated? | Maximises | When to Use |
|---|---|---|---|---|
| **Varimax** | Orthogonal | No | Variance within columns | Factors conceptually independent |
| **Quartimax** | Orthogonal | No | Variance within rows | Single general factor expected |
| **Promax** | Oblique | Yes | Kurtosis of loadings | Factors expected to correlate |
| **Direct Oblimin** | Oblique | Yes | Simplicity (oblique) | Correlated factors, flexible δ |

### 8.2 Bartlett's Test of Sphericity

Described in Step 4 (Eq. 5), Bartlett's test [34] checks whether the correlation matrix differs significantly from an identity matrix. A significant result ($p < 0.05$) is necessary. Note that with very large samples, Bartlett's test will always be significant; KMO becomes more diagnostic.

### 8.3 Communalities

The communality $h^2$ (Eq. 3) represents the proportion of variance shared with extracted factors. Low communalities (< 0.30) indicate a variable shares little variance with others and may be a candidate for removal [16].

### 8.4 Total Variance Explained

The eigenvalue for each factor indicates the amount of total variance explained. The cumulative percentage of variance explained by retained factors should reach ≥ 60-70% for an acceptable solution [23].

### 8.5 Scree Plot & Parallel Analysis

The scree plot (Cattell, 1966 [36]) graphs eigenvalues in descending order. The 'elbow' indicates where factors cease to explain meaningful variance. Parallel Analysis (Horn, 1965 [35]) is more rigorous retaining only factors with eigenvalues exceeding those from random data of the same size.

## 9. SOFTWARE FOR FACTOR ANALYSIS

All major statistical platforms provide FA routines [32]:

### 9.1 SPSS Implementation

The following figure illustrates the complete SPSS workflow for conducting Factor Analysis, from data import through to factor extraction.

*Table 4. Software platforms for Factor Analysis.*

| Software | Key Functions / Packages | Strengths | Limitations |
|---|---|---|---|
| **SPSS** | Analyse → Dimension Reduction → Factor | User-friendly GUI; standard output | Licensed; limited customisation |
| **R (psych, lavaan)** | fa(), principal(), cfa(), sem() | Free; extensive; publication graphics | Requires programming skill |
| **SAS** | PROC FACTOR, PROC CALIS | Robust for large datasets | Expensive; syntax-heavy |
| **Python (sklearn)** | sklearn.decomposition.FactorAnalysis | Integrates with ML pipelines | Less statistical depth |
| **AMOS** | GUI-based SEM/CFA | Intuitive CFA path diagrams | SPSS add-on; cost |
| **Mplus** | CFA, SEM, mixture models | Most comprehensive modelling | Expensive; steep learning curve |

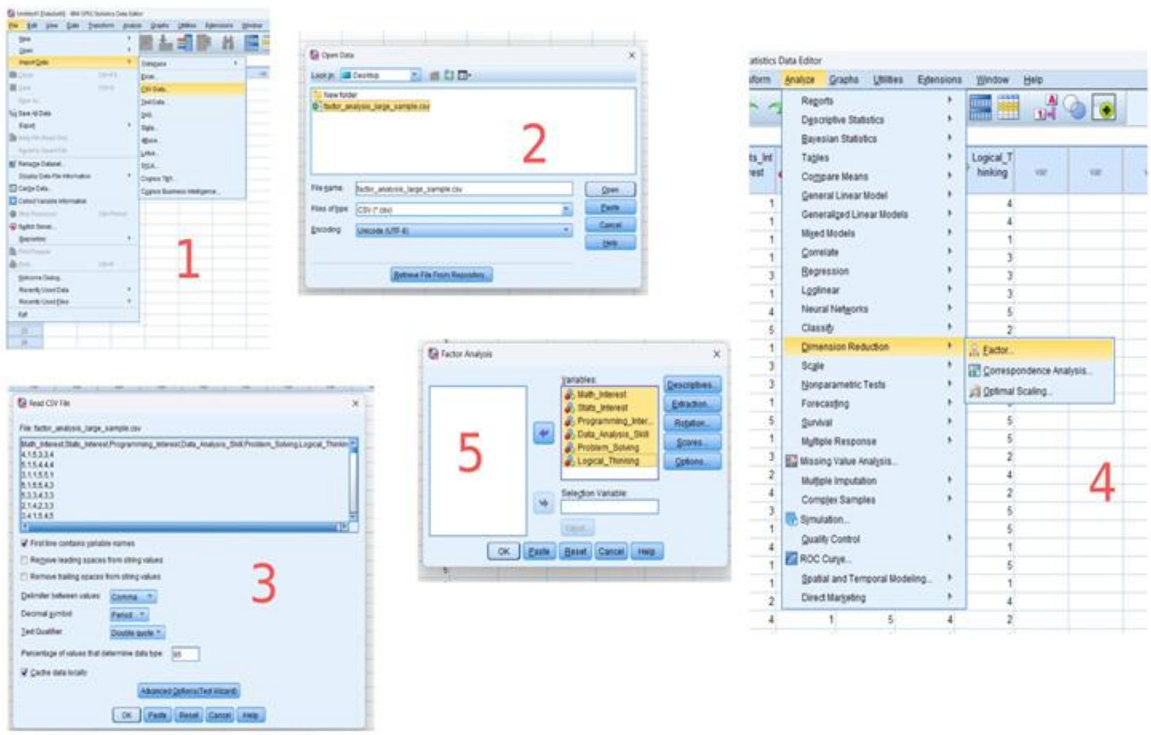


*Figure 1. SPSS workflow for Factor Analysis: (1) File → Import Data → CSV Data, (2) Select the data file, (3) Configure CSV import settings, (4) Navigate to Analyze → Dimension Reduction → Factor, (5) Assign variables in the Factor Analysis dialog.*

**9.2 R Code Example --- Complete EFA Workflow**

```
# Install and load required packages
install.packages(c('psych', 'GPArotation', 'corrplot'))
library(psych)        # fa(), fa.parallel(), KMO(), cortest.bartlett()
library(GPArotation)  # rotation algorithms
library(corrplot)     # correlation matrix visualisation
data <- read.csv('factor_analysis_large_sample.csv', header = TRUE)
data_matrix <- as.matrix(data)
describe(data)        # mean, sd, skewness, kurtosis per variable
cor_matrix <- cor(data_matrix, use = 'complete.obs')
print(round(cor_matrix, 3))
corrplot(cor_matrix, method = 'color', type = 'upper',
         tl.cex = 0.8, addCoef.col = 'black', number.cex = 0.7)
KMO(cor_matrix)                                   # KMO sampling adequacy
cortest.bartlett(cor_matrix, n = nrow(data)) # Bartlett's sphericity test
det(cor_matrix)                                   # determinant > 0.00001
fa.parallel(data_matrix, fa = 'fa', fm = 'pa',
            main = 'Parallel Analysis Scree Plot')
fa_result <- fa(r        = data_matrix,
                nfactors = 3,
                rotate   = 'varimax',
                fm       = 'pa',      # Principal Axis Factoring
                scores   = 'regression',
                missing  = FALSE,
                impute   = 'median')
print(fa_result$loadings, cutoff = 0.40) # rotated factor loadings
print(fa_result$communalities)          # communalities (h^2)
print(fa_result$Vaccounted)             # variance explained per factor
print(fa_result$RMSEA)                  # model fit (if MLE used)
fa.diagram(fa_result, cut = 0.40,
           main = 'Varimax Rotated Factor Structure')
factor_scores <- fa_result$scores       # regression-method scores
write.csv(factor_scores, 'factor_scores.csv', row.names = FALSE)
```

*Code Box 1. Complete EFA workflow in R using the psych package.*

### 9.3 Python Code Example --- Complete EFA Workflow

```
# Install and import required libraries
# pip install pandas numpy factor_analyzer matplotlib seaborn scipy

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from factor_analyzer import (FactorAnalyzer,
                             calculate_kmo,
                             calculate_bartlett_sphericity)

df = pd.read_csv('factor_analysis_large_sample.csv')
print(df.shape)    # (150, 6)
```

```
print(df.describe())
print('Skewness:',  df.skew().round(3).to_dict())
print('Kurtosis:',  df.kurt().round(3).to_dict())
cor_matrix = df.corr()
print(cor_matrix.round(3))
sns.heatmap(cor_matrix, annot=True, fmt='.2f', cmap='coolwarm',
            linewidths=0.5, vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.tight_layout(); plt.show()
kmo_all, kmo_model = calculate_kmo(df)
print(f'KMO (overall): {kmo_model:.3f}')
chi_sq, p_val = calculate_bartlett_sphericity(df)
print(f'Bartlett Chi-sq: {chi_sq:.3f},  p-value: {p_val:.4f}')
print(f'Determinant: {np.linalg.det(cor_matrix.values):.6f}')

fa_check = FactorAnalyzer(n_factors=6, rotation=None)
fa_check.fit(df)
eigenvalues, _ = fa_check.get_eigenvalues()
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(eigenvalues)+1), eigenvalues, 'o-', color='black')
plt.axhline(y=1, color='gray', linestyle='--', label='Kaiser criterion')
plt.xlabel('Factor Number'); plt.ylabel('Eigenvalue')
plt.title('Scree Plot'); plt.legend()
plt.tight_layout(); plt.show()
fa = FactorAnalyzer(n_factors=3,
                    method='principal',  # Principal Axis Factoring
                    rotation='varimax')
fa.fit(df)
loadings = pd.DataFrame(fa.loadings_,
                        index=df.columns,
                        columns=['Factor 1', 'Factor 2', 'Factor 3'])
print('Varimax Rotated Loadings:')
print(loadings.round(3))

communalities = pd.DataFrame(fa.get_communalities(),
                             index=df.columns, columns=['h^2'])
print('Communalities:'); print(communalities.round(3))

var_df = pd.DataFrame(fa.get_factor_variance(),
                      index=['SS Loadings','% Variance','Cumulative %'],
                      columns=['Factor 1','Factor 2','Factor 3'])
print('Variance Explained:'); print(var_df.round(4))

sns.heatmap(loadings, annot=True, fmt='.2f', cmap='RdBu_r',
            center=0, linewidths=0.5, vmin=-1, vmax=1)
plt.title('Varimax Rotated Factor Loadings')
plt.tight_layout(); plt.show()
scores_df = pd.DataFrame(fa.transform(df),
                         columns=['Factor_1','Factor_2','Factor_3'])
scores_df.to_csv('factor_scores_python.csv', index=False)
```

*Code Box 2. Complete EFA workflow in Python using the factor_analyzer library.*

## 10. ADVANTAGES OF FACTOR ANALYSIS

Factor Analysis offers well-documented advantages [1,5,23]:

- Dimension Reduction: Converts numerous correlated variables into interpretable factors, reducing complexity and multicollinearity
- Latent Construct Identification: Reveals hidden patterns not directly observable, bridging theory and measurement
- Improved Model Efficiency: Factor scores replace correlated variables in regression, improving coefficient stability
- Scale Development: Forms the scientific backbone of psychometric scale construction and validation
- Parsimony: Adheres to Occam's Razor---same information conveyed with fewer parameters
- Index Construction: Provides objective, data-driven weights for composite indices

## 11. LIMITATIONS OF FACTOR ANALYSIS

Despite its power, FA has important limitations [7,25,26]:

- Subjectivity in Factor Naming: Interpretation requires subjective judgement, raising reproducibility concerns
- Sample Size Sensitivity: FA is unreliable with small samples ($n < 100$). Solutions should be replicated
- Rotation Indeterminacy: Infinitely many rotations fit equally well; chosen rotation influences structure
- Factor Score Indeterminacy: Factor scores are estimates, not exact values. Multiple methods yield different results
- Assumption Violations: Non-normality, outliers, and non-linearity can distort results
- Requires Strong Statistical Background: Proper application demands familiarity with linear algebra and multivariate statistics

## 12. CASE STUDY - FACTOR ANALYSIS OF LEARNING INTEREST & SKILL VARIABLES

### 12.1 Dataset Description

A dataset of 150 observations was collected measuring six variables on a 1-5 Likert scale: Math Interest, Statistics Interest, Programming Interest, Data Analysis Skill, Problem Solving, and Logical Thinking. Sample size $n = 150$ satisfies the 25:1 subject-to-variable ratio, exceeding the recommended 5:1 minimum [28]. Variables exhibit approximately uniform distributions (means 2.76-3.15, SDs 1.38-1.49) with slight negative kurtosis (-1.24 to -1.42), indicating flatter-than-normal distributions.

## 12.2 Descriptive Statistics

*Table 5. Descriptive statistics for the six observed variables (n = 150).*

| Variable | n | Mean | SD | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| **Math_Interest** | 150 | 3.15 | 1.38 | 1.0 | 5.0 | −0.19 | −1.24 |
| **Stats_Interest** | 150 | 2.76 | 1.40 | 1.0 | 5.0 | 0.05 | −1.30 |
| **Programming_Interest** | 150 | 3.02 | 1.45 | 1.0 | 5.0 | 0.03 | −1.41 |
| **Data_Analysis_Skill** | 150 | 3.15 | 1.49 | 1.0 | 5.0 | −0.13 | −1.42 |
| **Problem_Solving** | 150 | 2.95 | 1.46 | 1.0 | 5.0 | 0.08 | −1.35 |
| **Logical_Thinking** | 150 | 2.92 | 1.42 | 1.0 | 5.0 | 0.09 | −1.29 |

## 12.3 Correlation Matrix & Preliminary Diagnostics

The correlation matrix (Table 6) reveals weak inter-correlations among variables. The largest correlation is 0.174 (Problem_Solving ↔ Logical_Thinking), with most correlations < |0.13|. The correlation matrix determinant = 0.8714 (well above 0.00001), indicating no multicollinearity issues.

**Interpretation:** The KMO = 0.488 falls in the 'Unacceptable' range (Kaiser 1974 [31]), and Bartlett's test is non-significant (p = 0.167), indicating the correlation matrix does not differ significantly from an identity matrix. These results suggest that Factor Analysis is NOT appropriate for this dataset. The variables are relatively independent and do not share substantial common variance. However, for pedagogical purposes, we proceed with the FA workflow to demonstrate the complete methodology.

*Table 6. Correlation matrix showing weak inter-variable relationships.*

| Variable | Math Int | Stats Int | Prog Int | Data Skill | Problem | Logical |
|---|---|---|---|---|---|---|
| **Math_Interest** | 1.000 | 0.022 | −0.028 | 0.093 | −0.003 | 0.013 |
| **Stats_Interest** | 0.022 | 1.000 | −0.117 | 0.120 | −0.122 | −0.094 |
| **Programming_Interest** | −0.028 | −0.117 | 1.000 | 0.020 | −0.025 | 0.161 |
| **Data_Analysis_Skill** | 0.093 | 0.120 | 0.020 | 1.000 | 0.007 | −0.131 |
| **Problem_Solving** | −0.003 | −0.122 | −0.025 | 0.007 | 1.000 | 0.174 |
| **Logical_Thinking** | 0.013 | −0.094 | 0.161 | −0.131 | 0.174 | 1.000 |

## 12.4 KMO and Bartlett's Test Results

*Table 7. Sampling adequacy tests indicate FA is NOT recommended for this dataset.*

| Test | Statistic | DF | p-value | Interpretation |
|---|---|---|---|---|
| **KMO Measure** | 0.488 | --- | --- | Unacceptable (< 0.50) |
| **Bartlett's Test** | $\chi^2 = 20.12$ | 15 | 0.167 | Non-significant (p > 0.05) |
| **Determinant** | \|R\| = 0.8714 | --- | --- | No multicollinearity |

## 12.5 Eigenvalue Analysis & Factor Retention

*Table 8. Eigenvalue analysis --- 3 factors retained by Kaiser Criterion (λ ≥ 1.0).*

| Factor | Eigenvalue | % Variance | Cumulative % | Kaiser Criterion |
|---|---|---|---|---|
| **1** | 1.381 | 23.01% | 23.01% | Retain (λ ≥ 1.0) |
| **2** | 1.079 | 17.98% | 41.00% | Retain (λ ≥ 1.0) |
| **3** | 1.028 | 17.13% | 58.13% | Retain (λ ≥ 1.0) |
| **4** | 0.935 | 15.58% | 73.70% | Drop (λ < 1.0) |
| **5** | 0.905 | 15.08% | 88.79% | Drop (λ < 1.0) |
| **6** | 0.673 | 11.21% | 100.00% | Drop (λ < 1.0) |

Three factors have eigenvalues ≥ 1.0, cumulatively explaining 58.13% of total variance. While this marginally meets the 60% threshold, the weak correlations suggest these factors may not be substantively meaningful.

## 12.6 Varimax Rotated Factor Loadings

*Table 9. Varimax rotated factor loadings (loadings ≥ |0.40| are substantive).*

| Variable | Factor 1 | Factor 2 | Factor 3 | Communality (h²) |
|---|---|---|---|---|
| **Math_Interest** | 0.230 | −0.678 | −0.005 | 0.513 |
| **Stats_Interest** | −0.398 | −0.304 | −0.303 | 0.343 |
| **Programming_Interest** | −0.062 | −0.038 | 0.917 | 0.846 |
| **Data_Analysis_Skill** | −0.144 | −0.755 | 0.054 | 0.594 |
| **Problem_Solving** | 0.812 | −0.059 | −0.178 | 0.694 |
| **Logical_Thinking** | 0.569 | 0.119 | 0.400 | 0.497 |

## 12.7 Factor Interpretation

Based on loadings ≥ |0.40|, the three-factor solution suggests:

**Factor 1: 'Problem-Solving Orientation'**

- Problem_Solving (0.812)
- Logical_Thinking (0.569)

This factor represents cognitive abilities related to analytical thinking and reasoning.

**Factor 2: 'Quantitative Aptitude'**

- Data_Analysis_Skill (−0.755)
- Math_Interest (−0.678)

Negative loadings indicate an inverse factor; this reflects mathematical and data-analytical capabilities.

**Factor 3: 'Programming Affinity'**

- Programming_Interest (0.917)

This factor is dominated by a single variable, suggesting it may represent a unique dimension not shared with others.

## 12.8 Discussion

While the three-factor solution is mathematically extractable, the low KMO (0.488) and non-significant Bartlett's test ($p = 0.167$) indicate that these factors are not statistically robust. The weak inter-correlations (most < |0.13|) suggest the six variables measure relatively independent constructs. In practice, this dataset should NOT undergo Factor Analysis.

This case study demonstrates an important lesson: not all datasets are suitable for FA. Researchers must rigorously assess KMO and Bartlett's test before proceeding. Data with KMO < 0.60 and non-significant Bartlett's test indicate that variables do not share substantial common variance---the fundamental prerequisite for FA [31,34].

## 13. CONCLUSION

### 13.1 Summary

This article provided a systematic treatment of Factor Analysis from historical origins (Spearman [3], Thurstone [17,18]) through mathematical foundations (common factor model, Eqs. 1-3), assumptions (linearity, normality, sample size), extraction methods (PCM, PAF, MLE), rotation techniques (Varimax, Quartimax, Promax, Oblimin), and model adequacy assessment (KMO, Bartlett's, communalities, scree plots). The distinction between Exploratory and Confirmatory FA was clarified, and software implementation was outlined.

The case study using a 150-observation dataset demonstrated the complete FA workflow---including critically important diagnostic steps. The dataset exhibited weak inter-correlations (KMO = 0.488, non-significant Bartlett's test), illustrating that not all datasets are suitable for FA. This underscores the necessity of rigorous preliminary diagnostics.

### 13.2 Practical Recommendations

For researchers applying Factor Analysis [27,32]:

- Always report KMO and Bartlett's test results to establish suitability
- Prefer PAF extraction for EFA; use MLE when model fit testing is needed
- Choose rotation based on theory: Varimax for independent factors; Promax for correlated factors
- Interpret factors collaboratively with domain experts---statistics alone does not replace subject knowledge
- Use factor scores as composite variables in subsequent regression, ANOVA, or clustering analyses
- Replicate factor solutions on independent samples to establish stability

### 13.3 Scope for Future Research

Extensions merit further investigation:

- Bifactor models for hierarchical factor structures
- Integration of FA with machine learning pipelines for predictive modelling
- Longitudinal Factor Analysis for time-series data
- Bayesian Factor Analysis incorporating informative priors

- Multi-group CFA for testing factorial invariance across populations

## REFERENCES


1. Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Lawrence Erlbaum Associates.
2. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272-299.
3. Spearman, C. (1904). 'General intelligence,' objectively determined and measured. American Journal of Psychology, 15(2), 201-292.
4. Thompson, B. (2004). Exploratory and confirmatory factor analysis. American Psychological Association.
5. Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Pearson Prentice Hall.
6. Mulaik, S. A. (2010). Foundations of factor analysis (2nd ed.). CRC Press / Taylor & Francis.
7. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.
8. Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson Education.
9. Harman, H. H. (1976). Modern factor analysis (3rd ed.). University of Chicago Press.
10. Bellman, R. (1961). Adaptive control processes: A guided tour. Princeton University Press.
11. Everitt, B. S., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer.
12. Dunteman, G. H. (1989). Principal components analysis. SAGE Publications.
13. Sharma, R., Prasad, K., & Rao, P. V. (2015). Factor analysis of yield and morphological traits in wheat genotypes. Journal of Wheat Research, 7(1), 55-61.
14. Singh, A. K., Kumar, P., & Nath, C. P. (2018). Soil quality assessment using factor analysis. Indian Journal of Horticulture, 75(2), 210-218.
15. Kumari, P., & Singh, R. P. (2014). Socio-economic status index construction for farming households using factor analysis. Indian Journal of Agricultural Economics, 69(2), 167-178.

**16.** Dass, A., Sudhishri, S., & Lenka, N. K. (2016). Factor analysis of soil physicochemical properties. Agricultural Research, 5(3), 251-259.

**17.** Thurstone, L. L. (1931). Multiple factor analysis. Psychological Review, 38(5), 406-427.

**18.** Thurstone, L. L. (1947). Multiple-factor analysis: A development and expansion of the vectors of mind. University of Chicago Press.

**19.** Lawley, D. N., & Maxwell, A. E. (1963). Factor analysis as a statistical method. Butterworths.

**20.** Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183-202.

**21.** Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons.

**22.** Loehlin, J. C. (2004). Latent variable models: An introduction to factor, path, and structural equation analysis (4th ed.). Lawrence Erlbaum Associates.

**23.** Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

**24.** Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

**25.** Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

**26.** Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis. Practical Assessment, Research & Evaluation, 10(7), 1-9.

**27.** MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. Psychological Methods, 4(1), 84-99.

**28.** Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

**29.** Comrey, A. L., & Lee, H. B. (1992). A first course in factor analysis (2nd ed.). Lawrence Erlbaum Associates.

**30.** Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44(4), 443-460.

**31.** Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31-36.

**32.** Revelle, W. (2023). psych: Procedures for psychological, psychometric, and personality research. Northwestern University.

**33.** Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). John Wiley & Sons.

**34.** Bartlett, M. S. (1950). Tests of significance in factor analysis. British

Journal of Psychology---Statistical Section, 3(2), 77-85.

**35.** Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185.

**36.** Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245-276.

**37.** Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. Multivariate Behavioral Research, 36(1), 111-150.

**38.** Widaman, K. F. (1990). Bias in pattern loadings represented by common factor analysis and component analysis. Multivariate Behavioral Research, 25(1), 89-95.

**39.** Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23(3), 187-200.

**40.** Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology, 17(1), 65-70.