EXAMINING THE ROBUSTNESS OF FACTORIAL TECHNIQUES APPLIED TO SURVEY DATA

1st Hind Bourass
FSJES Agdal
Mohammed V University
Rabat, Morocco
bourasshind1@gmail.com

2nd Outmane Soussi Noufail
FSJES Salé
Mohammed V University
Rabat, Morocco
n.soussi@um5r.ac.ma
0000-0002-0269-7935

***Abstract*—This article delves into the use of factor analysis with survey data, highlighting its robustness under diverse conditions unique to survey research. It investigates the performance of these techniques in practical scenarios, especially when confronted with challenges, aiming to ascertain their ability to consistently and accurately analyze survey data, despite its complexity. This study seeks to enhance the practical application of these methods and guarantee meaningful outcomes. Additionally, it addresses the critical aspect of the reliability of measurement instruments in this context.**

***Index Terms*—Cronbach's alpha coefficient, PCA, Bartlett test, KMO test**

I. INTRODUCTION

Factor analysis is a valuable tool for uncovering the underlying structures of multidimensional survey data. Yet, ensuring the reliability of results hinges on the robustness of these techniques under diverse conditions, including outliers, missing data, or non-normal distributions. This article investigates the robustness of factor analysis in survey data analysis, evaluating its performance with real-world data that frequently present challenges. By assessing these techniques' capability to handle survey data accurately and consistently amid these complexities, we seek to elucidate their practical utility and provide guidelines for reliable and meaningful analyses.

II. THE RELIABILITY OF MEASURING INSTRUMENTS

Assessing internal consistency aims to enhance data quality by identifying the most representative elements of the studied concepts. This evaluation occurs in two stages:

1) Calculating Cronbach's alpha coefficient allows for the measurement of internal consistency within a set of measurement indicators. This value evaluates the extent to which an item can compromise the overall consistency of a composite scale.
2) Removing items that weaken Cronbach's alpha coefficient by adhering to a predefined decision rule.

This process aims to enhance the reliability of survey measures by eliminating elements that could compromise the overall consistency of the data.

Cronbach's alpha is thus a measure of the internal consistency of a measurement scale, commonly employed in psychometrics.

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma_{ik}^2}{\sigma_T^2}\right) \tag{1}$$

With:

- α is the Cronbach's alpha coefficient,
- N is the number of elements (observations) in the scale,
- k is the number of elements in the scale (i.e., the number of questions or items in the questionnaire),
- σ_i^2 is the variance of each individual item, and
- σ_T^2 is the total variance of the set of item scores.

Cronbach's alpha is often used as a preliminary step before conducting factor analysis. While it's not a perfect conceptual fit, alpha is sometimes interpreted as the average correlation among all possible pairs of items within a group. A high alpha value suggests strong internal correlation among items and is typically used as a criterion to determine if further factor analysis is justified. This measure is critical for evaluating the internal consistency of items in a dataset, aiding in the decision of whether factor analysis is appropriate for exploring the data's underlying structures.

For instance, imagine you have an 11-question questionnaire designed to gauge customer satisfaction with a product. You've gathered responses from 12 customers to these questions. These responses, detailed in the database provided in Appendix 1, form the foundation for assessing customer satisfaction with the product.

The table (database) can be represented schematically as follows:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \ddots & & \\ & & \ddots & \\ x_{N1} & x_{N2} & \cdots & x_{Nk} \end{pmatrix} \tag{2}$$

We are dealing with a variable[1] that is measured through a series of questions, specifically 11 items, which together form a measurement scale for this latent variable. Thus:

[1] Not directly observed

- $N = 12$ represents the total number of observations, i.e., the number of customers who responded to the questionnaire.
- $k = 11$ corresponds to the number of items in the measurement scale, each contributing to assessing the latent variable.
- σ_{ik}^2 represents the variance of each item; with $k = 11$, we need to calculate 11 individual variances to assess the dispersion of responses for each question.
- σ_T^2 represents the total variance, i.e., the variance of the cumulative item scores. In this case, we have a single variance to calculate to evaluate the overall dispersion of responses across all the items on the measurement scale.

To calculate individual variances, we use the command: (Excel : =VAR.P(plage))

$$\boxed{\sigma_{ik}^2 = \frac{\sum_{i=1}^{N}(x_{ik} - \bar{x_k})^2}{N}} \qquad (3)$$

1) For example for $k = 1$:

$$\sigma_{i1}^2 = \frac{\sum_{i=1}^{12}(x_{i1} - \bar{x_1})^2}{N} = 0,139$$

2) Calculation of the sum of individual variances gives:

$$\sum_{i=1}^{n} \sigma_{ik}^2 = 2.1458 \approx 2.15$$

3) Calculation of score variance

$$\boxed{\sigma_T^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 6.521 \approx 6.52 \qquad (4)$$

With y_i is the sum of the responses of individual i

4) Calculating the alpha gives:

$$\alpha = \frac{11}{11-1}\left(1 - \frac{2.15}{6.52}\right) = 0.738 \boxed{\approx 0.74}$$

To evaluate the internal consistency of the latent variable measurement scale, the Cronbach's alpha parameter is used, where $0 < \alpha < 1$.

We apply the following **decision rule**:

Cronbach's alpha value	$\alpha > 0.8$	$0.6 < \alpha < 0.8$	$\alpha < 0.6$
Consistency	High	Moderately	Weakly
Decision	Acceptable	Acceptable	Unacceptable

Based on the calculations performed, the Cronbach's alpha coefficient falls between 0.6 and 0.8. This range indicates that the test is acceptable, confirming the reliability of the measurement scale for the latent variable under scrutiny. In essence, this result underscores an alpha value within an acceptable range, thereby affirming the internal consistency of the measurement scale employed to evaluate the latent variable.

$$\boxed{0.6 < \alpha = 0.74 < 0.8} \Rightarrow \text{Acceptable measurement scale}$$

This coefficient can also be calculated using another formula:

$$\boxed{\alpha = \frac{N \times \bar{r}}{1 + (N-1) \times \bar{r}}} \qquad (5)$$

With :

- $\bar{r}$ is the average correlation between all pairs of items.

$$\bar{r} = \frac{\sum r_{ij}}{Card(s)}$$

- $Card(s)$ is the number of correlation coefficients to calculate: it is a Combination without Repetition:

$$Card(s) = C_N^p = \binom{N}{p} = \binom{N}{2} = C_N^2$$

- Note that the correlation coefficient of PEARSON between the variables X and Y is given by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

With :

  - n is the number of observations.
  - x_i and y_i are the individual values of the variables X and Y.
  - $\bar{x}$ and $\bar{y}$ are the means of the variables X and Y respectively.

Based on the previous data, we calculate Cronbach's alpha using the latter formula by following these steps:

1) We have 55 distinct correlation coefficients to calculate:

$$\binom{11}{2} = \frac{11 \times 10}{2} = 55$$

2) We therefore calculate the correlation matrix and we are only interested in the inter-item coefficients (=COEFFICIENT.CORRELATION(Matrix1;Matrix2)) for example: $r_{12} = -0.26$
3) We calculate the average correlation: $\bar{r} = \frac{10.94}{55} = 0.20$
4) Alpha is given by:

$$\alpha = \frac{11 \times 0.20}{1 + (11-1) \times 0.20} = 0.733 \boxed{\approx 0.73}$$

III. PCA FACTORIZATION

A. Principle and methods

To refine the measuring instrument, factor analysis is essential. This method aims to reduce the data's dimensionality by identifying the principal components that capture most of the variance. By eliminating redundancy and distinguishing significant variables from those with little impact on the data's variance, this analysis refines the measurement instrument.

Factor analysis also allows for visualizing relationships between variables in a principal component space. This visualization facilitates understanding the data's underlying structure and helps identify patterns and connections between different variables.

Various factor analysis techniques exist, each offering unique nuances and approaches to better explore and interpret multidimensional data.

Factor analysis simplifies result interpretation by condensing many variables into a few. Often called the "Dimension Reduction Method," this approach reduces data dimensions into one or more "super-variables," also known as "constructs." This transformation is crucial for understanding the data's underlying structure and simplifying result interpretation.

The most common factor analysis technique is Principal Component Analysis (PCA). However, the choice between PCA and other multidimensional analyses depends largely on the data's nature. For example, PCA is suitable for quantitative data in a table format ($n \times p$), where rows represent individuals and columns represent variables, such as data from a coded questionnaire. This specific data structure is well-suited for PCA to extract meaningful variable relationships and efficiently reduce data dimensionality.

Principal Component Analysis (PCA) is a statistical method used for data reduction. It involves calculating the eigenvectors of the correlation or covariance matrix of the variables. These eigenvectors describe uncorrelated linear combinations of the variables, enabling data reduction while preserving most of the variance. Moreover, examining PCA eigenvectors helps in better understanding the data's underlying structure.

PCA enables the construction of a new representation system comprising linear combinations of the original variables, facilitating information synthesis.

When applying PCA to analyse a questionnaire, it is crucial to consider several questions:

- Proximity between individuals: Which individuals responded similarly to questions related to a specific variable or concept?
- Resemblance between individuals: What answers show similarities or differences among respondents?
- Relations between questions (items): What connections exist between the various questions?

Thus, the primary objectives, based on the questionnaire data, are to examine the similarity between responses related to a specific concept to ensure a degree of homogeneity, and to explore the variability among items to identify correlations between them.

For example, if a concept (or variable) is measured by a system of n items, PCA can construct a reduced representation of this concept (or variable) comprising ($p < n$) items. This new system will preserve the existing distances (relationships) (internal coherence) between these items.

When implementing a PCA, the main result is:

1) Construct a set of main components (C_1, C_2, ..., C_k, ..., C_p), defined as linear combinations of the original items (centered and scaled)[2], of which we can assess the quality of information retrieval through the reproduced inertia[3] (λ_k)[4].

$$\begin{cases} C_1 = & a_{11}z_1 + a_{21}z_1 + \cdots + a_{p1}z_p(\lambda_1) \\ \vdots & \\ C_k = & a_{1k}z_1 + a_{2k}z_2 + \cdots + a_{pk}z_k(\lambda_k) \\ \vdots & \\ C_p = & a_{1p}z_1 + a_{2p}z_2 + \cdots + a_{pp}z_p(\lambda_p) \end{cases}$$

With: z_k is the value of the variable (X_k after centering and scaling) specific to individual k.

2) We observe the decomposition of information into uncorrelated (orthogonal) components.
3) Retain the principal component(s) that maximize the square of their correlation with the variables in the database.

Given that PCA yields multiple results, especially in the context of analyzing questionnaire data measurement instruments, it is crucial to propose a precise approach.

B. PCA Procedure

To analyze the similarity within a dataset, a geometric approach involves studying the distances between individuals, typically measured by the Euclidean distance between two individuals (i, i'): $d^2(i, i') = \sum(x_{ij} - x_{i'j})^2$

Step 1 - Graphical study: This involves examining the graphical representation of the point cloud and includes the following stages:
- Center the data: This ensures that the center of inertia G is located at the origin.
- Standardize the data: This process aims to make variables comparable, especially if they are expressed in different scales or units[5].

Step 2. Analyze the centred and standardized data table: Visualizing the data directly is not possible. To address this, you need to find a more manageable representation (a better projection) of the data. This involves finding a subspace that summarizes the data, achieved through PCA.
- The goal is to project the cloud of 12 individuals onto the first two inertia axes, often referred to as the first factorial plane.
- This projection matrix helps identify the significant factorial axes.

To achieve this, we calculate the eigenvalues (which quantify the amount of information contained in each axis) and the associated eigenvectors. This analysis aids in determining the number of principal components that should be considered.

Step 3. To accomplish this, we calculate the covariance matrix.
- If the data are on the same measurement scale (binary, Likert), there is no need to center or scale; you can simply calculate the correlation matrix.

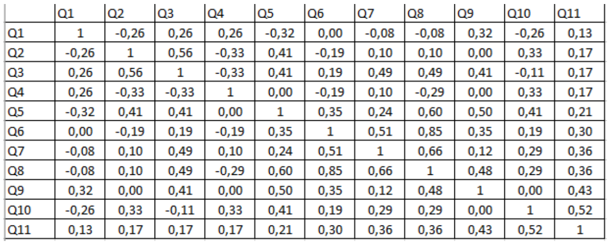
[2] normed PCA

[3] Dispersion around the barycenter; it is a multidimensional variance (calculated on p dimensions)

[4] Eigenvalues

[5] This refers to a standardized PCA

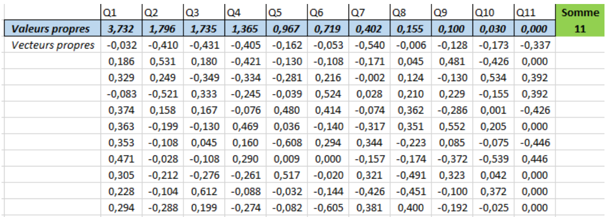
- In our case, we calculate the correlation matrix.



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Q1	1	-0,26	0,26	0,26	-0,32	0,00	-0,08	-0,08	0,32	-0,26	0,13
Q2	-0,26	1	0,56	-0,33	0,41	-0,19	0,10	0,10	0,00	0,33	0,17
Q3	0,26	0,56	1	-0,33	0,41	0,19	0,49	0,49	0,41	-0,11	0,17
Q4	0,26	-0,33	-0,33	1	0,00	-0,19	0,10	-0,29	0,00	0,33	0,17
Q5	-0,32	0,41	0,41	0,00	1	0,35	0,24	0,60	0,50	0,41	0,21
Q6	0,00	-0,19	0,19	-0,19	0,35	1	0,51	0,85	0,35	0,19	0,30
Q7	-0,08	0,10	0,49	0,10	0,24	0,51	1	0,66	0,12	0,29	0,36
Q8	-0,08	0,10	0,49	-0,29	0,60	0,85	0,66	1	0,48	0,29	0,36
Q9	0,32	0,00	0,41	0,00	0,50	0,35	0,12	0,48	1	0,00	0,43
Q10	-0,26	0,33	-0,11	0,33	0,41	0,19	0,29	0,29	0,00	1	0,52
Q11	0,13	0,17	0,17	0,17	0,21	0,30	0,36	0,36	0,43	0,52	1

Figure 1. Correlation matrix

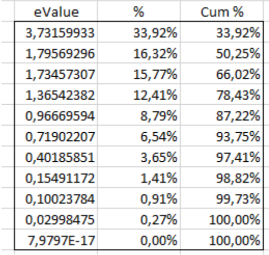
Step 4.1 Next, we reorganize the data into a new system. This is achieved through the diagonalization[6] of the covariance matrix.
We calculate eigenvalues and eigenvectors[7] from the correlation matrix (covariance matrix on centered values). (Excel =eVECTORS(C18:M28;100;FAUX)).



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Somme
Valeurs propres	***3,732***	***1,796***	***1,735***	***1,365***	***0,967***	***0,719***	***0,402***	***0,155***	***0,100***	***0,030***	***0,000***	**11**
Vecteurs propres	-0,032	-0,410	-0,431	-0,405	-0,162	-0,053	-0,540	-0,006	-0,128	-0,173	-0,337	
	0,186	0,531	0,180	-0,421	-0,130	-0,108	-0,171	0,045	0,481	-0,426	0,000	
	0,329	0,249	-0,349	-0,334	-0,281	0,216	-0,002	0,124	-0,130	0,534	0,392	
	-0,083	-0,521	0,333	-0,245	-0,039	0,524	0,028	0,210	0,229	-0,155	0,392	
	0,374	0,158	0,167	-0,076	0,480	0,414	-0,074	0,362	-0,286	0,001	-0,426	
	0,363	-0,199	-0,130	0,469	0,036	-0,140	-0,317	0,351	0,552	0,205	0,000	
	0,353	-0,108	0,045	0,160	-0,608	0,294	0,344	-0,223	0,085	-0,075	-0,446	
	0,471	-0,028	-0,108	0,290	0,009	0,000	-0,157	-0,174	-0,372	-0,539	0,446	
	0,305	-0,212	-0,276	-0,261	0,517	-0,020	0,321	-0,491	0,323	0,042	0,000	
	0,228	-0,104	0,612	-0,088	-0,032	-0,144	-0,426	-0,451	-0,100	0,372	0,000	
	0,294	-0,288	0,199	-0,274	-0,082	-0,605	0,381	0,400	-0,192	-0,025	0,000	

Figure 2. Calculation of eigenvalues and eigenvectors

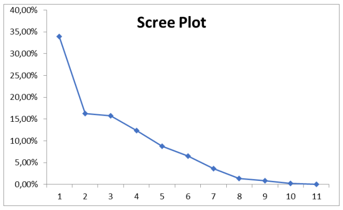
Step 4.2 Calculate the proportion of variation explained by each eigenvalue and the cumulative percentage explained. For example, 33.92% represents the contribution of C_1 (factorial representation) to the total variability. An alternative method is to construct a graph (Scree Plot) that illustrates the successive differences between the eigenvalues.

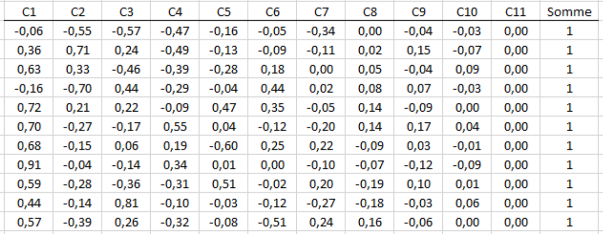


eValue	%	Cum %
3,73159933	33,92%	33,92%
1,79569296	16,32%	50,25%
1,73457307	15,77%	66,02%
1,36542382	12,41%	78,43%
0,96669594	8,79%	87,22%
0,71902207	6,54%	93,75%
0,40185851	3,65%	97,41%
0,15491172	1,41%	98,82%
0,10023784	0,91%	99,73%
0,02998475	0,27%	100,00%
7,9797E-17	0,00%	100,00%

The four components—C_1, C_2, C_3, and C_4—explain 78.43% of the variation, which is a relatively high percentage.

Step 5.1 Load the total weight matrix: This matrix will provide us with the loadings =VecteursP*RACINE(ABS(ValeursP)).





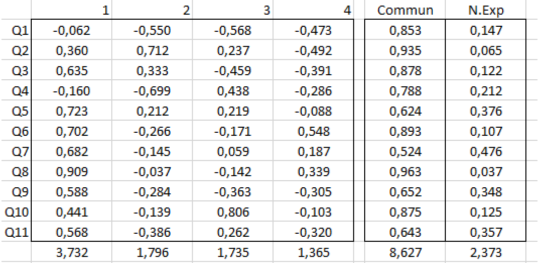
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Somme
-0,06	-0,55	-0,57	-0,47	-0,16	-0,05	-0,34	0,00	-0,04	-0,03	0,00	1
0,36	0,71	0,24	-0,49	-0,13	-0,09	-0,11	0,02	0,15	-0,07	0,00	1
0,63	0,33	-0,46	-0,39	-0,28	0,18	0,00	0,05	-0,04	0,09	0,00	1
-0,16	-0,70	0,44	-0,29	-0,04	0,44	0,02	0,08	0,07	-0,03	0,00	1
0,72	0,21	0,22	-0,09	0,47	0,35	-0,05	0,14	-0,09	0,00	0,00	1
0,70	-0,27	-0,17	0,55	0,04	-0,12	-0,20	0,14	0,17	0,04	0,00	1
0,68	-0,15	0,06	0,19	-0,60	0,25	0,22	-0,09	0,03	-0,01	0,00	1
0,91	-0,04	-0,14	0,34	0,01	0,00	-0,10	-0,07	-0,12	-0,09	0,00	1
0,59	-0,28	-0,36	-0,31	0,51	-0,02	0,20	-0,19	0,10	0,01	0,00	1
0,44	-0,14	0,81	-0,10	-0,03	-0,12	-0,27	-0,18	-0,03	0,06	0,00	1
0,57	-0,39	0,26	-0,32	-0,08	-0,51	0,24	0,16	-0,06	0,00	0,00	1

Figure 3. Total weight matrix

For instance, the first principal component $\hat{C}_1$ can be calculated using the elements of the first eigenvector.

$$\begin{cases} \hat{C}_1 = & -0.06Q_1 + 0.36Q_2 + \cdots + 0.57Q_{11} \\ \hat{C}_2 = & -0.55Q_1 + 0.71Q_2. + \cdots - 0.39Q_{11} \\ \quad\vdots & \qquad\qquad\vdots \\ \hat{C}_{11} = & 0.00Q_1 + 0.00Q_2. + \cdots + 0.00Q_{11} \end{cases}$$

Step 5.2 This PCA is termed ”Without rotation.” It involves retaining the $p = 4$ principal components, which account for 78.43% of the total variability.
It is now important to recalculate the eigenvector/score matrix *only* for the main p factors (components).



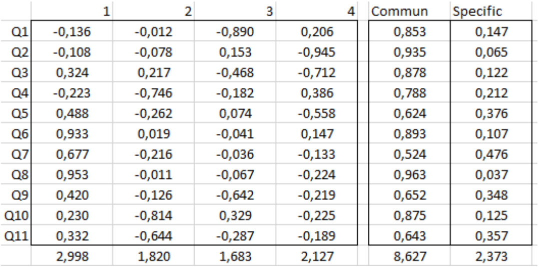
	1	2	3	4	Commun	N.Exp
Q1	-0,062	-0,550	-0,568	-0,473	0,853	0,147
Q2	0,360	0,712	0,237	-0,492	0,935	0,065
Q3	0,635	0,333	-0,459	-0,391	0,878	0,122
Q4	-0,160	-0,699	0,438	-0,286	0,788	0,212
Q5	0,723	0,212	0,219	-0,088	0,624	0,376
Q6	0,702	-0,266	-0,171	0,548	0,893	0,107
Q7	0,682	-0,145	0,059	0,187	0,524	0,476
Q8	0,909	-0,037	-0,142	0,339	0,963	0,037
Q9	0,588	-0,284	-0,363	-0,305	0,652	0,348
Q10	0,441	-0,139	0,806	-0,103	0,875	0,125
Q11	0,568	-0,386	0,262	-0,320	0,643	0,357
	3,732	1,796	1,735	1,365	8,627	2,373

Step 6.1 Perform a rotated factor analysis: Obtain a clearer representation of each item's contribution to the selected factor. Here, we opt for an orthogonal Varimax rotation (=VARIMAX(matrix of components;100))
There are several orthogonal rotation methods as well, such as Quartimax, Equamax, and Parsimax, among others.

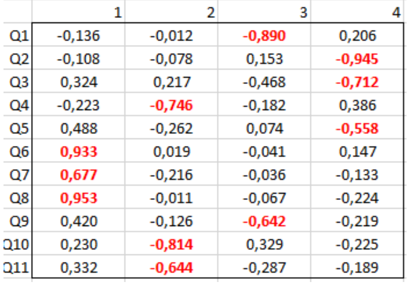
- The interpretation of principal components involves identifying the variables that are most strongly correlated with each component $\hat{C}_p$.
- A correlation greater than 0.5 in absolute value is considered significant.
- $\hat{\boldsymbol{C}}\mathbf{1}$ is the linear combination of variable Q that exhibits the maximum variance (among all linear combinations).

[6] The diagonal matrix is formed from the eigenvalues.

[7] An eigenvector of a linear transformation f is any vector x such that $f(x) = \lambda x$



	1	2	3	4	Commun	Specific
Q1	-0,136	-0,012	-0,890	0,206	0,853	0,147
Q2	-0,108	-0,078	0,153	-0,945	0,935	0,065
Q3	0,324	0,217	-0,468	-0,712	0,878	0,122
Q4	-0,223	-0,746	-0,182	0,386	0,788	0,212
Q5	0,488	-0,262	0,074	-0,558	0,624	0,376
Q6	0,933	0,019	-0,041	0,147	0,893	0,107
Q7	0,677	-0,216	-0,036	-0,133	0,524	0,476
Q8	0,953	-0,011	-0,067	-0,224	0,963	0,037
Q9	0,420	-0,126	-0,642	-0,219	0,652	0,348
Q10	0,230	-0,814	0,329	-0,225	0,875	0,125
Q11	0,332	-0,644	-0,287	-0,189	0,643	0,357
	2,998	1,820	1,683	2,127	8,627	2,373



	1	2	3	4
Q1	-0,136	-0,012	**-0,890**	0,206
Q2	-0,108	-0,078	0,153	**-0,945**
Q3	0,324	0,217	-0,468	**-0,712**
Q4	-0,223	**-0,746**	-0,182	0,386
Q5	0,488	-0,262	0,074	**-0,558**
Q6	**0,933**	0,019	-0,041	0,147
Q7	**0,677**	-0,216	-0,036	-0,133
Q8	**0,953**	-0,011	-0,067	-0,224
Q9	0,420	-0,126	**-0,642**	-0,219
Q10	0,230	**-0,814**	0,329	-0,225
Q11	0,332	**-0,644**	-0,287	-0,189

- More precisely, the coefficients following $\boldsymbol{a}11, \boldsymbol{a}12, \ldots, \boldsymbol{a}1p$ are determined to maximize the variance, while ensuring that the sum of the squares of the coefficients equals one, i.e., $\mathbf{A}_1'\mathbf{A}_1 = \sum_{j=1}^{p} A_{1j}^2 = 1$
- This constraint is necessary to obtain a unique solution.

$\mapsto$ The $\hat{\boldsymbol{C}}_1$ exhibits a strong correlation with three initial variables. It rises alongside the scores of Q_6, Q_8, and Q_7, indicating a simultaneous variation in these three criteria (items). An increase in one of them tends to coincide with increases in the others.

$\mapsto$ Moreover, $\hat{\boldsymbol{C}}_1$ exhibits the highest correlations with Q_8 (0.953) and Q_6 (0.933). Based on these correlations, we can assert that this principal component primarily reflects these two items, suggesting that it serves as a measure (or construct) of these variables.

- $\hat{\boldsymbol{C}}\mathbf{2}$ is the linear combination of variable Q that captures as much of the remaining variation as possible, with the additional constraint that its correlation with $\hat{\boldsymbol{C}}\mathbf{1}$ is 0.
- More precisely, we define $\boldsymbol{a}21, \boldsymbol{a}22, \ldots, \boldsymbol{a}2p$ to maximize the variance of this new component, while ensuring that the sum of the squared coefficients $\sum j = 1^p a_{2j}^2 = 1$, and with the additional constraint that these components are uncorrelated: $\operatorname{cov}(C_1, C_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} a_{1k} a_{2l} \sigma_{kl} = \mathbf{A}_1' \Sigma \mathbf{A}_2 = 0$.
  $\mapsto$ $\hat{\boldsymbol{C}}_2$ decreases as the scores of Q_4, $Q10$, and Q_{11} increase. This component can be interpreted as a construct that combines these three items.
- In this example, we assume that the variable Q is measured by different items ($Q_1, Q_2, \ldots, Q_{11}$).
- The application of PCA *without rotation* resulted in retaining 4 principal components that explain a total variability of 78.43% (Eigenvalues > 1).
- *Following rotation* (Varimax), we conclude that:
  - $\hat{\boldsymbol{C}}\mathbf{1}$ alone captures 33.92% of the variance (eigenvalue = 3.7315). This factor is primarily represented by: Q_6, Q_7, Q_8.
  - $\hat{\boldsymbol{C}}\mathbf{2}$ alone captures 16.32% of the variance (eigenvalue = 1.795) and contributes to a cumulative variance of 50.25%. This factor is mainly represented by: Q_4, Q_{10}, Q_{11}.
  - $\hat{\boldsymbol{C}}\mathbf{3}$ alone captures 15.77% of the variance (eigenvalue = 1.734) and contributes to a cumulative variance of 66.02%. This factor is mainly represented by: $Q_1, Q9$.
  - $\hat{\boldsymbol{C}}\mathbf{4}$ alone captures 12.41% of the variance (eigenvalue = 1.305) and contributes to a cumulative variance of 78.43%. This factor is mainly represented by: $Q_2, Q3, Q_5$.
- Q is measured by 4 constructed, non-collinear (orthogonal) components. Consequently, regressions can be performed.
- Each construct is a linear combination of the items that constitute it.

IV. Diagnostic Tools

When conducting Principal Component Analysis (PCA), it's often necessary to explore various approaches to reach a satisfactory solution. This may involve multiple analyses to assess result relevance.

A crucial step is reviewing the correlation matrix. This involves checking for excessively high correlations between variables and evaluating the overall quality of the data representation.

An important criterion is that each variable should have a factor loading greater than 0.30 for at least one factor, indicating a significant contribution to the data structure.

This process may require iteration, repeating the analysis until a simple and satisfactory solution is found that effectively summarizes the data structure.

The evaluation of PCA's relevance includes a subjective element: does grouping these elements make sense?

After rotation, there are many potential solutions, making it challenging to determine a single "correct" one. The solution should be seen as a plausible proposition consistent with the data, rather than an absolute "answer."

These considerations lead to two distinct situations:

1) When variables are perfectly correlated, a single factor axis can restore all (100%) of the available information.
2) Conversely, if variables are pairwise independent (i.e., orthogonal), the number of factors needed is equal to the number of variables.

To validate this data reduction, it is essential to calculate certain indicators:

- Bartlett's sphericity test, typically conducted before implementing PCA.
- The Kaiser-Meyer-Olkin (KMO) index, typically evaluated after PCA has been conducted.

These indicators verify the validity of the data reduction approach and ensure the robustness of the results obtained.

A. Sphericity test

This test aims to determine the extent to which the correlation matrix $\Re$ of the data (observed matrix) significantly deviates from the unit matrix (theoretical matrix under the null hypothesis H_0).

Decision rule :

$$\begin{cases} H_0: & |\Re| = 1 \\ H_1: & |\Re| \neq 1 \end{cases} \Leftrightarrow \begin{cases} H_0: & \text{Determinant of } \Re = 1 \\ H_1: & \text{Determinant of } \Re \neq 1 \end{cases}$$

Test statistic :

$$\chi^2 = -(n - 1 - \frac{2p+5}{6}) \times ln\,|\Re|$$

Under H_0, it follows a χ^2 distribution with $[p \times \frac{(p-1)}{2}]$ degrees of freedom.

If the test result leads to rejecting the null hypothesis H_0 (significantly different from H_0), this implies that there are very strong redundancies in the data, indicating that they only contain one type of information ($|\Re| \neq 1$). In this case, the test statistic is such that $\chi^2 < \chi^{2\theta}$.

If the test result incorrectly leads us to reject the ”accepted” null hypothesis H_0 ($|\Re| = 1$), PCA will not be very useful because the variables are nearly orthogonal pairwise. In this case, $\chi^{2c} > \chi^{2\theta}$.

Bartlett’s test for sphericity should not be confused with Bartlett’s test for equality of variances, as they are two distinct tests despite their similar names.

B. The KMO index

The Kaiser-Meyer-Olkin (KMO) test assesses whether it is feasible to find a meaningful factor analysis of the data. In this context, the index compares the raw correlations with the partial correlations.

The concept behind partial correlation is that the raw correlation between two variables is influenced by the other $(p-2)$ variables.

Partial correlation is used to assess the net relationship between two variables by eliminating the influence of the other variables.

Decision rule :

The index KMO takes values between 0 and 1.

Test statistic:

$$kmo = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

- If the KMO index is close to 0, partial correlations are similar to raw correlations, indicating that effective data reduction is not possible and that the variables are pairwise orthogonal.
- If the KMO index is close to 1, it indicates that we can obtain an excellent summary of the information on the first factorial axes.

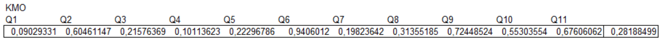
In practice : The decision rule adopted within the framework of ACP is as follows:

KMO	< 0.5	0.5 – 0.6	0.6 – 0.7
	Unacceptable	insufficient	poor
KMO	0.7 – 0.8	0.8 – 0.9	0.9 – 1
	Moderately	Good	Excellent

Table I
KMO Index Decision Rule

To perform this test in Excel, we first need to calculate the matrix of partial correlations by inverting the correlation matrix (denoted as F), which is then multiplied by the square root of the diagonal of matrix F with the addition of the identity matrix twice.

The KMO index is calculated using two matrices: the correlation matrix and the partial correlation matrix. Specifically, the KMO index is equal to the ratio of the sum of the squares of the correlations to the sum of the squares of the partial correlations.



KMO

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
0,09029331	0,60461147	0,21576369	0,10113623	0,22296786	0,9406012	0,19823642	0,31355185	0,72448524	0,55303554	0,67606062	0,28188499

Figure 4. Calculation of the KMO index

Based on the results from our hypothetical data, a KMO index of 0.28 indicates a relatively poor fit of the data for this technique. This suggests that the variables in the dataset are not highly interdependent and are not suitable for dimensional reduction using Factor Analysis.

In essence, a KMO value below 0.5 typically indicates that Factor Analysis may not be appropriate or reliable for that particular dataset. This suggests that alternative analysis methods might be more suitable, or that the data itself may need to be revised or transformed for better utilization in this analytical context.

C. Study of the distributions of constructed variables

The goal is to calculate the ”average” score obtained by each individual on the different items related to the same construct. This is done to test the normality of the constructs obtained from the PCA analysis.

In practice, we need to create one or more new variables (factors) called scores. These scores are then used to study the distribution of the constructs.

In Stata, we use the command predict score1, score to calculate these scores based on the results of the PCA (i.e., after rotating the PCA).

Analyzing the distributions of the newly generated variables is crucial. It is important to visually inspect and compare these distributions to a Gaussian distribution.

To do this, providing descriptive statistics of the scores is essential, as it offers a detailed overview of these new variables. Additionally, for visual representation, creating graphs that illustrate the distribution of scores is necessary. This typically involves generating a histogram of scores with a normal distribution curve and plotting a quantile-quantile plot (q-q plot) to assess the normality of the data.

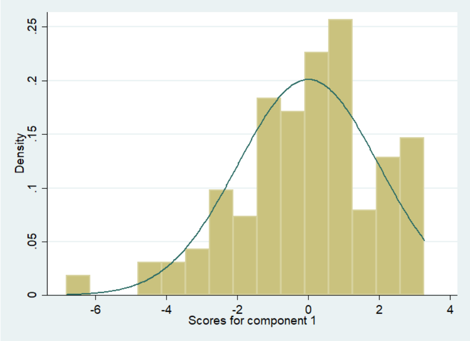


Figure 5. Distribution of the first principal component

The analysis of this graph confirms that the latent variable named ”score1” does not follow a normal distribution, as indicated by its lack of symmetry and its leptokurtic shape.

CONCLUSION

In this study, we investigated the application of factor analysis to survey data, emphasizing its robustness in dealing with various challenges encountered in real-world scenarios. Our findings indicate that despite the unique context of data collection, these techniques demonstrate a remarkable ability to accurately and coherently process complex data. While their use in such contexts requires careful attention, it remains viable and fruitful for uncovering the underlying structures of multidimensional data.

This study underscores the significance of considering the reliability of measurement instruments in surveys, as it directly impacts the quality of the analyzed data. By outlining key recommendations for conducting reliable analyses, we aim to assist practitioners and researchers in adopting robust methodologies, thereby ensuring the meaningful and reliable interpretation of results.

In conclusion, this comprehensive exploration of the robustness of factorial techniques in survey data sheds light on their practical utility. It also underscores the ongoing need for research aimed at enhancing these methods, while highlighting their relevance and potential for accurate and meaningful analyses in complex survey data scenarios.

REFERENCES

[1] Bourass H. & Soussi Noufail O. (2022) Principal component Analysis applied to servey data : Methodological scpects & application. International Journal of Optimisation and applications.

[2] Byrne, B. M. (2016). *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming* (3rd ed.). Routledge.

[3] Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1-9.

[4] Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates, Inc.

[5] Gorsuch, R. L. (1988). Exploratory Factor Analysis: Its Role in Item Analysis. *Journal of Personality Assessment*, 52(3), 355-369.

[6] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis*. Pearson Prentice Hall.

[7] Kaiser, H. F. (1974). An Index of Factorial Simplicity. *Psychometrika*, 39(1), 31-36.

[8] Norman, G. R., & Streiner, D. L. (2008). *Biostatistics: The Bare Essentials*. PMPH-USA.

[9] Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, 80(1), 99-103.

[10] Stevens, J. (2009). *Applied Multivariate Statistics for the Social Sciences*. Routledge.

[11] Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Pearson Education.

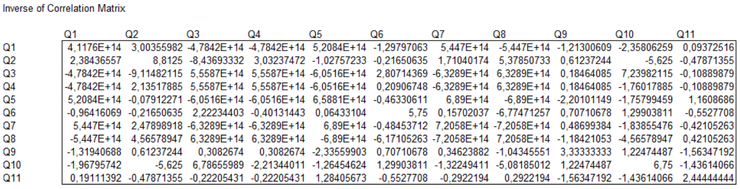
V. ANNEXE

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Score
1	1	1	1	1	1	1	1	1	1	1	1	11
2	1	1	1	1	1	1	1	1	0	1	0	9
3	1	0	1	1	1	1	1	1	1	0	0	8
4	1	1	1	0	1	1	0	1	1	0	0	7
5	1	1	1	1	1	0	0	0	1	0	0	6
6	0	1	1	0	1	1	1	1	0	0	0	6
7	1	1	1	1	0	0	1	0	0	0	0	5
8	1	1	1	1	1	0	0	0	0	0	0	5
9	0	1	0	1	1	0	0	0	0	1	0	4
10	1	0	0	1	0	1	0	0	0	0	0	3
11	1	1	1	0	0	0	0	0	0	0	0	3
12	1	0	0	1	0	0	0	0	0	0	0	2
$\sum$	10	9	9	9	8	6	5	5	4	3	1	**69**

Table II
EXAMPLE DATABASE

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
0,447	0,577	0,577	0,577	0,707	1,000	1,183	1,183	1,414	1,732	3,317
0,447	0,577	0,577	0,577	0,707	1,000	1,183	1,183	-0,707	1,732	-0,302
0,447	-1,732	0,577	0,577	0,707	1,000	1,183	1,183	1,414	-0,577	-0,302
0,447	0,577	0,577	-1,732	0,707	1,000	-0,845	1,183	1,414	-0,577	-0,302
0,447	0,577	0,577	0,577	0,707	-1,000	-0,845	-0,845	1,414	-0,577	-0,302
-2,236	0,577	0,577	-1,732	0,707	1,000	1,183	1,183	-0,707	-0,577	-0,302
0,447	0,577	0,577	0,577	-1,414	-1,000	1,183	-0,845	-0,707	-0,577	-0,302
0,447	0,577	0,577	0,577	0,707	-1,000	-0,845	-0,845	-0,707	-0,577	-0,302
-2,236	0,577	-1,732	0,577	0,707	-1,000	-0,845	-0,845	-0,707	1,732	-0,302
0,447	-1,732	-1,732	0,577	-1,414	1,000	-0,845	-0,845	-0,707	-0,577	-0,302
0,447	0,577	0,577	-1,732	-1,414	-1,000	-0,845	-0,845	-0,707	-0,577	-0,302
0,447	-1,732	-1,732	0,577	-1,414	-1,000	-0,845	-0,845	-0,707	-0,577	-0,302

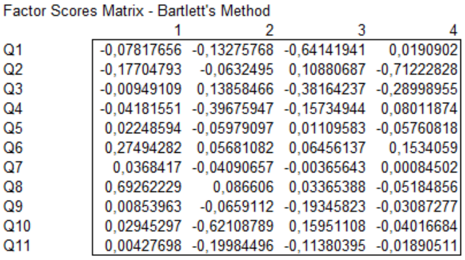
Table III
CALCULATION OF CENTERED VALUES



Inverse of Correlation Matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Q1	4,1176E+14	3,00355982	-4,7842E+14	-4,7842E+14	5,2084E+14	-1,29797063	5,447E+14	-5,447E+14	-1,21300609	-2,35806259	0,09372516
Q2	2,38436557	8,8125	-8,43693332	3,03237472	-1,02757233	-0,21650635	1,71040174	5,37850733	0,61237244	-5,625	-0,47871355
Q3	-4,7842E+14	-9,11482115	5,5587E+14	5,5587E+14	-6,0516E+14	2,80714369	-6,3289E+14	6,3289E+14	0,18464085	7,23982115	-0,10889879
Q4	-4,7842E+14	2,13517885	5,5587E+14	5,5587E+14	-6,0516E+14	0,20906748	-6,3289E+14	6,3289E+14	0,18464085	-1,76017885	-0,10889879
Q5	5,2084E+14	-0,07912271	-6,0516E+14	-6,0516E+14	6,5881E+14	-0,46330611	6,89E+14	-6,89E+14	-2,20101149	-1,75799459	1,1608686
Q6	-0,96416069	-0,21650635	2,22234403	-0,40131443	0,06433104	5,75	0,15702037	-6,77471257	0,70710678	1,29903811	-0,5527708
Q7	5,447E+14	2,47898918	-6,3289E+14	-6,3289E+14	6,89E+14	-0,48453712	7,2058E+14	-7,2058E+14	0,48699384	-1,83855476	-0,42105263
Q8	-5,447E+14	4,56578947	6,3289E+14	6,3289E+14	-6,89E+14	-6,17105263	-7,2058E+14	7,2058E+14	-1,18421053	-4,56578947	0,42105263
Q9	-1,31940688	0,61237244	0,3082674	0,3082674	-2,33559903	0,70710678	0,34623882	-1,04345551	3,33333333	1,22474487	-1,56347192
Q10	-1,96795742	-5,625	6,78655989	-2,21344011	-1,26454624	1,29903811	-1,32249411	-5,08185012	1,22474487	6,75	-1,43614066
Q11	0,19111392	-0,47871355	-0,22205431	-0,22205431	1,28405673	-0,5527708	-0,2922194	0,2922194	-1,56347192	-1,43614066	2,44444444

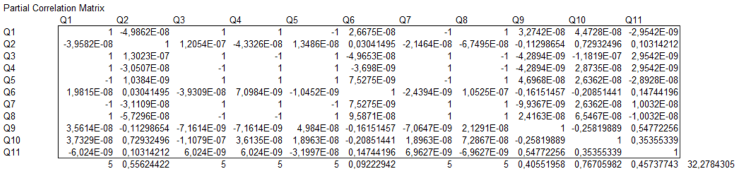
Figure 6. Inverse of the correlation matrix



Factor Scores Matrix - Bartlett's Method

	1	2	3	4
Q1	-0,07817656	-0,13275768	-0,64141941	0,0190902
Q2	-0,17704793	-0,0632495	0,10880687	-0,71222828
Q3	-0,00949109	0,13858466	-0,38164237	-0,28998955
Q4	-0,04181551	-0,39675947	-0,15734944	0,08011874
Q5	0,02248594	-0,05979097	0,01109583	-0,05760818
Q6	0,27494282	0,05681082	0,06456137	0,1534059
Q7	0,0368417	-0,04090657	-0,00365643	0,00084502
Q8	0,69262229	0,086606	0,03365388	-0,05184856
Q9	0,00853963	-0,0659112	-0,19345823	-0,03087277
Q10	0,02945297	-0,62108789	0,15951108	-0,04016684
Q11	0,00427698	-0,19984496	-0,11380395	-0,01890511

Figure 8. Score matrix



Partial Correlation Matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
Q1	1	-4,9862E-08	1	1	-1	2,6675E-08	-1	1	3,2742E-08	4,4728E-08	-2,9542E-09	
Q2	-3,9582E-08	1	1,2054E-07	-4,3326E-08	1,3486E-08	0,03041495	-2,1464E-08	-6,7495E-08	-0,11298654	0,72932496	0,10314212	
Q3	1	1,3023E-07	1	-1	1	-4,9653E-08	1	-1	-4,2894E-09	-1,1819E-07	2,9542E-09	
Q4	1	-3,0507E-08	-1	1	1	-3,698E-09	1	-1	-4,2894E-09	2,8735E-08	2,9542E-09	
Q5	-1	1,0384E-09	1	1	1	7,5275E-09	-1	1	4,6968E-08	2,6362E-08	-2,8928E-08	
Q6	1,9815E-08	0,03041495	-3,9309E-08	7,0984E-09	-1,0452E-09	1	-2,4394E-09	1,0525E-07	-0,16151457	-0,20851441	0,14744196	
Q7	-1	-3,1109E-08	1	1	-1	7,5275E-09	1	1	-9,9367E-09	2,6362E-08	1,0032E-08	
Q8	1	-5,7296E-08	-1	-1	1	9,5871E-08	1	1	2,4163E-08	6,5467E-08	-1,0032E-08	
Q9	3,5614E-08	-0,11298654	-7,1614E-09	-7,1614E-09	4,984E-08	-0,16151457	-7,0647E-09	2,1291E-08	1	-0,25819889	0,54772256	
Q10	3,7329E-08	0,72932496	-1,1079E-07	3,6135E-08	1,8963E-08	-0,20851441	1,8963E-08	7,2867E-08	-0,25819889	1	0,35355339	
Q11	-6,024E-09	0,10314212	6,024E-09	6,024E-09	-3,1997E-08	0,14744196	6,9627E-09	-6,9627E-09	0,54772256	0,35355339	1	
	5	0,55624422	5	5	5	0,09222942	5	5	0,40551958	0,76705982	0,45737743	32,2784305

Figure 7. Partial correlation matrix