Tutorials in Quantitative Methods for Psychology
2013, Vol. 9(2), p. 79-94.

A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis

An Gie Yong and Sean Pearce
University of Ottawa

The following paper discusses exploratory factor analysis and gives an overview of the statistical technique and how it is used in various research designs and applications. A basic outline of how the technique works and its criteria, including its main assumptions are discussed as well as when it should be used. Mathematical theories are explored to enlighten students on how exploratory factor analysis works, an example of how to run an exploratory factor analysis on SPSS is given, and finally a section on how to write up the results is provided. This will allow readers to develop a better understanding of when to employ factor analysis and how to interpret the tables and graphs in the output.

The broad purpose of factor analysis is to summarize data so that relationships and patterns can be easily interpreted and understood. It is normally used to regroup variables into a limited set of clusters based on shared variance. Hence, it helps to isolate constructs and concepts. Factor analysis uses mathematical procedures for the simplification of interrelated measures to discover patterns in a set of variables (Child, 2006). Attempting to discover the simplest method of interpretation of observed data is known as parsimony, and this is essentially the aim of factor analysis (Harman, 1976).

Factor analysis has its origins in the early 1900's with Charles Spearman's interest in human ability and his development of the Two-Factor Theory; this eventually lead to a burgeoning of work on the theories and mathematical principles of factor analysis (Harman, 1976). The method involved using simulated data where the answers were already known to test factor analysis (Child, 2006). Factor analysis is used in many fields such as behavioural and social sciences, medicine, economics, and geography as a result of the technological advancements of computers.

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). This tutorial will be focusing on EFA by providing fundamental theoretical background and practical SPSS techniques. EFA is normally the first step in building scales or a new metrics. Finally, a basic guide on how to write-up the results will be

---

Note that both Sean Pearce and An Gie Yong should be considered as first authors as they contributed equally and substantially in the preparation of this manuscript. The authors would like to thank Dr. Louise Lemyre and her team, Groupe d'Analyse Psychosociale Santé (GAP-Santé), for their generous feedback; in particular, Dr. Lemyre who took the time to provide helpful suggestions and real world data for the tutorial. The authors would also like to thank Levente Orbán and Dr. Sylvain Chartier for their guidance.

The original data collection was funded by PrioNet Center of Excellence, the McLaughlin Chair in Psychosocial Aspects of Health and Risk, and a SSHRC grant to Louise Lemyre, Ph.D., FRSC, with the collaboration of Dr. Daniel Krewski.

Address correspondence to An Gie Yong, Groupe d'Analyse Psychosociale Santé, GAP-Santé, University of Ottawa, Social Sciences Building, 120 University Street, room FSS-5006, Ottawa, Ontario, K1N 6N5, Canada. Email: ayong089@uottawa.ca.

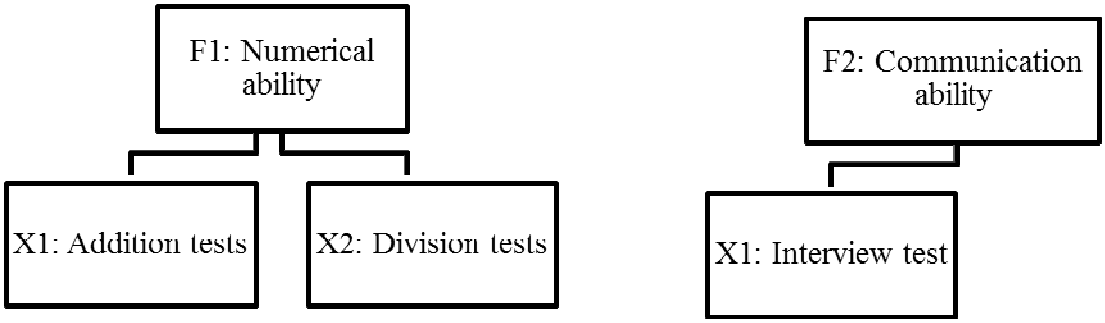

Figure 1. Graphical representation of the types of factor in factor analysis where numerical ability is an example of common factor and communication ability is an example of specific factor.

outlined.

A Look at Exploratory Factor Analysis

What is Factor Analysis?

Factor analysis operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable, which is known as reducing dimensionality (Bartholomew, Knott, & Moustaki, 2011). These un-observable factors are not directly measured but are essentially hypothetical constructs that are used to represent variables (Cattell, 1973). For example, scores on an oral presentation and an interview exam could be placed under a factor called 'communication ability'; in this case, the latter can be inferred from the former but is not directly measured itself.

EFA is used when a researcher wants to discover the number of factors influencing variables and to analyze which variables 'go together' (DeCoster, 1998). A basic hypothesis of EFA is that there are *m* common 'latent' factors to be discovered in the dataset, and the goal is to find the smallest number of common factors that will account for the correlations (McDonald, 1985). Another way to look at factor analysis is to call the dependent variables 'surface attributes' and the underlying structures (factors) 'internal attributes' (Tucker & MacCallum, 1997). Common factors are those that affect more than one of the surface attributes and specific factors are those which only affect a particular variable (see Figure 1; Tucker & MacCallum, 1997).

Why Use Factor Analysis?

Large datasets that consist of several variables can be reduced by observing 'groups' of variables (i.e., factors) – that is, factor analysis assembles common variables into descriptive categories. Factor analysis is useful for studies that involve a few or hundreds of variables, items from questionnaires, or a battery of tests which can be reduced to a smaller set, to get at an underlying concept, and to facilitate interpretations (Rummel, 1970). It is easier to focus on some key factors rather than having to consider too many variables that may be trivial, and so factor analysis is useful for placing variables into meaningful categories. Many other uses of factor analysis include data transformation, hypothesis-testing, mapping, and scaling (Rummel, 1970).

What are the Requirements for Factor Analysis?

To perform a factor analysis, there has to be univariate and multivariate normality within the data (Child, 2006). It is also important that there is an absence of univariate and multivariate outliers (Field, 2009). Also, a determining factor is based on the assumption that there is a linear relationship between the factors and the variables when computing the correlations (Gorsuch, 1983). For something to be labeled as a factor it should have at least 3 variables, although this depends on the design of the study (Tabachnick & Fidell, 2007). As a general guide, rotated factors that have 2 or fewer variables should be interpreted with caution. A factor with 2 variables is only considered reliable when the variables are highly correlated with each another ($r > .70$) but fairly uncorrelated with other variables.

The recommended sample size is at least 300 participants, and the variables that are subjected to factor analysis each should have at least 5 to 10 observations (Comrey & Lee, 1992). We normally say that the ratio of respondents to variables should be at least 10:1 and that the factors are considered to be stable and to cross-validate with a ratio of 30:1. A larger sample size will diminish the error in your data and so EFA generally works better with larger sample sizes. However, Guadagnoli and Velicer (1988) proposed that if the dataset has several high factor loading scores (> .80), then a smaller small size ($n > 150$) should be sufficient. A factor loading for a variable is a measure of how much the variable contributes to the factor; thus, high

factor loading scores indicate that the dimensions of the factors are better accounted for by the variables.

Next, the correlation *r* must be .30 or greater since anything lower would suggest a really weak relationship between the variables (Tabachnick & Fidell, 2007). It is also recommended that a heterogeneous sample is used rather than a homogeneous sample as homogeneous samples lower the variance and factor loadings (Kline, 1994). Factor analysis is usually performed on ordinal or continuous variables, although it can also be performed on categorical and dichotomous variables[1]. If your dataset contains missing values, you will have to consider the sample size and if the missing values occur at a nonrandom pattern. Generally speaking, cases with missing values are deleted to prevent overestimation (Tabachnick & Fidell, 2007). Finally, it is important that you check for an absence of multicollinearity and singularity within your dataset by looking at the Squared Multiple Correlation (SMC; Tabachnick & Fidell, 2007). Variables that have issues with singularity (i.e., SMC close to 0) and multicollinearity (SMC close to 1.0) should be removed from your dataset.

Limitations

One of the limitations of this technique is that naming the factors can be problematic. Factor names may not accurately reflect the variables within the factor. Further, some variables are difficult to interpret because they may load onto more than one factor which is known as split loadings. These variables may correlate with each another to produce a factor despite having little underlying meaning for the factor (Tabachnick & Fidell, 2007). Finally, researchers need to conduct a study using a large sample at a specific point in time to ensure reliability for the factors. It is not recommended to pool results from several samples or from the same sample at different points in time as these methods may obscure the findings (Tabachnick & Fidell, 2007). As such, the findings from factor analysis can be difficult to replicate.

Theoretical Background: Mathematical and Geometric Approach

Broadly speaking, there are many different ways to express the theoretical ideas behind factor analysis. Therefore, we will just focus on basic mathematical and geometric approaches.

Mathematical Models

In the 'classical factor analysis' mathematical model, p denotes the number of variables ($X_1, X_2,\ldots,X_p$) and m denotes the number of underlying factors ($F_1, F_2,\ldots,F_m$). X_j is the variable represented in latent factors. Hence, this model assumes that there are m underlying factors whereby each observed variables is a linear function of these factors together with a residual variate. This model intends to reproduce the maximum correlations.

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \ldots\ldots a_{jm}F_m + e_j \qquad (1)$$

where $j = 1, 2, \ldots, p$.

The factor loadings are $a_{j1}, a_{j2},\ldots,a_{jm}$ which denotes that a_{j1} is the factor loading of j^{th} variable on the 1st factor. The specific or unique factor is denoted by e_j. The factor loadings give us an idea about how much the variable has contributed to the factor; the larger the factor loading the more the variable has contributed to that factor (Harman, 1976). Factor loadings are very similar to weights in multiple regression analysis, and they represent the strength of the correlation between the variable and the factor (Kline, 1994).

Factor analysis uses matrix algebra when computing its calculations. The basic statistic used in factor analysis is the correlation coefficient which determines the relationship between two variables. Researchers cannot run a factor analysis until 'every possible correlation' among the variables has been computed (Cattell, 1973). The researcher examines if variables have some features in common and then computes a correlation or covariance matrix (Rummel, 1970). Generally, a factor analysis performed using a correlation matrix produces standardized data, thus it is recommended for variables that are not meaningfully comparable (e.g., items from different scales). On the other hand, factor analysis performed using a covariance matrix is conducted on variables that are similar (e.g., items from the same scales). The correlation matrix is often used because it is easier to interpret compared to the covariance tables, although there is not a strict requirement for which matrix to use (Fung, 1995).

The diagonal element of the matrix is always the value 1 (i.e., the correlation of a variable within itself). In principal components analysis, the diagonal values of the correlation matrix, 1s, are used for the analysis. Conversely, computation for the factor analysis techniques involves replacing the diagonal element of the matrix with the prior communality estimates (h^2). The communality estimate is the estimated proportion of variance of the variable that is free

[1] The limitations and special considerations required when performing factor analysis on categorical and dichotomous variables are beyond the scope of this paper. We suggest referring to 'Recent Developments in the Factor Analysis of Categorical Variables' by Mislevy (1986) and 'Factor Analysis for Categorical Data' by Bartholomew (1980) for further explanation.

Table 1. The sums of square of each factor loading (artificial data) for variable1 can be used to produce the communality score for that variable

	Factor1	Factor2	Factor3	Factor4
Variable1	0.56	0.43	0.41	0.33

of error variance and is shared with other variables in the matrix. These estimates reflect the variance of a variable in common with all others together. Factor analysis is also rooted in regression and partial correlation theory so analyzing it from this perspective may shed light on the theories behind this technique (McDonald, 1985).

To understand how factor analysis works, suppose that $X_i, X_j, \ldots, X_p$ are variables and $F_1, F_2, \ldots, F_m$ are factors. For all pairs X_i and X_j, we want to find factors such that when they are extracted, there is an absence of partial correlation between the tests – that is, the partial correlations are zero (Jöreskog & Sörbom, 1979). The basic idea behind this model is that factor analysis tries to look for factors such that when these factors are extracted, there remain no intercorrelations between any pairs X_i and X_j because the factors themselves will account for the intercorrelations. This means that for all pairs of any two elements, X_i, $X_{j,\ldots,}X_p$, are conditionally independent given the value of $F_1, F_{2,\ldots,}F_m$. Once a correlation matrix is computed, the factor loadings are then analyzed to see which variables load onto which factors. In matrix notation, factor analysis can be described by the equation $R = P\,C\,P' + U^2$, where R is the matrix of correlation coefficients among observed variables, P is the primary factor pattern or loading matrix (P' is the transpose), C is the matrix of correlations among common factors, and U^2 is the diagonal matrix or unique variances (McDonald, 1985).

The fundamental theorem of factor analysis, which is used in the common factor analysis model, is illustrated in the equation $R_{m\times m} - U^2_{m\times m} = F_{m\times p}F'_{p\times m}$, where R_{mxm} denotes the correlation matrix, U^2_{mxm} is the diagonal matrix of unique variances of each variable, and F_{mxp} represents the common factor loadings. The left-hand side of the equation represents the correlation matrix of the common parts. Since U^2 is the unique variances, when we subtract this out of R then it gives us the common variance (Rummel, 1970). Finding F_{mxp} can be solved by determining the eigenvalues and eigenvectors of the matrix. Essentially, this equation describes which variable is a linear combination of which common factors.

Geometrical Approach

Factor analysis can be examined through a geometrical approach to gain a better understanding of how the technique works. In a coordinate system, the factors are represented by the axes and the variables are lines or vectors (Cattell, 1973). When a variable is in close proximity to a certain factor, this means that the variable is associated with that particular factor. When there are more than three factors, this exceeds the three-dimensional space thus the dimensions are represented in hyperspace (Harman, 1976). Figure 2 shows two factors and the variables plotted as a function of the factors.

The factor axes act as a reference frame to determine where the data-variable vectors can be placed by giving factor loadings or coordinates – that is, the numerical labels on the axes represent factor loadings (Comrey & Lee, 1992). The length of the vector is equal to the square root of the communalities; variance explained by the common factors. Using Pythagorean Theorem ($a^2 + b^2 = c^2$) the squared hypotenuse can be found if the other two variables are known by the following formula: $c^2 = \sqrt{a^2 + b^2}$. The cosine of the angle between the variable and the factor gives insight to the correlation between each variable and each factor (Gorsuch, 1983). The correlation between a vector and one of the factors or with another variable (vector) can be determined as a function of the angle between them. In the equation $r_{12} = h_1 h_2 \cos\alpha_{12}$, the length of the vector is represented by h. The length of the first vector times the length of the second one times the cosine of the angle between the two vectors will give the correlation. Since all the variance in a factor is included in the dimension that it defines, its length is 1.0 (Gorsuch, 1983).

Variance

Factor analysis uses variances to produce communalities between variables. The variance is equal to the square of the factor loadings (Child, 2006). In many methods of factor analysis, the goal of extraction is to remove as much common variance in the first factor as possible (Child, 2006). The communality is the variance in the observed variables which are accounted for by a common factor or *common variance* (Child, 2006). The communality is denoted by h^2 and is the summation of the squared correlations of the variable with the factors (Cattell, 1973). The formula for deriving the communalities is $h_j^2 = a_{j1}^2 + a_{j2}^2 \ldots\cdots + a_{jm}^2$ where a equals the loadings for j variables. Using the factor loadings in Table 1, we then calculate the communalities using the aforementioned formula, thus $h_j^2 = 0.56^2 + 0.43^2 + 0.41^2 + 0.33^2$ = 0.78. The values in the table represent the factor loadings and how much the variable contributes to

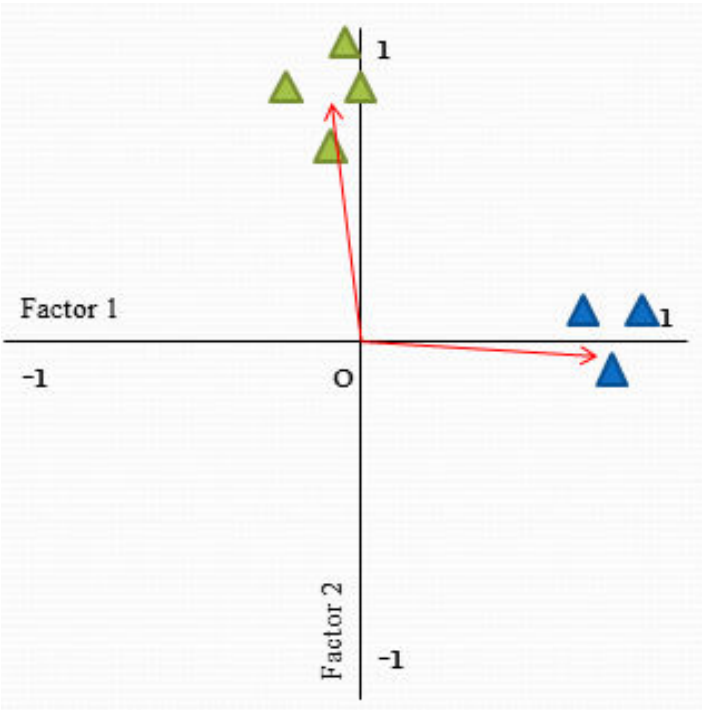

Figure 2. A geometrical representation of factor analysis in two-dimensional space where the blue triangles load onto factor 1 and the green triangles load onto factor 2.

each factor - in this case, it contributes the most to Factor1. The calculated communality shown above means that 78% of variable1 can be predicted based on the knowledge of the four factors; hence, the communality is the variance accounted for by the common factors. A particular set of factors is said to explain a lot of the variance of a variable if it has a high communality (Kline, 1994). Often times variables with low communalities (less than .20 so that 80% is unique variance) are eliminated from the analysis since the aim of factor analysis is to try and explain the variance through the common factors (Child, 2006).

A second type of variance in factor analysis is the *unique variance*. The unique variance is denoted by u^2 and is the proportion of the variance that excludes the common factor variance which is represented by the formula $u^2 = 1 - h^2$ (Child, 2006). In the case of the example above, if we know that the communality is 0.78, then $u^2 = 1 - 0.78 = 0.22$. Hence, we can say that 22% of the variance is specific to variable1. Unique variance can be split into specific variance and error variance, the latter referred to as the unreliability of the variance (Harman, 1976). The communality, the specificity and the unreliability comprise the total variance of a variable. The formula $V_{Total} = V_{Common} + V_{Specific} + V_{Error}$ is used to represent the total variance in factor analysis models.

In terms of the variance, the unique factors are never correlated with the common factors; however, the common factors may be uncorrelated or correlated with each other (Harman, 1976). Generally, the cumulative percentage of variance is extracted after each factor is removed from the matrix, and this cycle continues until approximately 75-85% of the variance is accounted for (Gorsuch, 1983). The percentage variance tells us how much each factor contributed to the total variance.

Components of Factor Analysis

Factor Extraction

Factor analysis is based on the 'common factor model' which is a theoretical model. This model postulates that observed measures are affected by underlying common factors and unique factors, and the correlation patterns need to be determined. There is an array of extraction methods[2] available, but we will briefly touch on a few commonly used techniques that are available on SPSS. Maximum Likelihood attempts to analyze the maximum likelihood of sampling the observed correlation matrix (Tabachnick & Fidell, 2007). Maximum Likelihood is more useful for confirmatory factor analysis and is used to estimate the factor loadings for a population. The Principal Axis Factor method is based on

[2] A useful summary of extraction methods can be found in Table 13.7 (p. 633) in 'Using Multivariate Statistics (5th ed.)' by Tabachnick and Fidell (2007).

the notion that all variables belong to the first group and when the factor is extracted, a residual matrix is calculated. Factors are then extracted successively until there is a large enough of variance accounted for in the correlation matrix (Tucker & MacCallum, 1997). Principal Axis Factor is recommended when the data violate the assumption of multivariate normality (Costello & Osborne, 2005).

Principal Components analysis is used to extract maximum variance from the data set with each component thus reducing a large number of variables into smaller number of components (Tabachnick & Fidell, 2007). Principal Components analysis is a data reduction technique and the issues of whether it is truly a factor analysis technique has been raised (Costello & Osborne, 2005). That is, Principal Components produces *components* whereas Principal Axis Factor produces *factors.* There are also differences in how the correlation matrix is constructed and how the communalities are calculated when comparing these techniques (Kline, 1994; Tucker & MacCallum, 1997). Researchers may use Principal Components analysis as the first step to reduce the data, then follow-up with a 'true' factor analysis technique. Overall, the factor loadings are fairly similar and you will need to perform rotation regardless of the extraction technique (Tabachnick & Fidell, 2007). It is best to pick the extraction technique based on your research question and the ease of interpretation.

Rotation Methods

Factors are rotated for better interpretation since unrotated factors are ambiguous. The goal of rotation is to attain an optimal simple structure which attempts to have each variable load on as few factors as possible, but maximizes the number of high loadings on each variable (Rummel, 1970). Ultimately, the simple structure attempts to have each factor define a distinct cluster of interrelated variables so that interpretation is easier (Cattell, 1973). For example, variables that relate to language should load highly on language ability factors but should have close to zero loadings on mathematical ability.

Broadly speaking, there are orthogonal rotation and oblique rotation[3]. Orthogonal rotation is when the factors are rotated 90° from each other, and it is assumed that the factors are uncorrelated (DeCoster, 1998; Rummel, 1970). This is less realistic since factors generally are correlated with each other to some degree (Costello & Osborne, 2005). Two common orthogonal techniques are Quartimax and Varimax rotation. *Quartimax* involves the minimization of the number of factors needed to explain each variable (Gorsuch, 1983). *Varimax* minimizes the number of variables that have high loadings on each factor and works to make small loadings even smaller.

Oblique rotation is when the factors are not rotated 90° from each other, and the factors are considered to be correlated. Oblique rotation is more complex than orthogonal rotation, since it can involve one of two coordinate systems: a system of primary axes or a system of reference axes (Rummel, 1970). Additionally, oblique rotation produces a pattern matrix that contains the factor or item loadings and factor correlation matrix that includes the correlations between the factors. The common oblique rotation techniques are Direct Oblimin and Promax. *Direct Oblimin* attempts to simplify the structure and the mathematics of the output, while *Promax* is expedient because of its speed in larger datasets. *Promax* involves raising the loadings to a power of four which ultimately results in greater correlations among the factors and achieves a simple structure (Gorsuch, 1983).

Interpretations of Factor Loadings

When interpreting the factors, you need to look at the loadings to determine the strength of the relationships. Factors can be identified by the largest loadings, but it is also important to examine the zero and low loadings in order to confirm the identification of the factors (Gorsuch, 1983). For example if you have a factor called 'anxiety' and variables that load high on this factor are 'heartbeat' and 'perspiration', you also need to make sure that a variable such as 'lethargy' does not load onto this factor. There should be few item crossloadings (i.e., split loadings) so that each factor defines a distinct cluster of interrelated variables. A crossloading is when an item loads at .32 or higher on two or more factors (Costello & Osborne, 2005). Depending on the design of the study, a complex variable (i.e., an item that is in the situation of crossloading) can be retained with the assumption that it is the latent nature of the variable, or the complex variable can be dropped when the interpretation is difficult. Another option is to choose a significant loading cut-off to make interpretation easier. The signs of the loadings show the direction of the correlation and do not affect the interpretation of the magnitude of the factor loading or the number of factors to retain (Kline, 1994).

Researchers will also need to determine the cut-off for a statistically meaningful rotated factor loading. A general rule to determine the reliability of the factor is to look at the relationship between the individual rotated factor loading and the magnitude of the absolute sample size. That is, the larger the sample size, smaller loadings are allowed for a

[3] A summary of the rotation techniques can be found in Table 13.9 (p. 639) in 'Using Multivariate Statistics (5th ed.)' by Tabachnick and Fidell (2007).

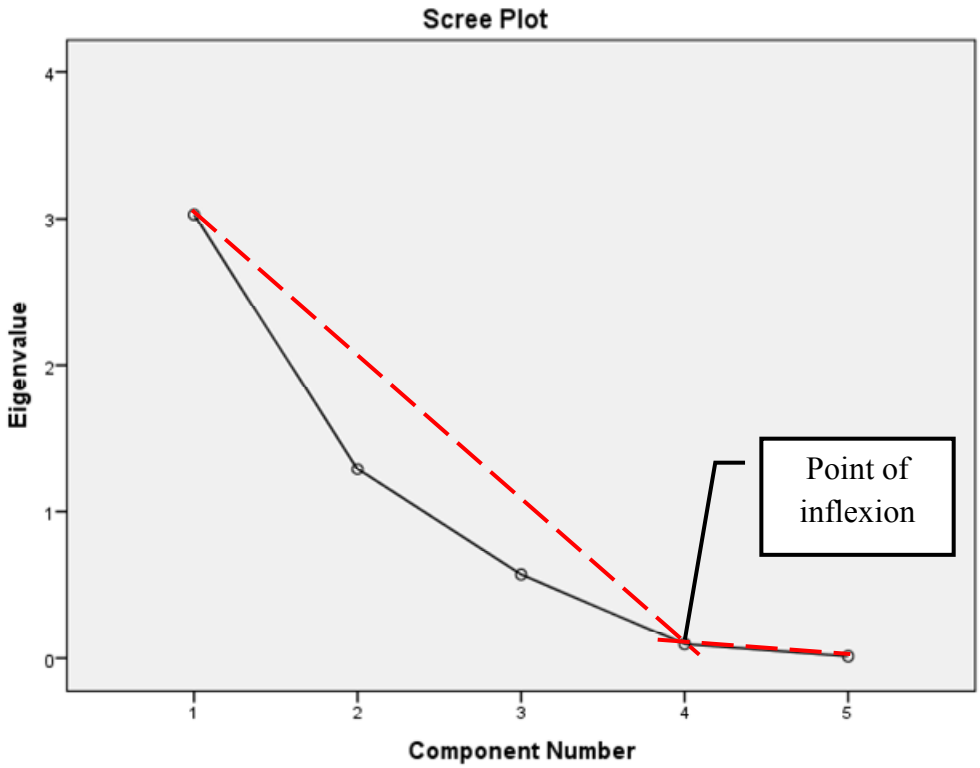

Figure 3. Example of scree test or scree plot for data that most likely have 3 underlying factors.

factor to be considered significant (Stevens, 2002). According to a rule of thumb, using an alpha level of .01 (two-tailed), a rotated factor loading for a sample size of at least 300 would need to be at least .32 to be considered statistically meaningful (Tabachnick & Fidell, 2007). A factor loading of .32 gives us approximately 10% of the overlapping variance *% overlapping variance = (Factor loading)*^{2}. The choice of cut-off may depend on the ease of interpretation including how complex variables are being handled.

Number of Factors to Retain

Extracting too many factors may present undesirable error variance but extracting too few factors might leave out valuable common variance. So it is important to select which criterion is most suitable to your study when deciding on the number of factors to extract. The eigenvalues and scree test (i.e., scree plot) are used to determine how many factors to retain. One criterion that can be used to determine the number of factors to retain is *Kaiser's criterion* which is a rule of thumb. This criterion suggests retaining all factors that are above the eigenvalue of 1 (Kaiser, 1960). Another criterion is based on *Jolliffe's criterion* which recommends retaining factors above .70 (Jolliffe, 1986). It has been argued that both criteria may result in overestimation in the number of factors extracted (Costello & Osborne, 2005; Field, 2009); therefore, it is suggested to use the scree test in conjunction with the eigenvalues to determine the number of factors to retain.

The scree test (see Figure 3) consists of eigenvalues and factors (Cattell, 1978). The number of factors to be retained is the data points that are above the break (i.e., point of inflexion). To determine the 'break', researchers draw a horizontal line and a vertical line starting from each end of the curve. The scree test is only reliable when you have a sample size of at least 200. In situations when the scree test is hard to interpret (e.g., clustered data points at the point of inflexion), you will need to rerun the analysis several times and manually set the number of factors to extract each time (Costello & Osborne, 2005). The number of factors to extract should be set once at the number based on the a priori factor structure, once at the number of factors predicted by the scree test, at the numbers above and below the number based on the a priori factor structure, and at the numbers above and below the number of factors suggested by the scree test. You would end up with a set of four numbers if the number of factors from the scree test is different from the predicted number of factors, or a set of three numbers if the number of factors from the scree test is identical to the predicted number of factors. To determine the number of factors to retain, you will need to pick the solution that

Table 2. Options available in the 'Factor Analysis: Descriptive' dialog box in SPSS, and the descriptions of each option

Options	Descriptions
Univariate descriptives	Mean and standard deviation
Initial solution	Communalities estimate for the factors
Coefficient	R-matrix
Significance levels	Significance value matrix for the R-matrix
Determinant	Test for multicollinearity or singularity
KMO and Bartlett's	Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test
Inverse	Provides inverse of the correlation matrix
Reproduced	Correlation matrix for the model
Anti-image	Anti-image matrix of covariance and correlation

provides the most desirable rotated factor structure. Factors that have less than three variables, many complex variables and item loadings that are less than .32 are generally viewed as undesirable.

Factor Scores

A factor score can be considered to be a variable describing how much an individual would score on a factor. One of the methods to produce factor score is called *Bartlett method* (or regression approach) which produces unbiased scores that are correlated only with their own factor. Another method is called the *Anderson-Rubin* method which produces scores that are uncorrelated and standardized. The method that you choose will depend on your research question, but the Bartlett method is the most easily understood (Tabachnick & Fidell, 2007). Factor scores can be treated as variables for further statistical analyses of variables (e.g., ANOVA) or can be used to overcome the issue of multicollinearity as uncorrelated variables can be produced.

SPSS Tutorial

We begin our tutorial with an example of a national survey[4] that investigates the perception of food risks amongst Canadians (see Figure 4). In this study, we intend to determine the underlying construct of how Canadians perceive food risk; hence, the research question is *'what are the underlying mechanisms or factors that can produce correlation amongst the different types of food risk perception within the Canadian population?'*

[4] The original data collection was funded by PrioNet Center of Excellence, the McLaughlin Chair in Psychosocial Aspects of Health and Risk, and a SSHRC grant to Louise Lemyre, Ph.D., FRSC, with the collaboration of Dr. Daniel Krewski.

Running Exploratory Factor Analysis on SPSS

Prior to running EFA, we confirmed that all the requirements were met for EFA. In the SPSS dialog box, go to *Analyze* → *Dimension Reduction* → *Factor…* to launch *the Factor Analysis* dialog box (see Figure 5). We will move the variables from the left-hand box to the right-hand *Variables* box.

Step 1: Descriptives. We will select all the options in the *Descriptives* dialog box (see Figure 5). The description of each option is provided in Table 2.

Step 2: Extraction. We will select *Principal axis factoring* and the following options in Figure 5. *Correlation matrix* is used by default whereas *Covariance matrix* is used when the variables are commensurable. We have the option of customizing the eigenvalue cut-off so we will use Kaiser's criterion of 1.0. If you have any theoretical reasoning that you should be able to extract a particular number of factors, then select the *Fixed number of factor* option. We will select *Unrotated factor solution* and *Scree plot* to aid our interpretation. The *Unrotated factor solution* gives you the *Unrotated pattern matrix* which can be used to compare the factors before and after rotation.

Step 3: Rotation. For *Rotation* (see Figure 5), we will select *Varimax* as it is a recommended rotation technique to use when you start exploring the dataset. You may select oblique rotation if there is pre-existing evidence that the factors are correlated. *Rotated Solution* gives you the output for rotated factor interpretation and the output varies depending on the type of rotation you pick. *Loading plot(s)* are selected to produce a factor loading plot. Finally, the *Maximum Iterations for Convergence* is used to determine the number of times SPSS will search for an optimal solution. The default value is 25 which is usually sufficient for most analyses. If the value is too low for your analysis, you can pick a larger value when you have a large dataset. In our example, we can pick a larger value since we have a large

In the second section, I am going to read you a list of items related to food safety. I would like to get your opinion about the potential risk it can represent for the Canadian public. Please respond to the following questions using the same 5-point scale used previously. **The question is what level of risk to Canadians would you say there is related to the following:**

	Not at all	A little	Moderately	Very Much	Extremely	Don't Know/ No Opinion	No Response
Bacteria in food (e.g., E. coli, Salmonella)							
Pesticides							
Imported food							
Tap water							
Food irradiation (to preserve food)							
Use of antibiotics in livestock							
Mad cow disease							
Disease in wild game							
Foot and Mouth disease							
Food additives (def: chemicals used to preserve or color food or improve its taste)							
Bottled water							
Genetically modified foods							
Improper food labeling							
Mercury in fish							
Growth hormones							
Artificial sweeteners (aspartame, saccharin)							
Food packaging materials (food wrapped in plastics)							
Agroterrorism (def: deliberate introduction of harmful agents into the food chain)							

Figure 4. Questions about Canadians' perception of potential food risk taken from the National Public Survey on Risk Perceptions and Risk Acceptability of Prion Disease and Food Safety (Lemyre et al., 2008). Reprinted with permission.

dataset.

Step 4: Factor score and options. We will ask SPSS to produce factor scores (see Figure 5) as new variables using *Anderson-Rubin* method by selecting *Save as variables*. SPSS will then create new columns with the factors scores in your dataset. The *Display factor score coefficient matrix* shows the correlation between factors and the coefficients used to produce the factor scores through multiplication – this is not a mandatory option for interpretation.

Step 5: Options. We will set the missing values option and the coefficient display format. Next, we will select *Exclude cases listwise* (see Figure 5) to prevent overestimation of factors within our large dataset. For the ease of interpretation, we will select *Sorted by size* to display the loadings in a descending order and *Suppress small coefficients* using an *Absolute value below* .32.

Finally, we go back to the main dialog box and click *OK* to run the analysis.

Interpretation of the SPSS Output

Preliminary Interpretation

We will need to determine if our dataset is suitable for EFA. If you notice issues at this stage, you should resolve the issue and rerun the analysis. First, we check if there is a patterned relationship amongst our variables by referring to the *Correlation matrix* (see Figure 6). Variables that have a

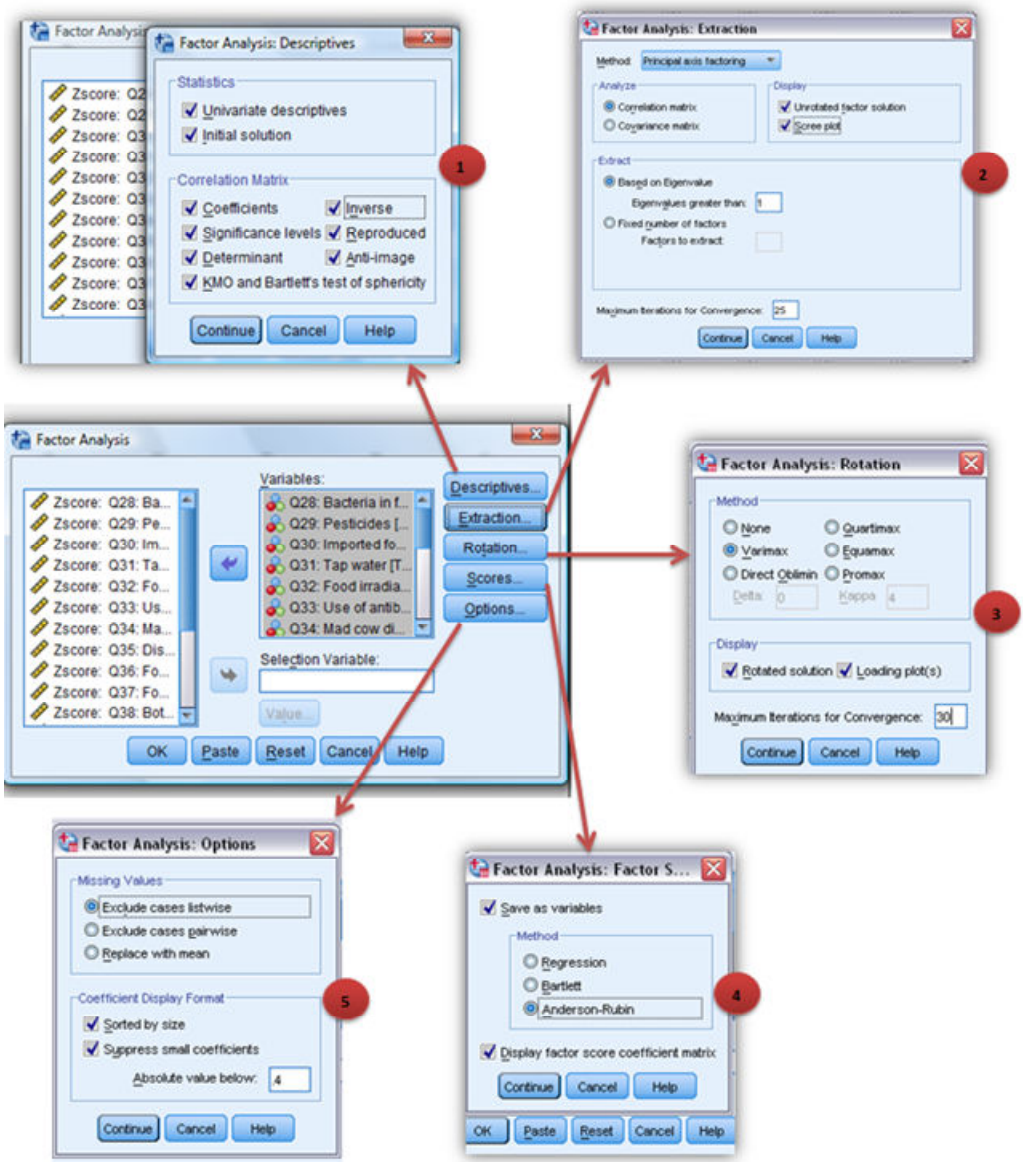

Figure 5. Sub-dialog options used in step 1 to 5 for running EFA on SPSS.

large number of low correlation coefficient (r < +/- .30) should be removed as they indicate a lack of patterned relationships. Furthermore, correlations that are above r = +/- .90 indicate that your data may have a problem of multicollinearity. As a follow-up, check if the *Determinant score* is above the rule of thumb of .00001 as this indicates an absence of multicollinearity. You may also use the Haitovsky's test (1969) to test if your *Determinant score* is significantly different from zero which indicates an absence of multicollinearity. If your data have an issue of multicollinearity, you will need to determine the item that is causing the problem and remove it from the analysis. We found that our example does not have an issue of multicollinearity and there seem to be patterned relationships amongst the variables.

Second, we will look at the *Bartlett's Test of Sphericity* (see Figure 7; significant level of p < .05) to confirm that our example has patterned relationships. Indeed, these tests show that we do have patterned relationships amongst the variables (p < .001). Finally, we will determine if our example is suitable for EFA by looking at the *Kaiser-Meyer-Olkin Measure (KMO) of Sampling Adequacy* (see Figure 7; cut-off above .50) and the diagonal element of the *Anti-Correlation* matrix that has the 'a' superscript (see Figure 8; cut-off of above .50). If this requirement is not met, this means that distinct and reliable factors cannot be produced. Hence, you may want to increase the sample size or remove the item that is causing diffused correlation patterns as indicated by the diagonal value in the Anti-Correlation matrix. Our example is suitable for EFA as the KMO is .94

		Bacteria	Pesticides	Imported_food
Correlation	Bacteria	1.000	.505	.334
	Pesticides	.505	1.000	.388
	Imported_food	.334	.388	1.000
	Tap_water	.294	.334	.301
	Food_irradiation	.317	.502	.340
	Antibiotics_food	.280	.510	.329
	Mad_cow	.357	.418	.276
	Wild_game	.378	.344	.245
	Foot_mouth	.397	.411	.274
	Food_additives	.269	.490	.395
	Bottled_water	.216	.298	.280
	GMO	.282	.497	.350
	Improper_label	.350	.475	.327
	Mercury_fish	.365	.476	.287
	Growth_hormones	.330	.540	.354
	Artificial_sweet	.257	.385	.301
	Food_packaging	.266	.442	.335

Figure 6. Truncated SPSS output for Correlation matrix. The Determinant score available below this matrix is not shown.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.935
Bartlett's Test of Sphericity	Approx. Chi-Square	8189.553
	df	153
	Sig.	.000

Figure 7. SPSS output for KMO and Bartlett's Test.

Anti-image Correlation	Bacteria	.928[a]	-.297	-.121
	Pesticides	-.297	.948[a]	-.062
	Imported_food	-.121	-.062	.967[a]
	Tap_water	-.078	-.030	-.097
	Food_irradiation	.018	-.124	-.003
	Antibiotics_food	.016	-.131	-.020
	Mad_cow	.009	-.030	-.005

Figure 8. Truncated SPSS output for the Anti-image correlation portion obtained from the Anti-image Matrices. The Anti-Image covariance portion is not shown

and the individual diagonal elements were > .90.

Factor Extraction and Rotation

We will look at the *Total Variance Explained* table (see Figure 9) to determine the number of significant factors. It is important to note that only extracted and rotated values are meaningful for interpretation. The factors are arranged in the descending order based on the most explained variance. The *Extraction Sums of Squared Loadings* is identical to the *Initial Eigenvalues* except factors that have eigenvalues less

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.493	41.625	41.625	6.999	38.882	38.882	3.165	17.586	17.586
2	1.593	8.850	50.475	1.178	6.547	45.429	2.957	16.430	34.016
3	1.072	5.955	56.430	.526	2.923	48.352	2.580	14.336	48.352
4	.905	5.026	61.456						
5	.772	4.287	65.743						

Figure 9. Truncated SPSS output for the total variance explained for extracted factors.

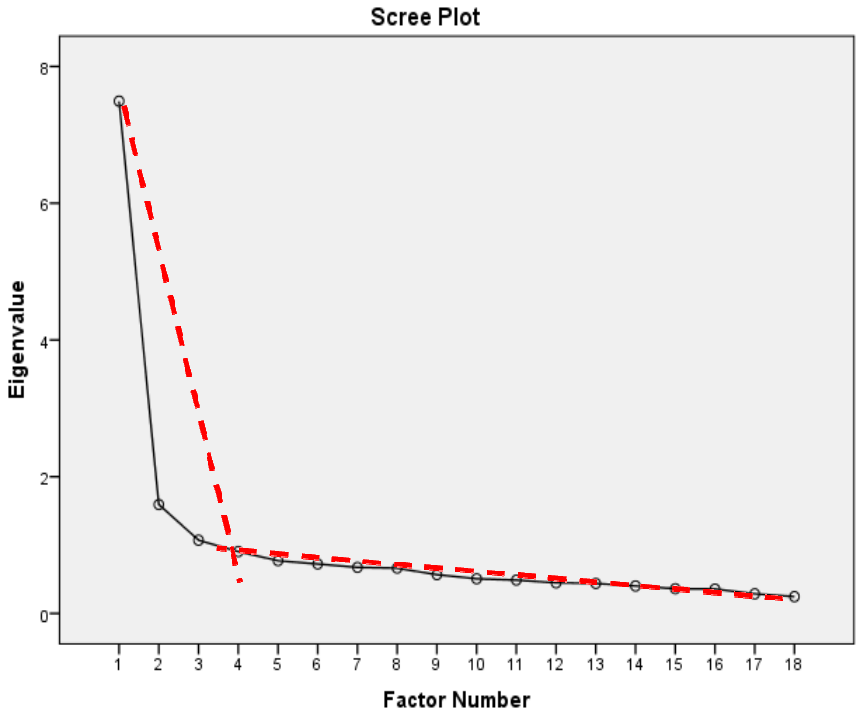

Figure 10. SPSS output for scree plot indicating that the data have three factors.

than 1 are not shown. These columns show you the eigenvalues and variance prior to rotation. The *Rotation Sums of Squared Loadings* show you the eigenvalues and variance after rotation. We will use the rotated eigenvalues and scree plot (see Figure 10) to determine the number of significant factors.

We can calculate the averaged extracted communalities[5] (see Figure 11) to determine the eigenvalue cut-off based on which criteria to follow. However, we will stick to Kaiser's criterion in this example for simplicity. Both methods indicate that we have 3 'meaningful' factors. Next, we will check if the model is a good fit by looking at the summary of the percentage of the non-redundant residuals at the *Reproduced Correlation Matrix* (see Figure 12). A model that is a good fit will have less than 50% of the non-redundant residuals with absolute values that are greater than .05 which is true for our example. We can also compare the *Reproduced Correlation Matrix* with the original *Correlation Coefficients Matrix*. If the model is a good fit, we should expect small residuals between the two matrices.

The *Factor Matrix* shows you the factor loadings prior to rotation whereas the *Rotated Factor Matrix* shows you the rotated factor loadings (see Figure 13). As illustrated in the figure, using rotation and suppressing small coefficients help with the interpretation. The factor loadings show that our factors are fairly desirable with at least 3 variables per factors that are above .32. However, our factors consist of many complex variables. At this step, we can choose a

[5] The Kaiser Criterion is said to be reliable when: a) the averaged extracted communalities is at least more than .70 and when there are less than 30 variables, or b) the averaged extracted communalities is equal or above .60 and the sample size is above 250 cases (Field, 2009).

Communalities

	Initial	Extraction
Bacteria	.347	.298
Pesticides	.520	.511
Imported_food	.270	.273
Tap_water	.292	.318
Food_irradiation	.479	.506
Antibiotics_food	.493	.503
Mad_cow	.595	.624
Wild_game	.490	.548
Foot_mouth	.641	.755
Food_additives	.495	.525
Bottled_water	.283	.334
GMO	.541	.565
Improper_label	.433	.432
Mercury_fish	.449	.458
Growth_hormones	.605	.739
Artificial_sweet	.427	.451
Food_packaging	.447	.483
Agroterrorism	.386	.381

Extraction Method: Principal Axis Factoring.

Figure 11. SPSS output for Communalities.

Extraction Method: Principal Axis Factoring.

a. Reproduced communalities
b. Residuals are computed between observed and reproduced correlations. There are 12 (7.0%) nonredundant residuals with absolute values greater than 0.05.

Figure 12. Truncated SPSS output for the summary of non-redundant residuals available below the Reproduced Correlation Matrix (not shown).

different significant loading cut-off of .40 based on pragmatic reasoning. To resolve the issue of non-significant loading item (e.g., imported food), we can rerun the analysis without that item or we can pick a lower cut-off if we cannot afford to exclude that item from our study.

Factor Plot produced using SPSS is only useful for interpretation when there are two or less factors (i.e., factors are represented in three-dimensional space when there are three factors whereas factors are represented in hyperspace when there are more than three factors). Hence, our *Factor Plot* (see Figure 14) is not useful for interpretation in this case because we have three factors. In sum, our example has three factors: a) industrial food processing risks (factor1), b) animal-borne food risks (factor2), and c) food packaging risks (factor3). Food additives and artificial sweet are complex variables as they load onto both factor1 and factor3 and factor1 and factor2, respectively.

Is the Rotation Technique Used Suitable?

To determine whether a rotation technique is suitable, we will look at the *Factor Transformation Matrix*'s off diagonal elements (see Figure 15). A suitable rotation technique will result in a nearly symmetrical off-diagonal element which is not true in this case. Hence, orthogonal rotation may not be a suitable rotation technique. This indicates that food risk perception factors may be correlated which is more theoretically realistic. Accordingly, we can repeat the analysis using an oblique rotation, but it will not be demonstrated in this tutorial for brevity. It is important to note that oblique rotations are more difficult to interpret. Therefore, Field (2009) suggests ignoring the *Factor Transformation Matrix* during interpretation of orthogonal rotation methods if you are not familiar with factor analysis techniques.

Final Steps: Naming the Factors, Writing the Results, and Factor Scores

Naming of factors is more of an 'art' as there are no rules for naming factors, except to give names that best represent the variables within the factors. An example of the write-up is outlined in Figure 16 with a truncated table reporting the rotated factor loading. Depending on your research questions, you may want to extend your findings. For instance, you may want to use the factor scores (see Figure 17) in a regression to predict behavioral outcomes using food risk perceptions. You could also run a Confirmatory Factor Analysis (CFA) to validate the factorial validity of the models derived from the results of your EFA. Finally, you

Factor Matrix[a]

	Factor		
	1	2	3
Growth_hormones	.741		-.352
Pesticides	.709		
GMO	.708		
Food_irradiation	.681		
Food_additives	.672		
Foot_mouth	.664	.560	
Antibiotics_food	.654		
Mad_cow	.653	.441	
Mercury_fish	.638		
Improper_label	.634		
Food_packaging	.632		
Artificial_sweet	.600		
Agroterrorism	.591		
Wild_game	.570	.472	
Bacteria	.523		
Imported_food	.510		
Bottled_water	.480		
Tap_water	.478		

Extraction Method: Principal Axis Factoring.

a. 3 factors extracted. 9 iterations required.

Rotated Factor Matrix[a]

	Factor		
	1	2	3
Growth_hormones	.802		
GMO	.614		.380
Antibiotics_food	.598		.344
Mercury_fish	.519	.399	
Pesticides	.514	.344	.359
Food_additives	.505		.497
Improper_label	.490	.385	
Foot_mouth		.833	
Mad_cow		.730	
Wild_game		.707	
Agroterrorism		.471	
Bacteria		.419	
Food_packaging	.341		.578
Food_irradiation	.349		.547
Bottled_water			.534
Artificial_sweet	.406		.525
Tap_water			.486
Imported_food			.377

Extraction Method: Principal Axis Factoring.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Figure 13. SPSS output for Factor Matrix before and after Varimax rotation to illustrate how rotation aids interpretation.

could perform reliability testing if you are using factor analysis to validate or construct a questionnaire.

Conclusion

Factor analysis is used to identify latent constructs or factors. It is commonly used to reduce variables into a smaller set to save time and facilitate easier interpretations. There are many extraction techniques such as Principal Axis Factor and Maximum Likelihood. Factor analysis is mathematically complex and the criteria used to determine the number and significance of factors are vast. There are two types of rotation techniques – orthogonal rotation and oblique rotation. Orthogonal rotation (e.g., Varimax and Quartimax) involves uncorrelated factors whereas oblique rotation (e.g., Direct Oblimin and Promax) involves correlated factors. The interpretation of factor analysis is based on rotated factor loadings, rotated eigenvalues, and scree test. In reality, researchers often use more than one extraction and rotation technique based on pragmatic reasoning rather than theoretical reasoning.

References

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society Series B (Methodological), 42*(3), 293-321.

Bartholomew, D., Knotts, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. (3rd ed.). West Sussex, UK: John Wiley & Sons.

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York, NY: Plenum Press.

Cattell, R.B. (1973). *Factor analysis.* Westport, CT: Greenwood Press.

Child, D. (2006). *The essentials of factor analysis*. (3rd ed.). New York, NY: Continuum International Publishing Group.

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

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. P*ractical Assessment, Research and Evaluation*, 10(7), 1-9.

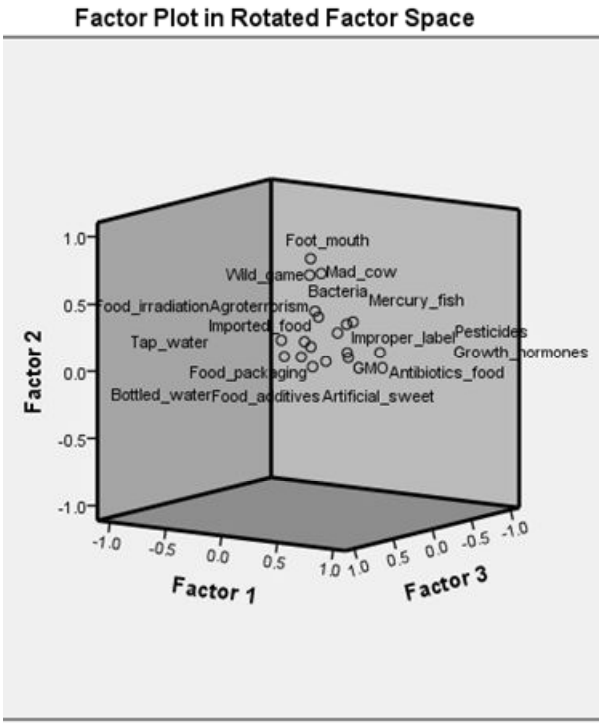

Figure 14. SPSS output for Factor Plot for three factors illustrated in two-dimensional space.

Factor Transformation Matrix

Factor	1	2	3
1	.623	.553	.553
2	-.438	.833	-.339
3	-.648	-.031	.761

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

Figure 15. SPSS output for Factor Transformation Matrix to determine if the chosen rotation technique is sufficient for this data.

DeCoster, J. (1998). *Overview of factor analysis.* Retrieved March 22, 2012 from http://www.stat-help .com/notes.html

Field, A. (2009). *Discovering Statistics Using SPSS: Introducing Statistical Method* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Fung, D. K. (1995). Sensitivity analysis in factor analysis: Difference between using covariance and correlation matrices. *Psychometrika, 60* (4), 607-617. doi: 10.1007/ BF02294330

Gorsuch, R.L. (1983). *Factor analysis (2nd ed.)*. Hillside, NJ: Lawrence Erlbaum Associates.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103* (2), 265-275. doi: 10.1037/0033-2909.103.2.265

Haitovsky, Y. (1969). Multicollinearity in regression analysis: A comment. *Review of Economics and Statistics*, *51* (4), 486-489.

Harman, H.H. (1976). *Modern factor analysis (3rd ed. revised)*. Chicago, IL: University of Chicago Press.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis, I: Artificial data. *Applied Statistics, 21*, 160-173.

Jöreskog, K.G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. doi: 10.1177/001316446002000116

Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

McDonald, R.P. (1985). *Factor analysis and related methods*. Hillside, NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*(1), 3-31.

Rummel, R.J. (1970). *Applied factor analysis.* Evanston, IL: Northwestern University Press.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NS: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston, MA: Allyn & Bacon.

Tucker, L.R., & MacCallum, R.C. (1997). *Exploratory factor analysis*. Retrieved March 27, 2012 from http://www.unc.edu/~rcm/book/ch7.pdf

Manuscript received 2 November 2012
Manuscript accepted 5 February 2013

Figures 16 and 17 follows.

Data were subjected to factor analysis using Principal Axis Factoring and orthogonal Varimax rotation. All KMO values for the individual items (> .90) were well above .5 and the Kaiser-Meyer-Olkin measure (KMO) was .94 indicating the data were sufficient for EFA. The Bartlett's test of sphericity χ^2 (153)= 8189.55, $p <$ .001 showed that there were patterned relationships between the items. Using an eigenvalue cut-off of 1.0, there were 3 factors that explain a cumulative variance of 48.35%. The scree plot confirmed the findings of retaining 3 factors. The table below shows the factor loadings after rotation using a significant factor criterion of .4. Accordingly, 'food additives' and 'artificial sweets' were complex variables, and 'imported food' was removed from the final analysis as it was not significant in our model.

	Industrial food processing risks	Animal-borne food risks	Food packaging risks
Growth-hormones	.80		
Food additives	.51		.50
Mercury fish	.52		
Pesticides	.51		
Foot mouth disease		.83	
Mad cow disease		.73	
Wild game disease		.71	
Agroterrorism		.47	
Food packaging			.58
Food irradiation			.55
Bottled water			.53
Artificial sweet			.53
Eigenvalues	3.17	2.96	2.58
% of variance	17.59	16.43	14.34

Figure 16. Results write-up for EFA with a truncated table. A truncated table is shown for conciseness and you are required to report the complete table in an actual write-up.

FAC1_1	FAC2_1	FAC3_1
.38929	1.59791	2.18091
.39394	1.47419	2.14977
.83403	1.60692	1.26346
1.29988	-.13440	1.55356
.68396	.69284	2.32667
.87638	.33378	1.81150
1.43731	-.21533	.98173
.53357	-.41061	2.22741
1.33768	-.84647	2.19052
.60496	.80947	1.80631
.50793	1.63305	1.91531
.67792	1.58518	1.43320
.72803	.49271	2.11326
1.31008	-.08146	1.25071
.97195	.25233	1.97649

Figure 17. SPSS output for factor scores derived from the example.