ROUTLEDGE

APPLIED MULTIVARIATE STATISTICAL CONCEPTS

Debbie L. Hahs-Vaughn

Chapter 9

EXPLORATORY FACTOR ANALYSIS

CHAPTER OUTLINE

KEY CONCEPTS

1. Factor
2. Eigenvalue
3. Communality

4. Factor extraction
5. Orthogonal rotation
6. Oblique rotation
7. Factor retention

Up to this point, we have generally concerned our analyses with procedures that have as the goal the examination of one or more a priori outcomes. With this chapter, we begin to deviate from this method of examination in that we are now doing just as the name of this procedure implies—exploring the data. Actually, some would say it is not even that but rather "it is reconnaissance" (Kaiser, 1970, p. 402). Rather than having one or more a priori outcomes, we are using exploratory factor analysis to reduce a large number of variables into identifiable clusters of variables to better understand the structure of the data.

Our objectives are that, by the end of this chapter, you will be able to (a) understand the concepts underlying exploratory factor analysis, (b) determine and interpret the results of exploratory factor analysis, and (c) understand and evaluate how to screen data prior to conducting exploratory factor analysis.

9.1 WHAT EXPLORATORY FACTOR ANALYSIS IS AND HOW IT WORKS

As we visit the statistics lab today, we find that Addie Venture and Oso Wyse have been tasked with an exploration analysis of data.

> As graduate student researchers in the stats lab, Addie and Oso have become quite accustomed to working with their teammates on data analyses that examine one or more outcomes of interest. Many times, these outcomes have been computed as composite variables from psychological assessments. While Addie and Oso have appreciated the ability to group together individual items to form various constructs, they had never really been concerned with the process underlying that construction—until today, that is. Dr. Wesley, a faculty member from the Higher Education program, is interested in examining the factor structure of measures of perceived use of skills at home and at the workplace for a select group of individuals who participated in the Survey of Adult Skills, a large data collection effort from the Organization for Economic Cooperation and Development's Programme for the International Assessment of Adult Competencies (PIAAC). Addie and Oso suggest the following research question to Dr. Wesley: *What is the underlying factor structure for perceived use of skills at home and work?* Given that dimension reduction is the goal of the project, the team recommends exploratory factor analysis to answer Dr. Wesley's question. Always up for adventure and armed with statistical knowledge, Addie and Oso are excited to embark on this task.

Globally, exploratory factor analysis (EFA) is a statistical procedure that allows us to cluster together variables into what we'll refer to in this chapter as factors (but are also known as constructs or latent constructs, a term we will use when we discuss confirmatory factor analysis). These variables may be, as just one example, a number of indices designed to measure general skills. Examining each of the variables individually may provide useful information simply by reviewing descriptive statistics of the individual measures. However, even *more* useful information may be provided by examining the underlying constructs from the variables, those variables that group together and make the number of measures parsimonious and more manageable. In essence, what exploratory factor analysis allows us to do is to work with all variables simultaneously, but at the same time know something about their underlying data structure. Exploratory factor analysis is therefore often used to provide evidence of construct validity. The underlying focus of factor analysis deals with finding common variance (distributed among the factors) and eliminating the unique variance that is not of interest (where total variance = common variance + specific variance + error variance).

Although confirmatory factor analysis will be introduced in detail in a later chapter, it is important to broach the topic here so there is a good understanding of when each is most appropriate. Confirmatory factor analysis is a statistical technique that can be used to identify the factor structure of observed variables and to test the hypothesis that a relationship exists between the respective observed variables and one or more underlying latent constructs. Additionally, much of the terminology and concepts that we will discuss in relation to EFA generalize to CFA. The titles of the procedures may give some indication of when one is more appropriate than the other is. By nature of exploration, EFA is appropriate when there is a lack of theory to dictate relationships between the variables. Brown refers to this as a "data-driven approach" (2006, p. 14). In comparison, CFA is appropriate when a strong theoretical base exists such that the relationships between variables are known and can be specified in the modeling process. In fact, it is very common for researchers to first conduct EFA prior to CFA so that there is a better understanding of how the items relate to each other and the underlying constructs or factors.

9.1.1 Characteristics

9.1.1.1 Principal Components Versus Exploratory Factor Analysis

Before we delve into this chapter, it is important to understand the difference between principal components analysis (PCA, sometimes also known as 'component factor analysis' or 'component analysis') and exploratory factor analysis (EFA, sometimes known as 'common factor analysis'). There is a difference, although in reading published literature, it seems that many authors understand them to be used interchangeably (and they should not be). If your goal is to estimate underlying factors and attach some meaning to those factors (as a form of construct validity, for example), then EFA is required. PCA, on the other hand, can be used to estimate and understand the contributions of the variables to the linear components within the data, but PCA is simply a method of decomposition—a technique for data reduction only. As stated by Borsboom, "the extraction of a principal components structure, by itself, will not ordinarily shed much light on the correspondence with a putative latent variable structure" (Borsboom, 2006, p. 426). If the interest is in placing substantive meaning on the factors extracted, EFA is the procedure needed. Throughout the chapter, it will be assumed that EFA is the goal. However, keep in mind that generating PCA or EFA is as simple as a toggle menu option in SPSS. While the results are mathematically different, the solutions you see *may* actually be quite similar. This is, again, one of those times when you must be a responsible researcher and understand the goal of your research (decomposition only, PCA, or extraction of meaning, EFA) so that you can select the appropriate method.

9.1.1.2 Exploratory Factor Analysis Specification Conditions and Decisions

There are a number of conditions that must be understood and decisions that must be made when selecting to use and implement either PCA or EFA. These are related to

a. determining factorability
b. fitting the factor model
c. selecting the factor(s)
d. rotating the factor solution

As we'll learn, researchers must first determine if factor analysis is appropriate for both their research question and their data. Within factorability, we will discuss measurement scale, sample size, and sample homogeneity, followed by tools for determining initial factorability. Second, researchers must select procedures to fit the model and estimate the model parameters. Within this realm, factor extraction and factor rotation will be reviewed. Third, the number of common factors to specify when fitting the model has to be determined. Lastly, whether or not to rotate, and how to rotate if needed, must be determined.

9.1.1.3 Factorability

Measurement scale and sample homogeneity are important considerations for determining factorability. Sample size (discussed later in the chapter) is also a consideration. In this section, we will also discuss tools for determining initial factorability.

Measurement Scale of Variables

It is important to remember that factor analysis (PCA and EFA) has as the primary requirement that a correlation matrix (denoted in statistical terms as uppercase bold **R**, the input correlation matrix with unities—or 1.0—in the diagonal, which is also referred to as the unreduced correlation matrix) be calculated from the variables in the model. (Note that a covariance matrix can also be applied in EFA; interpretation tends to be much easier with a correlation matrix. The remainder of the chapter will focus on a correlation matrix.) With conventional factor analysis, the computed correlation matrix is a Pearson matrix. This, therefore, suggests that the variables applied must be metric (at least interval in scale) so that a linear relationship exists between the variables. (However, this does not guarantee that linearity will be met.) A bit more will be added to this discussion as we talk about factor loadings later in the chapter. Even though one of the conditions of conventional factor analysis is measurement that is at least interval in scale, it is quite common to find factor analysis applied to Likert-type items which are ordinal in scale (e.g., five-point scale ranging from strongly agree to strongly disagree), particularly as the number of levels of the items increases. And should you find that your ordinal items meet the assumption of linearity, then proceeding with the factor analysis is fine (assuming other conditions and assumptions are satisfactorily met). However, items with small numbers of levels (less than seven categories in particular) are often not good candidates for conventional factor analysis, and the factors may be more difficult to interpret. Technically, binary (i.e., dichotomous) items can be factor analyzed with conventional methods, however the interpretation can be problematic as the results can reflect variation in the endorsement rate of the variables rather than the underlying construct (Fabrigar & Wegener, 2012). Categorical variables that have similar splits will tend to correlate even if the context of correlation of the variables doesn't make sense (see Gorsuch, 1983). This problem is augmented with binary data where correlations tend to reflect similar 'difficulty' as evidenced in a testing type of environment. If you do decide to proceed with conventional factor analysis using categorical variables, the factor loadings should be examined with extreme care to determine if they reflect 'difficulty' (where difficulty is defined as approximately the proportion of individuals with a '1' for their item score, as opposed to a '0') as compared to a substantive relationship. The use of binary data in conventional factor analysis can also result in a factor solution with too many factors. In the case of categorical variables, dichotomous in particular, it is highly recommended that a specialized factor analytic program that is designed for that type of data be applied to it. Later in this chapter, SPSS categorical principal components analysis (CATPCA), an add-on in SPSS, will be used to illustrate the application of ordinal data with factor analysis.

Homogeneity of the Sample in Relation to the Underlying Factor Structure

An important condition of factor analysis is that the sample of cases from which the variables were measured must be homogenous in respect to the underlying factor structure. In other words, if your collective sample of cases is known to differ, based on some characteristic, on the set of variables for which you are factor analyzing, then separate factor analysis should be performed for the groups that are anticipated to differ. For example, say that all employees of a company have been surveyed about their perceptions of the work environment, and previous empirical research suggests that those in management positions have different perceptions as compared to nonmanagement positions. The factor analysis should be conducted separately for those groups (i.e., management and nonmanagement) that are expected to differ.

Initial Factorability Assessment

There are a number of indices that should be reviewed prior to conducting the factor analysis that will help you gauge the extent to which the variables and the matrices produced from them are factorable. These include (1) correlation coefficient values, (2) Bartlett's test of sphericity, (3) anti-image correlation matrix, and (4) Kaiser-Meyer-Olkin measure of sampling adequacy.

Correlation coefficient values between the variables being factor analyzed should be .30 (in absolute value terms) or greater. This will ensure sufficient relationships to justify examination of the potential underlying components. Correlations lower than .30 may be due to low variance, which can result when samples are homogenous (but does not necessarily imply homogeneity in the sample). (However, correlations of more complex scores, such as difference scores, may have correlations between .20 and .30 and still have variables that are extremely factorable.) If there are correlation coefficient values that are not satisfactory and that are not theoretically critical, remove the variable with the lowest individual correlation value and rerun (doing so until, collectively, the correlation values reach what you deem acceptable).

Bartlett's test of sphericity is conducted to determine if the observed correlation matrix is statistically significantly different from an identity matrix (i.e., diagonal elements are 1 and off-diagonal elements are 0). Statistically significant results for Bartlett's test are desirable, as they allow you to reject the null hypothesis, which states that the observed correlation matrix equals the identity matrix. We want to see redundant variance, overlapping variance among variables, in order to reduce the variables into a fewer number of latent factors, and this is accomplished with a statistically significant Bartlett's test. Should the null hypothesis not be rejected, this provides evidence that the correlation matrix produced from the variables cannot be factor analyzed.

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) is an index of shared variance in the variables and compares the magnitudes of the observed to those of the partial correlation coefficients. MSA values range from zero to one, and large values are another form of evidence to suggest that the variables are factorable. In its early origination, Kaiser (Kaiser & Rice, 1974, p. 112) proposed the following guidelines for interpreting the index: in the .90s = marvelous; in the .80s = meritorious; in the .70s = middling; in the .60s = mediocre; in the .50s = miserable; below .50 = unacceptable. As we'll see when we compute our factor analysis, an MSA for each individual item and an overall MSA will be generated. If the overall MSA is not satisfactory, remove the variable with the lowest individual MSA value and rerun (doing so until the MSA value reaches what you deem acceptable). The overall KMO-MSA numerator is the sum of squared correlations of all variables, and the denominator is the numerator value (i.e., the sum of squared correlations of all variables) plus the sum of squared partial correlations of each variable i with each variable j, controlling for the other variables. The idea behind the MSA is that the partial correlations (reflected in the denominator) should not be relatively small if one is to expect distinct factors to emerge from factor analysis (i.e., creating a small denominator that will then provide for a larger MSA value). The size of the MSA can therefore be expected to increase as the following increase—sample size, average correlation, number of variables—and as the number of factors decrease. In SPSS, the MSA values are provided on the diagonal of the anti-image correlation matrix.

9.1.1.4 Fitting the Factor Model

Factor Extraction

Assuming you have made it through the previous examination and have determined that factor analysis is appropriate for your data, the next level of decisions has to deal with implementing or actually computing the factor analysis. Beginning with this section, we will discuss a number of concepts and procedures that should be understood to fit the model appropriately. Fitting the model, in reference to factor analysis, is also known as factor extraction. Although many times the algorithms will produce similar results, this is not always the case. Therefore, understanding how they operate and situations where they are most effective is needed.

Factor analytic models that generate two or more factors will have an infinite number of ways that the factors can be oriented in multidimensional space, each with an equally best-fitting solution (Fabrigar & Wegener, 2012). Let's say that we have a factor model where two factors are suggested. If we think about our items in two-dimensional space, the axes represent the factors and the space between the individual observed variables represents their intercorrelations—variables closer together have stronger relationships with each other. This implies that one single unique best-fitting solution does not exist when the model generates more than one factor. Therefore, this puts the burden on the researchers to select one solution. This decision process of factors to retain is the factor extraction process. Good model fit is achieved when the mathematical model for converting physical distance into predicted correlations between variables is similar to the correlations among observed variables.

A number of different algorithms can be used to fit factor analytic models, all of which calculate orthogonal factors that combine to reproduce the correlation matrix. Our discussion will focus on a few of the most common. Those commonly found in standard statistical software include principal components, unweighted least squares, generalized (weighted) least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring. Of these, principal components, principal axis, and maximum likelihood are likely the most common and are those on which our discussion will focus. In addition to its common use in EFA, maximum likelihood is also the most commonly applied estimation method in CFA (Brown, 2006).

Which extraction method selected is the researcher's choice. Generally, any extraction method will require rotation in order for the solution to be interpretable. The solutions from the different extraction methods will converge in situations where you have a large number of cases and variables and communality estimates that are similar. Evidence of the stability of your factor solution can be seen in cases where there is convergence of factor analytic solutions when using different extraction methods. While applying every estimation procedure to your data would be akin to a fishing expedition, it is quite common to select a small handful of estimation techniques to test the stability of your factor analytic model under different estimation methods—for example, first applying principal axis factoring then proceeding with maximum likelihood and ceasing the analyses when a sound solution is achieved. Generating factor analysis using two different estimation methods has been recommended (Child, 2006). Should the solutions result in discrepancies, an attempt to determine the reason(s) for the discrepancies is appropriate, followed by generation of the factor model with a third estimation technique (Child, 2006).

Principal Components

We have already broached the topic of principal components analysis as compared to common factor analysis, thus we will not delve further into that difference other than to mention a few notables as it relates to how the data is extracted. In a nutshell, the *variance* is analyzed in PCA whereas the *covariance* (communality) is analyzed in common factor analysis. In PCA, the goal is to extract the most variance from the variables with each factor.

Unweighted and Generalized (Weighted) Least Squares

Both unweighted and weighted least squares methods of factor extraction attempt to minimize the squared differences between the observed and reproduced (off-diagonal) correlation matrices. The difference between the two is that variables that share substantial variance with other variables are weighted more heavily, and variables that have more unique variance (i.e., less shared variance) receive less weight. The heavily weighted items thus contribute more to the solution than the items with lesser weight.

Maximum Likelihood (ML)

Maximum likelihood estimation calculates factor loadings that maximize the probability that the observed correlation matrix would be sampled from the population. ML is the most statistically advanced extraction method and one of the most commonly applied.

Principal Axis Factoring

Principal axis factoring has communality estimates, which are estimated through an iterative process, in the diagonal of the correlation matrix. The goal of principal axis factoring is to extract maximum variance from the variables with each factor, and this makes principal axis factoring less desirable in some situations as compared to other extraction methods that can be more effective in reproducing the correlation matrix. Principal axis factoring is one of the most commonly applied extraction methods (Child, 2006).

Alpha Factoring

Alpha factoring uses an iterative procedure to estimate communalities that then maximize coefficient alpha (i.e., an index of reliability). Unlike score reliability in psychometric research (i.e., consistency of subjects), alpha factoring focuses on determining consistency of *variables*, in other words, extracting factors that are consistently found when repeated samples of *variables* (not subjects) are drawn from a population of variables (not subjects).

Image Factoring

Image factoring uses multiple regression, with each variable serving as the dependent variable and the remaining as the independent variables, to predict image scores that are then used to compute a covariance matrix. The communalities in this extraction method are the variances from the image score covariance matrix. Factor loadings represent covariance values (as compared to correlation values seen in the other estimation procedures) between the factors and variables.

Communalities

The communality, h^2, interpreted as the *reliability of the variable*, measures the percent of variance (squared multiple correlation) of a given variable explained by all the factors jointly. The total communality is calculated by adding the squares of all the loadings of a variable across the common factors. It is the sum of all the common variance—the proportion of common variance within a variable. Computationally, the communality is the sum of squared factor loadings for a variable across all the factors.

A variable that has a low communality (.20 or below) has low common variance and high specific and error variance. A variable with a low communality may be a candidate for removal from the model, as this suggests that the factor model may not be working well for that variable. Low communalities across the set of variables indicate that the variables have weak relationships with each other. However, please note the following: A low communality can still be meaningful if the variable is contributing to a well-defined factor. The communality coefficient is not the critical element per se, but rather it is the extent to which the variable plays a role in the interpretation of the factor that is key.

It is also possible to have communalities that are too large. A communality that exceeds 1.0 is evidence of a spurious solution and may reflect a sample size that is too small or a factor model that has too few or too many factors. If you find yourself in this situation, and it is unfeasible to collect more data (either more cases and/or more variables), then remove the variable with the largest communality and rerun, repeating this process until the communality estimates are less than one. It is important to note that communalities are unaffected by rotation but are impacted by extraction method, thus the only communalities provided in standard statistical software such as SPSS are the initial and extracted estimates. Extracted communalities represent the percent of variance in a given variable explained by the extracted factors, which are often fewer in number than all the possible factors, resulting in coefficients less than 1.0 (as a side note, the communalities will be less than one even initially, with exceptions noted previously). Assuming most of the common variance is contained within those extracted factors, then the unique variance can be calculated as $1 - h^2$.

9.1.1.5 Factor Retention

Once variables are factored, the researcher must determine how many factors to retain. While this may hold only a small fraction of this chapter, the number of factors to retain has been characterized as "the crucial decision" in the EFA process as when the optimal number of factors are retained, other EFA results will generally be similar (O'Connor, 2000, p. 396). When too few factors are extracted, important information is lost, potentially important factors are neglected, error in factor loadings increases, and other problematic issues arise (Zwick & Velicer, 1986). When too many factors are retained, factors are unnecessarily split resulting in low loadings and the attribution of importance to factors which really are not (Zwick & Velicer, 1986).

Theoretically, there are as many potential factors as there are variables. For example, in a case where 12 variables are being factor analyzed, theoretically, there are 12 factors. Obviously, a researcher would be ill-guided to retain that many factors, as the goal of factor analysis is parsimony (at least in respect to data reduction)—retention of the smallest number of factors that explains the most variance of the observed variables. Historically, the number of factors to retain from a factor analytic solution have relied more often on visual (and subjective) inspection and subjective rules rather than empirical evidence, and there is not one single tool recommended. Rather, multiple decision rules are recommended and deemed desirable, as is the application of more sophisticated factor retention strategies such as parallel analysis and bootstrapping (Thompson & Daniel, 1996). Despite this recommendation, much published research exists that does not adhere (e.g., Gaskin & Happell, 2014; Henson & Roberts, 2006). Never fear, by the end of the chapter you will have the skills to call yourself a sophisticated researcher!

Scree Plots

Scree plots, where the number of factors to retain is based on where the elbow bends in the plot, are a visual tool that can be used to decide on the number of factors to retain. We see an example of a scree plot in Table 9.4. The factor numbers are plotted on the *X* axis and the eigenvalues are on the *Y* axis. In interpreting the scree plot, we look for the clearest delineation where the line goes from being diagonal to being horizontal. Then, to determine the number of factors suggested by the scree plot, we count the number of straight lines (not dots), stopping at the point where the line becomes more horizontal than diagonal. As with all visual tools, however, there is a certain degree of subjectivity that comes with making this decision. Even among experts, the reliability of scree plot interpretation is low (Streiner, 1998). When used as a decision rule to determine the number of factors to retain, scree plots generally perform better than the eigenvalue greater than one rule but are less accurate than parallel analysis (Zwick & Velicer, 1986).

Kaiser's Rule (Eigenvalues Greater Than One)

This rule is also known as the Unity Rule or the Kaiser-Guttman Criterion as it was proposed by Guttman and modified by Kaiser. Determining the number of factors to retain using Kaiser's Rule is quite simple—only those factors with eigenvalues greater than 1.0 are retained and factors with eigenvalues that are less than 1.0 are dropped. The value of one is the cut point given that the total variance contributed by each variable is one, and the variance of the factors retained should be greater than the contribution of only one variable. Eigenvalues, also known as characteristic roots or latent roots, are a measure of variance that are computed from the input (i.e., unreduced) correlation matrix. More specifically, eigenvalues measure the amount of variance in the total sample that is accounted for by each factor, and eigenvectors summarize this variance for the respective correlation or variance-covariance matrix (Brown, 2006). Factors with small eigenvalues suggest the respective factor is contributing little to explaining the variance in the variables.

Despite its widespread, and often sole, application to determining the number of factors to retain, the application of eigenvalues greater than one consistently misestimates the number of factors (either over- or underestimating) (Zwick & Velicer, 1982, 1986). Other criticisms are that an overestimation of the number of factors occurs when there are low communalities and a large number of variables and an underestimation of the number of factors to retain occurs when there are a small number of variables or when the sample size is very large. Kaiser's Rule tends to work best in conditions of moderate to large communalities, modest sample sizes, and 20–50 variables. Given these limiting conditions within which Kaiser's Rule tends to produce fairly accurate estimates, applying Kaiser's rule should only be done as a starting point (if at all) when generating your factor model. When appropriate, the results should be reviewed and the model recomputed based on a fixed number of factors.

Parallel Analysis

In comparison to the eigenvalue greater than one rule and visual examination of scree plots, there are statistically based procedures that exist for determining the number of factors to retain. Parallel analysis is one such procedure that is considered superior for determining optimal solutions for factor retention, and with 92% accuracy, has been considered the most accurate of the common methods used for retaining factors (including Kaiser's rule, Velicer's minimum average partial—MAP, scree plots, and Bartlett's test) (Zwick & Velicer, 1986). Introduced by Horn (Horn, 1965), parallel analysis is a method by which the cut-off point for factor retention can be judged, where below the cutoff, the factors possess generally trivial error variance. In simple terms, parallel analysis generates numerous replications of analyses that are drawn from random, normally distributed data with sample size *N* and number of variables *V*, concentrating on the number of factors that account for more variance than the factors derived from the random data (O'Connor, 2000). In other words, eigenvalues are extracted from the random data sets that reflect the same number of cases and variables as the observed data (thus the random data parallels the observed data in cases and variables). In the example we will later work with, we have 191 cases and 8 variables. In parallel analysis, there would then be 191 multiplied by 8 random data matrices generated with eigenvalues computed for both the observed correlation matrix and each random data matrix. Decisions on the number of factors to retain are based on comparing the eigenvalues from the original data to the eigenvalues of the random data. Factors are retained when the *i*th eigenvalue from the observed data is greater than the *i*th eigenvalue from the random data (O'Connor, 2000). Current practice recommends the use of the eigenvalue which corresponds to the percentile selected by the researcher (e.g., 95th) (Glorfeld, 1995).

Although parallel analysis is not currently available within the point-and-click user interface of popular statistical software such as SPSS, there is user-friendly syntax that has been written that allows users to perform this procedure with their own data within software such as SPSS and SAS (O'Connor, 2000). The syntax can be copied (alleviating potential error in rewriting the code), and the user has to specify only a few simple elements: (a) number of cases, (b) number of variables, (c) location of the data, and (d) the percentile at which the researcher wishes the analysis to be generated.

Number of Variables per Factor

Researchers also need to consider the number of observed variables per factor in their solution. Three variables per factor is the absolute minimum needed to define a factor (Child, 2006). Why at least three variables are needed can be understood by considering a straight line with only two points as estimation of a linear relationship. We can imagine how our line may change if error is introduced by drawing two small circles around each point. Rather than simply two unique points, now these points can be placed anywhere within that circle—this is our margin of error. We can quickly see how different our line may be depending on where the points are placed within the circle. This illustrates that two points are insufficient for estimation of a linear relationship (Child, 2006). Factors that are defined by very few variables (e.g., two or three) may be underdetermined and very unstable when the model is replicated (Brown, 2006).

9.1.1.6 Factor Rotation

Once the data are extracted, it is most always the case that the solution be rotated in order for it to be interpretable. In the world of factor analysis, rotation simply means that the axes (i.e., factor vectors or reference axes) are placed in a different position by turning about the origin (Child, 2006). If the factors are not rotated, axes will lay between the clusters of variables and the variables will not clearly differentiate to a primary factor. *It is important to note that rotation does nothing to the mathematical fit between the observed and reproduced correlation matrices, as there is mathematical equivalence between solutions prior to rotation and orthogonally rotated solutions.* Rather, rotation serves only to clarify and improve the ability to interpret the solution—there will be clearer differentiation by factor of the factor loadings of the variables. As we discussed with extraction methods, data that has clear correlational patterns will likely produce similar results regardless of rotation method. There are only two types of rotations: orthogonal and oblique, although there are quite a few methods available in standard statistical software that will accomplish the rotation.

Orthogonal Rotation

Variables that are orthogonal are unrelated, and perfect orthogonality is characterized by a correlation value of zero. In orthogonal rotation, axes are rotated at 90-degree angles. Going back to our general understanding of relationships, a correlation of zero means that knowledge of one variable in no way enhances our knowledge of the second. Thus, in the context of factor rotation, orthogonal rotation will produce uncorrelated factors. Considering many situations where factor analysis is applied in the social sciences in particular (and more specifically as we think about human behavior and attributes), however, it is likely the case to anticipate some correlation between factors as (more often than not) the constructs being measured are seldom completely independent of the others. In cases where some correlation between factors does exist, orthogonal rotation will result in a less interpretable solution than oblique rotation. Even if there are substantial correlations between factors, orthogonal rotation will constrain the solution to produce uncorrelated variables, thereby resulting in misleading solutions (Brown, 2006). In cases where there is indeed a lack of relationship between factors, orthogonal and oblique rotations will produce quite similar results.

There are a number of different types of orthogonal rotation techniques available in standard statistical software. These include varimax, quartimax, and equamax, each of which works with a different statistic to maximize or minimize it. Varimax rotation, the most common orthogonal rotation, maximizes the variance of the factor loadings within the factors and across the variables to simplify the factors. Quartimax rotation, on the other hand, maximizes the variance of the factor loadings within the variables and across the factor loadings to simplify the variables. Equamax attempts to bridge varimax and quartimax by simultaneously simplifying both the factors and the variables. Equamax is the least preferred orthogonal rotation, as research suggests it is unstable in situations other than when the number of factors can be specified with confidence.

Oblique Rotation

Factors that are oblique are related, and perfect obliqueness is characterized by a correlation value of one (in absolute value terms). When oblique rotation is applied, the factor axes are rotated independently of each other at different angles (i.e., not just

90 degrees, as is the case with orthogonal rotation). Going back to our general understanding of relationships, a correlation of one means that knowledge of one factor (in this case) perfectly enhances our knowledge of the second factor. Thus, in the context of factor rotation, oblique rotation will produce correlated *factors* (not correlated variables). While oblique seems to be the most defensible option of the two rotations (given that it is reasonable to assume there would be correlation between constructs), be prepared for the possibility that it may increase the difficulty in attaching meaning to your factors. This is because there will likely be an increased number of cross-loading variables in the oblique, as compared to orthogonal, rotated solution. Cross-loading variables are variables that have similar factor loadings for multiple factors. If you are unsure which rotation to select, you may wish to test oblique rotation first and review the factor correlation matrix. Small factor correlations (e.g., less than .30) may warrant orthogonal rotation. There are a few different types of oblique rotation techniques available in standard statistical software, including direct oblimin and promax.

Associated Matrices

The type of rotation selected will alter the matrices generated in your factor solution. In an orthogonal rotation solution, the structure matrix is simply the factor-loading matrix (and is the only matrix that requires review). Oblique rotations will result in generation of both a structure and a pattern matrix. The structure matrix coefficients represent the variance in the observed variables explained by a factor, both a unique (i.e., relationship between the variable and the factor, as with the pattern matrix) and common (i.e., relationship between the variable and the shared variance among the factors) contribution. In oblique rotations, the structure matrix is the product of the pattern and factor correlation matrices, and the loadings in the structure matrix will often be larger than those in the pattern matrix because they reflect overlap in the factors (i.e., are inflated due to this), unless there is a weak relationship between the factors. The pattern matrix coefficients or loadings represent unique contributions only, i.e., unique relationships between the variables and factors. Generally, the larger number of factors, the lower the coefficients in the pattern matrix since there is more common contribution to the variance explained. Because both a structure and pattern matrix are generated with oblique rotation, this requires examination of both when interpreting the meaning of the factors. Of the two matrices, the pattern matrix is the one that is most often reported and interpreted (Brown, 2006).

9.1.1.7 Factor Loadings

In simple terms, the factor loading is the coordinate of a variable along a classification axis. It reflects the relationship between a factor and an observed variable and is the slope of increase (when positive) or decrease (when negative) in the observed variable for each unit of increase or decrease in the factor. The factor-loading value is interpreted in the same units as the measured variables. Now, this is where consideration of the measurement scale of items in the factor analysis come into play. . . . This type of index that measures the relationship between the factor and observed variables is meaningful if, and only if, the observed variables are measured in such a way that the units can be ordered or ranked and there is equal distance between the units—this implies the variables must be interval or ratio scale. Nominal and ordinal items (even with three or more categories) are usually insufficient to meet this condition (Fabrigar & Wegener, 2012).

The factor loading provides information on the relative contribution that an individual variable makes to a factor, and the researcher must decide which variables load onto which factor. An often-followed, 'moderately rigorous' guideline is that a variable should have a factor loading of at least .30 in order to be retained with that factor; however, this rule should be applied only in models with samples of 80 or more (as with samples of this size, a correlation coefficient is statistically significant at an alpha of .01) (Child, 2006, p. 63). A variable with a factor loading of .30, when squared, is interpreted as variance, and this would mean that variable accounts for slightly less than 10% (9% specifically) of the common variance of the factor.

A squared factor loading is a measure of variance accounted for, similar to *R* squared. More specifically, it estimates the amount of variance in a factor that is accounted for by the individual variable—the proportion of variance in the item response or variable scores that are explained by a factor. EFA allows the decomposition of observed variance into both common/shared variance and unique variance. In the ideal situation, a variable will have a large coordinate for only one axis and low coordinates for all other axes—providing evidence to suggest that the variable relates to one and only one factor. It is possible to have negative factor loadings. Factors that are defined by variables with both positive and negative factor loadings are called bipolar factors (Child, 2006). The percent of variance in all the variables accounted for by each factor is computed as the sum of the squared factor loadings for that factor divided by the number of variables—which is also the same as dividing the eigenvalue of a factor by the number of variables in the model. Box 9.1 summarizes the process of fitting the factor model.

BOX 9.1 FITTING THE FACTOR MODEL

Element	Options
Factor Extraction	Select an algorithm: • Principal components • Unweighted and generalized (weighted) least squares • Maximum likelihood (ML) • Principal axis factoring • Alpha factoring • Image factoring
Communalities	Review communalities: • Low communalities (< 2.0): consider removing unless inclusion of the variable is key to interpreting the factor • High communalities (> 1.0): may be evidence of a spurious solution, and may reflect a sample size that is too small or a factor model that has too few or too many factors. Remove the variable with the largest communality and rerun the EFA—repeating this process until the communality estimates are less than one.

Factor Retention	Determine the number of factors to retain: • Scree plots: Subjective visual tool; can be used as a guide but do not rely on this absent other more objective means • Eigenvalues greater than one: Kaiser's Rule works best and produces fairly accurate results in conditions of moderate to large communalities, modest sample sizes, and 20–50 variables. Given these limiting conditions, applying Kaiser's rule should only be done as a starting point (if at all) when generating your factor model. • Parallel analysis: Most accurate option for determining the number of factors to retain. Decisions on the number of factors to retain are based on statistical analysis, comparing the eigenvalues from the original data to the eigenvalues of randomly generated data.
Number of Variables per Factor	Review the number of variables per factor: • Minimum: 3 per factor
Factor Rotation	Determine how to rotate the factors: • Orthogonal (uncorrelated): varimax, quartimax, and equamax • Oblique (correlated): direct oblimin and promax
Factor Loadings	Review factor loadings: • Ideally, a variable will have a large factor loading for only one factor • A 'moderately rigorous' recommendation: a variable should have a factor loading of at least .30 in order to be retained with that factor ◦ This rule should be applied only in models with samples > 80

9.1.2 Sample Size

Unlike traditional statistical procedures, there is not a power calculation to suggest appropriate sample size for factor analysis. What exists are a number of sample size recommendations for factor analysis that have been made throughout the years, with none reaching consensus as the absolute criterion that must be followed and all later being determined invalid (MacCallum, Widaman, Zhang, & Hong, 1999). These recommendation are generally based on a subject-to-variable ratio (STV) or absolute sample size per number of cases (*N*).

Case or subject-to-variable ratio (STV) recommendations range from two times the number of cases (Kline, 1979) to five or more times the number of items with a case-to-item ratio greater than or equal to 5 and a minimum of 100 cases, regardless of the case-to-item ratio (Bryant & Yarnold, 1995; Suhr, 2006), more than 5 times the number of items to allow for missing data (Suhr, 2006), 10 times the number of items (Nunally, 1978), and 51 more cases than the number of variables (Lawley & Maxwell, 1971);

Other criterion are based on an absolute number of *cases* (*N*), with 100 cases being the suggested bare minimum sample size (Gorsuch, 1983; MacCallum et al., 1999), at least 150–300 (tending toward 150 if items are not highly correlated) (Hutcheson & Sofroniou, 1999), at least 200 (Guilford, 1954), at least 250 (Cattell, 1978), and a sliding scale ranging from 100 to 1,000 (with 100 = poor, 200 = fair, 300 = good, 500 = very good, and 1,000 or greater = excellent) (Comrey & Lee, 1992).

All this to be said, many researchers today would likely agree that these recommendation for STV and absolute number of cases are weak criteria to follow to estimate the sample size for EFA, and there is research to suggest the invalidity of these rules (MacCallum et al., 1999). What is more important is the factorability of the model, as seen through communalities (percent of variance in an variable that is explained jointly by all factors), the degree of overdetermination (ratio of factors to variables), the size of the factor loading, and general model fit. Simulation research suggests that estimating factor structure is achievable, even with small sample sizes (particularly $N > 20$), given the following conditions are met: (a) high communalities (approximately .8 to .9), (b) small number of *expected* factors to be retained (2 to 4), and (c) low model error (which is likely evidenced in situations where communalities are high; RMSR = .00 to .06) (Preacher & MacCallum, 2002). Other simulation research has shown that factors with four or more variables with factor loadings of .60 or greater are interpretable regardless of the sample size (Guadagnoli & Velicer, 1988). Solutions with lower factor loadings (.40) can still be interpreted if the number of cases is at least 150 and the number of variables per factor is larger (> 10) (Guadagnoli & Velicer, 1988).

The take-home message for sample size with EFA is this: Do not adhere to a recommendation criterion for STV or absolute number of cases. Rather, design your study so that you collect the largest sample size that resources will allow. In some cases, this may mean that the sample size will be unnecessarily small. In those instances—and all others, as a matter of fact—be prepared to defend your sample size using previous empirical research, such as the simulation research presented here. And if you are a researcher so inclined to study methodological issues, this is an area ripe for continued examination.

9.1.3 Power

There are no power calculations to suggest appropriate sample size for exploratory factor analysis given a priori or post hoc power. What exists are a number of sample size recommendations as presented previously.

9.1.4 Effect Size

Factor analytic solutions, in and of themselves, do not produce effect size results. Once composite variables are created based on the factor analytic solutions and then those composite variables are applied in an inferential procedure, effect sizes can then be generated.

9.1.5 Assumptions

As with most multivariate statistical procedures, there are a number of assumptions that must be considered with factor analysis, either EFA or PCA. These include

(a) independence, (b) linearity, (c) absence of outliers (both bivariate and multivariate) in cases and variables, and (d) lack of extreme multicollinearity and singularity. As previously discussed, a condition required for conventional factor analysis is continuous data (assuming the factor analytic procedure is computed from a Pearson correlation, as we will assume in this chapter). A large sample size is not necessarily required (as detailed previously) but may be helpful depending on the factor model. Factor analysis is actually robust to violations of the assumption of normality and normality is really not applicable in EFA as it is with many other multivariate procedures. The only exception to this is in the situation where tests of inference are used to determine the number of factors to retain (e.g., when using ML estimation), and in this case, multivariate normality *is* an assumption. Examination of univariate normality, which is not overly sensitive as are multivariate normality tests, can be done through examination of skewness and kurtosis, formal tests of normality, and plots (e.g., Q-Q plots). In terms of multivariate normality, a macro in SPSS (DeCarlo, 1997) (illustrated with MANOVA in chapter 4) can be used to examine a number of multivariate normality indices that include (a) multivariate kurtosis (Mardia, 1970), (b) multivariate skewness and kurtosis based on Small's (1980) multivariate extension of univariate skewness and kurtosis (Looney, 1995), (c) multivariate normality omnibus test (Looney, 1995), (d) largest squared and plot of squared Mahalanobis distance, and (e) critical values for hypothesis test for a single multivariate outlier using Mahalanobis distance (Penny, 1996).

9.1.5.1 Independence

The first assumption is concerned with **independence** of the observations. Violations of this assumption can detrimentally impact standard error values and thus any resulting hypothesis tests. Testing for this assumption is a bit nebulous in exploratory factor analysis, as there are no independent and dependent variables that allow for this type of examination. In the absence of statistical evidence, we will rely on theoretical evidence: If the units have been randomly sampled from a population, there is evidence that the assumption of independence has been met.

9.1.5.2 Linearity

As you recall, factor analysis uses relationships among the variables as the basis for determining factors with conventional factor analysis doing so via a Pearson correlation matrix. Therefore, it is assumed there is a linear relationship among the variables. Bivariate scatterplots can be examined to determine the extent to which this assumption is held.

9.1.5.3 Absence of Outliers in Cases and Variables

Outliers in factor analysis operate in an unfavorable fashion, just as they do in other procedures. One or more outlying cases (either univariate or multivariate) can have undue and unwanted influence on the factor model. In addition to the ways we've screened for outliers in previous procedures (e.g., boxplots), they can also be screened by reviewing standard scores of the variables. Standardized scores with absolute values of 3.29 or greater (which equates to values more than 3–1/4 standard deviation units from the mean; about .05% of cases are above and below this point in a standardized normal distribution) should be flagged as outliers. Multivariate outliers can be determined by Mahalanobis distance values, which can be calculated using multiple regression, discriminant analysis, or logistic regression (or via simple matrix algebra, without generating other analyses). Multivariate outliers are evidenced by statistically significant Mahalanobis distance scores (alpha = .001 if you tend toward the liberal edge, which is appropriate with EFA), evaluated using a chi-square distribution with degrees of freedom equal to the number of variables. To generate Mahalanobis distance, apply all the variables as independent variables with the dependent variable being a binary variable coded 1 for potential outliers and 0 for all other variables. The process for examining outliers is therefore to look for univariate outliers first. If any are detected, then screen for multivariate outliers.

In factor analysis, it is also possible to have outlying variables, that is, variables that are unrelated to others in the factor model. These outlying variables can be determined by reviewing the following: (a) squared multiple correlations with all other variables and (b) weak correlations with the factors that are identified in the factor analytic model. Outlying variables that are identified can be disregarded.

9.1.5.4 Lack of Extreme Multicollinearity and Singularity

In other procedures, we have discussed how multicollinearity can be problematic because it makes the matrix inversion process unstable. As you recall, multicollinearity is a very strong linear relationship between two or more of the predictors. You may be wondering how it is the case that this can be problematic in factor analysis, as one of the indices we use to determine the ability to factor analyze is the relationship between variables and there is no matrix inversion. In factor analysis, we are concerned with *severe and extreme multicollinearity*, which can be problematic in factor analysis. Singularity is a special case of multicollinearity; it is perfect multicollinearity and occurs when two or more variables perfectly predict and are therefore perfectly redundant. This can occur in factor analysis (just as it did in multiple regression), for example, when a composite variable as well as its component variables are used as predictors in the same factor analytic model.

How do we detect violations of this assumption? Remember that we are looking only for extreme multicollinearity, so we will limit our detection methods quite a bit as compared to our data examination in multiple regression. For EFA, the simplest method is to conduct a series of multiple regression models, one regression model for each variable where that variable is the dependent variable and all remaining variables are the independent variables. If any of the resultant R_k^2 values are close to one (greater than .9 is a good guideline to go by), then there may be an extreme multicollinearity problem. However, large R^2 values may also be due to small sample sizes, so be cautious in interpreting cases where the number of cases is small. If the number of variables is greater than or equal to n, then perfect multicollinearity is a possibility.

TABLE 9.1

Assumptions and Violation of Assumptions: Exploratory Factor Analysis

Assumption	Effect of Assumption Violation
Independence	Influences standard errors of the model and thus hypothesis tests
Linearity	Reduces interpretability of the factor analytic solution
Absence of outlying cases and variables	Exerts undue influence on and distorts the factor analytic solution
Lack of extreme multi-collinearity	Reduces ability to separate effects of variables
Multivariate normality	Minimal effect when violated with exceptions including (a) when hypothesis testing is conducted as part of the EFA, (b) when maximum likelihood is used to estimate the factor model, and (c) with small sample sizes

9.1.5.5 Concluding Thoughts on Assumptions

As mentioned in previous chapters, there is no rule stating that research that violates assumptions must be scrapped. However, researchers who face violations of assumptions must handle these situations on a case-by-case basis, considering both the goal of the analyses and the extent to which the assumptions were violated and the resulting effect of violation. It is also important that researchers present the evidence found, along with justification for decisions that were made. The assumptions are summarized in Table 9.1.

9.2 MATHEMATICAL INTRODUCTION SNAPSHOT

Now that we understand the conditions and decision points, there are a few additional foundational topics related to the underlying mathematics of exploratory factor analysis that may be helpful with which to become acquainted. Note that this is not meant to be a primer on the mathematical proofs nor is it meant to serve as a foundation for which hand calculations can be made. Rather, it is meant to provide a bit more of the mathematical representation for those who are interested in delving deeper into this aspect.

Using matrix algebra, we can express the correlational structure of the common factor model as follows:

$$P = \Lambda\Phi\Lambda^T + D_\Psi$$

In this equation, ***P*** refers to the population correlation matrix of observed variables. The factor-loading matrix, Λ (lambda), represents the linear influence strength and direction of the latent or component factors on the observed variables. In this matrix, the columns represent the factors and the rows represent the observed variables. Thus, $\Lambda_{3\cdot 1}$ refers to the factor loading for (the value of which is the path between) the effect or influence of common factor one on observed variable three.

The transpose of the factor-loading matrix is represented by lambda superscript *T*, Λ^T. As reviewed in the material on matrix algebra in the appendix, transposing means that what was originally in the rows now become columns (and what was originally in the columns now become rows).

The covariance matrix among the unique factors is represented by D_Ψ (the subscript for which is psi). The diagonals of this matrix are the variances of the unique factors. The off-diagonals are the covariances and are zero when orthogonality is assumed.

The correlation matrix between the factors is represented by Φ (phi). When orthogonality of errors is assumed (i.e., the factors are uncorrelated), the population correlation matrix is simply: $P = \Lambda\Lambda^T + D_\Psi$.

Because our interest is in the conceptual understanding of EFA, we'll end our mathematical discussion at this point. The summary of the underlying mathematics of EFA was drawn from Fabrigar and Wegener (2012), which provides a very accessible account. Readers interested in learning more of the mathematics are referred to that source, among others.

TABLE 9.2

Factor-Loading Matrix Example

Factor Matrix[a]		
	Factor	
	[Common Factor 1]	[Common Factor 2]
Index of use of numeracy skills at home	$\Lambda_{1\cdot1} = .843$	$\Lambda_{1\cdot2} = -.175$
Index of use of ICT skills at home	$\Lambda_{2\cdot1} = .673$	$\Lambda_{2\cdot2} = .066$
Index of use of reading skills at home	$\Lambda_{3\cdot1} = .528$	$\Lambda_{3\cdot2} = .153$
Index of use of numeracy skills at work	$\Lambda_{4\cdot1} = .330$	$\Lambda_{4\cdot2} = .086$
Index of readiness to learn	$\Lambda_{5\cdot1} = .412$	$\Lambda_{5\cdot2} = .504$
Index of use of task discretion at work	$\Lambda_{6\cdot1} = .059$	$\Lambda_{6\cdot2} = .311$
Index of learning at work	$\Lambda_{7\cdot1} = .062$	$\Lambda_{7\cdot2} = .300$
Index of use of planning skills at work	$\Lambda_{8\cdot1} = -.183$	$\Lambda_{8\cdot2} = .296$

Extraction Method: Maximum Likelihood.
a. Two factors extracted. Six iterations required.

TABLE 9.3

Example of Correlation Matrix of Common Factors

Factor Correlation Matrix		
Factor	[Common Factor 1]	[Common Factor 2]
1	1.000	
2	$\Phi_{2\cdot1} = .165$	1.000

Extraction Method: Maximum Likelihood.
Rotation Method: Promax with Kaiser Normalization.

9.3 COMPUTING EFA USING SPSS

As we know by now, conventional factor analysis requires continuous data. There are many situations, however, where EFA of ordinal survey (e.g., Likert) items is desirable. Thus, our use of SPSS will first illustrate EFA with continuous data, and this will be followed by an illustration of the use of parallel analysis for factor retention. Next, we will illustrate how to use one of the SPSS add-ons for conducting EFA with ordinal data.

9.3.1 Computing EFA With Continuous Data Using SPSS

Next, we consider SPSS for conducting exploratory factor analysis with data that is continuous in scale (should you have only ordinal items, please see the following SPSS section, "Computing EFA With Ordinal Data"). Before we conduct the analysis, let us talk about the data. The data we are using is the 2013 Survey of Adult Skills (http://www.oecd.org/site/piaac/surveyofadultskills.htm), available through the Organisation for Economic Co-operation and Development (OECD). Thank you to OECD for making this data publicly available.

The Survey of Adult Skills, conducted in 33 countries, is part of the Programme for the International Assessment of Adult Competencies (PIAAC), and the first results from the survey were released in 2013. Measured in the survey are "key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper" (see http://www.oecd.org/site/piaac/surveyofadultskills.htm). Adults ages 16 to 65 were interviewed in their homes, with 5,000 individuals from each country participating. It is important to note that the Survey of Adult Skills is a complex sample (i.e., not a simple random sample). Although each country was allowed to create their own sampling design and selection plan (for example, some countries oversampled some groups of individuals), it had to adhere to technical standards published by the PIAAC. For example, the U.S. sampling design was a four-stage stratified probability proportional to size design. If you access the full dataset, you will find the last few variables are various weights as well as stratum and unit variables. We won't get into the technical aspects of this, but when the data are analyzed to adjust for the sampling design (including nonsimple random sampling procedure and disproportionate sampling), the end results are then representative of the intended population. The purpose of the text is not to serve as a primer for understanding complex samples, and thus readers interested in learning more about complex survey designs are referred to any number of excellent resources (Hahs-Vaughn, 2005; Hahs-Vaughn, McWayne, Bulotskey-Shearer, Wen, & Faria, 2011a, 2011b; Lee, Forthofer, & Lorimor, 1989; Skinner, Holt, & Smith, 1989). Additionally, so as to not complicate matters any more than necessary (learning EFA is generally complicated enough!), the applications in this textbook do not illustrate how to adjust for the complex sample design. As such, the results that we see should not be interpreted to represent any larger population but only that select sample of individuals who actually completed the survey. I want to stress that the reason why the sampling design has not been illustrated in the textbook applications is because the point of this section of the textbook is to illustrate how to use statistical software to generate various procedures and how to interpret the output and not to ensure the results are representative of the intended population. Please do not let this discount or diminish the need to apply this critical step in your own analyses when using complex survey data, as quite a large body of research exists that describes the importance of effectively analyzing complex samples and provides evidence of biased results when the complex sample design is not addressed in the analyses (Hahs-Vaughn, 2005, 2006a, 2006b; Hahs-Vaughn et al., 2011a, 2011b; Kish & Frankel, 1973, 1974; Korn & Graubard, 1995; Lee et al., 1989; Lumley, 2004; Pfeffermann, 1993; Skinner et al., 1989).

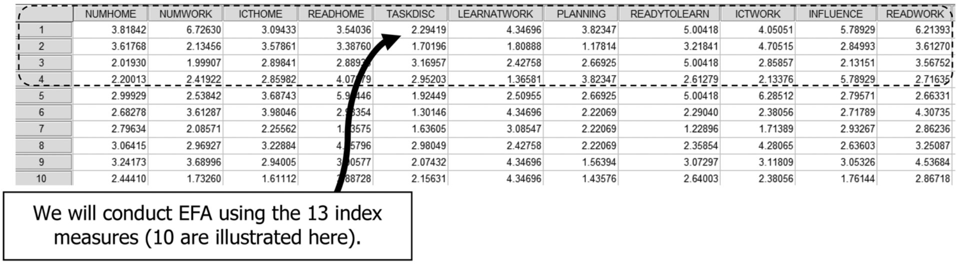
Now, let's review the data. We are using the **PIAAC_EFA.sav** file. This is data from the U.S., and the data file has been delimited to include only individuals who were between the ages of 25–29 [AGEG5LFS = 3], who were employed or participated in education or training during the 12 months prior to completing the survey [NEET = 0], and who reported having 'above high school' education [B_Q01a_T = 3] (*n* = 288). The size of this sample is more than sufficient to generate EFA, but at the same time small enough to work with for readers who may be using a version of SPSS that limits the number of cases. Additionally, it creates at least an intuitively homogenous sample that would be anticipated to respond similarly on the items. (Note: The complete PIAAC Survey of Adult Skills data file, which includes 5,010 cases, is available from the textbook's companion website and is titled PIAAC_SurveyOfAdultSkills.sav.)

Before we run the data, it's always important to examine frequency distributions of the variables that will be used in the model to assess missing data, potential data entry problems, and similar. With this data, we have some missing data (it has already been coded by the survey collectors as 9996), and thus I've taken the liberty to perform listwise deletion on the missing items (resulting in *n* = 191); however, the remaining variables in the data file have been left as is so that you may practice your data cleaning skills in working with 'real data.'

Let's look at the data. For the EFA illustration, we'll be working with 13 indices (variables 1–13 in your SPSS file), each of which is measured on a continuous scale.

1. Index of use of numeracy skills at home (basic and advanced—derived)
2. Index of use of numeracy skills at work (basic and advanced—derived)
3. Index of use of ICT skills at home (derived)
4. Index of use of reading skills at home (prose and document texts—derived)
5. Index of use of task discretion at work (derived)
6. Index of learning at work (derived)
7. Index of use of planning skills at work (derived)
8. Index of readiness to learn (derived)
9. Index of use of ICT skills at work (derived)
10. Index of use of influencing skills at work (derived)
11. Index of use of reading skills at work (prose and document texts—derived)
12. Index of use of writing skills at work (derived)
13. Index of use of writing skills at home (derived)

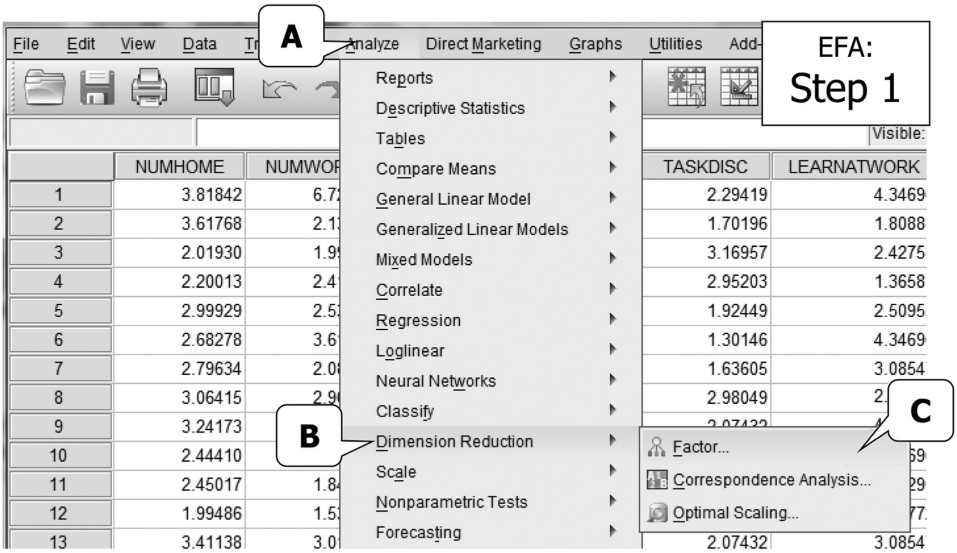
The first 13 variables are the indices for EFA. The next three variables in the SPSS dataset were used to delimit the sample. A few variables used for data screening are included (outlier and MAH_1, Mahalanobis distance, which we will discuss as we test assumptions). This is followed by three variables in the dataset that represent the country and participant ID variables. I've left those in the data file just in case you are interested in merging variables from the full dataset with this smaller, delimited file. Each row in the data set still represents one individual. As seen in the screenshot below, the SPSS data is in the form of multiple columns that represent the variables on which the respondents were measured. For the EFA illustration, we will work with the 13 continuous index measures.



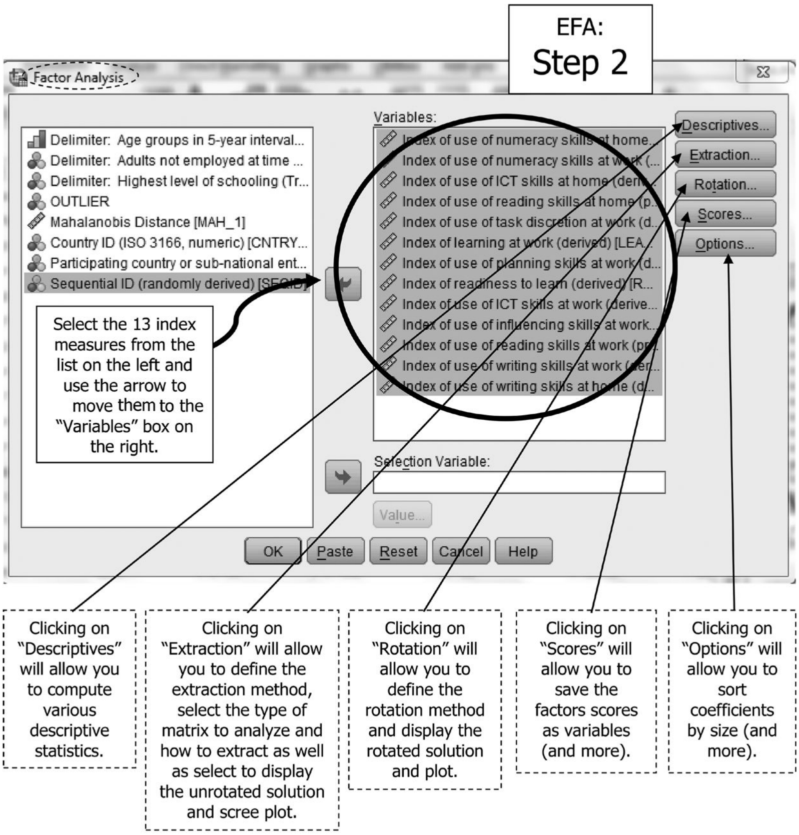
	NUMHOME	NUMWORK	ICTHOME	READHOME	TASKDISC	LEARNATWORK	PLANNING	READYTOLEARN	ICTWORK	INFLUENCE	READWORK
1	3.81842	6.72630	3.09433	3.54036	2.29419	4.34696	3.82347	5.00418	4.05051	5.78929	6.21393
2	3.61768	2.13456	3.57861	3.38760	1.70196	1.80888	1.17814	3.21841	4.70515	2.84993	3.61270
3	2.01930	1.99907	2.89841	2.889[illegible]	3.16957	2.42758	2.66925	5.00418	2.85857	2.13151	3.56752
4	2.20013	2.41922	2.85982	4.07[illegible]79	2.95203	1.36581	3.82347	2.61279	2.13376	5.78929	2.71635
5	2.99929	2.53842	3.68743	5.9[illegible]446	1.92449	2.50955	2.66925	5.00418	6.28512	2.79571	2.66331
6	2.68278	3.61287	3.98046	2.[illegible]3354	1.30146	4.34696	2.22069	2.29040	2.38056	2.71789	4.30735
7	2.79634	2.08571	2.25562	1.[illegible]3575	1.63605	3.08547	2.22069	1.22896	1.71389	2.93267	2.86236
8	3.06415	2.96927	3.22884	4.[illegible]5796	2.98049	2.42758	2.22069	2.35854	4.28065	2.63603	3.25087
9	3.24173	3.68996	2.94005	3.[illegible]0577	2.07432	4.34696	1.56394	3.07297	3.11809	3.05326	4.53684
10	2.44410	1.73260	1.61112	[illegible]88728	2.15631	4.34696	1.43576	2.64003	2.38056	1.76144	2.86718

We will conduct EFA using the 13 index measures (10 are illustrated here).

Step 1. To conduct exploratory factor analysis, go to "Analyze" in the top pull-down menu, then select "Dimension Reduction," and then select "Factor." Following the screenshot below (Step 1) produces the "Factor Analysis" dialog box.



Step 2. Click the 13 index measures and move into the "Variables" box by clicking the arrow button (see screenshot Step 2).



EFA: Step 2

Factor Analysis

- Delimiter: Age groups in 5-year interval...
- Delimiter: Adults not employed at time ...
- Delimiter: Highest level of schooling (Tr...
- OUTLIER
- Mahalanobis Distance [MAH_1]
- Country ID (ISO 3166, numeric) [CNTRY...
- Participating country or sub-national ent...
- Sequential ID (randomly derived) [SEQID]

Variables:

- Index of use of numeracy skills at home...
- Index of use of numeracy skills at work (...
- Index of use of ICT skills at home (deri...
- Index of use of reading skills at home (p...
- Index of use of task discretion at work (d...
- Index of learning at work (derived) [LEA...
- Index of use of planning skills at work (d...
- Index of readiness to learn (derived) [R...
- Index of use of ICT skills at work (derive...
- Index of use of influencing skills at work...
- Index of use of reading skills at work (pr...
- Index of use of writing skills at work (der...
- Index of use of writing skills at home (d...

Selection Variable:

Select the 13 index measures from the list on the left and use the arrow to move them to the "Variables" box on the right.

Clicking on "Descriptives" will allow you to compute various descriptive statistics.

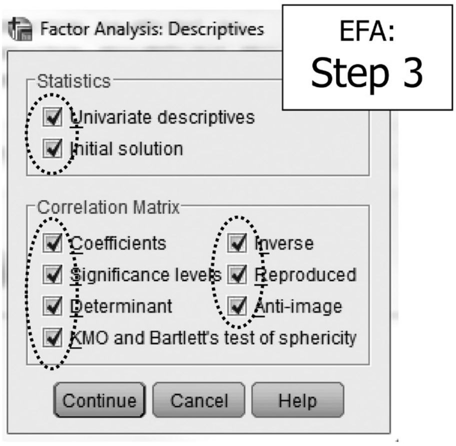
Clicking on "Extraction" will allow you to define the extraction method, select the type of matrix to analyze and how to extract as well as select to display the unrotated solution and scree plot.

Clicking on "Rotation" will allow you to define the rotation method and display the rotated solution and plot.

Clicking on "Scores" will allow you to save the factors scores as variables (and more).

Clicking on "Options" will allow you to sort coefficients by size (and more).

Step 3. From the Factor Analysis dialog box (see screenshot Step 2), clicking on "Descriptives" will provide the option to compute various descriptive statistics (see screenshot Step 3). From the Factor Analysis: Descriptives dialog box, place a checkmark in all the boxes. Click on "Continue" to return to the Factor Analysis dialog box.



EFA: Step 3

Factor Analysis: Descriptives

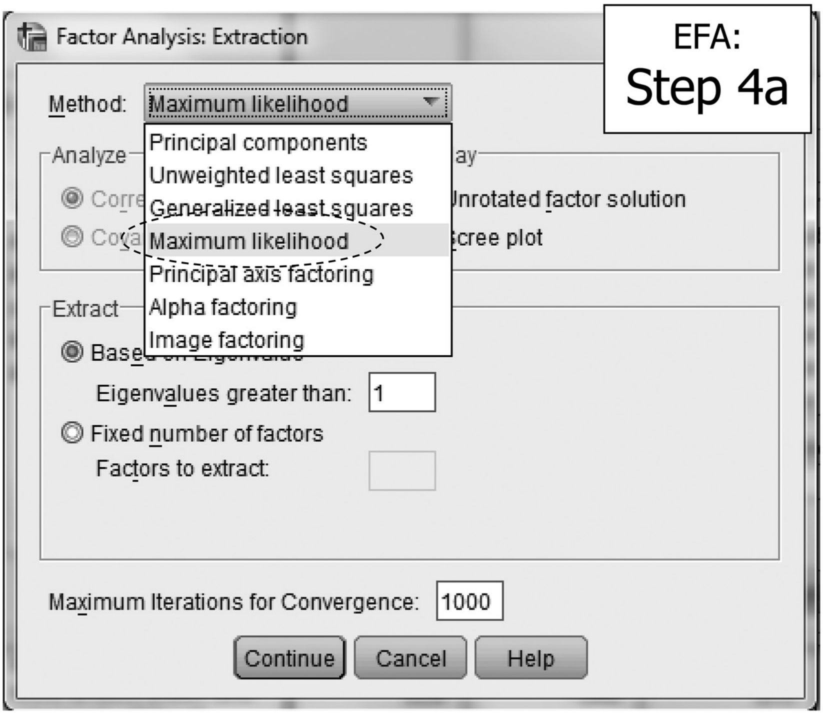
Statistics

- ☑ Univariate descriptives
- ☑ Initial solution

Correlation Matrix

- ☑ Coefficients
- ☑ Significance levels
- ☑ Determinant
- ☑ KMO and Bartlett's test of sphericity
- ☑ Inverse
- ☑ Reproduced
- ☑ Anti-image

Step 4a. From the Factor Analysis dialog box (see screenshot Step 2), clicking on "Extraction" will provide the option to select various options related extraction methods and what is displayed (see screenshot Step 4a). Using the pull-down menu, click on "Maximum likelihood." Recall that we discussed how solutions from the different extraction methods will converge in situations where you have a large number of cases and variables and communality estimates that are similar. We also stated that evidence of the stability of the factor solution can be seen in cases where there is convergence of factor analytic solutions when using different extraction methods. Thus, you may want to select a small handful of estimation techniques to test the stability of your factor analytic model under different estimation methods, although for this illustration, we will apply only one.



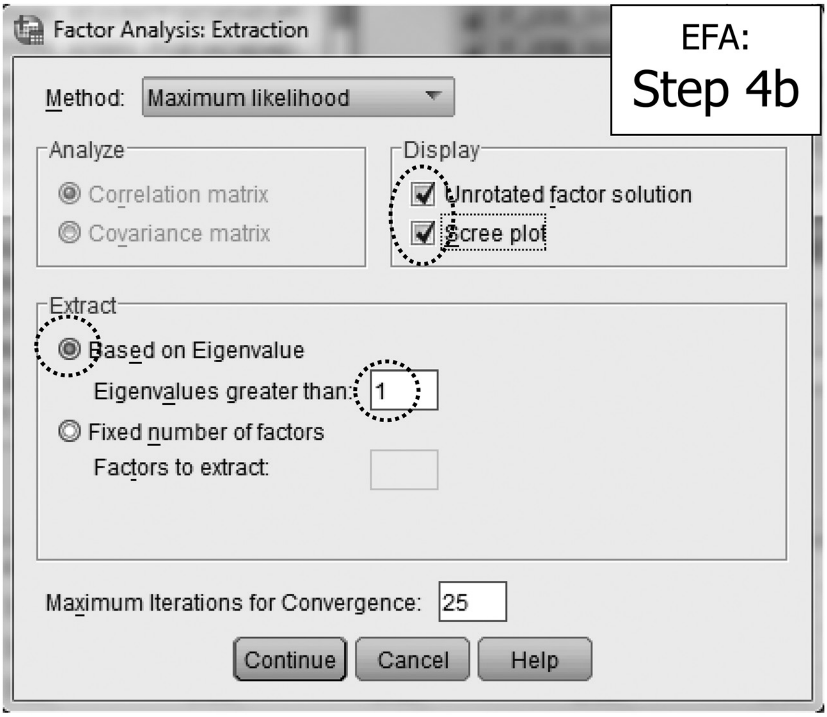
Factor Analysis: Extraction

Method: Maximum likelihood

Eigenvalues greater than: 1

Maximum Iterations for Convergence: 1000

Step 4b. Also from the Factor Analysis: Extraction dialog box, place a checkmark in the box next to the following: (1) unrotated factor solution and (2) scree plot (see screenshot Step 4b). Under the heading for 'Extract,' click the radio button for 'based on eigenvalue' and then enter 1 in the box for 'eigenvalues greater than:'. Recall that the application of Kaiser's rule consistently (often substantially) overestimates the number of factors (Zwick & Velicer, 1982, 1986), thus we won't base our factor solution interpretation on it as an important piece of the results. Knowing that it usually overestimates the number of factors to retain, however, it does give us a starting point from which to work. Depending on the solution, we may choose to rerun the model and base the number of factors to extract on a 'fixed number of factors.' Leave the default setting for 'Maximum Iterations for Convergence' at 25. Click on "Continue" to return to the Factor Analysis dialog box.



Method: Maximum likelihood

Analyze

● Correlation matrix

○ Covariance matrix

Display

☑ Unrotated factor solution

☑ Scree plot

Extract

● Based on Eigenvalue

Eigenvalues greater than: 1

○ Fixed number of factors

Factors to extract:

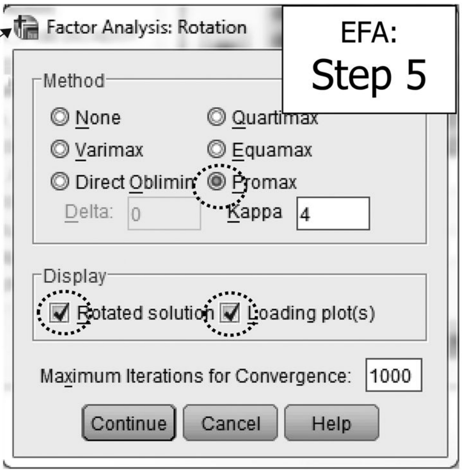
Maximum Iterations for Convergence: 25

Step 5. From the Factor Analysis dialog box (see screenshot Step 2), clicking on "Rotation" will provide the option to select various options related to rotation methods. Place a checkmark in the box next to the following: (1) rotated solution and (2) loading plot(s) (see screenshot Step 5). In terms of the factor-loading plot, in the event that only one factor is extracted, no plot will be displayed. When two factors are extracted, a two-dimensional plot will be displayed. When three or more factors are extracted, a three-dimensional factor-loading plot of only the first three factors extracted is displayed. Under the heading for 'Method,' click the radio button for 'Promax' and then enter 4 in the box for 'Kappa' (which is the default). (Other values of kappa can be introduced, with the ideal kappa value being one that results in the simplest factor structure with low correlations among the factors; higher kappa values lead to larger correlations among factor and simpler loading structures. The default of 4 is based on previous research which suggests this value produces a generally good solution (Hendrickson & White, 1964).) Change the default setting for 'Maximum Iterations for Convergence' to 1000. The number of iterations to convergence simply defines how many iterations the algorithm can take to perform the rotation. It is likely the case that 1000 is overkill, but it doesn't hurt to set it at a large value just in case it's required. Click on "Continue" to return to the Factor Analysis dialog box.

Clicking on "Rotation" will allow you to define the rotation method, select what to display, and define iterations.

Direct oblimin and promax are oblique rotation methods, allowing the factors to be correlated. Varimax, quartimax, and equamax are orthogonal rotation methods, assuming unrelated factors and maintaining the axes at 90 degrees.

What is displayed in the output is dependent on the method of rotation. The rotated pattern and factor transformation matrices are displayed with orthogonal rotations. The pattern, structure, and factor correlation matrices are displayed with oblique rotations.



EFA: Step 5

Factor Analysis: Rotation

Method

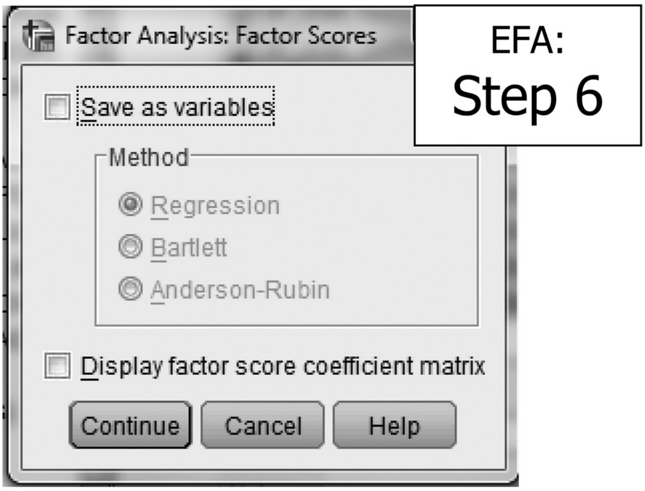
○ None ○ Quartimax
○ Varimax ○ Equamax
○ Direct Oblimin ● Promax
Delta: 0 Kappa 4

Display

☑ Rotated solution ☑ Loading plot(s)

Maximum Iterations for Convergence: 1000

Step 6. From the Factor Analysis dialog box (see screenshot Step 2), clicking on "Scores" will provide the option to save the variables created as composite scores and to display the factor score coefficient matrix (see screenshot Step 6). Many times, researchers select to skip this step and use methods such as the mean sum (i.e., adding all the items together and dividing by the number of items) as a method to create the composite score. If you do choose to allow the software to create your composite score, there are three methods from which to choose to estimate the factor score coefficients. The *regression method* produces factor scores that have a mean of 0 and a variance that equals the squared multiple correlation between the estimated factor scores and the true factor values. The factor scores estimated from the regression method may be correlated even if the factors are orthogonal. The *Bartlett score* produces factor scores that have a mean of 0. This method minimizes



EFA: Step 6

Factor Analysis: Factor Scores

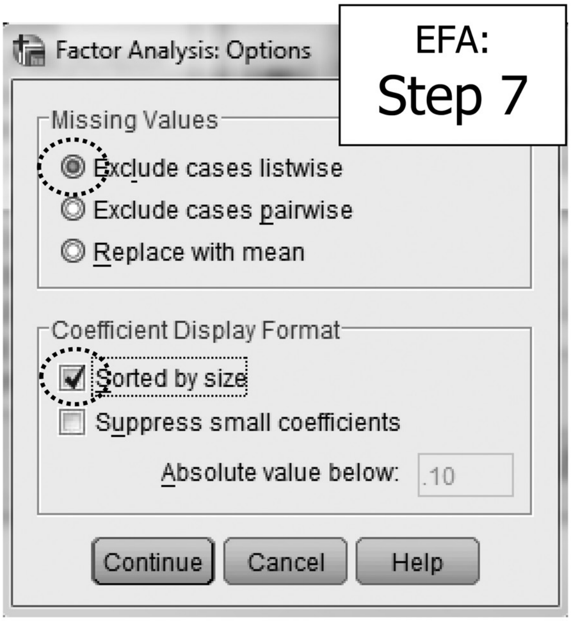
□ Save as variables

Method

● Regression
○ Bartlett
○ Anderson-Rubin

□ Display factor score coefficient matrix the sum of squares of the unique factors over the range of variables. The *Anderson-Rubin method* is a modified Bartlett method that produces factor scores with a mean of 0 and standard deviation of 1 and that maintains orthogonality of the estimated factors. Thus, the scores produced are uncorrelated. At this time, do not make any selections on this screen, as we will adhere to the mean sum method for creating a composite score. Click on "Continue" to return to the Factor Analysis dialog box.

Step 7. From the Factor Analysis dialog box (see screenshot Step 2), clicking on "Options" will bring up the dialog box that allows various options for dealing with missing values, as well as options for displaying the coefficients (see screenshot Step 7). We will leave the default setting for the Missing Values as 'exclude cases listwise.' For our purposes, because we have already dealt with missing values, which selection is made for missing values is moot. As you conduct your own research, however, should you have missing values, it should be dealt with prior to generating the factor analysis and not within the EFA, as none of the three options provided are acceptable means for which to address missing values (the exception may be if you have an extremely small percentage of missing, such as 5% or less). Under the heading for Coefficient Display Format, place a checkmark in the box for 'sorted by size.' This will make it much easier to see the clusters of variables produced in the factor solution, as it groups the items by factor in descending order of factor-loading size. Then click on "Continue" to return to the Factor Analysis dialog box. From the "Factor Analysis" dialog box, click on "OK" to generate the output.



Factor Analysis: Options

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Coefficient Display Format

☑ Sorted by size

□ Suppress small coefficients

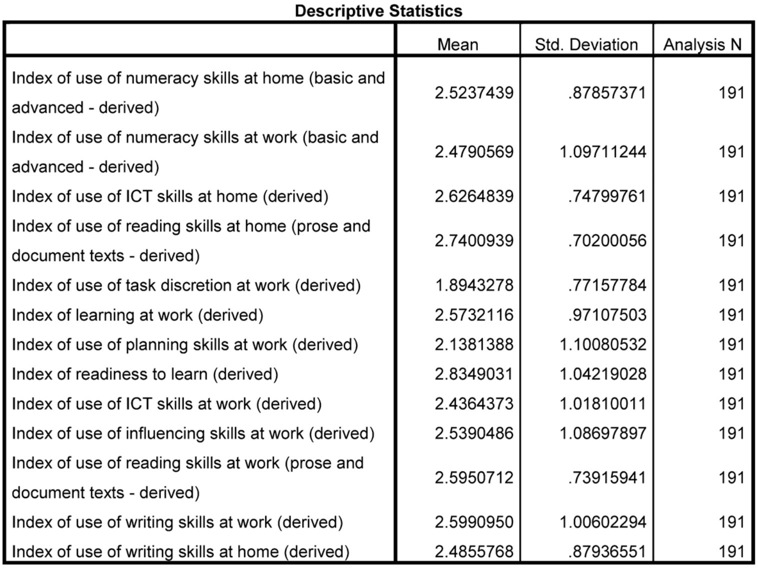
Absolute value below: .10

EFA: Step 7

Interpreting the output. Annotated results are presented in Table 9.4.

TABLE 9.4

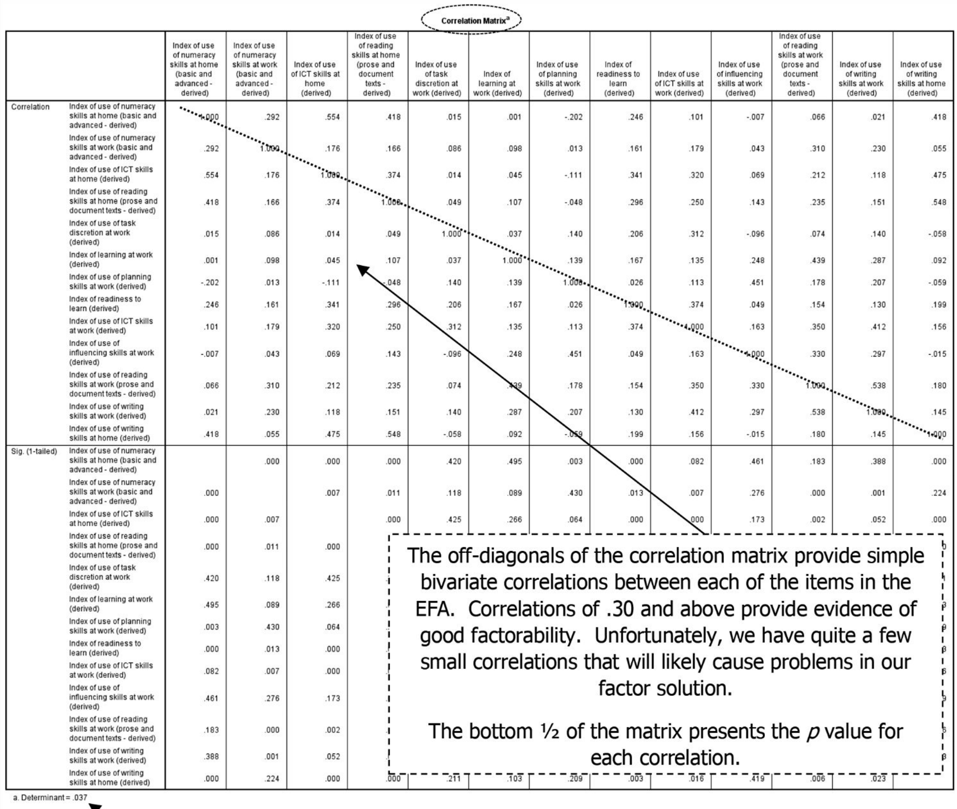
SPSS Results for the Exploratory Factor Analysis Example



Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Index of use of numeracy skills at home (basic and advanced - derived)	2.5237439	.87857371	191
Index of use of numeracy skills at work (basic and advanced - derived)	2.4790569	1.09711244	191
Index of use of ICT skills at home (derived)	2.6264839	.74799761	191
Index of use of reading skills at home (prose and document texts - derived)	2.7400939	.70200056	191
Index of use of task discretion at work (derived)	1.8943278	.77157784	191
Index of learning at work (derived)	2.5732116	.97107503	191
Index of use of planning skills at work (derived)	2.1381388	1.10080532	191
Index of readiness to learn (derived)	2.8349031	1.04219028	191
Index of use of ICT skills at work (derived)	2.4364373	1.01810011	191
Index of use of influencing skills at work (derived)	2.5390486	1.08697897	191
Index of use of reading skills at work (prose and document texts - derived)	2.5950712	.73915941	191
Index of use of writing skills at work (derived)	2.5990950	1.00602294	191
Index of use of writing skills at home (derived)	2.4855768	.87936551	191

This table provides information on the mean, standard deviation, and sample size for each variable in our EFA. No missing data is reflected by the 'analysis *N*' being the same for each item.



Correlation Matrix[a]

		Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of ICT skills at home (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)	Index of learning at work (derived)	Index of use of planning skills at work (derived)	Index of readiness to learn (derived)	Index of use of ICT skills at work (derived)	Index of use of influencing skills at work (derived)	Index of use of reading skills at work (prose and document texts - derived)	Index of use of writing skills at work (derived)	Index of use of writing skills at home (derived)
Correlation	Index of use of numeracy skills at home (basic and advanced - derived)	1.000	.292	.554	.418	.015	.001	-.202	.246	.101	-.007	.066	.021	.418
	Index of use of numeracy skills at work (basic and advanced - derived)	.292	1.000	.176	.166	.086	.098	.013	.161	.179	.043	.310	.230	.055
	Index of use of ICT skills at home (derived)	.554	.176	1.000	.374	.014	.045	-.111	.341	.320	.069	.212	.118	.475
	Index of use of reading skills at home (prose and document texts - derived)	.418	.166	.374	1.000	.049	.107	-.048	.296	.250	.143	.235	.151	.548
	Index of use of task discretion at work (derived)	.015	.086	.014	.049	1.000	.037	.140	.206	.312	-.096	.074	.140	-.058
	Index of learning at work (derived)	.001	.098	.045	.107	.037	1.000	.139	.167	.135	.248	.439	.287	.092
	Index of use of planning skills at work (derived)	-.202	.013	-.111	-.048	.140	.139	1.000	.026	.113	.451	.178	.207	-.059
	Index of readiness to learn (derived)	.246	.161	.341	.296	.206	.167	.026	1.000	.374	.049	.154	.130	.199
	Index of use of ICT skills at work (derived)	.101	.179	.320	.250	.312	.135	.113	.374	1.000	.163	.350	.412	.156
	Index of use of influencing skills at work (derived)	-.007	.043	.069	.143	-.096	.248	.451	.049	.163	1.000	.330	.297	-.015
	Index of use of reading skills at work (prose and document texts - derived)	.066	.310	.212	.235	.074	.439	.178	.154	.350	.330	1.000	.538	.180
	Index of use of writing skills at work (derived)	.021	.230	.118	.151	.140	.287	.207	.130	.412	.297	.538	1.000	.145
	Index of use of writing skills at home (derived)	.418	.055	.475	.548	-.058	.092	-.059	.199	.156	-.015	.180	.145	1.000
Sig. (1-tailed)	Index of use of numeracy skills at home (basic and advanced - derived)		.000	.000	.000	.420	.495	.003	.000	.082	.461	.183	.388	.000
	Index of use of numeracy skills at work (basic and advanced - derived)	.000		.007	.011	.118	.089	.430	.013	.007	.276	.000	.001	.224
	Index of use of ICT skills at home (derived)	.000	.007		.000	.425	.266	.064	.000	.000	.173	.002	.052	.000
	Index of use of reading skills at home (prose and document texts - derived)	.000	.011	.000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of task discretion at work (derived)	.420	.118	.425	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of learning at work (derived)	.495	.089	.266	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of planning skills at work (derived)	.003	.430	.064	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of readiness to learn (derived)	.000	.013	.000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of ICT skills at work (derived)	.082	.007	.000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of influencing skills at work (derived)	.461	.276	.173	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of reading skills at work (prose and document texts - derived)	.183	.000	.002	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of writing skills at work (derived)	.388	.001	.052	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Index of use of writing skills at home (derived)	.000	.224	.000	.000	.211	.103	.209	.003	.016	.419	.006	.023	

a. Determinant = .037

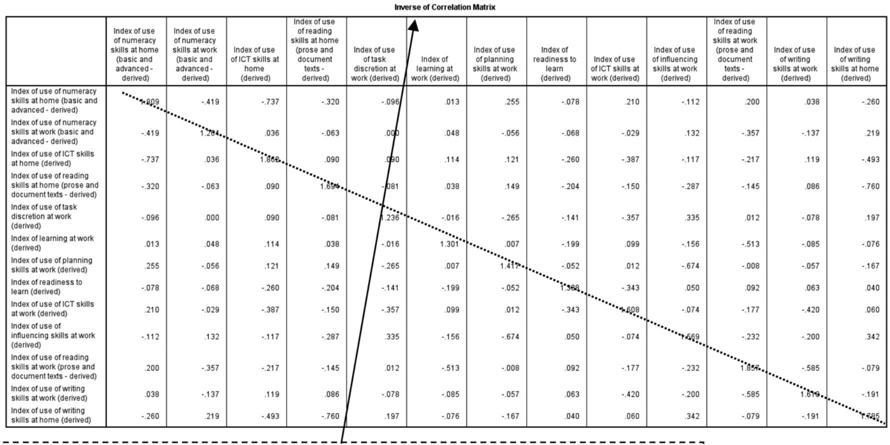
The off-diagonals of the correlation matrix provide simple bivariate correlations between each of the items in the EFA. Correlations of .30 and above provide evidence of good factorability. Unfortunately, we have quite a few small correlations that will likely cause problems in our factor solution.

The bottom ½ of the matrix presents the *p* value for each correlation.

The footer provides information on the determinant. Non-zero determinant values will help ensure the factor solution can be computed.

TABLE 9.4 (continued)

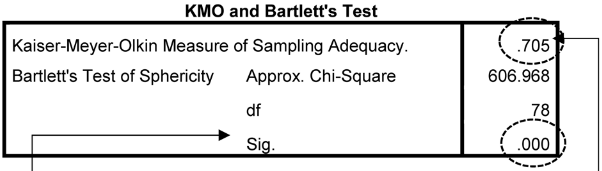
SPSS Results for the Exploratory Factor Analysis Example



Inverse of Correlation Matrix

	Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of ICT skills at home (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)	Index of learning at work (derived)	Index of use of planning skills at work (derived)	Index of readiness to learn (derived)	Index of use of ICT skills at work (derived)	Index of use of influencing skills at work (derived)	Index of use of reading skills at work (prose and document texts - derived)	Index of use of writing skills at work (derived)	Index of use of writing skills at home (derived)
Index of use of numeracy skills at home (basic and advanced - derived)	1.809	-.419	-.737	-.320	-.096	.013	.255	-.078	.210	-.112	.200	.038	-.260
Index of use of numeracy skills at work (basic and advanced - derived)	-.419	1.264	.036	-.063	.000	.048	-.056	-.068	-.029	.132	-.357	-.137	.219
Index of use of ICT skills at home (derived)	-.737	.036	1.862	.090	.090	.114	.121	-.260	-.387	-.117	-.217	.119	-.493
Index of use of reading skills at home (prose and document texts - derived)	-.320	-.063	.090	1.694	-.081	.038	.149	-.204	-.150	-.287	-.145	.086	-.760
Index of use of task discretion at work (derived)	-.096	.000	.090	-.081	1.236	-.016	-.265	-.141	-.357	.335	.012	-.078	.197
Index of learning at work (derived)	.013	.048	.114	.038	-.016	1.301	.007	-.199	.099	-.156	-.513	-.085	-.076
Index of use of planning skills at work (derived)	.255	-.056	.121	.149	-.265	.007	1.417	-.052	.012	-.674	-.008	-.057	-.167
Index of readiness to learn (derived)	-.078	-.068	-.260	-.204	-.141	-.199	-.052	1.388	-.343	.050	.092	.063	.040
Index of use of ICT skills at work (derived)	.210	-.029	-.387	-.150	-.357	.099	.012	-.343	1.608	-.074	-.177	-.420	.060
Index of use of influencing skills at work (derived)	-.112	.132	-.117	-.287	.335	-.156	-.674	.050	-.074	1.669	-.232	-.200	.342
Index of use of reading skills at work (prose and document texts - derived)	.200	-.357	-.217	-.145	.012	-.513	-.008	.092	-.177	-.232	1.857	-.585	-.079
Index of use of writing skills at work (derived)	.038	-.137	.119	.086	-.078	-.085	-.057	.063	-.420	-.200	-.585	1.619	-.191
Index of use of writing skills at home (derived)	-.260	.219	-.493	-.760	.197	-.076	-.167	.040	.060	.342	-.079	-.191	1.785

We won't spend a lot of time examining the inverted correlation but this is another method that can be used to determine factorability. The off-diagonals should be close to zero (Guttman, 1953). Kaiser (1970) extended this work, and thus we'll review the KMO MSA as a technique for factorability.



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.705
Bartlett's Test of Sphericity	Approx. Chi-Square	606.968
	df	78
	Sig.	.000

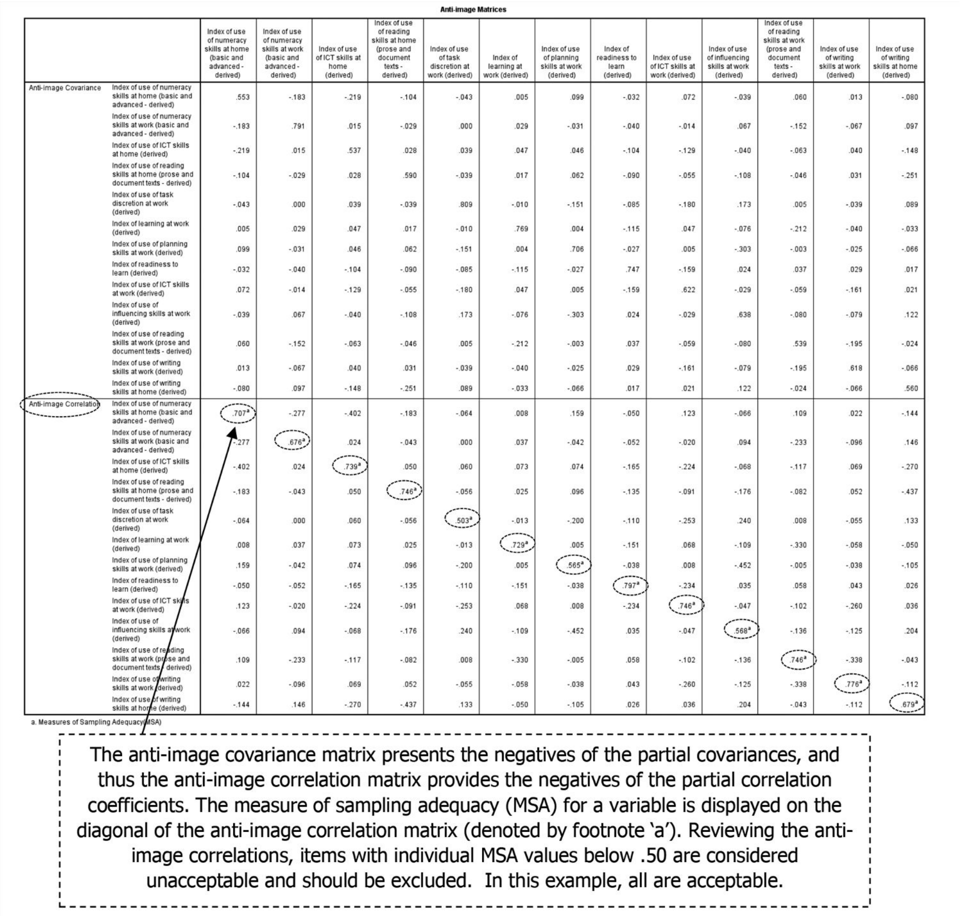
Measure of Sampling Adequacy (MSA): An overall MSA of .50 or above should be achieved before proceeding with factor analysis. According to Kaiser and Rice (1974), our MSA is 'middling.'

Bartlett's Test of Sphericity: This is a statistical test to determine if the overall correlation matrix is an identity matrix (i.e., the null hypothesis is that our overall correlation matrix is equal to an identity matrix), and thus we want to find statistical significance here—suggesting that we do not have an identity matrix. In practical terms, a statistically significant Bartlett's test indicates at least some of the variables have significant correlations. *A statistically significant Bartlett's test is desirable and suggests factor analysis is appropriate.*

In this example, we've met both criteria suggesting it is appropriate to factor analyze our variables.

TABLE 9.4 (continued)

SPSS Results for the Exploratory Factor Analysis Example



Anti-image Matrices

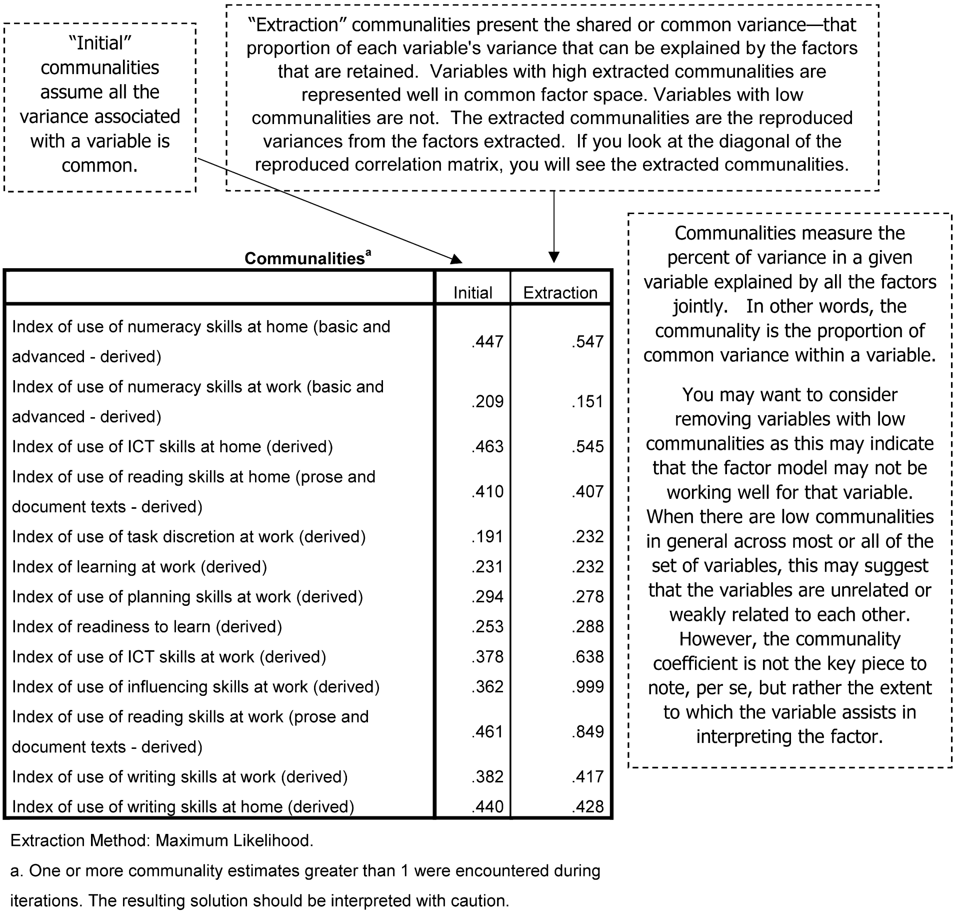
		Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of ICT skills at home (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)	Index of learning at work (derived)	Index of use of planning skills at work (derived)	Index of readiness to learn (derived)	Index of use of ICT skills at work (derived)	Index of use of influencing skills at work (derived)	Index of use of reading skills at work (prose and document texts - derived)	Index of use of writing skills at work (derived)	Index of use of writing skills at home (derived)
Anti-image Covariance	Index of use of numeracy skills at home (basic and advanced - derived)	.553	-.183	-.219	-.104	-.043	.005	.099	-.032	.072	-.039	.060	.013	-.080
	Index of use of numeracy skills at work (basic and advanced - derived)	-.183	.791	.015	-.029	.000	.029	-.031	-.040	-.014	.067	-.152	-.067	.097
	Index of use of ICT skills at home (derived)	-.219	.015	.537	.028	.039	.047	.046	-.104	-.129	-.040	-.063	.040	-.148
	Index of use of reading skills at home (prose and document texts - derived)	-.104	-.029	.028	.590	-.039	.017	.062	-.090	-.055	-.108	-.046	.031	-.251
	Index of use of task discretion at work (derived)	-.043	.000	.039	-.039	.809	-.010	-.151	-.085	-.180	.173	.005	-.039	.089
	Index of learning at work (derived)	.005	.029	.047	.017	-.010	.769	.004	-.115	.047	-.076	-.212	-.040	-.033
	Index of use of planning skills at work (derived)	.099	-.031	.046	.062	-.151	.004	.706	-.027	.005	-.303	-.003	-.025	-.066
	Index of readiness to learn (derived)	-.032	-.040	-.104	-.090	-.085	-.115	-.027	.747	-.159	.024	.037	.029	.017
	Index of use of ICT skills at work (derived)	.072	-.014	-.129	-.055	-.180	.047	.005	-.159	.622	-.029	-.059	-.161	.021
	Index of use of influencing skills at work (derived)	-.039	.067	-.040	-.108	.173	-.076	-.303	.024	-.029	.638	-.080	-.079	.122
	Index of use of reading skills at work (prose and document texts - derived)	.060	-.152	-.063	-.046	.005	-.212	-.003	.037	-.059	-.080	.539	-.195	-.024
	Index of use of writing skills at work (derived)	.013	-.067	.040	.031	-.039	-.040	-.025	.029	-.161	-.079	-.195	.618	-.066
	Index of use of writing skills at home (derived)	-.080	.097	-.148	-.251	.089	-.033	-.066	.017	.021	.122	-.024	-.066	.560
Anti-image Correlation	Index of use of numeracy skills at home (basic and advanced - derived)	.707[a]	-.277	-.402	-.183	-.064	.008	.159	-.050	.123	-.066	.109	.022	-.144
	Index of use of numeracy skills at work (basic and advanced - derived)	-.277	.676[a]	.024	-.043	.000	.037	-.042	-.052	-.020	.094	-.233	-.096	.146
	Index of use of ICT skills at home (derived)	-.402	.024	.739[a]	.050	.060	.073	.074	-.165	-.224	-.068	-.117	.069	-.270
	Index of use of reading skills at home (prose and document texts - derived)	-.183	-.043	.050	.746[a]	-.056	.025	.096	-.135	-.091	-.176	-.082	.052	-.437
	Index of use of task discretion at work (derived)	-.064	.000	.060	-.056	.503[a]	-.013	-.200	-.110	-.253	.240	.008	-.055	.133
	Index of learning at work (derived)	.008	.037	.073	.025	-.013	.729[a]	.005	-.151	.068	-.109	-.330	-.058	-.050
	Index of use of planning skills at work (derived)	.159	-.042	.074	.096	-.200	.005	.565[a]	-.038	.008	-.452	-.005	-.038	-.105
	Index of readiness to learn (derived)	-.050	-.052	-.165	-.135	-.110	-.151	-.038	.797[a]	-.234	.035	.058	.043	.026
	Index of use of ICT skills at work (derived)	.123	-.020	-.224	-.091	-.253	.068	.008	-.234	.746[a]	-.047	-.102	-.260	.036
	Index of use of influencing skills at work (derived)	-.066	.094	-.068	-.176	.240	-.109	-.452	.035	-.047	.568[a]	-.136	-.125	.204
	Index of use of reading skills at work (prose and document texts - derived)	.109	-.233	-.117	-.082	.008	-.330	-.005	.058	-.102	-.136	.746[a]	-.338	-.043
	Index of use of writing skills at work (derived)	.022	-.096	.069	.052	-.055	-.058	-.038	.043	-.260	-.125	-.338	.776[a]	-.112
	Index of use of writing skills at home (derived)	-.144	.146	-.270	-.437	.133	-.050	-.105	.026	.036	.204	-.043	-.112	.679[a]

a. Measures of Sampling Adequacy(MSA)

The anti-image covariance matrix presents the negatives of the partial covariances, and thus the anti-image correlation matrix provides the negatives of the partial correlation coefficients. The measure of sampling adequacy (MSA) for a variable is displayed on the diagonal of the anti-image correlation matrix (denoted by footnote 'a'). Reviewing the anti-image correlations, items with individual MSA values below .50 are considered unacceptable and should be excluded. In this example, all are acceptable.

TABLE 9.4 (continued)

SPSS Results for the Exploratory Factor Analysis Example



"Initial" communalities assume all the variance associated with a variable is common.

"Extraction" communalities present the shared or common variance—that proportion of each variable's variance that can be explained by the factors that are retained. Variables with high extracted communalities are represented well in common factor space. Variables with low communalities are not. The extracted communalities are the reproduced variances from the factors extracted. If you look at the diagonal of the reproduced correlation matrix, you will see the extracted communalities.

Communalities[a]

	Initial	Extraction
Index of use of numeracy skills at home (basic and advanced - derived)	.447	.547
Index of use of numeracy skills at work (basic and advanced - derived)	.209	.151
Index of use of ICT skills at home (derived)	.463	.545
Index of use of reading skills at home (prose and document texts - derived)	.410	.407
Index of use of task discretion at work (derived)	.191	.232
Index of learning at work (derived)	.231	.232
Index of use of planning skills at work (derived)	.294	.278
Index of readiness to learn (derived)	.253	.288
Index of use of ICT skills at work (derived)	.378	.638
Index of use of influencing skills at work (derived)	.362	.999
Index of use of reading skills at work (prose and document texts - derived)	.461	.849
Index of use of writing skills at work (derived)	.382	.417
Index of use of writing skills at home (derived)	.440	.428

Extraction Method: Maximum Likelihood.

a. One or more communality estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.

Communalities measure the percent of variance in a given variable explained by all the factors jointly. In other words, the communality is the proportion of common variance within a variable.

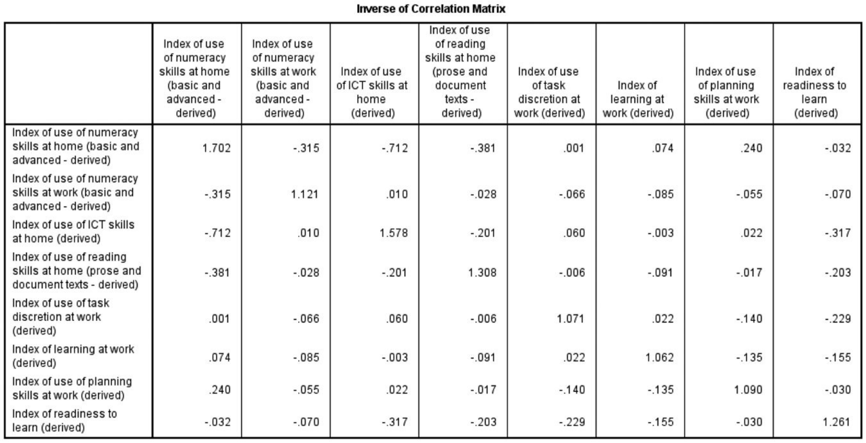
You may want to consider removing variables with low communalities as this may indicate that the factor model may not be working well for that variable. When there are low communalities in general across most or all of the set of variables, this may suggest that the variables are unrelated or weakly related to each other. However, the communality coefficient is not the key piece to note, per se, but rather the extent to which the variable assists in interpreting the factor.

For the sake of brevity, the output from the additional models generated are not presented. However, in addition to removing 'Index of use of influencing skills at work,' the following variables also had communalities greater than 1.0 and were removed through the iterative process of running the factor solution, reviewing communalities, and removing the one variable with the largest communality: 'index of using writing skills at home,' 'index of using reading skills at work (prose and document text),' 'index of using writing skills at work,' 'index of using ICT skills at work.' *Thus a total of 5 of our original 13 variables were removed from the model as they suggested they were not factorable, leaving 8 variables to factor analyze.*

Because the output prior to the communalities table has been annotated in detail, it will not be presented again. The descriptive statistics and bivariate correlation coefficients do not change because variables are removed from the set that are factor analyzed, and thus they are not presented again. Rather, we'll pick up with the tables that do reflect new, adjusted values as a result of removing variables from the set.

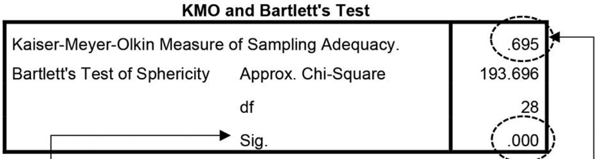
TABLE 9.4 (continued)

SPSS Results for the Exploratory Factor Analysis Example



Inverse of Correlation Matrix

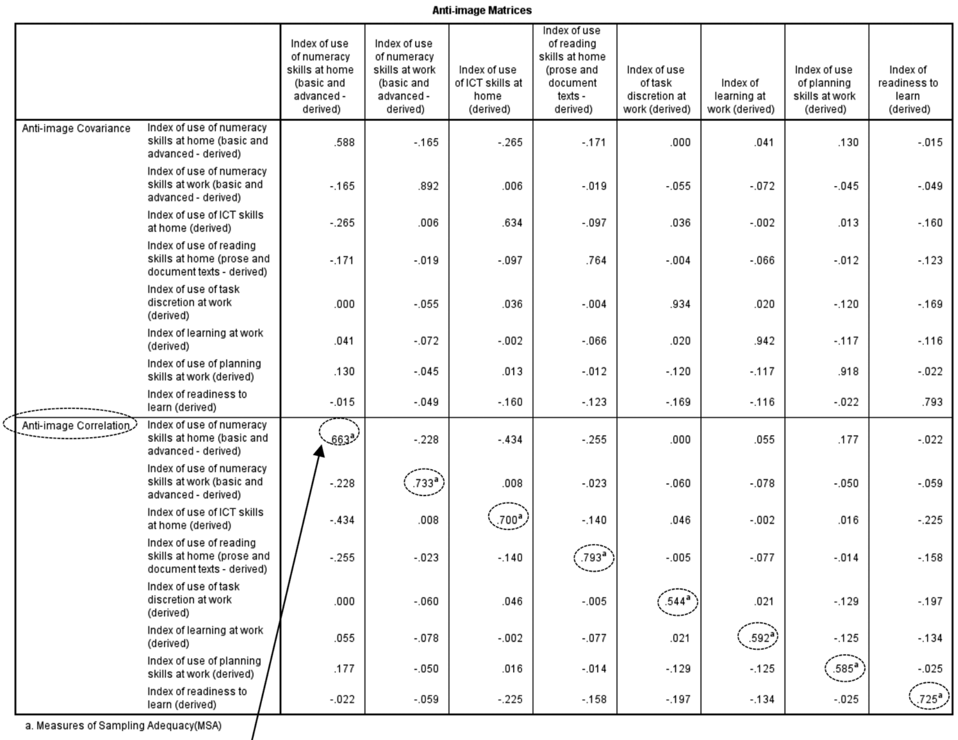
	Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of ICT skills at home (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)	Index of learning at work (derived)	Index of use of planning skills at work (derived)	Index of readiness to learn (derived)
Index of use of numeracy skills at home (basic and advanced - derived)	1.702	-.315	-.712	-.381	.001	.074	.240	-.032
Index of use of numeracy skills at work (basic and advanced - derived)	-.315	1.121	.010	-.028	-.066	-.085	-.055	-.070
Index of use of ICT skills at home (derived)	-.712	.010	1.578	-.201	.060	-.003	.022	-.317
Index of use of reading skills at home (prose and document texts - derived)	-.381	-.028	-.201	1.308	-.006	-.091	-.017	-.203
Index of use of task discretion at work (derived)	.001	-.066	.060	-.006	1.071	.022	-.140	-.229
Index of learning at work (derived)	.074	-.085	-.003	-.091	.022	1.062	-.135	-.155
Index of use of planning skills at work (derived)	.240	-.055	.022	-.017	-.140	-.135	1.090	-.030
Index of readiness to learn (derived)	-.032	-.070	-.317	-.203	-.229	-.155	-.030	1.261



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.695
Bartlett's Test of Sphericity	Approx. Chi-Square	193.696
	df	28
	Sig.	.000

With an overall MSA of .50 and a statistically significant Bartlett's test, we've again met both criteria suggesting it is appropriate to factor analyze our variables.



Anti-image Matrices

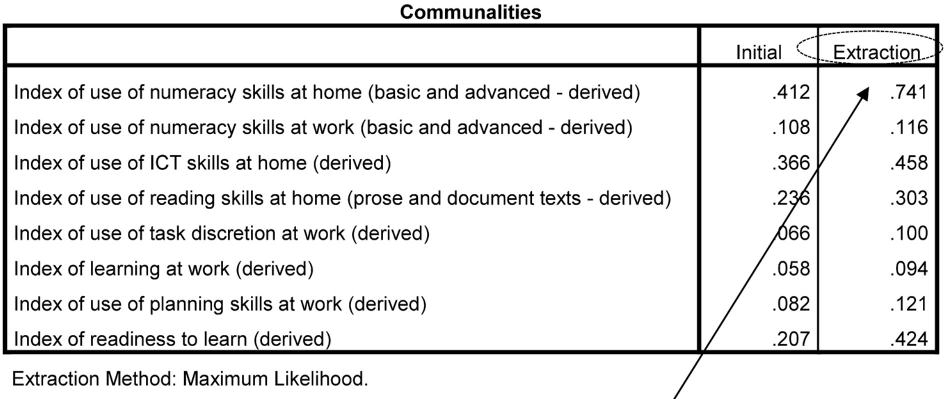
		Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of ICT skills at home (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)	Index of learning at work (derived)	Index of use of planning skills at work (derived)	Index of readiness to learn (derived)
Anti-image Covariance	Index of use of numeracy skills at home (basic and advanced - derived)	.588	-.165	-.265	-.171	.000	.041	.130	-.015
	Index of use of numeracy skills at work (basic and advanced - derived)	-.165	.892	.006	-.019	-.055	-.072	-.045	-.049
	Index of use of ICT skills at home (derived)	-.265	.006	.634	-.097	.036	-.002	.013	-.160
	Index of use of reading skills at home (prose and document texts - derived)	-.171	-.019	-.097	.764	-.004	-.066	-.012	-.123
	Index of use of task discretion at work (derived)	.000	-.055	.036	-.004	.934	.020	-.120	-.169
	Index of learning at work (derived)	.041	-.072	-.002	-.066	.020	.942	-.117	-.116
	Index of use of planning skills at work (derived)	.130	-.045	.013	-.012	-.120	-.117	.918	-.022
	Index of readiness to learn (derived)	-.015	-.049	-.160	-.123	-.169	-.116	-.022	.793
Anti-image Correlation	Index of use of numeracy skills at home (basic and advanced - derived)	.663[a]	-.228	-.434	-.255	.000	.055	.177	-.022
	Index of use of numeracy skills at work (basic and advanced - derived)	-.228	.733[a]	.008	-.023	-.060	-.078	-.050	-.059
	Index of use of ICT skills at home (derived)	-.434	.008	.700[a]	-.140	.046	-.002	.016	-.225
	Index of use of reading skills at home (prose and document texts - derived)	-.255	-.023	-.140	.793[a]	-.005	-.077	-.014	-.158
	Index of use of task discretion at work (derived)	.000	-.060	.046	-.005	.544[a]	.021	-.129	-.197
	Index of learning at work (derived)	.055	-.078	-.002	-.077	.021	.592[a]	-.125	-.134
	Index of use of planning skills at work (derived)	.177	-.050	.016	-.014	-.129	-.125	.585[a]	-.025
	Index of readiness to learn (derived)	-.022	-.059	-.225	-.158	-.197	-.134	-.025	.725[a]

a. Measures of Sampling Adequacy(MSA)

Reviewing the anti-image correlations, items with individual MSA values below .50 are considered unacceptable and should be excluded. In this example, all are again acceptable.

TABLE 9.4 (continued)

SPSS Results for the Exploratory Factor Analysis Example

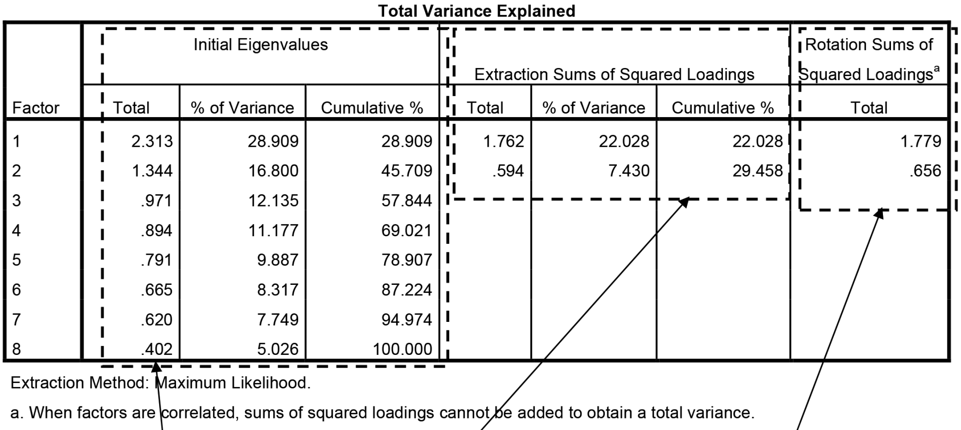


Communalities

	Initial	Extraction
Index of use of numeracy skills at home (basic and advanced - derived)	.412	.741
Index of use of numeracy skills at work (basic and advanced - derived)	.108	.116
Index of use of ICT skills at home (derived)	.366	.458
Index of use of reading skills at home (prose and document texts - derived)	.236	.303
Index of use of task discretion at work (derived)	.066	.100
Index of learning at work (derived)	.058	.094
Index of use of planning skills at work (derived)	.082	.121
Index of readiness to learn (derived)	.207	.424

Extraction Method: Maximum Likelihood.

Although all our communalities are now under the threshold of 1.0 so as not to generate a warning, we do see that we still have some communalities that are relatively small (under .30), however we will retain them in our set of variables given that we have already removed quite a few variables.



Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.313	28.909	28.909	1.762	22.028	22.028	1.779
2	1.344	16.800	45.709	.594	7.430	29.458	.656
3	.971	12.135	57.844				
4	.894	11.177	69.021				
5	.791	9.887	78.907				
6	.665	8.317	87.224				
7	.620	7.749	94.974				
8	.402	5.026	100.000				

Extraction Method: Maximum Likelihood.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

'Initial eigenvalues' presents the variance explained by the initial solution. Only two factors in the initial solution have eigenvalues greater than 1, accounting for about 46% of the variability in the original variables. This suggests much unexplained variation. While we do not suggest applying Kaiser's rule to determine the number of factors to extract, if you do apply that rule to your data, the 'initial eigenvalues' should be eigenvalues reviewed to make the decision as these eigenvalues are derived from the unreduced input correlation matrix.

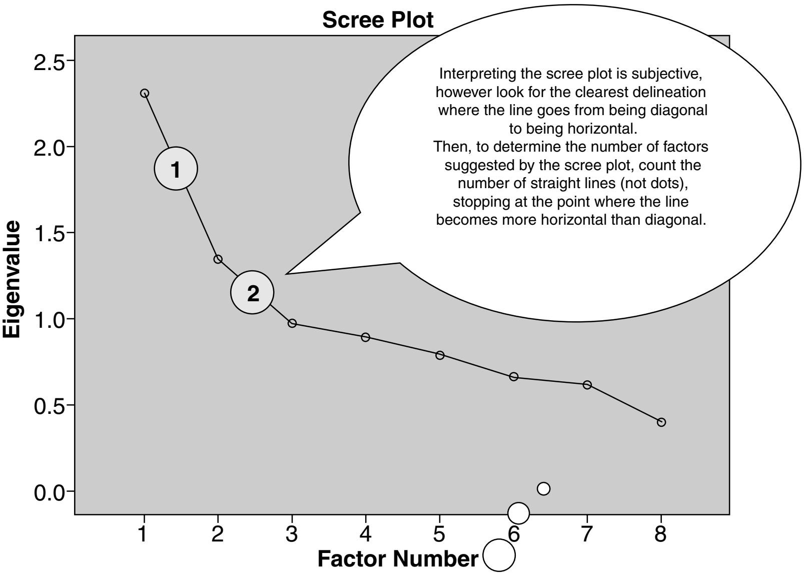
'Extraction sum of squared loadings' presents the variance explained by the extracted factors *before rotation.* The cumulative variability explained by the two factor extracted solution is about 29%, about 17% less than the initial solution. This means about 17% of the variation explained by the initial solution is lost as a result of factors unique to the original variables and variability unexplained by the factor solution.

In the social and behavioral sciences, it is common to find around 60% of the total variance explained in factor analytic models (Child, 2006).

'Rotation sums of squared loadings' provides the variance explained by the extracted factors *after rotation.* Note the footer regarding our oblique rotation (i.e., sums of squared factor loadings) cannot be added together to reflect total variance.

TABLE 9.4 (continued)

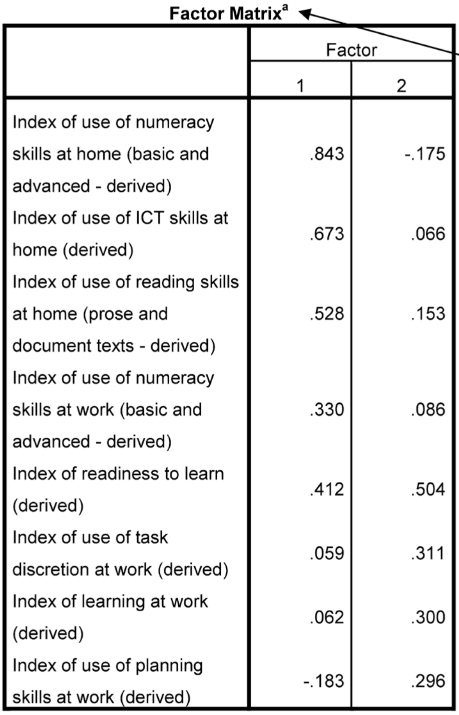
SPSS Results for the Exploratory Factor Analysis Example



Interpreting the scree plot is subjective, however look for the clearest delineation where the line goes from being diagonal to being horizontal.
Then, to determine the number of factors suggested by the scree plot, count the number of straight lines (not dots), stopping at the point where the line becomes more horizontal than diagonal.

NOTE!
Remember that there is a fair amount of subjectivity in interpreting the scree plot. With this graph, there appear to clearly be two factors with much less variance accounted for by factors three through five.

Factor Matrix[a]



	Factor	
	1	2
Index of use of numeracy skills at home (basic and advanced - derived)	.843	-.175
Index of use of ICT skills at home (derived)	.673	.066
Index of use of reading skills at home (prose and document texts - derived)	.528	.153
Index of use of numeracy skills at work (basic and advanced - derived)	.330	.086
Index of readiness to learn (derived)	.412	.504
Index of use of task discretion at work (derived)	.059	.311
Index of learning at work (derived)	.062	.300
Index of use of planning skills at work (derived)	-.183	.296

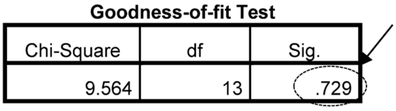
Extraction Method: Maximum Likelihood.

a. 2 factors extracted. 6 iterations required.

The factor matrix presents the unrotated solution. We rotated our solution (and this will nearly always be the case), thus we are not interested in these results.

TABLE 9.4 (continued)

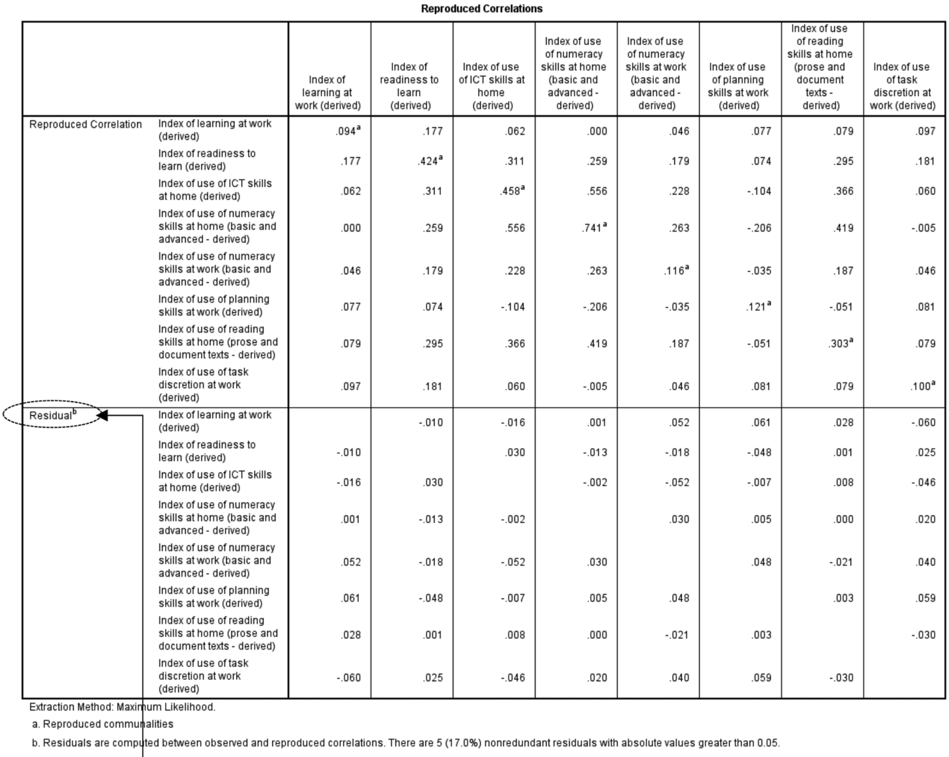
SPSS Results for the Exploratory Factor Analysis Example



Goodness-of-fit Test

Chi-Square	df	Sig.
9.564	13	.729

The null hypothesis for the goodness-of-fit test is that the factor model sufficiently describes the data. The results of this test provide evidence of the extent to which our factor solution reproduces the variance-covariance matrix. In this example, we fail to reject the null hypothesis providing evidence that the factor model does indeed describe the data—in other words, the relationships among the variables is sufficiently described by the factor model and good fit is suggested.



Reproduced Correlations

		Index of learning at work (derived)	Index of readiness to learn (derived)	Index of use of ICT skills at home (derived)	Index of use of numeracy skills at home (basic and advanced - derived)	Index of use of numeracy skills at work (basic and advanced - derived)	Index of use of planning skills at work (derived)	Index of use of reading skills at home (prose and document texts - derived)	Index of use of task discretion at work (derived)
Reproduced Correlation	Index of learning at work (derived)	.094[a]	.177	.062	.000	.046	.077	.079	.097
	Index of readiness to learn (derived)	.177	.424[a]	.311	.259	.179	.074	.295	.181
	Index of use of ICT skills at home (derived)	.062	.311	.458[a]	.556	.228	-.104	.366	.060
	Index of use of numeracy skills at home (basic and advanced - derived)	.000	.259	.556	.741[a]	.263	-.206	.419	-.005
	Index of use of numeracy skills at work (basic and advanced - derived)	.046	.179	.228	.263	.116[a]	-.035	.187	.046
	Index of use of planning skills at work (derived)	.077	.074	-.104	-.206	-.035	.121[a]	-.051	.081
	Index of use of reading skills at home (prose and document texts - derived)	.079	.295	.366	.419	.187	-.051	.303[a]	.079
	Index of use of task discretion at work (derived)	.097	.181	.060	-.005	.046	.081	.079	.100[a]
Residual[b]	Index of learning at work (derived)		-.010	-.016	.001	.052	.061	.028	-.060
	Index of readiness to learn (derived)	-.010		.030	-.013	-.018	-.048	.001	.025
	Index of use of ICT skills at home (derived)	-.016	.030		-.002	-.052	-.007	.008	-.046
	Index of use of numeracy skills at home (basic and advanced - derived)	.001	-.013	-.002		.030	.005	.000	.020
	Index of use of numeracy skills at work (basic and advanced - derived)	.052	-.018	-.052	.030		.048	-.021	.040
	Index of use of planning skills at work (derived)	.061	-.048	-.007	.005	.048		.003	.059
	Index of use of reading skills at home (prose and document texts - derived)	.028	.001	.008	.000	-.021	.003		-.030
	Index of use of task discretion at work (derived)	-.060	.025	-.046	.020	.040	.059	-.030	

Extraction Method: Maximum Likelihood.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 5 (17.0%) nonredundant residuals with absolute values greater than 0.05.

The correlation matrix based on the extracted factors is the **reproduced correlation matrix,** and these coefficients should be very close to the values in the original correlation matrix. When that happens (i.e., the reproduced and original coefficients are very close in value), the values in the residual matrix will be close to zero (as the residual values reflect the difference—as simple subtraction—between the original and the reproduced matrix) *and* the extracted factors account for much of the variance in the original correlation matrix—therefore, the extracted factors represent the original data well. The values on the diagonal of the reproduced correlation matrix are also the values presented as extracted communalities.

In this example, our residuals are quite small, the largest being about .50 in absolute value, suggesting the extracted factors are representing the original data well.

TABLE 9.4 (continued)

SPSS Results for the Exploratory Factor Analysis Example

Pattern Matrix[a]

	Factor	
	1	2
Index of use of numeracy skills at home (basic and advanced - derived)	.869	-.219
Index of use of ICT skills at home (derived)	.670	.034
Index of use of reading skills at home (prose and document texts - derived)	.514	.129
Index of use of numeracy skills at work (basic and advanced - derived)	.322	.071
Index of readiness to learn (derived)	.355	.490
Index of use of task discretion at work (derived)	.022	.312
Index of use of planning skills at work (derived)	-.219	.309
Index of learning at work (derived)	.027	.301

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

The rotation we selected was oblique rotation (assuming correlated factors) using promax rotation. Oblique rotations will produce *both* a factor **pattern matrix** (which are the coefficients for the linear combination of the variables; the factor loadings of each variable onto the factor) and a factor **structure matrix** (which are correlations between the variables and the factors—the product of the pattern and factor correlation matrices—thus taking into account the relationship between factors). It has been suggested that both the pattern and structure matrices be used to interpret the factors (Gorsuch, 1983)

The pattern and structure matrices will be identical with orthogonal rotations (recall that orthogonal means that the factors are not assumed to correlate).

Structure Matrix

	Factor	
	1	2
Index of use of numeracy skills at home (basic and advanced - derived)	.833	-.076
Index of use of ICT skills at home (derived)	.676	.
Index of use of reading skills at home (prose and document texts - derived)	.535	.214
Index of use of numeracy skills at work (basic and advanced - derived)	.333	.124
Index of readiness to learn (derived)	.436	.549
Index of use of task discretion at work (derived)	.074	.315
Index of learning at work (derived)	.077	.305
Index of use of planning skills at work (derived)	-.168	.273

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

The factors are identified by reviewing the coefficients. Variables that are associated with factor 1 have high values for factor 1 but not factor 2. *(Note that this is where sorting by size is important as the software automatically arranges the variables in descending value, making it easy to see the set of variables that load most strongly on each factor.)*

Although the values in the pattern and structure matrices are slightly different, both suggest that the same variables load on the same factors.

Factor Correlation Matrix

Factor	1	2
1	1.000	.165
2	.165	1.000

Extraction Method: Maximum Likelihood.

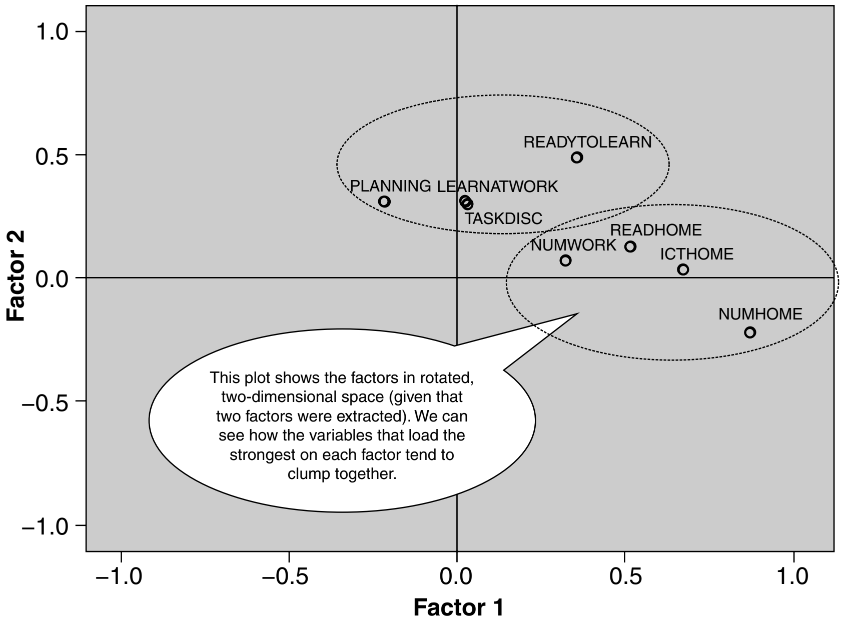
Rotation Method: Promax with Kaiser Normalization.

The factor correlations are generated only when oblique rotation is selected (correlations with orthogonal rotations are set to zero). In this example, the correlations between factors are relatively weak which may warrant re-analysis assuming orthogonality.

TABLE 9.4 (continued)

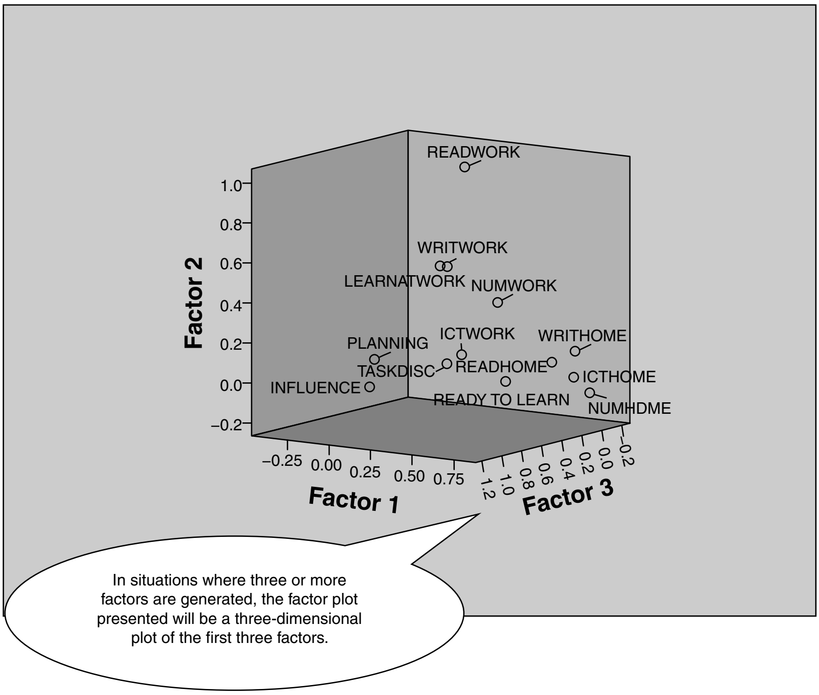
SPSS Results for the Exploratory Factor Analysis Example

Factor Plot in Rotated Factor Space



This plot shows the factors in rotated, two-dimensional space (given that two factors were extracted). We can see how the variables that load the strongest on each factor tend to clump together.

Factor Plot in Rotation Factor Space

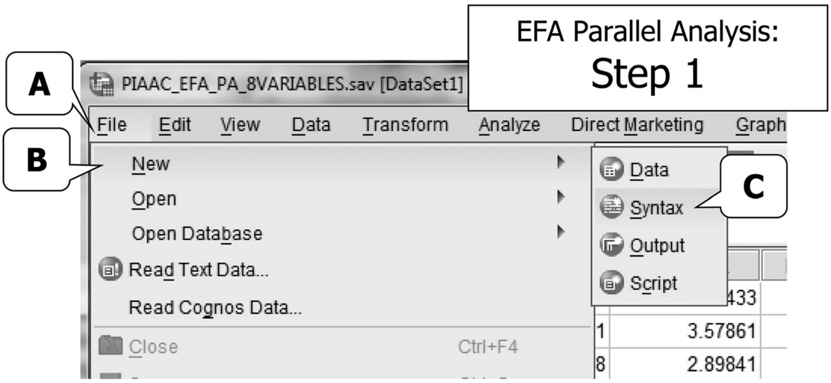


In situations where three or more factors are generated, the factor plot presented will be a three-dimensional plot of the first three factors.

9.3.1.1 SPSS Parallel Analysis for Determining Factor Retention

Next, we consider SPSS for conducting parallel analysis (PA). When you run the parallel analysis program, it is important that the data file is open so that the program will recognize that data file as the one with which to generate the PA. We will continue to work with the PIAAC_EFA.sav dataset.

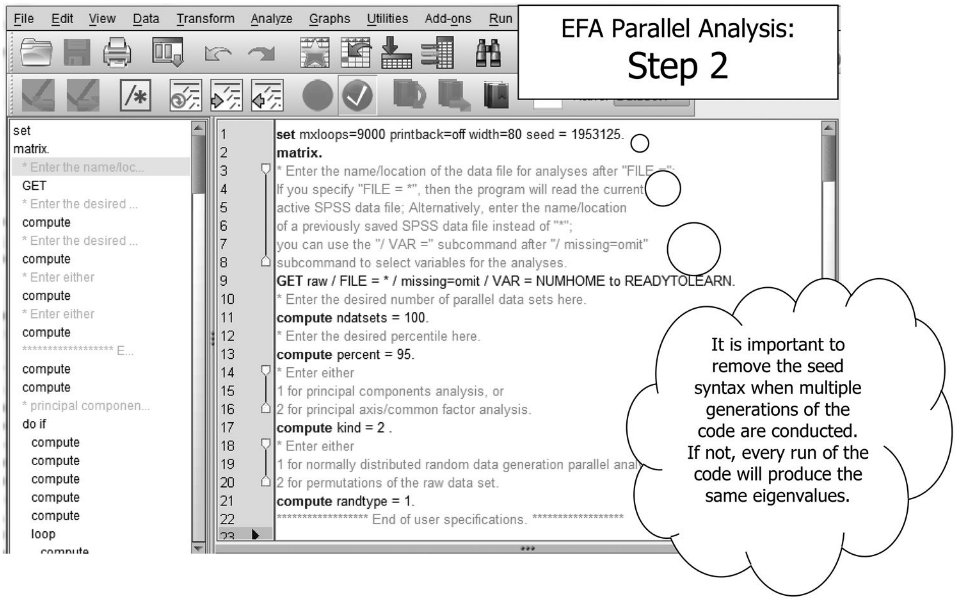
PA Step 1. As mentioned previously, this is not available in the point-and-click user interface but can easily be performed with syntax available from O'Connor (2000). [Additional annotated code, along with syntax to generate artificial raw data that may be helpful for getting a feel for how it works, is accessible online at https://people.ok.ubc.ca/brioconn/nfactors/rawpar.sps.] To open a new syntax file, click on "File" then "New" then "Syntax." Following the screenshot below (see screenshot EFA Parallel Analysis: Step 1) produces the "Syntax Editor."



Step 2. It is most helpful to access an electronic copy of the article or the supplementary online material (see https://people.ok.ubc.ca/brioconn/nfactors/rawpar.sps) so that the syntax can be copied and pasted directly into the SPSS syntax viewer (however, it has also been provided in Table 9.5). When the syntax is copied into the syntax editor, the code that needs your input will clearly be displayed (see screenshot EFA Parallel Analysis: Step 2). These include the following:

- The GET line tells SPSS that the file currently open is the one that should be used to generate the parallel analysis. For ease, the dataset we are using, PIAAC_EFA.sav, is organized so that the eight variables we will factor analyze are grouped together. In this instance, the GET syntax is GET raw / FILE = * / missing= omit / VAR = NUMHOME to READYTOLEARN. Specifying FILE = * tells the program to read the SPSS data file that is open (thus make sure there is only one dataset open when you run the program, and the one that is open is the one from which you want the parallel analysis generated). The VAR = NUMHOME to READYTOLEARN. tells the program only to generate the parallel analysis on the variables within this range (in this illustration, it happens to be the first eight variables in the data file).

- The number of parallel datasets to compute needs to be defined (100 is the default and is an appropriate starting place): compute ndatsets = 100.
- The percentile must be specified (95th is common): compute percent = 95.
- The kind of parallel analysis to compute must be specified with 1 referring to PCA and 2 referring to principal axis/common factor analysis (which is what we will generate in this illustration): compute kind = 2.
- The type of distribution must be specified with 1 being normally distributed and 2 being permutations of the raw data: compute randtype = 1. It is important to note that the distributions of the observed variables remain the same during the parallel analysis procedure. As noted by O'Connor, "Permutations of the raw data set are thus highly accurate and most relevant, especially in cases where the raw data are not normally distributed or when they do not meet the assumption of multivariate normality" (see https://people.ok.ubc.ca/brioconn/nfactors/raw-par.sps). O'Connor recommends specifying normally distributed data first (i.e., compute randtype = 1.) to get a general idea of the number of factors that the parallel analysis suggests retaining. Then specify distributions as permutations of the raw data (i.e., compute randtype = 2.) with a small number of datasets (e.g., 100) to see how long the program takes to run. Assuming the time for running the program is doable, then run the parallel analysis program with the number of parallel data sets desired for your analyses (with 1,000 generally being sufficient).



EFA Parallel Analysis:
Step 2

```
set mxloops=9000 printback=off width=80 seed = 1953125.
matrix.
* Enter the name/location of the data file for analyses after "FILE =";
If you specify "FILE = *", then the program will read the current
active SPSS data file; Alternatively, enter the name/location
of a previously saved SPSS data file instead of "*";
you can use the "/ VAR =" subcommand after "/ missing=omit"
subcommand to select variables for the analyses.
GET raw / FILE = * / missing=omit / VAR = NUMHOME to READYTOLEARN.
* Enter the desired number of parallel data sets here.
compute ndatsets = 100.
* Enter the desired percentile here.
compute percent = 95.
* Enter either
1 for principal components analysis, or
2 for principal axis/common factor analysis.
compute kind = 2 .
* Enter either
1 for normally distributed random data generation parallel anal
2 for permutations of the raw data set.
compute randtype = 1.
****************** End of user specifications. ******************
```

It is important to remove the seed syntax when multiple generations of the code are conducted. If not, every run of the code will produce the same eigenvalues.

TABLE 9.5

SPSS Syntax for Generating Parallel Analysis

```
set mxloops=9000 printback=off width=80 seed = 1953125.
matrix.
* Enter the name/location of the data file for analyses after "FILE =";
If you specify "FILE = *", then the program will read the current,
active SPSS data file; Alternatively, enter the name/location
of a previously saved SPSS data file instead of "*";
you can use the "/ VAR =" subcommand after "/ missing=omit"
subcommand to select variables for the analyses.
GET raw / FILE = * / missing=omit / VAR = NUMHOME to READYTOLEARN.
```

This tells SPSS to use the SPSS data set that is currently open and to generate the parallel analysis only on variables NUMHOME through READYTOLEARN.

```
* Enter the desired number of parallel data sets here.
compute ndatsets = 100.
```

This tells SPSS to generate 100 parallel datasets (100 is the default and is a good starting place).

```
* Enter the desired percentile here.
compute percent = 95.
```

This tells SPSS to compute the 95th percentile.

```
* Enter either
1 for principal components analysis, or
2 for principal axis/common factor analysis.
compute kind = 2 .
```

This tells SPSS to compute principal axis/common factor analysis.

```
* Enter either
1 for normally distributed random data generation parallel analysis, or
2 for permutations of the raw data set.
compute randtype = 1.
```

This tells SPSS to generate normally distributed random data.

```
****************** End of user specifications. ******************

compute ncases   = nrow(raw).
compute nvars    = ncol(raw).


* principal components analysis & random normal data generation.
do if (kind = 1 and randtype = 1).
compute nm1 = 1 / (ncases-1).
compute vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute realeval = eval(d * vcv * d).
```

TABLE 9.5 (continued)

SPSS Syntax for Generating Parallel Analysis

```
compute evals = make(nvars,ndatsets,-9999).
loop #nds = 1 to ndatsets.
compute x = sqrt(2 * (ln(uniform(ncases,nvars)) * -1) ) &*
            cos(6.283185 * uniform(ncases,nvars) ).
compute vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute evals(:,#nds) = eval(d * vcv * d).
end loop.
end if.

* principal components analysis & raw data permutation.
do if (kind = 1 and randtype = 2).
compute nm1 = 1 / (ncases-1).
compute vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute realeval = eval(d * vcv * d).
compute evals = make(nvars,ndatsets,-9999).
loop #nds = 1 to ndatsets.
compute x = raw.
loop #c = 1 to nvars.
loop #r = 1 to (ncases -1).
compute k = trunc( (ncases - #r + 1) * uniform(1,1) + 1 )  + #r - 1.
compute d = x(#r,#c).
compute x(#r,#c) = x(k,#c).
compute x(k,#c) = d.
end loop.
end loop.
compute vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/ncases)).
```

TABLE 9.5 (continued)

SPSS Syntax for Generating Parallel Analysis

```
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute evals(:,#nds) = eval(d * vcv * d).
end loop.
end if.

* PAF/common factor analysis & random normal data generation.
do if (kind = 2 and randtype = 1).
compute nm1 = 1 / (ncases-1).
compute vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute cr = (d * vcv * d).
compute smc = 1 - (1 &/ diag(inv(cr)) ).
call setdiag(cr,smc).
compute realeval = eval(cr).
compute evals = make(nvars,ndatsets,-9999).
compute nm1 = 1 / (ncases-1).
loop #nds = 1 to ndatsets.
compute x = sqrt(2 * (ln(uniform(ncases,nvars)) * -1) ) &*
            cos(6.283185 * uniform(ncases,nvars) ).
compute vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute r = d * vcv * d.
compute smc = 1 - (1 &/ diag(inv(r)) ).
call setdiag(r,smc).
compute evals(:,#nds) = eval(r).
end loop.
end if.

* PAF/common factor analysis & raw data permutation.
do if (kind = 2 and randtype = 2).
compute nm1 = 1 / (ncases-1).
compute vcv = nm1 * (sscp(raw) - ((t(csum(raw))*csum(raw))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute cr = (d * vcv * d).
```

TABLE 9.5 (continued)

SPSS Syntax for Generating Parallel Analysis

```
compute smc = 1 - (1 &/ diag(inv(cr)) ).
call setdiag(cr,smc).
compute realeval = eval(cr).
compute evals = make(nvars,ndatsets,-9999).
compute nm1 = 1 / (ncases-1).
loop #nds = 1 to ndatsets.
compute x = raw.
loop #c = 1 to nvars.
loop #r = 1 to (ncases -1).
compute k = trunc( (ncases - #r + 1) * uniform(1,1) + 1 )  + #r - 1.
compute d = x(#r,#c).
compute x(#r,#c) = x(k,#c).
compute x(k,#c) = d.
end loop.
end loop.
compute vcv = nm1 * (sscp(x) - ((t(csum(x))*csum(x))/ncases)).
compute d = inv(mdiag(sqrt(diag(vcv)))).
compute r = d * vcv * d.
compute smc = 1 - (1 &/ diag(inv(r)) ).
call setdiag(r,smc).
compute evals(:,#nds) = eval(r).
end loop.
end if.

* identifying the eigenvalues corresponding to the desired percentile.
compute num = rnd((percent*ndatsets)/100).
compute results = { t(1:nvars), realeval, t(1:nvars), t(1:nvars) }.
loop #root = 1 to nvars.
compute ranks = rnkorder(evals(#root,:)).
loop #col = 1 to ndatsets.
do if (ranks(1,#col) = num).
compute results(#root,4) = evals(#root,#col).
break.
end if.
```

TABLE 9.5 (continued)

SPSS Syntax for Generating Parallel Analysis

```
end loop.
end loop.
compute results(:,3) = rsum(evals) / ndatsets.

print /title="PARALLEL ANALYSIS:".
do if (kind = 1 and randtype = 1).
print /title="Principal Components & Random Normal Data Generation".
else if (kind = 1 and randtype = 2).
print /title="Principal Components & Raw Data Permutation".
else if (kind = 2 and randtype = 1).
print /title="PAF/Common Factor Analysis & Random Normal Data Generation".
else if (kind = 2 and randtype = 2).
print /title="PAF/Common Factor Analysis & Raw Data Permutation".
end if.
compute specifs = {ncases; nvars; ndatsets; percent}.
print specifs /title="Specifications for this Run:"
 /rlabels="Ncases" "Nvars" "Ndatsets" "Percent".
print results
 /title="Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues"
 /clabels="Root" "Raw Data" "Means" "Prcntyle"  /format "f12.6".

do if   (kind = 2).
print / space = 1.
print /title="Warning: Parallel analyses of adjusted correlation matrices".
print /title="eg, with SMCs on the diagonal, tend to indicate more factors".
print /title="than warranted (Buja, A., & Eyuboglu, N., 1992, Remarks on
parallel".
print /title="analysis. Multivariate Behavioral Research, 27, 509-540.).".
print /title="The eigenvalues for trivial, negligible factors in the real".
print /title="data commonly surpass corresponding random data eigenvalues".
print /title="for the same roots. The eigenvalues from parallel analyses".
print /title="can be used to determine the real data eigenvalues that are".
print /title="beyond chance, but additional procedures should then be used".
print /title="to trim trivial factors.".
print / space = 2.
```

TABLE 9.5 (continued)

SPSS Syntax for Generating Parallel Analysis

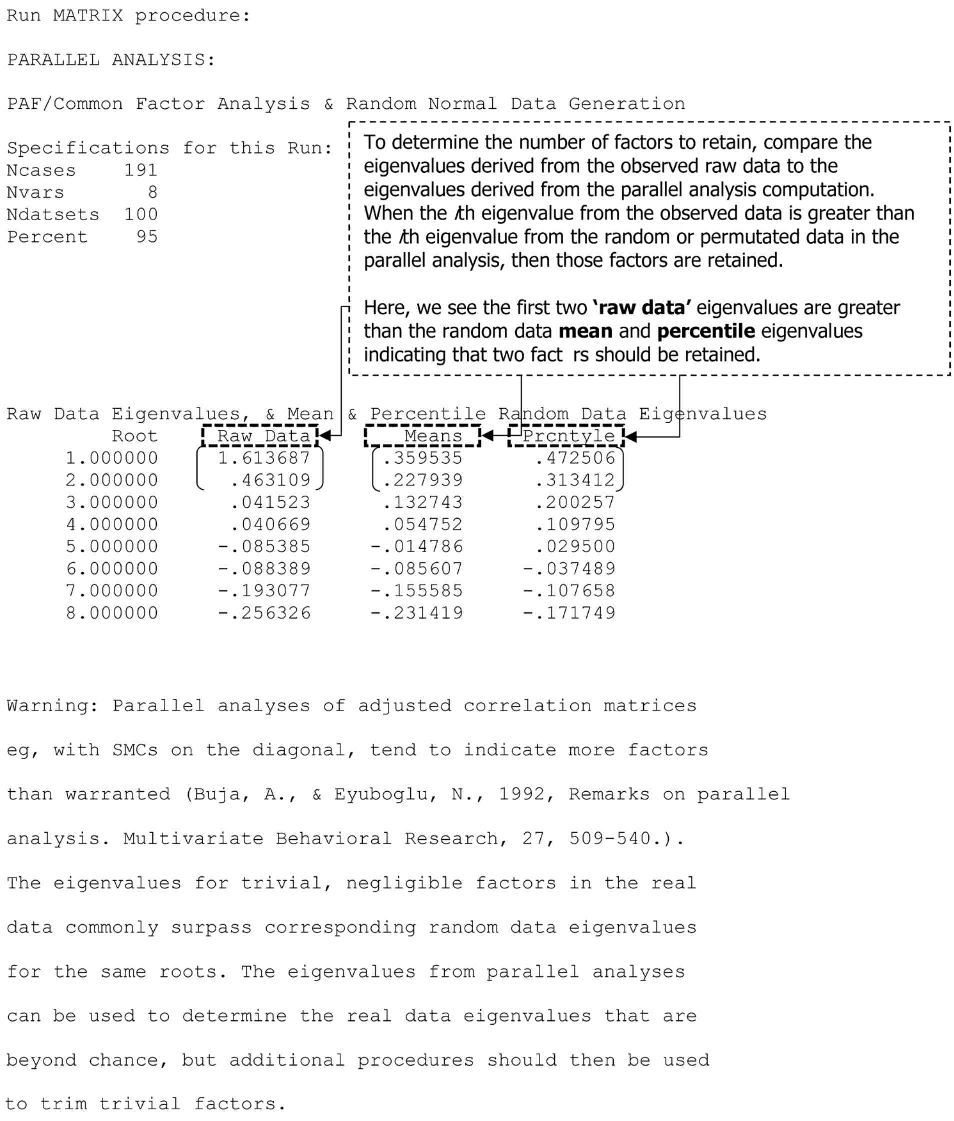
```
print /title="Principal components eigenvalues are often used to determine".
print /title="the number of common factors. This is the default in most".
print /title="statistical software packages, and it is the primary practice".
print /title="in the literature. It is also the method used by many factor".
print /title="analysis experts, including Cattell, who often examined".
print /title="principal components eigenvalues in his scree plots to
determine".
print /title="the number of common factors. But others believe this common".
print /title="practice is wrong. Principal components eigenvalues are based".
print /title="on all of the variance in correlation matrices, including both".
print /title="the variance that is shared among variables and the variances".
print /title="that are unique to the variables. In contrast, principal".
print /title="axis eigenvalues are based solely on the shared variance".
print /title="among the variables. The two procedures are qualitatively".
print /title="different. Some therefore claim that the eigenvalues from one".
print /title="extraction method should not be used to determine".
print /title="the number of factors for the other extraction method.".
print /title="The issue remains neglected and unsettled.".
end if.

compute root     = results(:,1).
compute rawdata = results(:,2).
compute percntyl = results(:,4).

save results /outfile= 'screedata.sav' / var=root rawdata means percntyl .

end matrix.
```

Step 3. Now that the syntax is created (see PA_PIAAC_n191.sps), run the program. For this data, we first generate 100 datasets using normally distributed data. Then we generate 1000 datasets using permutations of the raw data.

Interpreting the PA output. Annotated results are presented in Tables 9.6 and 9.7. More specifically, in Table 9.6 the results were generated using 100 datasets (compute ndatsets = 100.) with normally distributed random data (compute randtype = 1.). Table 9.7 results were generated using 1000 datasets (compute ndatsets = 1000.) with normally distributed random data (compute randtype = 2.). In both cases, we arrive at the same conclusion—two factors should be retained.

TABLE 9.6

SPSS Parallel Analysis Results for the Exploratory Factor Analysis Example: 100 Datasets with Normally Distributed Random Data



```
Run MATRIX procedure:

PARALLEL ANALYSIS:

PAF/Common Factor Analysis & Random Normal Data Generation

Specifications for this Run:
Ncases    191
Nvars       8
Ndatsets  100
Percent    95

Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues
        Root     Raw Data      Means      Prcntyle
    1.000000     1.613687    .359535      .472506
    2.000000      .463109    .227939      .313412
    3.000000      .041523    .132743      .200257
    4.000000      .040669    .054752      .109795
    5.000000     -.085385   -.014786      .029500
    6.000000     -.088389   -.085607     -.037489
    7.000000     -.193077   -.155585     -.107658
    8.000000     -.256326   -.231419     -.171749


Warning: Parallel analyses of adjusted correlation matrices

eg, with SMCs on the diagonal, tend to indicate more factors

than warranted (Buja, A., & Eyuboglu, N., 1992, Remarks on parallel

analysis. Multivariate Behavioral Research, 27, 509-540.).

The eigenvalues for trivial, negligible factors in the real

data commonly surpass corresponding random data eigenvalues

for the same roots. The eigenvalues from parallel analyses

can be used to determine the real data eigenvalues that are

beyond chance, but additional procedures should then be used

to trim trivial factors.
```

To determine the number of factors to retain, compare the eigenvalues derived from the observed raw data to the eigenvalues derived from the parallel analysis computation. When the *i*th eigenvalue from the observed data is greater than the *i*th eigenvalue from the random or permutated data in the parallel analysis, then those factors are retained.

Here, we see the first two **'raw data'** eigenvalues are greater than the random data **mean** and **percentile** eigenvalues indicating that two fact rs should be retained.

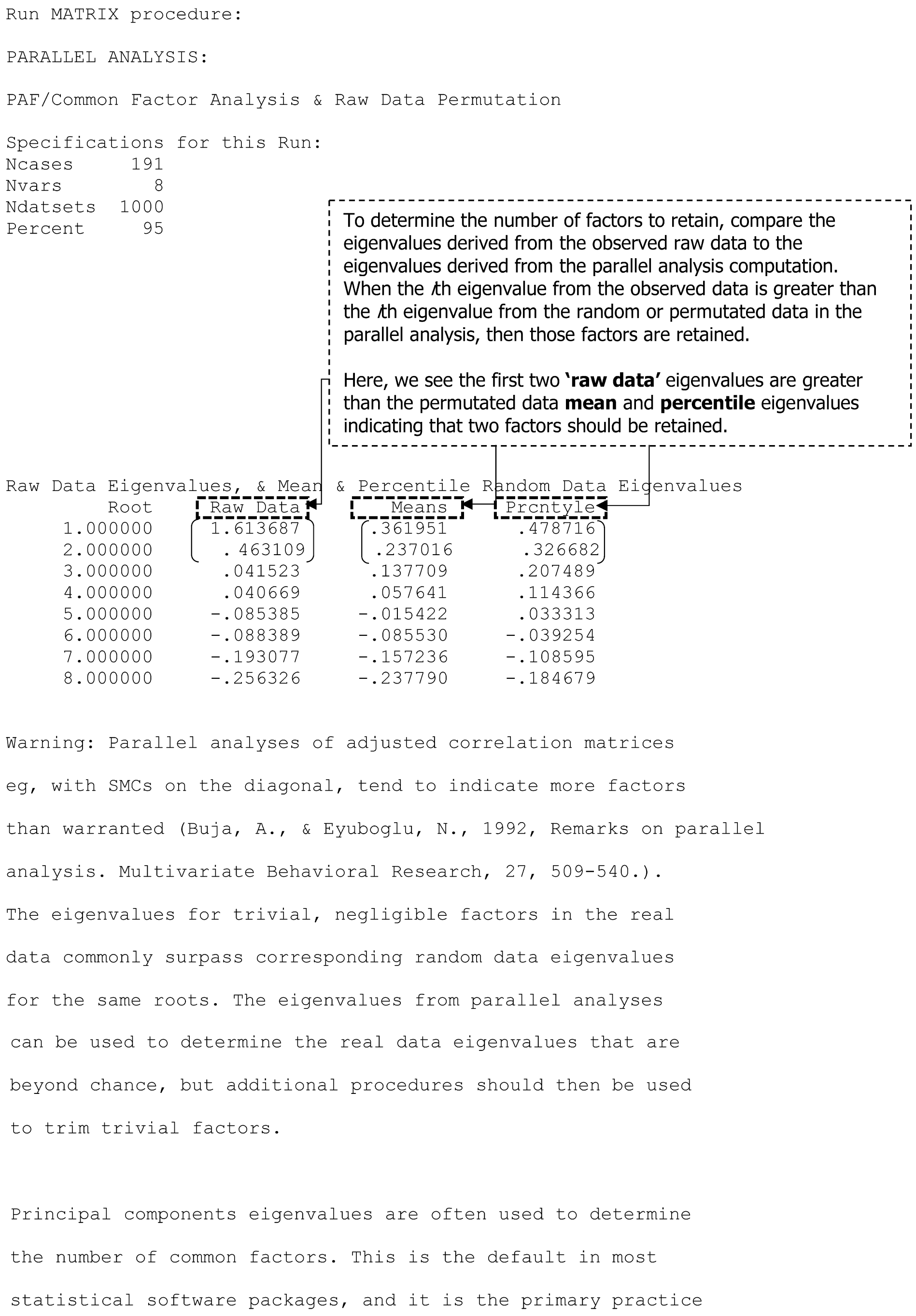
TABLE 9.6 (continued)

SPSS Parallel Analysis Results for the Exploratory Factor Analysis Example: 100 Datasets with Normally Distributed Random Data

```
Principal components eigenvalues are often used to determine
the number of common factors. This is the default in most
statistical software packages, and it is the primary practice
in the literature. It is also the method used by many factor
analysis experts, including Cattell, who often examined
principal components eigenvalues in his scree plots to determine
the number of common factors. But others believe this common
practice is wrong. Principal components eigenvalues are based
on all of the variance in correlation matrices, including both
the variance that is shared among variables and the variances
that are unique to the variables. In contrast, principal
axis eigenvalues are based solely on the shared variance
among the variables. The two procedures are qualitatively
different. Some therefore claim that the eigenvalues from one
extraction method should not be used to determine
the number of factors for the other extraction method.
The issue remains neglected and unsettled.
------ END MATRIX -----
```

TABLE 9.7

SPSS Parallel Analysis Results for the Exploratory Factor Analysis Example: 1000 Datasets with Permutations of the Raw Data



```
Run MATRIX procedure:

PARALLEL ANALYSIS:

PAF/Common Factor Analysis & Raw Data Permutation

Specifications for this Run:
Ncases     191
Nvars        8
Ndatsets  1000
Percent     95

Raw Data Eigenvalues, & Mean & Percentile Random Data Eigenvalues
         Root     Raw Data      Means      Prcntyle
     1.000000     1.613687     .361951     .478716
     2.000000      . 463109    .237016      .326682
     3.000000      .041523     .137709     .207489
     4.000000      .040669     .057641     .114366
     5.000000     -.085385    -.015422     .033313
     6.000000     -.088389    -.085530    -.039254
     7.000000     -.193077    -.157236    -.108595
     8.000000     -.256326    -.237790    -.184679

Warning: Parallel analyses of adjusted correlation matrices

eg, with SMCs on the diagonal, tend to indicate more factors

than warranted (Buja, A., & Eyuboglu, N., 1992, Remarks on parallel

analysis. Multivariate Behavioral Research, 27, 509-540.).

The eigenvalues for trivial, negligible factors in the real

data commonly surpass corresponding random data eigenvalues

for the same roots. The eigenvalues from parallel analyses

 can be used to determine the real data eigenvalues that are

 beyond chance, but additional procedures should then be used

 to trim trivial factors.


 Principal components eigenvalues are often used to determine

 the number of common factors. This is the default in most

 statistical software packages, and it is the primary practice
```

To determine the number of factors to retain, compare the eigenvalues derived from the observed raw data to the eigenvalues derived from the parallel analysis computation. When the *i*th eigenvalue from the observed data is greater than the *i*th eigenvalue from the random or permutated data in the parallel analysis, then those factors are retained.

Here, we see the first two **'raw data'** eigenvalues are greater than the permutated data **mean** and **percentile** eigenvalues indicating that two factors should be retained.

TABLE 9.7 (continued)

SPSS Parallel Analysis Results for the Exploratory Factor Analysis Example: 1000 Datasets with Permutations of the Raw Data

```
in the literature. It is also the method used by many factor

analysis experts, including Cattell, who often examined

principal components eigenvalues in his scree plots to determine

the number of common factors. But others believe this common

practice is wrong. Principal components eigenvalues are based

on all of the variance in correlation matrices, including both

the variance that is shared among variables and the variances

that are unique to the variables. In contrast, principal

axis eigenvalues are based solely on the shared variance

among the variables. The two procedures are qualitatively

different. Some therefore claim that the eigenvalues from one

extraction method should not be used to determine

the number of factors for the other extraction method.

The issue remains neglected and unsettled.


------ END MATRIX -----
```

9.3.2 Computing EFA With Ordinal Data Using SPSS

Next we consider an SPSS add-on, categorical principal components analysis (CATPCA), for conducting exploratory factor analysis in the case where our data is ordinal. I felt it critically important to provide this illustration in the textbook for two reasons: (1) there is an abundance of data collected and secondary data available that is ordinal in scale—specifically Likert items that measure attitude, perceptions, etc.—as well as nominal (which can also be handled with CATPCA); and (2) yet few resources are available that transparently help researchers select and use an appropriate EFA procedure and thereby avoid the pitfall of applying conventional EFA techniques to data for which it is really not appropriate. As an optimal scaling approach, nonlinear relationships between categorical variables can be modeled within CATPCA via optimal quantification in a specified dimension. *Unfortunately, CATPCA is only available as an add-on with SPSS.* Should you be renting a copy of SPSS, you likely have it. If you are accessing SPSS from an institution that purchases a finite number of SPSS licenses, you may or may not have it.

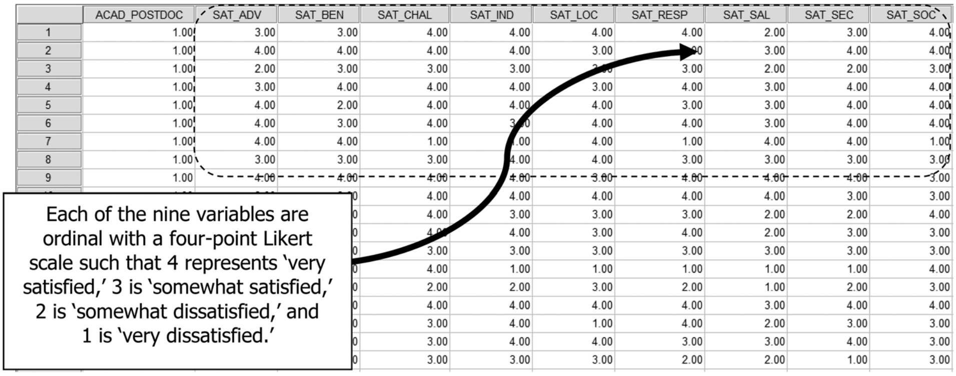
Before we conduct the analysis, let us talk about the data. The data we are using is the 2010 Survey of Doctorate Recipients (SDR, http://www.nsf.gov/statistics/srvydoctoratework/), available through the National Science Foundation (NSF) (2010). Thank you to NSF for making this data publicly available. This is only one of many secondary data sources available through NSF as well as other federal and nonfederal agencies, and I encourage you to explore these extremely rich resources (particularly for multivariate research) for your own research.

First conducted in 1973, the SDR is a “longitudinal biennial survey that provides demographic and career history information about individuals with a research doctoral degree in a science, health, or engineering (SHE) field from a U.S. academic institution. The survey follows a sample of individuals with SHE doctorates throughout their careers from the year of their degree until age 76 . . . Results are used to make decisions related to the educational and occupational achievements and career movements of the nation’s doctoral scientists and engineers” (see http://www.nsf.gov/statistics/srvydoctoratework/). It is important to note that the SDR is a complex sample (i.e., not a simple random sample). More specifically, it employs a stratified probability sampling design. As you will see in the dataset, the very last variable is a weight variable. When this weight is applied to the analysis, the results are adjusted for unequal selection probabilities and nonresponse (which also includes respondents who could not be located or whose eligibility was unknown) and aligned with poststratification (http://www.nsf.gov/statistics/srvydoctoratework/). The end results are then representative of the intended population. As stated previously, the purpose of the text is not to serve as a primer for understanding complex samples, and thus readers interested in learning more about complex survey designs are referred to resources noted earlier in the chapter.

Now, let’s review the data. We are using the SDR2010_POSTDOC.sav file. This data file has been delimited to include only individuals who completed the SDR in 2010 who were employed in a post-doctoral position during the week they responded to the survey (*n* = 1080). The size of this sample is more than sufficient to generate EFA but at the same time small enough to work with for readers who may be using a version of SPSS that limits the number of cases to 1,500. (Note: The complete SDR data file, which includes 31,362 cases, is available from the textbook’s companion website and is titled SDR2010_NSF.sav.)

You’ll notice that in both the post-doc (SDR2010_POSTDOC.sav) and full data file (SDR2010_NSF.sav) there is quite a bit of recoding that will need to be performed in order to get the data in shape for analysis. This includes defining the missing values, recoding the string variables to numeric, and where applicable, reverse coding. I’ve taken the liberty to perform this data cleaning for the variables with which we’ll be working; however, the remaining variables in the data file have been left as is so that you may practice your data cleaning skills in working with ‘real data.’

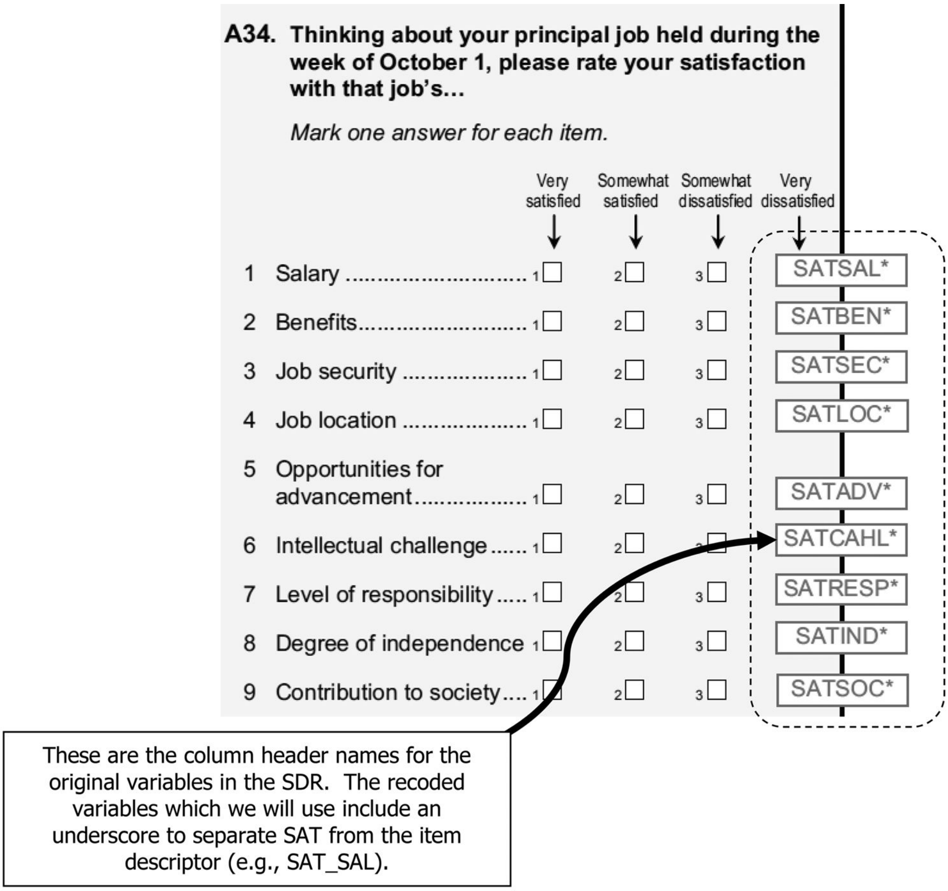
Let’s look at the data. The first variable in our dataset was the variable used to delimit the cases to only respondents who were employed in post-doctoral positions during the week of the survey. The next nine variables are those variables with which we will be analyzing for the EFA. Each row in the data set still represents one individual. As seen in the screenshot below, the SPSS data is in the form of multiple columns that represent the variables on which the respondents were measured. For the EFA illustration, we will work with the nine ordinal satisfaction measures.



	ACAD_POSTDOC	SAT_ADV	SAT_BEN	SAT_CHAL	SAT_IND	SAT_LOC	SAT_RESP	SAT_SAL	SAT_SEC	SAT_SOC
1	1.00	3.00	3.00	4.00	4.00	4.00	4.00	2.00	3.00	4.00
2	1.00	4.00	4.00	4.00	4.00	3.00	[illegible]	3.00	4.00	4.00
3	1.00	2.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00	3.00
4	1.00	3.00	4.00	4.00	4.00	3.00	4.00	3.00	4.00	4.00
5	1.00	4.00	2.00	4.00	4.00	4.00	3.00	3.00	4.00	4.00
6	1.00	4.00	3.00	4.00	3.00	4.00	4.00	3.00	4.00	4.00
7	1.00	4.00	4.00	1.00	1.00	4.00	1.00	4.00	4.00	1.00
8	1.00	3.00	3.00	3.00	4.00	4.00	3.00	3.00	3.00	3.00
9	1.00	4.00	4.00	4.00	4.00	3.00	4.00	4.00	4.00	3.00
				4.00	4.00	4.00	4.00	4.00	3.00	3.00
				4.00	3.00	3.00	3.00	2.00	2.00	4.00
				4.00	4.00	3.00	4.00	2.00	3.00	3.00
				3.00	3.00	3.00	3.00	3.00	3.00	3.00
				4.00	1.00	1.00	1.00	1.00	1.00	4.00
				2.00	2.00	3.00	2.00	1.00	2.00	3.00
				4.00	4.00	4.00	4.00	3.00	4.00	4.00
				3.00	4.00	1.00	4.00	2.00	3.00	3.00
				3.00	4.00	4.00	3.00	3.00	4.00	3.00
				3.00	3.00	3.00	2.00	2.00	1.00	3.00

Each of the nine variables are ordinal with a four-point Likert scale such that 4 represents 'very satisfied,' 3 is 'somewhat satisfied,' 2 is 'somewhat dissatisfied,' and 1 is 'very dissatisfied.'

Reviewing the annotated SDR questionnaire available from NSF (see screenshot of questionnaire), this is a question set responding to the item, "Thinking about your principal job held during the week of October 1, please rate your satisfaction with that job's . . ." The components of the job to which they responded included (1) salary, (2) benefits, (3) job security, (4) job location, (5) opportunities for advancement, (6) intellectual challenge, (7) level of responsibility, (8) degree of independence, and (9) contribution to society.



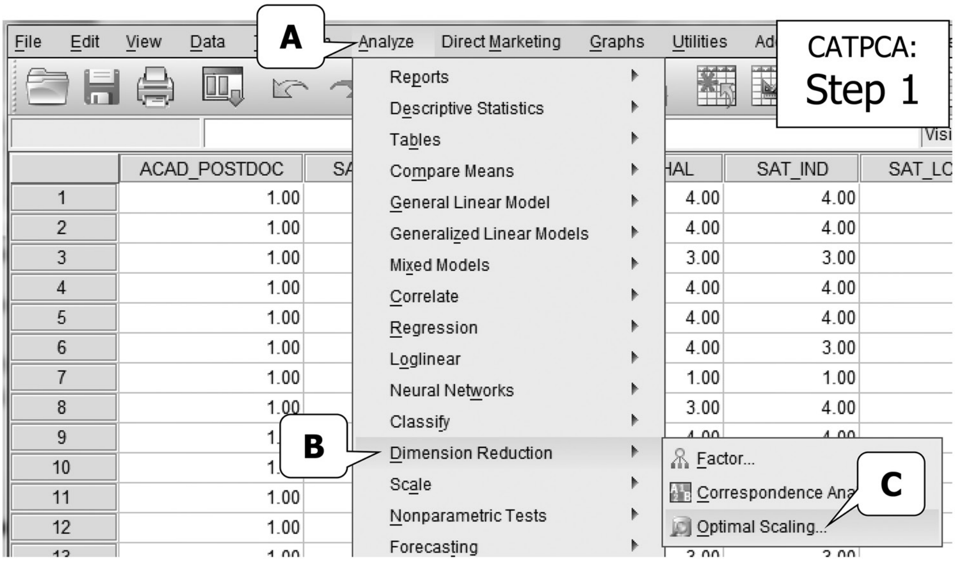
A34. Thinking about your principal job held during the week of October 1, please rate your satisfaction with that job's...

Mark one answer for each item.

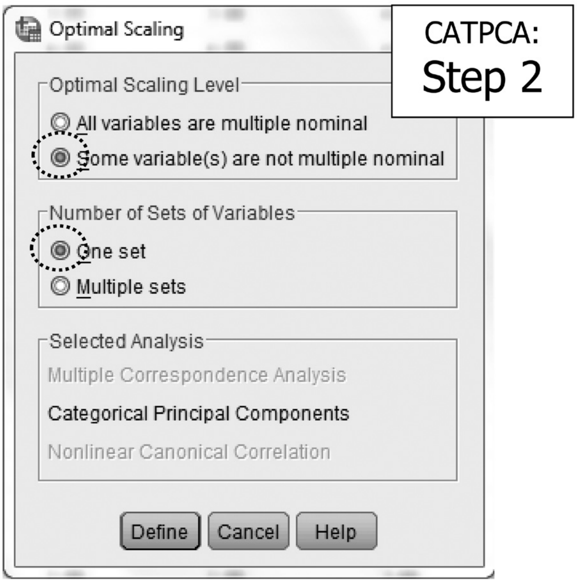
	Very satisfied	Somewhat satisfied	Somewhat dissatisfied	Very dissatisfied
1 Salary	1□	2□	3□	SATSAL*
2 Benefits..........................	1□	2□	3□	SATBEN*
3 Job security	1□	2□	3□	SATSEC*
4 Job location	1□	2□	3□	SATLOC*
5 Opportunities for advancement.................	1□	2□	3□	SATADV*
6 Intellectual challenge......	1□	2□	3□	SATCAHL*
7 Level of responsibility.....	1□	2□	3□	SATRESP*
8 Degree of independence	1□	2□	3□	SATIND*
9 Contribution to society....	1□	2□	3□	SATSOC*

These are the column header names for the original variables in the SDR. The recoded variables which we will use include an underscore to separate SAT from the item descriptor (e.g., SAT_SAL).

Step 1. To conduct categorical principal components analysis, go to "Analyze" in the top pull-down menu, then select "Dimension Reduction," and then select "Optimal Scaling." *Again, please remember that if you do not have this SPSS add-on, you will not see an option for "Optimal Scaling."* Following the screenshot below (CATPCA: Step 1) produces the "Factor Analysis" dialog box.



Step 2. Select the radio button for "Some variable(s) are not multiple nominal" (had all the variables been nominal, we would have selected the first option) and "one set" (see screenshot CATPCA: Step 2).

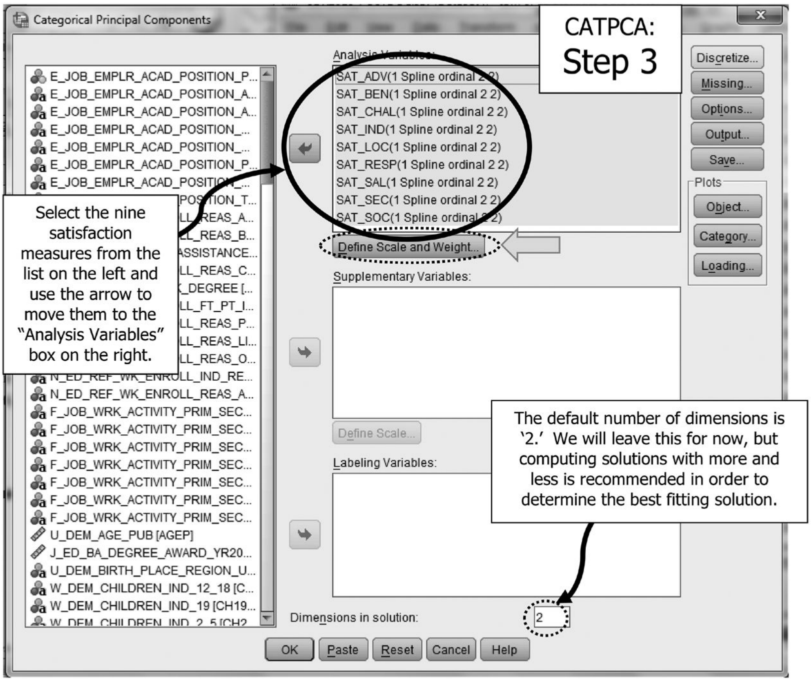


Optimal Scaling Level
○ All variables are multiple nominal
● Some variable(s) are not multiple nominal

Number of Sets of Variables
● One set
○ Multiple sets

Selected Analysis
Multiple Correspondence Analysis
Categorical Principal Components
Nonlinear Canonical Correlation

Step 3. Click the nine satisfaction measures and move into the "Analysis Variables" box by clicking the arrow button (see screenshot CATPCA: Step 3).


Categorical Principal Components
E_JOB_EMPLR_ACAD_POSITION_P...
E_JOB_EMPLR_ACAD_POSITION_A...
E_JOB_EMPLR_ACAD_POSITION_A...
E_JOB_EMPLR_ACAD_POSITION_...
E_JOB_EMPLR_ACAD_POSITION_...
E_JOB_EMPLR_ACAD_POSITION_P...
E_JOB_EMPLR_ACAD_POSITION_...
N_ED_REF_WK_ENROLL_IND_RE...
N_ED_REF_WK_ENROLL_REAS_A...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
F_JOB_WRK_ACTIVITY_PRIM_SEC...
U_DEM_AGE_PUB [AGEP]
J_ED_BA_DEGREE_AWARD_YR20...
U_DEM_BIRTH_PLACE_REGION_U...
W_DEM_CHILDREN_IND_12_18 [C...
W_DEM_CHILDREN_IND_19 [CH19...
W_DEM_CHILDREN_IND_2_5 [CH2...
Discretize...
Missing...
Options...
Output...
Save...
Plots
Object...
Category...
Loading...
Define Scale and Weight...
Supplementary Variables:
Define Scale...
Labeling Variables:
OK
Paste
Reset
Cancel
Help

CATPCA: Step 3

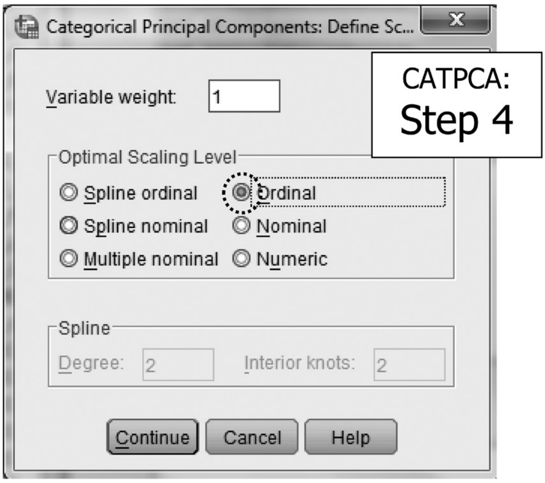
Analysis Variables:
SAT_ADV(1 Spline ordinal 2 2)
SAT_BEN(1 Spline ordinal 2 2)
SAT_CHAL(1 Spline ordinal 2 2)
SAT_IND(1 Spline ordinal 2 2)
SAT_LOC(1 Spline ordinal 2 2)
SAT_RESP(1 Spline ordinal 2 2)
SAT_SAL(1 Spline ordinal 2 2)
SAT_SEC(1 Spline ordinal 2 2)
SAT_SOC(1 Spline ordinal 2 2)

Dimensions in solution: 2

Select the nine satisfaction measures from the list on the left and use the arrow to move them to the "Analysis Variables" box on the right.

The default number of dimensions is '2.' We will leave this for now, but computing solutions with more and less is recommended in order to determine the best fitting solution.

Step 4. Click in "Define scale and weight" (displayed under the Analysis Variable box) to change the optimal scaling level (see screenshot CATPCA: Step 4). The default is 'spline ordinal.' We will select the radio button for 'ordinal.' Spline ordinal and ordinal optimal scaling levels are similar in that they both preserve the order of the categories in the optimally scaled variable. Ordinal optimal scaling results in a better fitting transformation than spline ordinal but is less smooth. Click Continue to return to the main CATPCA page.



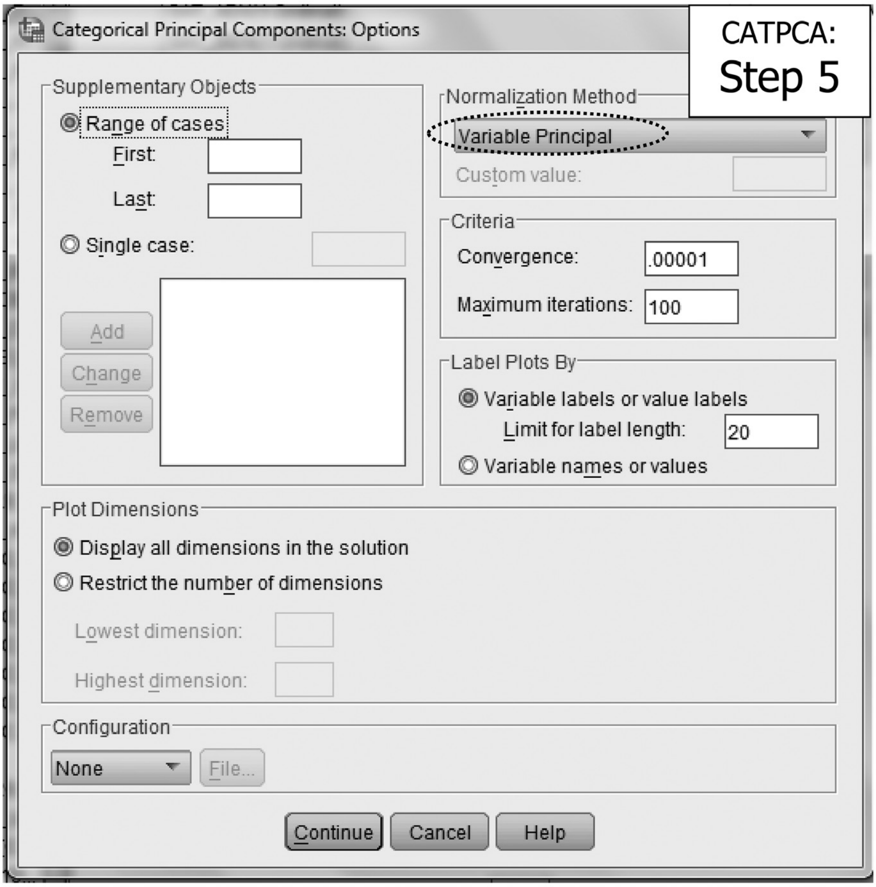
CATPCA: Step 4

Variable weight: 1

Optimal Scaling Level
○ Spline ordinal ● Ordinal
○ Spline nominal ○ Nominal
○ Multiple nominal ○ Numeric

Spline
Degree: 2 Interior knots: 2

Step 5. From the CATPCA page (see screenshot CATPCA: Step 3), click on Options to bring up the Options dialog box. We will leave all default selections as is on this page (see screenshot CATPCA: Step 5). In terms of the Normalization Method, the default selection is Variable Principal. This method optimizes the relationship between variables and is an appropriate selection if the correlation between variables is your primary interest. Click Continue to return to the main CATPCA page.



CATPCA: Step 5

Categorical Principal Components: Options

Supplementary Objects

● Range of cases

First:

Last:

○ Single case:

Normalization Method: Variable Principal

Custom value:

Criteria

Convergence: .00001

Maximum iterations: 100

Label Plots By

● Variable labels or value labels

Limit for label length: 20

○ Variable names or values

Plot Dimensions

● Display all dimensions in the solution

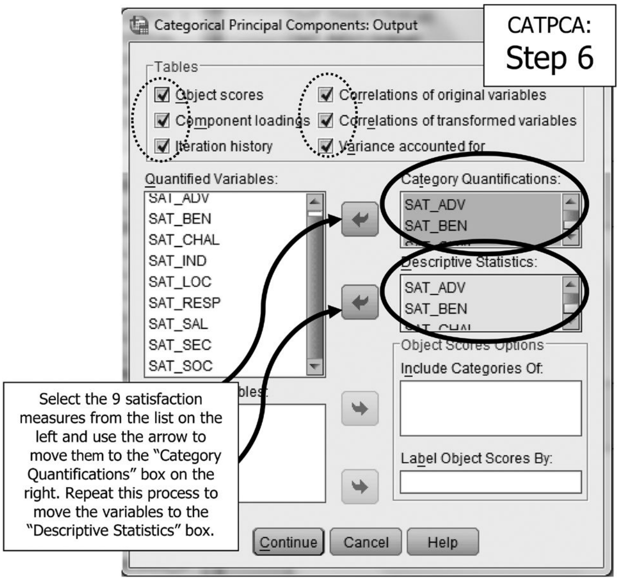
○ Restrict the number of dimensions

Lowest dimension:

Highest dimension:

Configuration: None

Step 6. From the CATPCA page (see screenshot CATPCA: Step 3), click on Output to bring up the Output dialog box (see screenshot CATPCA: Step 6). Object scores and Component loadings should already be selected, and we will keep those selected. Place a checkmark for the remaining tables including Iteration history, Correlations of original variables, Correlations of transformed variables, and Variance accounted for. Click Continue to return to the main CATPCA page. Move all the satisfaction variables from the Quantified Variables list to the Category Quantifications box by clicking the arrow in the middle. Repeat this process to move the variables to the Descriptive Statistics box. Click Continue to return to the main CATPCA page.



Categorical Principal Components: Output

CATPCA: Step 6

Tables

☑ Object scores
☑ Component loadings
☑ Iteration history
☑ Correlations of original variables
☑ Correlations of transformed variables
☑ Variance accounted for

Quantified Variables:
SAT_ADV
SAT_BEN
SAT_CHAL
SAT_IND
SAT_LOC
SAT_RESP
SAT_SAL
SAT_SEC
SAT_SOC

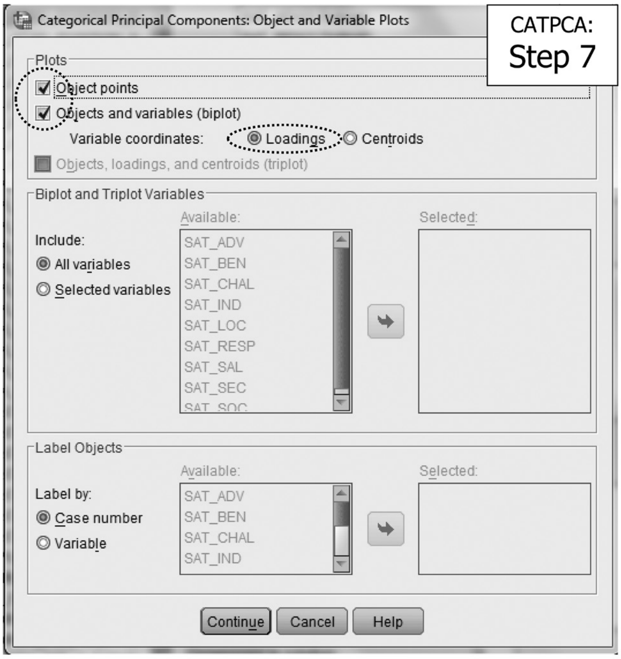
Category Quantifications:
SAT_ADV
SAT_BEN

Descriptive Statistics:
SAT_ADV
SAT_BEN

Object Scores Options
Include Categories Of:
Label Object Scores By:

Select the 9 satisfaction measures from the list on the left and use the arrow to move them to the "Category Quantifications" box on the right. Repeat this process to move the variables to the "Descriptive Statistics" box.

Step 7. From the CATPCA page (see screenshot CATPCA: Step 3), click on Object (listed under Plots in the right navigational menu) to bring up the Object and Variable Plots dialog box (see screenshot CATPCA: Step 7). Object points should already be selected, and we will keep that option selected. We will place a checkmark for Objects and variables (biplot) with variable coordinates Loadings. We will keep the default options selected for Biplot and Triplot Variables and Label Objects. Click Continue to return to the main CATPCA page.



Categorical Principal Components: Object and Variable Plots

CATPCA: Step 7

Plots

☑ Object points
☑ Objects and variables (biplot)
Variable coordinates: ● Loadings ○ Centroids
□ Objects, loadings, and centroids (triplot)

Biplot and Triplot Variables

Include:
● All variables
○ Selected variables

Available:
SAT_ADV
SAT_BEN
SAT_CHAL
SAT_IND
SAT_LOC
SAT_RESP
SAT_SAL
SAT_SEC
SAT_SOC

Selected:

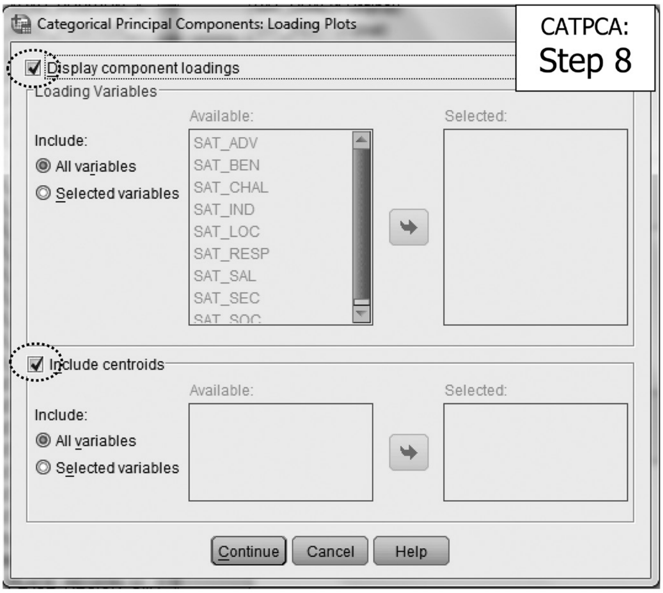
Label Objects

Label by:
● Case number
○ Variable

Available:
SAT_ADV
SAT_BEN
SAT_CHAL
SAT_IND

Selected:

Step 8. From the CATPCA page (see screenshot CATPCA: Step 3), click on Loading (listed under Plots in the right navigational menu) to bring up the Loading Plots dialog box (see screenshot CATPCA: Step 8). Display component loadings should already be selected, and we will keep that option selected. We will place a checkmark for Include Centroids. Click Continue to return to the main CATPCA page.



☑ Display component loadings

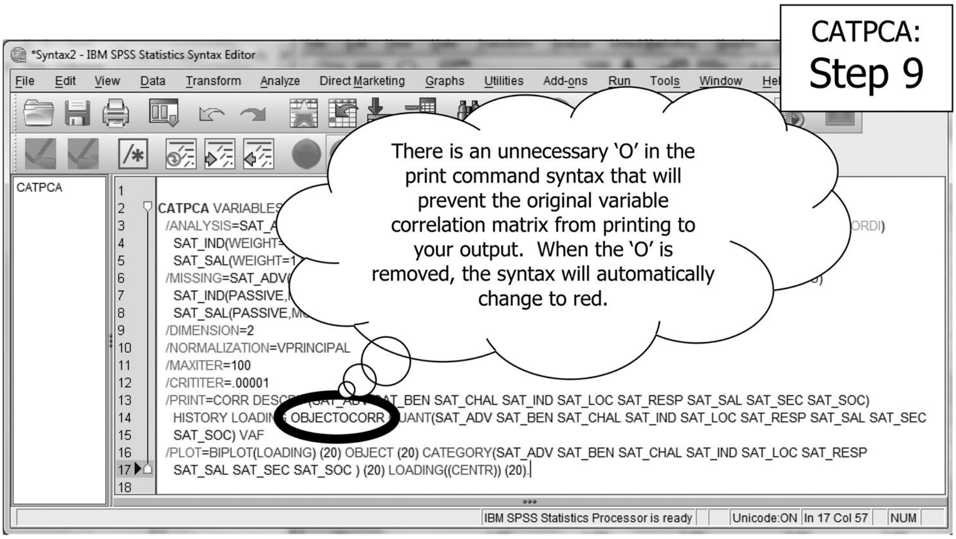
Include: ● All variables ○ Selected variables

Available: SAT_ADV, SAT_BEN, SAT_CHAL, SAT_IND, SAT_LOC, SAT_RESP, SAT_SAL, SAT_SEC, SAT_SOC

☑ Include centroids

Include: ● All variables ○ Selected variables

Step 9. From the CATPCA page (see screenshot CATPCA: Step 3), click 'paste' to open the syntax created from the commands just generated (see screenshot CATPCA:


CATPCA:
Step 9
*Syntax2 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Run Tools Window Hel
CATPCA
IBM SPSS Statistics Processor is ready
Unicode:ON
In 17 Col 57
NUM

There is an unnecessary 'O' in the print command syntax that will prevent the original variable correlation matrix from printing to your output. When the 'O' is removed, the syntax will automatically change to red.

```

CATPCA VARIABLES
  /ANALYSIS=SAT_A                                                                    ORDI)
   SAT_IND(WEIGHT=
   SAT_SAL(WEIGHT=1
  /MISSING=SAT_ADV(                                                                  )
   SAT_IND(PASSIVE,
   SAT_SAL(PASSIVE,M
  /DIMENSION=2
  /NORMALIZATION=VPRINCIPAL
  /MAXITER=100
  /CRITITER=.00001
  /PRINT=CORR DESC    AT_BEN SAT_CHAL SAT_IND SAT_LOC SAT_RESP SAT_SAL SAT_SEC SAT_SOC)
   HISTORY LOADIN OBJECTOCORR UANT(SAT_ADV SAT_BEN SAT_CHAL SAT_IND SAT_LOC SAT_RESP SAT_SAL SAT_SEC
   SAT_SOC) VAF
  /PLOT=BIPLOT(LOADING) (20) OBJECT (20) CATEGORY(SAT_ADV SAT_BEN SAT_CHAL SAT_IND SAT_LOC SAT_RESP
   SAT_SAL SAT_SEC SAT_SOC ) (20) LOADING((CENTR)) (20).

```

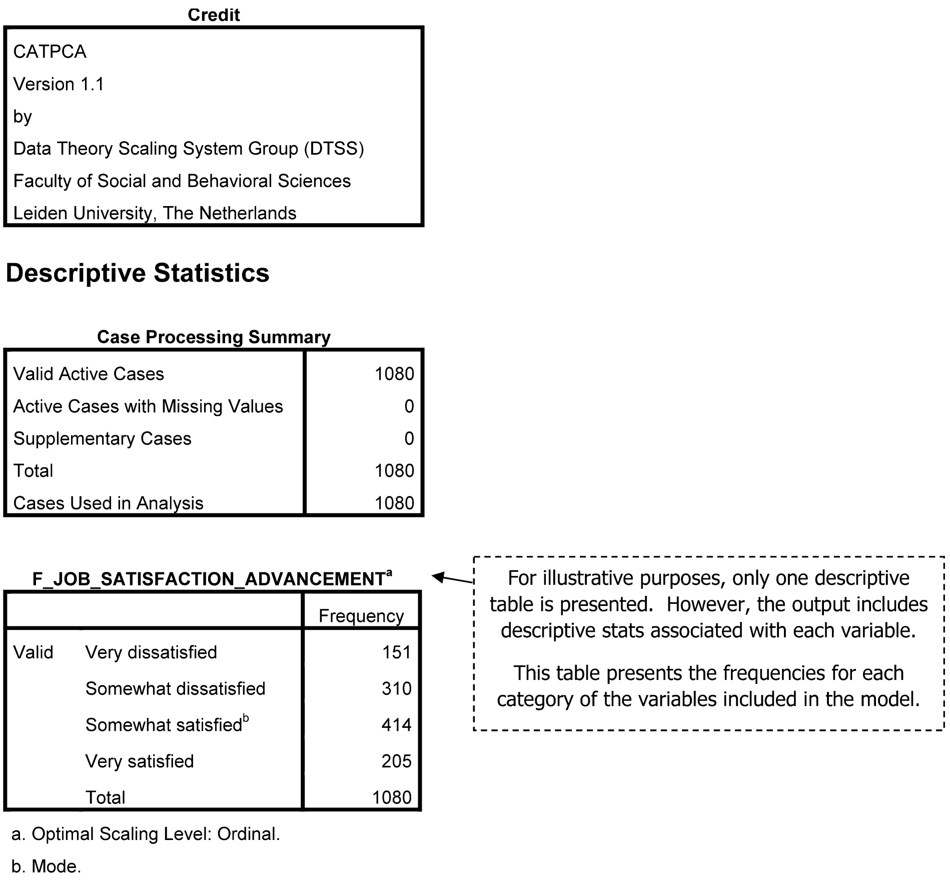
Step 9). In some versions of SPSS, an error occurs in the print command line in specifying to print the original variable correlation matrix such that an ‘O’ is included rather than a space. If this error occurs in your syntax, remove the ‘O’ and then run the syntax to generate the output.

When the erroneous ‘O’ is removed (you must manually do this using your delete or backspace key), ‘OBJECT’ and ‘CORR’ will appear in red, indicating that the original variable correlation matrix will be printed in the output.

Interpreting the CATPCA output. Annotated results are presented in Table 9.8.

TABLE 9.8

SPSS Results for the Categorical Principal Components Analysis Example



Credit

CATPCA Version 1.1 by Data Theory Scaling System Group (DTSS) Faculty of Social and Behavioral Sciences Leiden University, The Netherlands

Descriptive Statistics

Case Processing Summary

Valid Active Cases	1080
Active Cases with Missing Values	0
Supplementary Cases	0
Total	1080
Cases Used in Analysis	1080

F_JOB_SATISFACTION_ADVANCEMENT[a]

		Frequency
Valid	Very dissatisfied	151
	Somewhat dissatisfied	310
	Somewhat satisfied[b]	414
	Very satisfied	205
	Total	1080

a. Optimal Scaling Level: Ordinal.

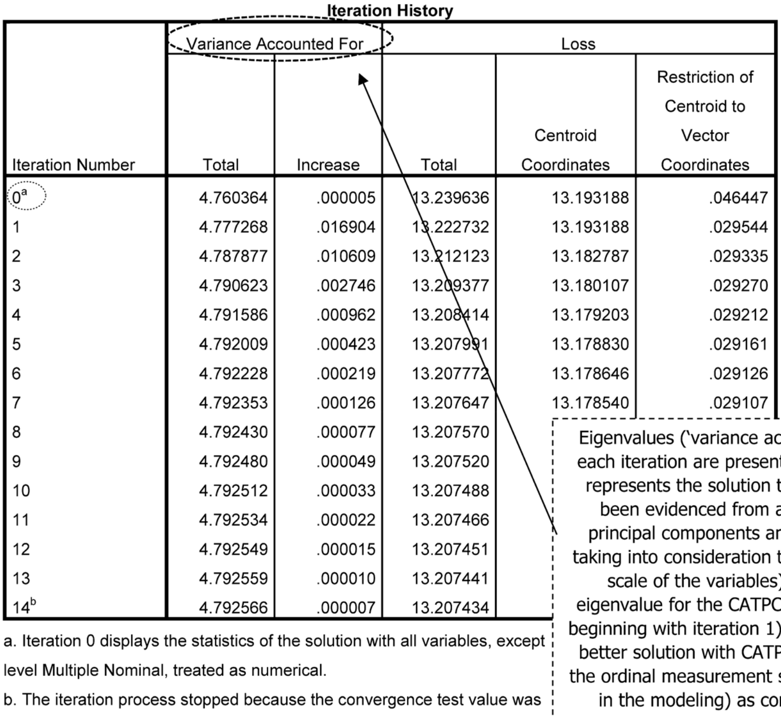
b. Mode.

For illustrative purposes, only one descriptive table is presented. However, the output includes descriptive stats associated with each variable.

This table presents the frequencies for each category of the variables included in the model.

TABLE 9.8 (continued)

SPSS Results for the Categorical Principal Components Analysis Example



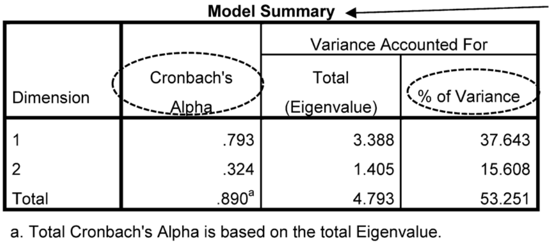
Iteration History

Iteration Number	Variance Accounted For		Loss		
	Total	Increase	Total	Centroid Coordinates	Restriction of Centroid to Vector Coordinates
0[a]	4.760364	.000005	13.239636	13.193188	.046447
1	4.777268	.016904	13.222732	13.193188	.029544
2	4.787877	.010609	13.212123	13.182787	.029335
3	4.790623	.002746	13.209377	13.180107	.029270
4	4.791586	.000962	13.208414	13.179203	.029212
5	4.792009	.000423	13.207991	13.178830	.029161
6	4.792228	.000219	13.207772	13.178646	.029126
7	4.792353	.000126	13.207647	13.178540	.029107
8	4.792430	.000077	13.207570		
9	4.792480	.000049	13.207520		
10	4.792512	.000033	13.207488		
11	4.792534	.000022	13.207466		
12	4.792549	.000015	13.207451		
13	4.792559	.000010	13.207441		
14[b]	4.792566	.000007	13.207434		

a. Iteration 0 displays the statistics of the solution with all variables, except level Multiple Nominal, treated as numerical.

b. The iteration process stopped because the convergence test value was

Eigenvalues ('variance accounted for') for each iteration are presented. 'Iteration 0' represents the solution that would have been evidenced from a *conventional* principal components analysis (i.e., not taking into consideration the measurement scale of the variables). The larger eigenvalue for the CATPCA solution (4.77, beginning with iteration 1) reflects a slightly better solution with CATPCA (which takes the ordinal measurement scale into account in the modeling) as compared to the conventional PCA.



Model Summary

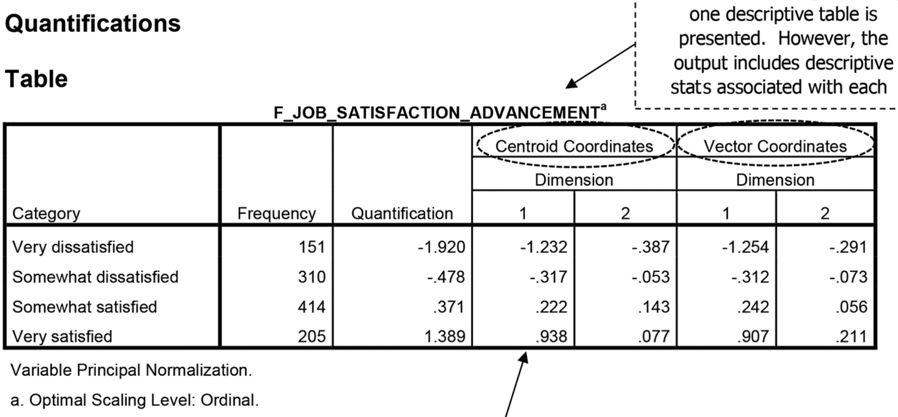
Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.793	3.388	37.643
2	.324	1.405	15.608
Total	.890[a]	4.793	53.251

a. Total Cronbach's Alpha is based on the total Eigenvalue.

Cronbach's alpha is a measure of internal consistency, and this value is provided for each dimension (i.e., factor; labeled '1' and '2') as well as the total which represents the combination of all factors. The '% of variance' is presented for each factor and for the combination of factors (i.e., total). Both factors account for 53% of the variance in the optimally scaled items.

TABLE 9.8 (continued)

SPSS Results for the Categorical Principal Components Analysis Example



Quantifications

Table

For illustrative purposes, only one descriptive table is presented. However, the output includes descriptive stats associated with each

F_JOB_SATISFACTION_ADVANCEMENT[a]

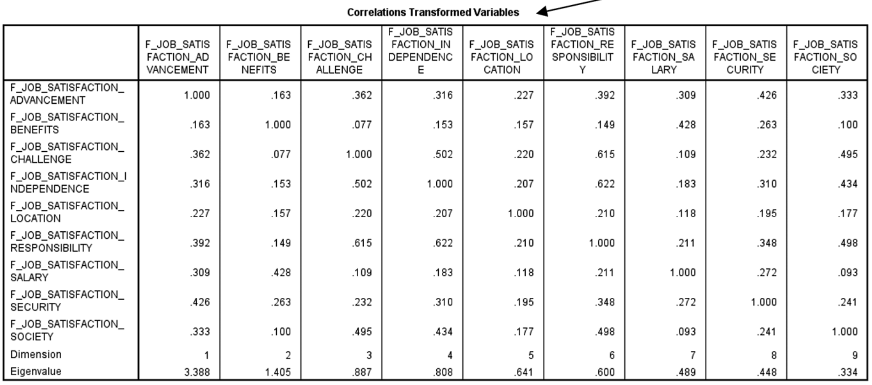
			Centroid Coordinates		Vector Coordinates	
			Dimension		Dimension	
Category	Frequency	Quantification	1	2	1	2
Very dissatisfied	151	-1.920	-1.232	-.387	-1.254	-.291
Somewhat dissatisfied	310	-.478	-.317	-.053	-.312	-.073
Somewhat satisfied	414	.371	.222	.143	.242	.056
Very satisfied	205	1.389	.938	.077	.907	.211

Variable Principal Normalization.

a. Optimal Scaling Level: Ordinal.

'Centroid coordinates' reflect the average of all object scores for cases for the respective category on each factor (labeled 'dimension'). 'Vector coordinates' are the coordinates for each category when the categories are represented by a straight line between factor 1 (X axis) and factor 2 (Y axis) in a scatterplot.

The correlation matrix reflects coefficients *after* optimal scaling has been performed. These coefficients are those used in the CATPCA. If you imputed data during the CATPCA procedure, these values would reflect correlations of imputed values.



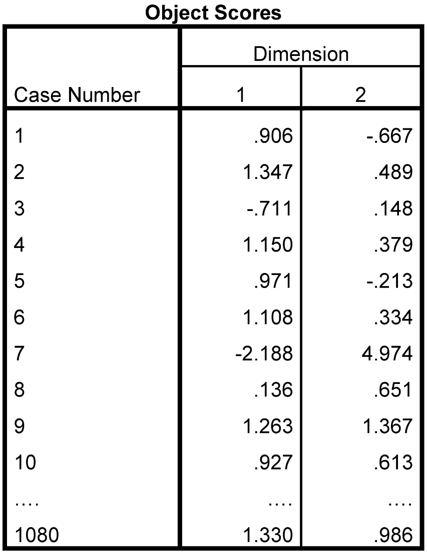
Correlations Transformed Variables

	F_JOB_SATISFACTION_ADVANCEMENT	F_JOB_SATISFACTION_BENEFITS	F_JOB_SATISFACTION_CHALLENGE	F_JOB_SATISFACTION_INDEPENDENCE	F_JOB_SATISFACTION_LOCATION	F_JOB_SATISFACTION_RESPONSIBILITY	F_JOB_SATISFACTION_SALARY	F_JOB_SATISFACTION_SECURITY	F_JOB_SATISFACTION_SOCIETY
F_JOB_SATISFACTION_ADVANCEMENT	1.000	.163	.362	.316	.227	.392	.309	.426	.333
F_JOB_SATISFACTION_BENEFITS	.163	1.000	.077	.153	.157	.149	.428	.263	.100
F_JOB_SATISFACTION_CHALLENGE	.362	.077	1.000	.502	.220	.615	.109	.232	.495
F_JOB_SATISFACTION_INDEPENDENCE	.316	.153	.502	1.000	.207	.622	.183	.310	.434
F_JOB_SATISFACTION_LOCATION	.227	.157	.220	.207	1.000	.210	.118	.195	.177
F_JOB_SATISFACTION_RESPONSIBILITY	.392	.149	.615	.622	.210	1.000	.211	.348	.498
F_JOB_SATISFACTION_SALARY	.309	.428	.109	.183	.118	.211	1.000	.272	.093
F_JOB_SATISFACTION_SECURITY	.426	.263	.232	.310	.195	.348	.272	1.000	.241
F_JOB_SATISFACTION_SOCIETY	.333	.100	.495	.434	.177	.498	.093	.241	1.000
Dimension	1	2	3	4	5	6	7	8	9
Eigenvalue	3.388	1.405	.887	.808	.641	.600	.489	.448	.334

TABLE 9.8 (continued)

SPSS Results for the Categorical Principal Components Analysis Example

Objects



Object Scores

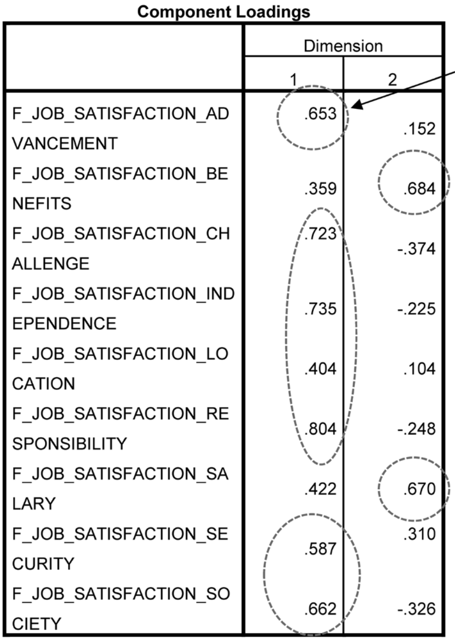
Case Number	Dimension 1	Dimension 2
1	.906	-.667
2	1.347	.489
3	-.711	.148
4	1.150	.379
5	.971	-.213
6	1.108	.334
7	-2.188	4.974
8	.136	.651
9	1.263	1.367
10	.927	.613
....	....	....
1080	1.330	.986

Variable Principal Normalization.

For illustrative purposes, only a portion of the output is presented.

The object scores represent the coordinates associated with the respective case for each factor. Thus, there will be as many *cases* listed in the table as your sample size.

Component Loadings



Component Loadings

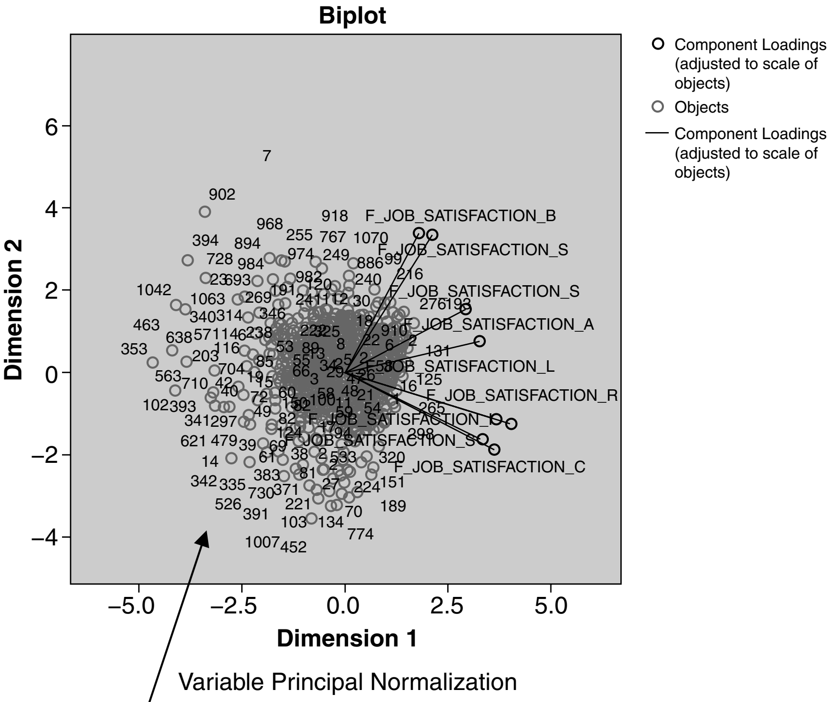
	Dimension 1	Dimension 2
F_JOB_SATISFACTION_ADVANCEMENT	.653	.152
F_JOB_SATISFACTION_BENEFITS	.359	.684
F_JOB_SATISFACTION_CHALLENGE	.723	-.374
F_JOB_SATISFACTION_INDEPENDENCE	.735	-.225
F_JOB_SATISFACTION_LOCATION	.404	.104
F_JOB_SATISFACTION_RESPONSIBILITY	.804	-.248
F_JOB_SATISFACTION_SALARY	.422	.670
F_JOB_SATISFACTION_SECURITY	.587	.310
F_JOB_SATISFACTION_SOCIETY	.662	-.326

Variable Principal Normalization.

The component loadings scatterplot graphs the coordinates for each variable on each factor, allowing us to see how the variables relate to each other as well as the factors. Variables that lump together suggest distinguishable factors. In this case, we see that seven of the variables coalesce together on factor 1 and only two on factor 2.

■ **TABLE 9.8 (continued)**

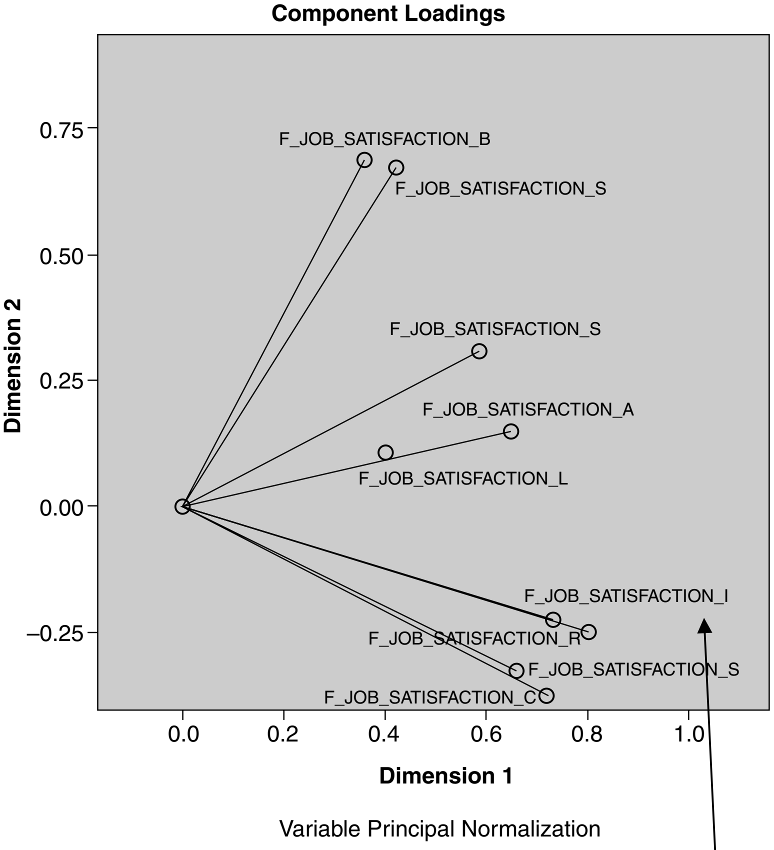
SPSS Results for the Categorical Principal Components Analysis Example



Finally, we get a scatterplot that you will see in color but is presented in grayscale here. Each variable is black and each case is green (grayscale here). Factor 1 is able to capture a bit more of the variance among the variables and cases and thus can explain the variance a bit better than factor 2 we see the variables and cases to be more tightly grouped (-4 to 4 for factor 2 as compared to -5 to 5 for factor 1). This suggests less variable variance captured.

TABLE 9.8 (continued)

SPSS Results for the Categorical Principal Components Analysis Example

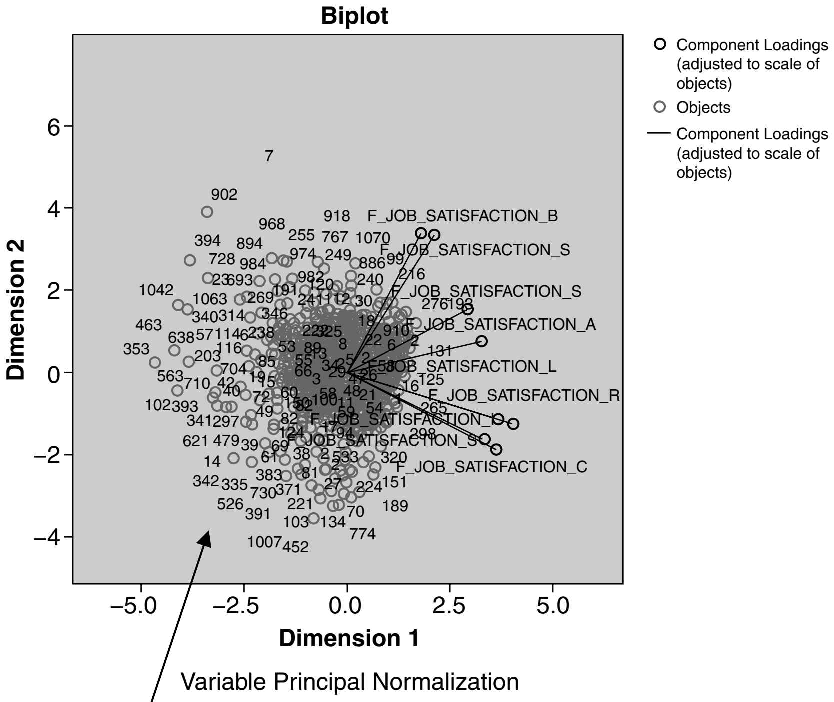


The component loadings scatterplot graphs the coordinates for each variable on each factor, allowing us to see how the variables relate to each other as well as the factors. Variables that lump together suggest distinguishable factors. In this case, we see the variables vary substantially along dimension 2 (i.e., factor 2) but tend to fall within a more narrow range of dimension 1 (i.e., factor 1) (between about .40 and .80). Here we can see how the two variables with large loadings on factor 2 are differentiating from those of factor 1. This may be where a decision is made to remove the two variables that appear to load on factor 2 and re-run. If the model improves without those items, there will be a clearer, tighter grouping of the variables on their respective factor(s).

The lines from the centroid to each variable are eigenvectors and the variable is at the eigenvalue for its vector. Thus the eigenvalue is a distance point along an eigenvector. With conventional EFA, a rotation strategy is applied to make interpretation easier. Here, we can imagine rotation such that both dimensions are rotated counterclockwise 45 degrees. In doing so, the axis of each factor (or dimension) would be going through a cloud of points (which represent the variables).

TABLE 9.8 (continued)

SPSS Results for the Categorical Principal Components Analysis Example



Finally, we get a scatterplot that you will see in color but is presented in grayscale here. Each variable is black and each case is green (grayscale here). Factor 1 is able to capture a bit more of the variance among the variables and cases and thus can explain the variance a bit better than factor 2 we see the variables and cases to be more tightly grouped (-4 to 4 for factor 2 as compared to -5 to 5 for factor 1). This suggests less variable variance captured.

9.4 DATA SCREENING

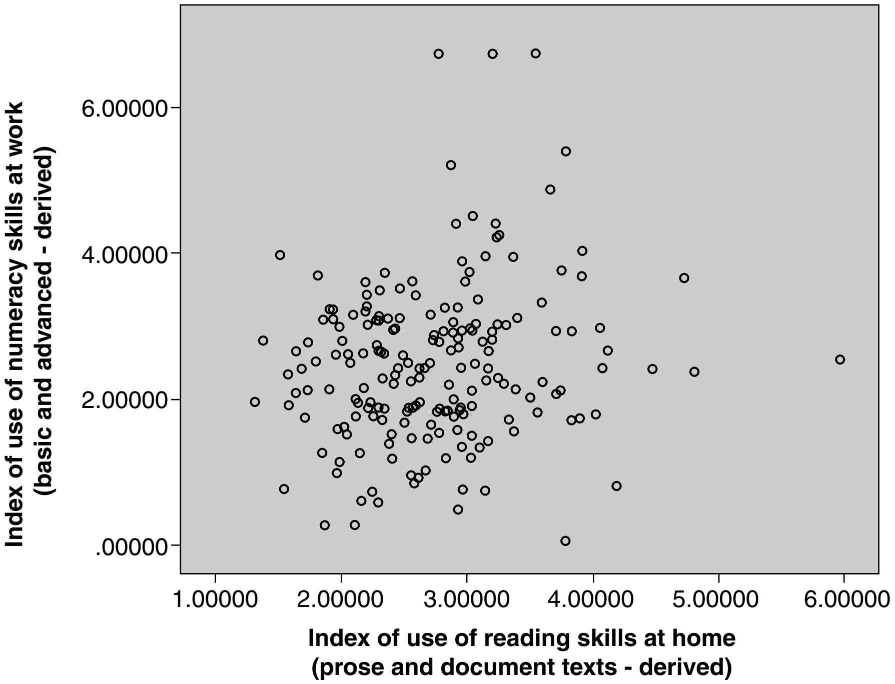
As you may recall, there were a number of assumptions associated with conventional exploratory factor analysis. These included (a) independence, (b) linearity, (c) absence of outliers (both univariate and multivariate), and (d) lack of extreme multicollinearity and singularity. Although fixed values of X were discussed in assumptions, this is not an assumption that will be tested, but is instead related to the use of the results (i.e., extrapolation and interpolation).

9.4.1 Independence

Testing for this assumption is a bit nebulous in exploratory factor analysis, as there are no independent and dependent variables that allow for this type of examination. In the absence of statistical evidence, we will rely on theoretical evidence: If the units have been randomly sampled from a population, there is evidence that the assumption of independence has been met.

9.4.2 Linearity

Linearity is an important assumption since correlation matrices underlie conventional EFA. You may recall that when you studied bivariate correlations, as well as simple and multiple regression, that scatterplots were one way that linearity could be examined. We will again use scatterplots to visually assess linearity. The challenge with EFA, as compared to other procedures where scatterplots have been applied, is the large (and often very large) number of variables, which makes review of all possible pairs of variables quite daunting and an inefficient use of your time. For example, with 10 variables (which tends to be toward the lower limit of the number of variables often applied to EFA), there are $[10(10 - 1)]/2$ or 45 different pairwise combinations of the variables, and with 20 variables there are nearly 200 combinations! One work-around for this is to generate and examine a few random scatterplots, assuming that these are representative of the entire population of scatterplots. Don't be surprised if the scatterplots do not provide picture-perfect linear relationships, and don't be ready to discard or transform variables if that is indeed the case—those consequences should be reserved only for cases where obvious curvilinearity is observed. For the PIAAC data, I ran a number of bivariate scatterplots and while not all scatterplots suggested a strong linear relationship, there does not appear to be evidence to suggest curvilinear associations. One example, graphing 'index of use of numeracy skills at work (basic and advanced)' with 'index of use of reading skills at home (prose and document texts),' is presented here:



9.4.3 Absence of Outliers

As discussed previously, factor analysis is quite robust to violations of the assumption of normality except where tests of inference are used to determine the number of factors to retain, and in this case, multivariate normality *is* an assumption. For this illustration, we are using maximum likelihood and thus will be thorough in our examination of multivariate normality.

We can examine univariate normality tests, which are less sensitive than multivariate normality tests, through skewness and kurtosis, formal tests of normality, and plots (e.g., Q-Q plots). Multivariate outliers are evidenced by statistically significant Mahalanobis distance scores (alpha = .001 if you tend toward the liberal edge, which is appropriate with EFA), evaluated using a chi-square distribution with degrees of freedom equal to the number of variables. To generate Mahalanobis distance, we will generate multiple regression, applying all the variables as independent variables with the dependent variable being a binary variable coded 1 for potential outliers and 0 for all other variables within the model. The process for examining outliers is therefore to look for univariate outliers first. If any are detected, then screen for multivariate outliers. In terms of multivariate normality, a macro in SPSS (DeCarlo, 1997) (illustrated in the MANOVA chapter) can also be used to examine a number of multivariate normality indices including

(a) multivariate kurtosis (Mardia, 1970), (b) multivariate skewness and kurtosis based on Small's (1980) multivariate extension of univariate skewness and kurtosis (Looney, 1995), (c) multivariate normality omnibus test (Looney, 1995), (d) largest squared and plot of squared Mahalanobis distance, and (e) critical values for hypothesis test for a single multivariate outlier using Mahalanobis distance (Penny, 1996).

Additionally, not only are we concerned with outlying cases, but we are also concerned with outlying variables and will need to examine our data for both. These outlying variables, which can be removed from the model if/when identified, can be determined by examination of the following: (a) squared multiple correlations with all other variables and (b) weak correlations with the factors that are identified in the factor analytic model.

Reviewing univariate normality for the PIAAC data, skewness for all measures are within the range of +/− 2.0 and kurtosis for all measures are within +/− 7.0, suggesting evidence of normality.

Descriptive Statistics

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
Index of use of numeracy skills at home (basic and advanced—derived)	191	1.220	.176	5.663	.350
Index of use of numeracy skills at work (basic and advanced—derived)	191	.929	.176	2.516	.350
Index of use of ICT skills at home (derived)	191	1.048	.176	3.149	.350
Index of use of reading skills at home (prose and document texts—derived)	191	.837	.176	1.896	.350
Index of use of task discretion at work (derived)	191	1.431	.176	2.249	.350
Index of learning at work (derived)	191	.449	.176	-.450	.350
Index of use of planning skills at work (derived)	191	.479	.176	-1.121	.350
Index of readiness to learn (derived)	191	.878	.176	-.110	.350
Index of use of ICT skills at work (derived)	191	.457	.176	.630	.350
Index of use of influencing skills at work (derived)	191	1.018	.176	1.894	.350
Index of use of reading skills at work (prose and document texts—derived)	191	.924	.176	2.960	.350
Index of use of writing skills at work (derived)	191	1.116	.176	3.656	.350
Index of use of writing skills at home (derived)	191	.642	.176	4.601	.350
Valid N (listwise)	191				

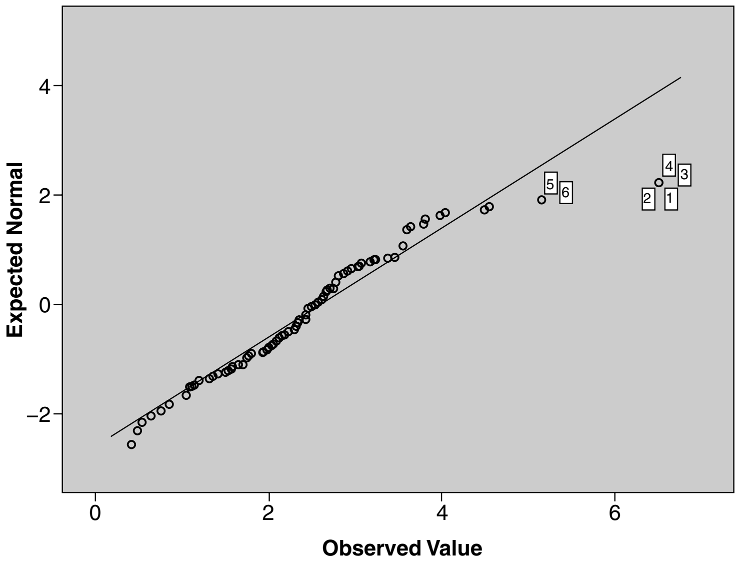
Shapiro-Wilk's formal test of normality was statistically significant for all variables suggesting evidence of nonnormality.

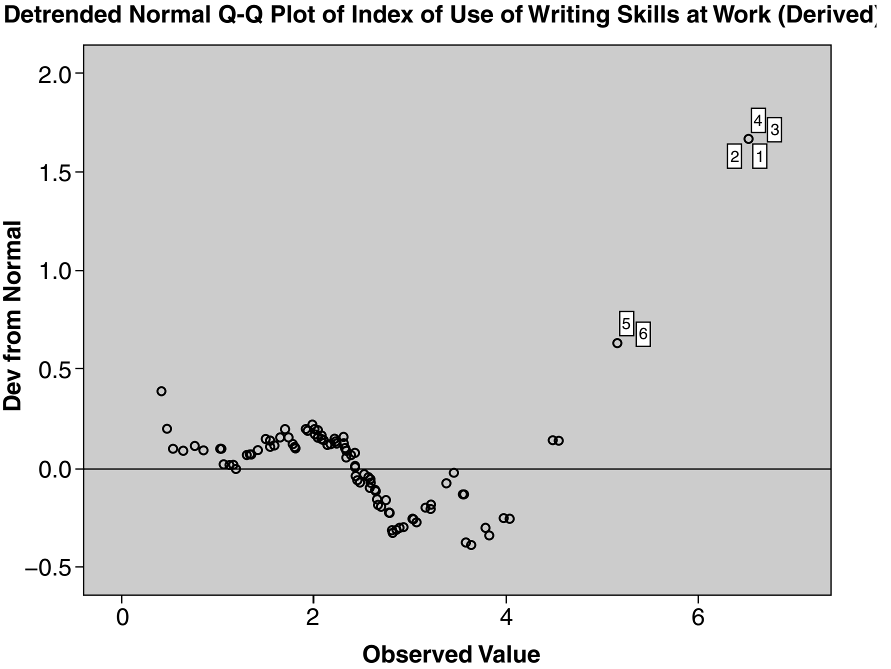
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Index of use of numeracy skills at home (basic and advanced—derived)	.098	191	.000	.920	191	.000
Index of use of numeracy skills at work (basic and advanced—derived)	.075	191	.011	.950	191	.000
Index of use of ICT skills at home (derived)	.068	191	.030	.951	191	.000
Index of use of reading skills at home (prose and document texts—derived)	.062	191	.075	.964	191	.000
Index of use of task discretion at work (derived)	.161	191	.000	.878	191	.000
Index of learning at work (derived)	.118	191	.000	.938	191	.000
Index of use of planning skills at work (derived)	.183	191	.000	.874	191	.000
Index of readiness to learn (derived)	.142	191	.000	.895	191	.000
Index of use of ICT skills at work (derived)	.077	191	.007	.983	191	.021
Index of use of influencing skills at work (derived)	.103	191	.000	.927	191	.000
Index of use of reading skills at work (prose and document texts—derived)	.107	191	.000	.951	191	.000
Index of use of writing skills at work (derived)	.124	191	.000	.915	191	.000
Index of use of writing skills at home (derived)	.106	191	.000	.937	191	.000

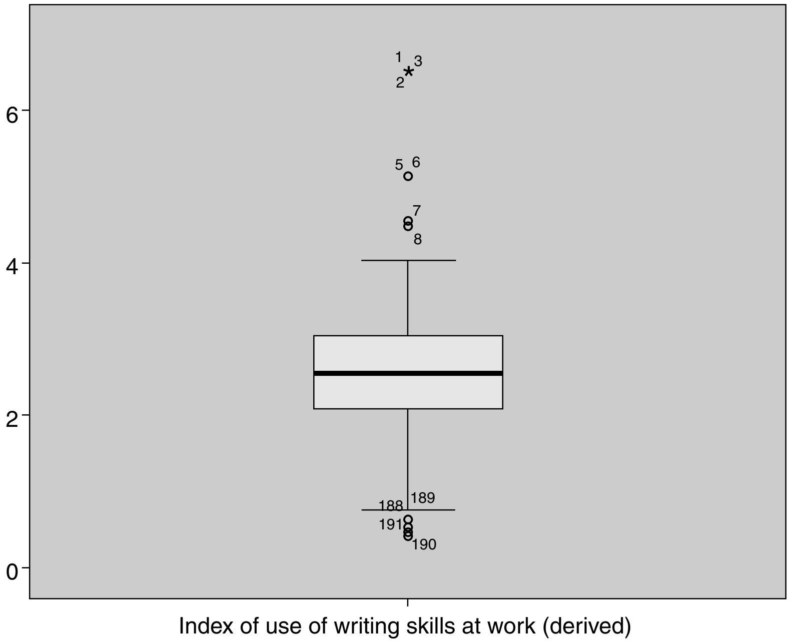
a. Lilliefors Significance Correction

The normal and detrended Q-Q plots suggest at least one potential outlying case for all 13 variables. For example, the ‘index of use of writing skills at work’ suggests that cases 1–6 may be outliers.


Normal Q-Q Plot of Index of Use of Writing Skills at Work (Derived)



Reviewing boxplots, there are quite a few variables that have outliers suggested by the graph. For many (but not all) of the variables, these are at least some of the same cases that showed up as potential outliers in the Q-Q plots. The boxplot for the 'index of use of writing skills at work' suggests additional outliers that were not as evident in reviewing the Q-Q plot—not only cases 1–6 but also cases 7–8 and 188–191.



Now that we've screened for univariate outliers and have identified cases that are suggestive of outliers, we need to screen for multivariate outliers. To do so, we create a new binary variable with '1' denoting that it showed up as an outlier and '0' denoting nonoutlying cases. This has been saved in the PIAAC.EFA.sav data file and is labeled 'OUTLIER.' This binary variable will be our dependent variable in a multiple regression model with all 13 of the index variables as the independent variables. (By this point in your statistics career, it is assumed that you are familiar with creating a new variable, thus the process for doing so is not presented. Should you need a refresher on generating multiple regression, please review the earlier chapter in this text.) When generating the multiple regression model, we are not interested in the results of the analysis. Rather, we run it simply to save the Mahalanobis distance values (saved as MAH_1 in the data file). Multivariate outliers are evidenced by statistically significant Mahalanobis distance values, evaluated using a chi-square distribution with degrees of freedom equal to the number of variables. With alpha of .001, our chi-square critical value is 34.53, and our Mahalanobis distance values range from 1.84 to 56.11. Fortunately, there are only five cases with statistically significant Mahalanobis distance values.

Mahalanobis Distance

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	37.54589	1	20.0	20.0	20.0
	39.71767	1	20.0	20.0	40.0
	42.84545	1	20.0	20.0	60.0
	43.86517	1	20.0	20.0	80.0
	56.11096	1	20.0	20.0	100.0
	Total	5	100.0	100.0	

We will retain these cases for the illustration given that factor analysis is relatively robust to violations with the exception of tests of inference. (Had we filtered them out, we would see that we still end up with a two-factor solution, however only seven variables remain in the model due to the communalities greater than 1 error. For practice, you may want to try this yourself!) In this illustration, we have used maximum likelihood so we are concerned with multivariate normality. As we present our results, we will caution readers to this limitation of our data.

In terms of outlying variables, our final factor model did not suggest this was problematic (i.e., both factors had multiple items).

9.4.4 Extreme Multicollinearity and Singularity

For EFA, the simplest method to detect extreme multicollinearity and singularity is to conduct a series of multiple regression models, one regression model for each variable where that variable is the dependent variable and all remaining variables are the independent variables. If any of the resultant R_k^2 values are close to one (greater than

.9 is a good guideline to go by), then there may be an extreme collinearity problem. However, large R^2 values may also be due to small sample sizes; thus, be cautious in interpretation in cases where the number of cases is small. If the number of variables is greater than or equal to n, then perfect collinearity is a possibility. The results are not presented here for brevity; however, the largest multiple R squared values were under .50, suggesting no problems with extreme multicollinearity.

To prevent singularity, none of the variables that are being used is a composite variable for which the component variables are also included in the EFA model.

9.5 RESEARCH QUESTION TEMPLATE AND EXAMPLE WRITE-UP

Finally, here is an example paragraph for the results of the exploratory factor analysis. Recall that our graduate research assistants, Addie and Oso, were assisting Dr. Wesley, a faculty member in higher education. Specifically, Dr. Wesley was interested in better understanding the underlying constructs of measures of perceived use of skills. The research question presented to Dr. Wesley from Addie and Oso included the following: What is the underlying factor structure for perceived use of skills at home and work?

Addie and Oso then assisted Dr. Wesley in conducting exploratory factor analysis, and a template for writing the research question for exploratory factor analysis is presented below.

What is the underlying factor structure for [variable set]?

It may be helpful to preface the results of the exploratory factor analysis with information on an examination of the extent to which the data were thoroughly screened.

> Prior to conducting the exploratory factor analysis, the data were screened to determine the extent to which the assumptions associated with exploratory factor analysis were met. These assumptions included (a) independence, (b) linearity, (c) absence of outliers (both univariate and multivariate), and (d) lack of extreme multicollinearity and singularity. Because the data were not randomly sampled, there is a possibility that the assumption of independence has not been met. Scatterplots of each combination of variables were generated and generally suggested that the assumption of linearity was feasible, as there was no evidence of curvilinear or other nonlinear relationships. Normal and detrended Q-Q plots and boxplots suggest the presence of a few univariate outliers. These outlying points were examined as potential multivariate outliers. Mahalanobis distance values were computed using the outlying points coded as binary dependent variables and all other variables as independent variables in a multiple regression model. Five of the

> cases had statistically significant Mahalanobis distance values. These items were retained, as EFA is relatively robust to violations of normality with the exception of tests of inference. However, given that maximum likelihood was the estimation method, multivariate normality was a concern. Given there is some evidence to suggest multivariate nonnormality, the model was rerun excluding the potential multivariate outliers. A two-factor solution with seven of the eight variables was achieved. Because the variable was theoretically important, it was retained in the model and the solution reflects all eight variables. Extreme multicollinearity was screened for by conducting a series of multiple regression models, one regression model for each variable where that variable is the dependent variable and all other variables are the independent variables. There were no multiple *R* squared values that were close to one; all were under .50, suggesting no problems with multicollinearity. To prevent singularity, none of the variables used are composite variables for which the component variables are also included.

Here is an example write-up of how the results for exploratory factor analysis can be presented (remember that this will be prefaced by the previous paragraph reporting the extent to which the data were thoroughly screened).

> Evidence for construct validity of indices of home and work skills from the PIAAC was obtained using exploratory factor analysis.
>
> Criteria that is often used to determine factorability of variables was applied in this analysis. These initial factorability criteria included examination of the following: (1) bivariate correlations, (2) Kaiser-Meyer-Olkin measure of sampling adequacy (overall and individual), (3) Bartlett's test of sphericity, and (4) communalities. Based on communalities above 1.0, there were five variables that were removed during this initial stage of determining factorability. The removal of these variables was done through an iterative process of removing the index with the highest communality, rerunning the EFA, and then examining the communalities. This process was repeated for each of the indices removed. The analysis presented is based on the remaining eight items.
>
> Three of the eight items correlated at least .30 with at least one other item and an additional variable was nearly .30 (see Table 1). The overall Kaiser-Meyer-Olkin measure of sampling adequacy was .695, larger than the recommended value of .50. In addition, the measure of sampling adequacy values for the individual items were all larger than the recommended value of .50. Bartlett's test of sphericity was statistically significant [χ^2 (28) = 193.696, $p <$ .001]. An additional criterion commonly used to determine factorability is that communalities should be above the recommended value of .30, providing evidence of shared variance among the items. In reviewing extracted communalities of the eight items, one-half of the variables (4 of the 8 variables) were below .30 (see Table 2). However, given the other criteria for determining factorability were met, it was determined that it was reasonable to proceed with determining the factor structure of the eight items.
>
> Maximum likelihood estimation with promax rotation was used to extract the factors from the data. Parallel analysis was used to determine the number

TABLE 1

Correlation Matrix for Cognitive and Work Ability Indices (*N* = 191)

Item	1	2	3	4	5	6	7
1. Index of use of numeracy skills at home (basic and advanced)	—						
2. Index of use of numeracy skills at work (basic and advanced)	.292	—					
3. Index of use of ICT skills at home (derived)	**.554**	.176	—				
4. Index of use of reading skills at home (prose and document texts)	**.418**	.166	**.374**	—			
5. Index of use of task discretion at work	.015	.086	.014	.049	—		
6. Index of learning at work	.001	.098	.045	.107	.037	—	
7. Index of use of planning skills at work	−.202	.013	−.111	−.048	.140	.139	—
8. Index of readiness to learn	.246	.161	**.341**	.296	.206	.167	.026

TABLE 2

Factor Loadings and Communalities Based on Maximum Likelihood Analysis for Cognitive and Work Ability Indices (*N* = 191)

Item	Indices of Cognitive Skills	Indices of Work Abilities	Communality
1. Index of use of numeracy skills at home (basic and advanced)	.833	−.076	.741
2. Index of use of numeracy skills at work (basic and advanced)	.333	.124	.116
3. Index of use of ICT skills at home (derived)	.676	.145	.458
4. Index of use of reading skills at home (prose and document texts)	.535	.214	.303
5. Index of use of task discretion at work	.074	.315	.100
6. Index of learning at work	.077	.305	.094
7. Index of use of planning skills at work	−.168	.273	.121
8. Index of readiness to learn	.436	.549	.424

of factors to retain. Both 100 parallel datasets using artificial normally distributed raw data and 1,000 parallel datasets using permutated data suggested a two-factor model was appropriate (i.e., the first two raw data eigenvalues were greater than the random and permutated mean and 95th percentile eigenvalues; all other raw data eigenvalues were less in value). Although a more subjective tool for determining the number of factors, the scree plot indicated the eigenvalues leveled off after two factors, again supporting a two-factor solution. Interpretation of a two-factor solution was also plausible and was a consideration in retaining two factors. The two-factor solution represented about 30% of the variance explained when extracted. The correlation between the two extracted factors was .165.

All items contributed to a simple factor structure and had a primary factor loading above the recommended .30 with one exception—index of use of planning skills at work—which had a primary factor loading in the structure matrix of .273. One variable (index of readiness to learn) had similar factor loadings for each factor but loaded slightly stronger on factor two. All other variables had a strong primary loading with only one of the two factors in the factor structure. However, for interpretative purposes, this item was grouped with factor two. Table 2 provides the factor loading pattern matrix for the final solution. The names for the two factors are (1) Indices of Cognitive Skills and (2) Indices of Work Abilities. The results of the factor analysis lend support to internal structure validity evidence supporting the conclusion that the scores from this instrument are a valid assessment of skills and abilities, specifically Indices of Cognitive Skills and Indices of Work Abilities. Composite scores were created for the two factors by computing the mean sum of the items that loaded most strongly on each of the factors.

PROBLEMS

Conceptual Problems

1. If your research goal is to attach meaning to the identified factors, which form of factor analysis is needed?
   a. Common factor analysis
   b. Principal component analysis

2. What is the recommended sample size for EFA?
   a. At least 100
   b. At least 300
   c. At least 500
   d. Current research does not recommend adhering to an absolute number of cases threshold

3. Which one of the following commonly held recommendations has been shown by simulation research to often overestimate the number of factors?
   a. Bartlett's test
   b. Kaiser's rule
   c. Measure of sampling adequacy
   d. Scree plot

4. A researcher calculates KMO measure of sampling adequacy and finds a value of .60. Does this provide one form of acceptable evidence to continue the factor analysis?
   a. Yes
   b. No

5. Which one of the following is not used as an index to determine the initial factorability of items?
   a. Correlations among observed items
   b. Communalities
   c. Measure of sampling adequacy
   d. Scree plot

6. A researcher assumes the items they are factoring are related. Which one of the following rotation methods should be applied?
   a. Oblique
   b. Orthogonal
7. A researcher generates factor analysis and finds that the various indices all suggest different numbers of factors. How should the researcher determine the number of factors?
   a. Select the fewest number of factors suggested by the results.
   b. Select the number of factors based on where the elbow bends in the scree plot.
   c. Apply Kaiser's rule, selecting the number of factors with eigenvalues greater than one.
   d. Use theory to interpret results from all indices, selecting the number of factors supported statistically and defensible by theory.
8. Which one of the following is not an assumption of factor analysis?
   a. Absence of outliers
   b. Homogeneity of variances
   c. Linearity
   d. Noncollinearity
9. The measurement scale for conventional factor analysis should be at least which one of the following?
   a. Nominal
   b. Ordinal
   c. Interval
   d. Ratio
10. What factor loading is recommended for retaining a variable in a factor?
    a. .10
    b. .30
    c. .60
    d. .80

Computational Problems

1. Using the CH9_HW1_PRESCHOOL.sav dataset, conduct exploratory factor analysis following the steps in this chapter, using maximum likelihood estimation and promax rotation. Determine initial factorability using overall MSA, Bartlett's test of sphericity, and communalities. Review the pattern and structure matrix for the initial solution, and determine the variables that appear to cluster together based on the pattern matrix.
2. Using the CH9_HW2_PIAAC_NORWAY.sav dataset, conduct exploratory factor analysis following the steps in this chapter, using maximum likelihood estimation and promax rotation. Determine initial factorability using overall MSA, Bartlett's test of sphericity, and communalities. Review the pattern and structure matrix for the initial solution, and determine the variables that appear to cluster together based on the pattern matrix. *(Note: This data has been delimited to*

individuals who indicated their highest level of school was 'above high school' [B_Q01a_T = 3] and who were employed the year prior to completing the survey [B_Q15a = 1].)

Interpretive Problem

1. Use SPSS to conduct exploratory factor analysis with the continuous PIAAC index variables from Italy (CH9_HW_INTERPRETATIVE_ITALY.sav). The data file has been delimited to include only individuals who reported having 'above high school' education [B_Q01a_T = 3] and who had complete data on the index variables. Write up the results. Just for fun, compare the results using maximum likelihood estimation as compared to other estimation results. For even further fun, conduct CATPCA using the categorized index variables.

REFERENCES

Borsboom, D. (2006). The attack of the psychometrician. *Psychometrika, 71*(3), 425–440.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York, NY: Plenum.

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

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis: Understanding statistics*. New York, NY: Oxford University Press.

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies, 51*, 511–521.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*, 377–393.

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

Guadagnoli, E., & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265–275.

Guilford, J. P. (1954). *Psychometric methods* (Vol. 2nd ed.). New York, NY: McGraw-Hill.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *Journal of Experimental Education, 73*(3), 221–248.

Hahs-Vaughn, D. L. (2006a). Analysis of data from complex samples. *International Journal of Research & Method in Education, 29*(2), 163–181.

Hahs-Vaughn, D. L. (2006b). Weighting omissions and best practices when using large-scale data in educational research. *Association for Institutional Research Professional File, 101*, 1–9.

Hahs-Vaughn, D. L., McWayne, C. M., Bulotskey-Shearer, R. J., Wen, X., & Faria, A. (2011a). Complex sample data recommendations and troubleshooting. *Evaluation Review, 35*(3), 304–313. doi: 10.1177/0193841X11412070

Hahs-Vaughn, D. L., McWayne, C. M., Bulotskey-Shearer, R. J., Wen, X., & Faria, A. (2011b). Methodological considerations in using complex survey data: An applied example with the head start family and child experiences survey. *Evaluation Review, 35*(3), 269–303.

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

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393–416.

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

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage

Kaiser, H. K. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415.

Kaiser, H. K., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*, 111–117.

Kish, L., & Frankel, M. R. (1973, October 17). *Inference from complex samples.* Paper presented at the annual meeting of the Royal Statistical Society.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B, 36*, 1–37.

Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *American Statistician, 49*, 291–305.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworths.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. Newbury Park, CA: Sage.

Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *American Statistician, 49*, 64–70.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(8), 1–19.

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

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519–530.

National Science Foundation. (2010). Survey of doctorate recipients 2010. Retrieved from http://www.nsf.gov/statistics/srvydoctoratework/

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NJ: McGraw Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments & Computers, 32*(3), 396–402.

Penny, K. I. (1996). Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Applied Statistics, 45*, 73–81.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*(2), 317–337.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics, 32*(2), 153–161.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex samples*. New York: Wiley.

Small, N. J. H. (1980). Marginal skewness and kurtosis in testing multivariate normality. *Applied Statistics, 29*, 85–87.

Streiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports, 83*, 687–694.

Suhr, D. D. (2006). *Exploratory or confirmatory factor analysis?* Paper presented at the SAS User's Group International 31 (SUGI), San Francisco, CA.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197–208.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17*, 253–269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432–442.