CHAPTER 9

FACTOR ANALYSIS

In this chapter, we shift the focus of our attention—this time to the third group of advanced/multivariate statistical techniques previously discussed in this text. Specifically, we present a discussion of a procedure known as *factor analysis*, which is used to describe the underlying structure that "explains" a set of variables. It is a technique—similar to correlation and regression—that capitalizes on shared variability. Factor analysis has many practical applications within social science settings, as the reader will see as a result of our discussion.

SECTION 9.1 PRACTICAL VIEW

Purpose

Generally speaking, factor analysis is a procedure used to determine the extent to which *measurement overlap* (Williams, 1992)—that is, shared variance—exists among a set of variables. Its underlying purpose is to determine if measures for different variables are, in fact, measuring something in common. The mathematical procedure essentially takes the variance, as defined by the intercorrelations among a set of measures, and attempts to allocate it in terms of a smaller number of underlying hypothetical variables (Williams, 1992). These underlying, hypothetical—and unobservable—variables are called ***factors***. Factor analysis, then, is essentially a process by which the number of variables is reduced by determining which variables "cluster" together, and factors are the groupings of variables that are measuring some common entity or construct.

The main set of results obtained from a factor analysis consists of ***factor loadings***. A factor loading is interpreted as the Pearson correlation coefficient of an original variable with a factor. Like correlations, loadings range in value from -1.00 (representing a perfect negative association with the factor) through 0 to +1.00 (indicating perfect positive association). Variables typically will have loadings on all factors but will usually have high loadings on only one factor (Aron & Aron, 1999).

Another index provided in the results of a factor analysis is the list of ***communalities*** for each variable. Communalities represent the proportion of variability for a given variable that is explained by the factors (Agresti & Finlay, 1997) and allows the researcher to examine how individual variables reflect the sources of variability (Williams, 1992). Communalities may also be interpreted as the squared multiple correlation of the variable as predicted from the combination of factors, or as the sum of squared loadings across all factors for that variable.

The process by which the factors are determined from a larger set of variables is called ***extraction***. There are actually several types of factor extraction techniques, although the most commonly used empirical approaches are principal components analysis and factor analysis (Stevens, 1992). (It should

be noted that the term "factor analysis" is *commonly* used to represent the *general* process of variable reduction, regardless of the actual method of extraction utilized. For a detailed description of the various additional extraction techniques available, including maximum likelihood, unweighted least squares, generalized least squares, image factoring, and alpha factoring, the reader should refer to Tabachnick and Fidell, 1996.) In both principal components analysis and factor analysis, linear combinations (the factors) of original variables are produced, and a small number of these combinations typically account for the majority of the variability within the set of intercorrelations among the original variables.

In ***principal components analysis***, *all* sources of variability—unique, shared, and error variability—are analyzed for each observed variable. However, in ***factor analysis***, only *shared* variability is analyzed—both unique and error variability are ignored. This is based on the belief that unique and error variance serve only to "confuse" the picture of the underlying structure of a set of variables (Tabachnick & Fidell, 1996). In other words, principal components analysis analyzes variance; factor analysis analyzes covariance. Principal components analysis is usually the preferred method of factor extraction, especially when the focus of an analysis searching for underlying structure is truly exploratory, which is typically the case. Its goal is to extract the maximum variance from a data set, resulting in a few orthogonal (uncorrelated) components. When principal components analysis is used for extraction, the resulting linear combinations are often referred to as "components," as opposed to "factors." For the remainder of this chapter, we will limit our discussion to principal components analysis.

Since principal components analysis is an exploratory procedure, the first—and probably most important—decision required by the researcher is deciding how many components or factors to retain and, thus, interpret. The most widely accepted criterion was developed in 1960 by Kaiser, and has become known appropriately as "Kaiser's rule." The rule states that only those components whose eigenvalues are greater than 1 should be retained. An ***eigenvalue*** is defined as the amount of total variance explained by each factor, with the total amount of variability in the analysis equal to the number of original variables in the analysis (i.e., each variable contributes one unit of variability to the total amount due to the fact that the variance has been standardized).

A second, graphical method for determining the number of components is called the *scree test* and involves the examination of a scree plot. A ***scree plot*** is a graph of the magnitude of each eigenvalue (vertical axis) plotted against their ordinal numbers (horizontal axis). In order to determine the appropriate number of components to retain and interpret, one should look for the "knee," or bend, in the line. A typical scree plot will show the first one or two eigenvalues to be relatively large in magnitude, with the magnitude of successive eigenvalues dropping off rather drastically. At some point, the line will appear to level off. This is indicative of the fact that these successive eigenvalues are relatively small and, for the most part, of equal size. The recommendation is to retain all components with eigenvalues in the sharp descent of the line *before* the first one where the leveling effect occurs (Stevens, 1992). (If you are curious about the origin of the name for this type of plot, "scree" is formally defined as the rock debris located at the bottom of a cliff—an image one could envision in an actual scree plot.)

A third criterion used to determine the number of factors to keep in a factor or principal components analysis is to retain and interpret as many factors as will account for a certain amount of total variance. A general rule of thumb is to retain the factors that account for at least 70% of the total variability (Stevens, 1992), although there may be situations where the researcher will desire an even greater amount of variability to be accounted for by the components. However, this may not always be feasible. For example, assume we wanted to reduce the number of variables in an analysis containing 20 original variables. If it takes 15 components (or factors) to achieve the 70% criterion, we have not

gained much with respect to variable reduction and some underlying structure. Realize that in this situation, some factors will undoubtedly be variable-specific (i.e., only one variable will load on a given factor). Therefore, we have not uncovered any underlying structure for the *combination* of original variables.

A final criterion for retaining components is the assessment of model fit. Recall in Chapter 8 that we discussed the assessment of model fit for a path analysis model. The assessment of model fit involved the computation of the reproduced correlations (i.e., those that would occur assuming the model represents reality) and comparing them to the original, observed correlations. If the number of correlations that are reasonably close (again, within .05 of each other) is small, it can be assumed that the model is consistent with the empirical data. One advantage of all factor analytic procedures over path analysis is that computer analysis programs—including SPSS—will calculate the reproduced correlations for you, and even provide a percentage of the total that exceed the cutoff value of .05.

With four different criteria to evaluate, how can one be sure of the number of components to retain and interpret, especially if examination of the four criteria results in different decisions regarding the number of components? Stevens (1992) offers several suggestions when faced with this often occurring dilemma. He states that Kaiser's rule has been shown to be quite accurate when the number of original variables is < 30 and the communalities are $> .70$, or when $N > 250$ and the mean communality is $\geq .60$ (p. 379). In other situations, use of the scree test with an $N > 250$ will provide fairly accurate results, provided that most of the communalities are somewhat large (i.e., $> .30$). Our recommendation is to examine all four criteria for alternative factor solutions and weigh them against the overall goal of any multivariate analysis—parsimony. *It is our belief that the principle of parsimony is more important in factor or principal components analysis than in any other analysis technique.*

Let us examine these various criteria for deciding how many components to keep through the development of an example that we will submit to a principal components analysis. Assume we wanted to determine what, if any, underlying structure exists for measures on ten variables, consisting of:

- male life expectancy (*lifexpm*),
- female life expectancy (*lifexpf*),
- births per 1,000 people (*birthrat*),
- infant mortality rate (*infmr*),
- fertility rate per woman (*fertrate*),
- natural log of doctors per 10,000 people (*lndocs*),
- natural log of radios per 100 people (*lnradio*),
- natural log of telephones per 100 people (*lnphone*),
- natural log of gross domestic product (*lngdp*), and
- natural log of hospital beds per 10,000 people (*lnbeds*).

We would first examine the number of eigenvalues greater than 1.00. The table of total variance accounted for in the initial factor solution for these ten variables is shown in Figure 9.1. With an eigenvalue equal to 8.161, only the first component has an eigenvalue that exceeds the criterion value of 1.00. The second component (.590) does not even approach the criterion value. Additionally, the first component accounts for nearly 82% of the total variability in the original variables, while the second component only accounts for 6%.

Figure 9.1 Initial Eigenvalues and Percentage of Variance Explained by Each Component for Initial Example.

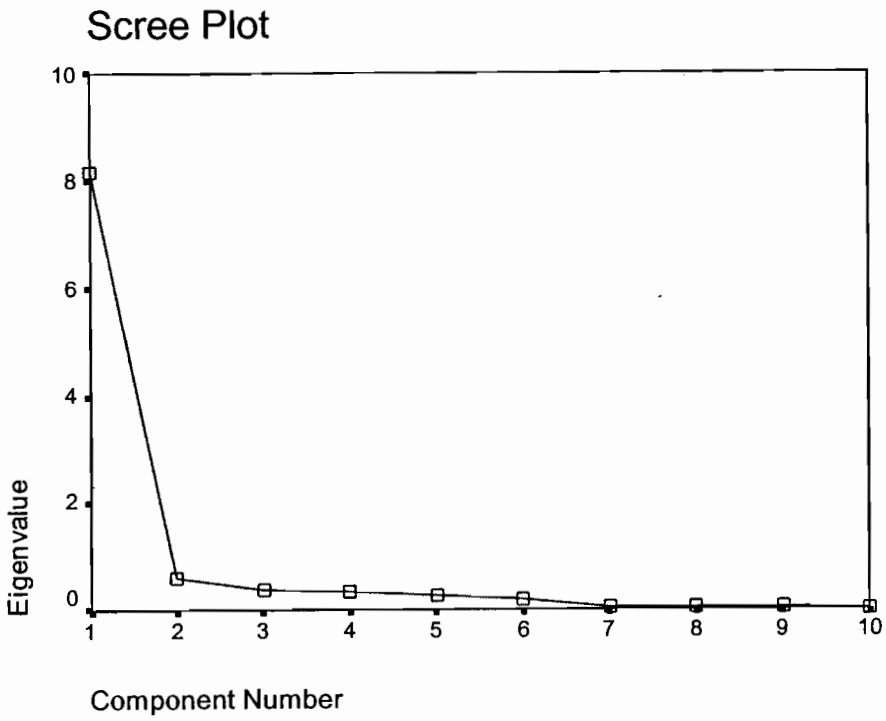
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608
2	.590	5.902	87.510			
3	.372	3.717	91.227			
4	.338	3.384	94.611			
5	.246	2.460	97.071			
6	.168	1.677	98.748			
7	5.491E-02	.549	99.297			
8	3.319E-02	.332	99.629			
9	2.978E-02	.298	99.927			
10	7.316E-03	7.316E-02	100.000			

Examination of the scree plot for this solution (see Figure 9.2) provides us with similar results. The first component is much larger than subsequent components in terms of eigenvalue magnitude; eigenvalues of successive components drop off quite drastically. Clearly, the line begins to level off at the second component. Based on this plot, it appears that we should retain and interpret only the first component.

Although we do not provide the output here, the reproduced correlation matrix indicates that 18 (40%) of the residuals (i.e., differences between observed and reproduced correlations) have absolute values greater than .05. Since this indicates that the one-component model does not fit the empirical correlations very well, we might want to investigate the possibility of a two-component model.

The reader should realize that when attempting a revised model with a different number of factors, the values for the initial eigenvalues, percent of variance explained, and the appearance of the scree plot will not change—these are based solely on the original correlations. The only substantive difference can be noticed in the numbers of residuals that exceed the criterion value. In the two-component model, only 11 (24%) of the residuals exceed our .05 criterion—a substantial improvement over our previous percentage, equal to 40%. We now compare the two possible models side-by-side in order to determine which we will interpret.

Figure 9.2 Scree Plot (Example number 1).



Number of Components	% Variance Accounted For	Number of Residuals > .05
1	81.6%	24 (40%)
2	5.9%	11 (24%)

Based on the variance explained and the scree plot, it would appear that one component should be interpreted; however, this may be an oversimplification of the reduction of our original data. Furthermore, the addition of a second component certainly improved the fit of the model to our empirical data. For this latter reason, we will proceed with the interpretation of the two-component solution.

Before we attempt to interpret the components based on the values of the loadings, it is imperative that we discuss the topic of factor (component) rotation. ***Rotation*** is a process by which a factor solution is made more interpretable without altering the underlying mathematical structure. Rotation is a complex mathematical procedure and it is sometimes helpful to consider it from a geometric perspective. For the sake of simplicity, let us assume that we have four variables we have submitted to a factor analysis. The analysis returned to us two components and the associated hypothetical loadings are presented below:

Variable	Loading on Component 1	Loading on Component 2
A	.850	.120
B	.700	.210
C	–.250	.910
D	.210	–.750

If we were to plot these combinations of loadings in a scatterplot of Component 1 by Component 2, the result would appear as in Figure 9.3. Notice that the possible loadings on each component range from -1.00 to +1.00 and that the locations of the four variables in that geometric space have been plotted according to the combination of loadings on the two components in Part (a) of Figure 9.3—for example, Variable A is located at X=.850, Y=.120. Although the points are generally near the lines, if we were able to rotate the axes, we would notice a better "fit" of the loadings to the actual components. In Part (b), we now see how three of the four loadings line up nearly perfectly with the two components. This process alters the original values of the component loadings without changing their mathematical properties. This gives us the ability to name the components with greater ease since three of the variables correlate nearly perfectly with the two components, and the rotated factor loadings would change accordingly.

The researcher must decide whether to use an orthogonal or oblique rotation. ***Orthogonal rotation*** is a rotation of factors that results in factors being uncorrelated with each other; the resultant computer output is a *loading* matrix (i.e., a matrix of correlations between all observed variables and factors) where the size of the loading reflects the extent of the relationship between each observed variable and each factor. Since the goal of factor analysis is to obtain underlying factors which are uncorrelated (thereby representing some *unique* aspect of the underlying structure), it is recommended that orthogonal rotation be used instead of oblique. There are three types of orthogonal rotation procedures—varimax, quartimax, and equamax—of which varimax is the most commonly used. Orthogonal rotation methods are described further in Section 9.5.

Oblique rotation results in factors being correlated with each other. Several matrices are produced from an oblique rotation: a *factor correlation* matrix (i.e., a matrix of correlations between all factors); a loading matrix that is separated into a *structure* matrix (i.e., correlations between factors and variables); and a *pattern* matrix (i.e., unique relationships with no overlapping among factors and each observed variable). The interpretation of factors is based on loadings found in the pattern matrix. One would use an oblique rotation only if there were some prior belief that the underlying factors are correlated. Several types of oblique rotations exist, including direct oblimin, direct quartimin, orthoblique, and promax. Direct oblimin is arguably the most frequently used form of oblique rotation. Oblique rotation methods are described further in Section 9.5.

Once we have rotated the initial solution, we are ready to *attempt* interpretation. We emphasize the *attempt* at interpretation because, by its very nature, interpretation of components or factors involves much subjective decision making on the part of the researcher (Williams, 1992). The rotated component loadings for our working example are presented in Figure 9.4. Initially, the reader should notice that each variable has a loading on each component, although in most cases each has a high loading on only one component. Some of the variables in our example have loaded relatively high on both compo-

nents, but we will "assign" a given variable to the component with the higher loading (as shown by the shaded boxes) and attempt to interpret them in that fashion.

Figure 9.3 Illustration of Geometric Interpretation of Rotation of Components.

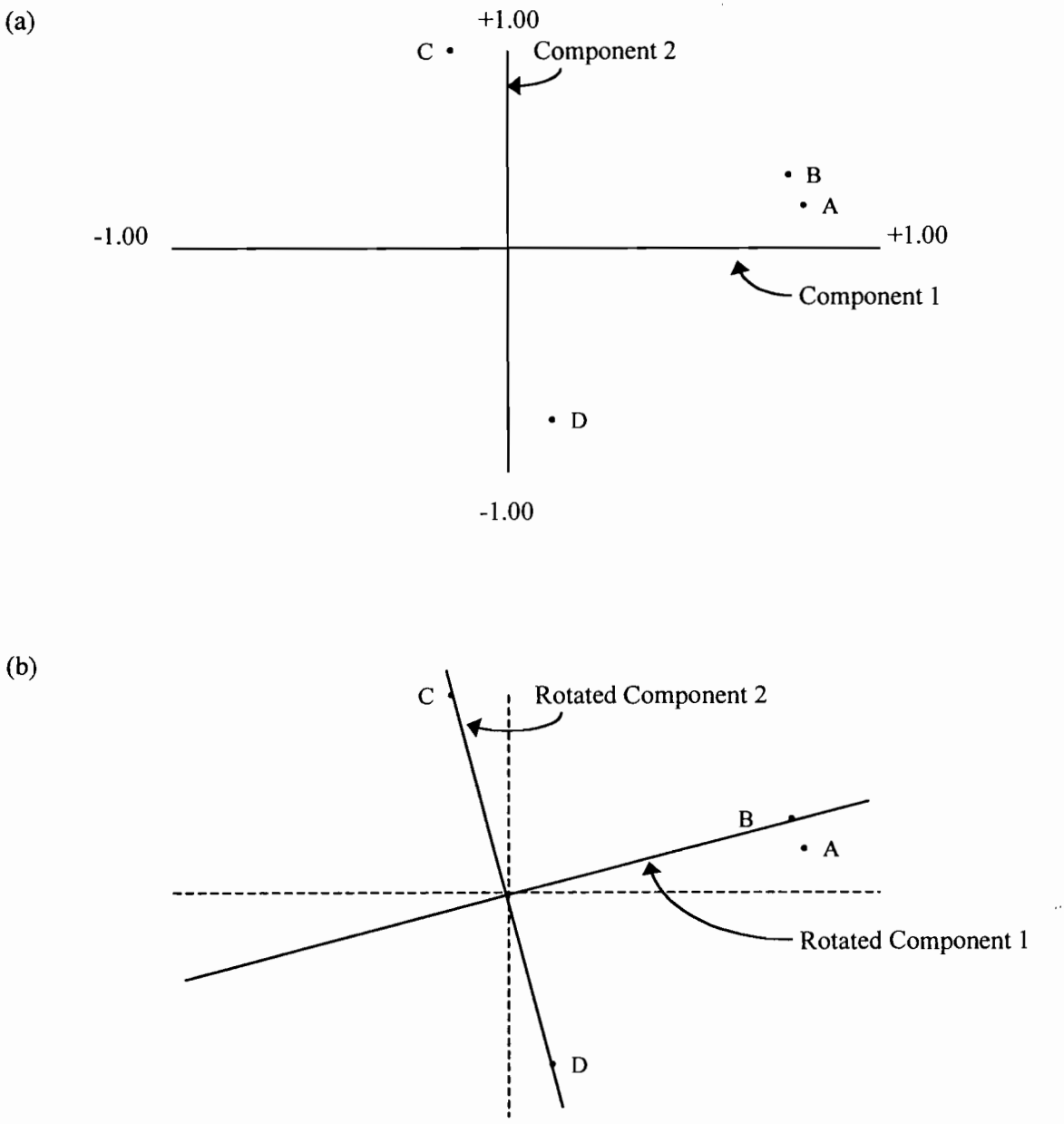


Figure 9.4 Component Loadings for the Rotated Solution.

	Component	
	1	2
Fertility rate per woman 1990	-.879	-.305
Births per 1,000 population 1992	-.858	-.393
Male life expectancy 1992	.846	.450
Infant mortality rate 1992 (per 1,000 live births)	-.839	-.461
Female life expectancy 1992	.829	.509
Natural log of doctors per 10,000	.763	.502
Natural log of radios per 100 people	.279	.879
Natural log hospital beds per 10,000	.480	.719
Natural log of GDP	.627	.688
Natural log of phones per 100 people	.675	.678

When interpreting or naming components, one should pay particular attention to two aspects—the size and direction—of each loading. The name attached to each component should reflect the relative sizes and directions of the loadings. Notice that Component 1 contains both high positive and high negative loadings; this is referred to as a *bipolar* factor. In other words, the name we assign to Component 1 should *primarily* reflect the strong positive loadings for male life expectancy (*lifexpm*) and female life expectancy (*lifexpf*), and subsequently the loading for *lndocs*. Due to negative loadings, the name should also reflect the *opposite* of fertility rate (*fertrat*), birth rate (*birthrat*), and infant mortality rate (*infmr*). Bipolar components or factors are usually more difficult to interpret. Since this component seems to address the *end* of an individual's life span as opposed to the beginning of that lifetime, and factoring in the number of docs (*lndocs*), we might consider attaching the label of *Healthy Lifespan* to this component. The second component—on which all variables have positive loadings—addresses the numbers of radios, hospital beds, and phones, as well as the nation's gross domestic product. We might consider attaching the label of *Economic Stature* to this component.

It is important to note that principal components analysis may be used as a variable reducing scheme for further analyses (Stevens, 1992). We have already examined the main application of the analysis—to determine empirically how many underlying constructs account for the majority of the variability among a set of variables. Principal components analysis may also be used as a precursor to multiple regression as a means of reducing the number of predictors, especially if the number of predictor variables is quite large relative to the number of subjects. Additionally, components analysis may be used to reduce the number of criterion variables in a multivariate analysis of variance. It is often recommended that a large number of DVs not be used in a MANOVA procedure; if you have a large number of DVs that you are interested in analyzing, reducing the overall number of variables through a principal components analysis would allow you to accomplish this task rather efficiently.

If a principal components analysis is to be followed by a subsequent analytic procedure, factor scores are often used. ***Factor scores*** are estimates of the scores subjects would have received on each of the factors had they been measured directly (Tabachnick & Fidell, 1996). Many different procedures may be used to estimate factor scores, the most basic of which is to simply sum the values across all variables that load on a given factor or component. Alternatively, a mean could be calculated across all variables that would then represent the score on a factor. Factor scores could also be estimated or predicted through a regression analysis. Those values are then entered into the subsequent analysis as if they were "raw" variables.

Finally, there are two basic types of factor analytic procedures, based on their overall intended function: exploratory and confirmatory factor analyses. In ***exploratory factor analysis***, the goal is to describe and summarize data by grouping together variables that are correlated. The variables included in the analysis may or may not have been chosen with these underlying structures in mind (Tabachnick & Fidell, 1996). Exploratory factor analysis usually occurs during the early stages of research, when it often proves useful to consolidate numerous variables.

Confirmatory factor analysis is much more advanced and sophisticated than exploratory factor analysis. It is often used to test a theory about latent (i.e., underlying, unobservable) processes that might occur among variables. A major difference between confirmatory and exploratory factor analyses is that in a confirmatory analysis, variables are painstakingly and specifically chosen in order to adequately represent the underlying processes (Tabachnick & Fidell, 1996). The main purpose of confirmatory factor analysis is to confirm—or disconfirm—some *a priori* theory. LISREL, as previously discussed in Chapter 8, is often used as the analytical computer program in such studies.

Sample Research Questions

Building on the example we began discussing in the previous section, we now specify the main research questions to be addressed by our principal components analysis. Using the ten original variables, the appropriate research questions would be:

(1) How many reliable and interpretable components are there among the following ten variables: male life expectancy, female life expectancy, births per 1,000 people, infant mortality rate, fertility rate per woman, number of doctors per 10,000 people, number of radios per 100 people, number of telephones per 100 people, gross domestic product, and number of hospital beds per 10,000 people?

(2) If reliable components are identified, how might we interpret those components?

(3) How much variance in the original set of variables is accounted for by the components?

SECTION 9.2 ASSUMPTIONS AND LIMITATIONS

If principal components analysis and factor analysis are being used in a descriptive fashion as a method of summarizing the relationships among a large set of variables, assumptions regarding the distributions of variables in the population are really not in force and, therefore, do not need to be assessed (Tabachnick & Fidell, 1996). This is usually the case since, as previously mentioned, principal components and factor analyses are almost always exploratory and descriptive in nature. It should be noted, however, that if the variables are normally distributed, the resultant factor solution will be enhanced; to the extent to which normality fails, the solution is degraded (although it still may be worthwhile) (Tabachnick & Fidell, 1996).

Previous versions of SPSS provided a statistical test of model fit—a value for a test statistic was provided and subsequently evaluated using a chi-square criterion. In situations where this test statistic was evaluated—in an inferential manner—and used to determine the number of factors or components, assessment of model assumptions takes on much greater importance. Since recent revisions of SPSS have omitted the chi-square test of model fit, this criterion can obviously no longer be used to determine the number of factors to interpret. Therefore, it is not necessary to test the assumptions of multivariate normality and linearity. However, we recommend that both of these assumptions be evaluated and any necessary transformations be made—*ensuring the quality of data will only improve the quality of the resulting factor or component solution.*

As a reminder to the reader, these two aforementioned assumptions are formally stated as follows:

(1) All variables, as well as all linear combinations of variables, must be normally distributed (assumption of multivariate normality).
(2) The relationships among all pairs of variables must be linear.

Factor analyses, in general, are subject to a potentially severe limitation. Recall that the basis for any underlying structure that is obtained from a factor analysis are the relationships among all original variables in the analysis. Correlation coefficients have a tendency to be less reliable when estimated from small samples. If unreliable—or at least, *less* reliable—correlations exist among variables, and those variables are subjected to a factor analysis, the resultant factors will also not be very reliable. Tabachnick and Fidell (1996) offer the following guidelines for sample sizes and factor analyses:

Approximate Sample Size	Estimated Reliability
50	very poor
100	poor
200	fair
300	good
500	very good
1000	excellent

As a general rule of thumb, they suggest that a data set include *at least* 300 cases for a factor analysis to return reliable factors. If a solution contains several high-loading variables (> .80), a smaller sample (e.g., $n = 150$) would be sufficient.

Stevens (1992) has offered a different, although somewhat similar, set of recommendations, based on the number of variables (with minimum/maximum loadings) per component (p. 384). Specifically, these recommendations are as follows:

(1) Components with four or more loadings above .60 in absolute value (i.e., |.60|) are reliable, regardless of sample size.
(2) Components with about 10 or more low loadings (i.e., < |.40|) are reliable as long as the sample size is greater than 150.

(3) Components with only a few low loadings should not be interpreted unless the sample size is at least 300.

It should be noted that the above constitute *general* guidelines, *not* specific criteria which *must* be met by the applied researcher. If researchers are planning a factor analysis with small sample sizes, it is recommended that they apply *Bartlett's sphericity test.* This procedure tests the null hypothesis that the variables in the population correlation matrix are uncorrelated. If one fails to reject this hypothesis, there is no reason to do a principal components analysis since the variables are already uncorrelated (i.e., they have nothing in common) (Stevens, 1992).

Methods of Testing Assumptions

The reader will recall that the assessment of multivariate normality is not easily accomplished through the use of standard statistical software packages. The most efficient method of assessing multivariate normality is to assess univariate normality—recall that univariate normality is a necessary condition for multivariate normality. Normality among individual variables may be evaluated by examining the values for skewness and kurtosis, normal probability plots, and/or bivariate scatterplots.

The assumption of linearity is best tested through the inspection of bivariate scatterplots obtained for each pair of original variables. The reader will recall that if a relationship is in fact linear, a general elliptical shape should be apparent in the scatterplot.

SECTION 9.3 PROCESS AND LOGIC

The Logic Behind Factor Analysis

The underlying, mathematical objective in principal components analysis is to obtain uncorrelated linear combinations of the original variables that account for as much of the total variance in the original variables as possible (Johnson & Wichern, 1998). These uncorrelated linear combinations are referred to as the ***principal components***. The logic behind principal components analysis involves the partitioning of this total variance by initially finding the first principal component (Stevens, 1992). The ***first principal component*** is the linear combination that accounts for the *maximum* amount of variance and is defined by the equation:

$$PC_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1p}x_p \qquad \text{(Equation 9.1)}$$

where PC_1 is the first principal component, x_i refers to the measure on the original variable, and a_{1i} refers to the weight assigned to a given variable for the first principal component (the first subscript following the a identifies the specific principal component, and the second subscript identifies the original variable)—e.g., the term $a_{11}x_1$ refers to the product of the weight for variable 1 on PC_1 and the original value for an individual on variable 1. The subscript $_p$ is equal to the total number of original variables. This linear combination, then, accounts for the maximum amount of variance within the original set of variables—the variance of the first principal component is equal to the largest eigenvalue (i.e., the eigenvalue for the first component).

The analytic operation then proceeds to find the second linear combination— *uncorrelated* with the first linear combination—that accounts for the next largest amount of variance (after that which has been attributed to the first component has been removed). The resulting equation would be:

$$PC_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2p}x_p \qquad \text{(Equation 9.2)}$$

It is important to note that the extracted principal components are not related. In other words,

$$r_{PC_1 \bullet PC_2} = 0$$

The third principal component is constructed so that it is uncorrelated with the first two and accounts for the next largest amount of variance in the system of original variables, after the two largest amounts have been removed. This process continues until all variance has been accounted for by the extracted principal components.

Interpretation of Results

The process of interpreting factor analysis results focuses on the determination of the number of factors to retain. As mentioned earlier, there are several methods/criteria to utilize in this process:

1) Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criteria is fairly reliable when the number of variables is < 30 and communalities are > .70, or the number of individuals is > 250 and the mean communality is ≥ .60.
2) Variance—Retain components that account for at least 70% total variability.
3) Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criteria is fairly reliable when the number of individuals is > 250 and communalities are > .30.
4) Residuals—Retain the components generated by the model if only a few residuals (the difference between the empirical and reproduced correlations) exceed .05. If several reproduced correlations differ, you may want to include more components.

Since the sample size and number of variables can impact the number of factors generated in the analysis as well as the assessment of these four criteria, we recommend utilizing all four. Another reason to examine all four criteria is that within an exploratory factor analysis, the eigenvalue is the default criteria for determining the number of factors, which can lead to an inaccurate number of factors retained. For example, if an analysis determines that only two components have eigenvalues greater than 1, the model generated will include only those two components. However, the researcher may examine the other three criteria and determine that one more component should be included. In such an instance, the analysis would have to be conducted again to override the eigenvalue criteria so that three components instead of two would be generated.

Once criteria have been evaluated and you have determined the appropriate number of components to retain (the reader should note that this decision may lead to further analysis in order to include the appropriate number of components), the nature of each component must be assessed in order to interpret/name it. This is done by noting positive and negative loadings, ordering variables with respect to loading strength, and examining the content of variables that composes each component.

Although this interpretation process has been somewhat applied to our initial example, we will describe this process more in-depth in conjunction with the output. Our example seeks to determine what, if any, underlying structure exists for measures on the following ten variables: male life expectancy (*lifexpm*), female life expectancy (*lifexpf*), births per 1,000 people (*birthrat*), infant mortality rate (*infmr*), fertility rate per woman (*fertrate*), natural log of doctors per 10,000 people (*lndocs*), natural log of radios per 100 people (*lnradio*), natural log of telephones per 100 people (*lnphone*), natural log of gross domestic product (*lngdp*), and natural log of hospital beds per 10,000 people (*lnbeds*). Data were

first screened for missing data and outliers. No outliers were found when utilizing Mahalanobis distance. Univariate linearity and normality were analyzed by creating a scatterplot matrix (see Figure 9.5). The elliptical shapes indicate normality and linearity. The reader should note that the following variables were previously transformed variables by taking the natural log: number of doctors per 10,000 people, number of radios per 100 people, number of telephones per 100 people, gross domestic product, and number of hospital beds per 10,000 people. A factor analysis was then conducted using **`Data Reduction`**, which utilized the eigenvalue criteria and varimax rotation. Applying the four methods of interpretation, we first examined the eigenvalues in the table of total variance (see Figure 9.6). Only one component had an eigenvalue greater than 1; however, the eigenvalue criteria is only reliable when the number of variables is less than 30 and communalities are greater than .7. Figure 9.7 presents the communalities and indicates that two variables have values less than .7. Consequently, the application of the eigenvalue criteria is questionable. The next criteria to assess is variance, also displayed in Figure 9.6. The first component accounts for nearly 82% of the total variance in the original variables, whereas the second component accounts for only 5.9%. The reader should note that since only one component was retained, the factor solution was not rotated. The scree plot was then assessed and indicates that the eigenvalues after the first component drop off drastically (see Figure 9.2). These last two methods imply that only the first component should be retained. However, evaluation of residuals (differences between the reproduced correlations and the original correlations) indicate that two components should be investigated. Figure 9.8 presents the reproduced correlations as well as the residuals. Of the 45 residuals, 18 (40%) exceed the .05 criteria. Since two of the four criteria are in question, we will investigate retaining two components to improve model fit.

Figure 9.5 Scatterplot Matrix of Variables (Example number 1).

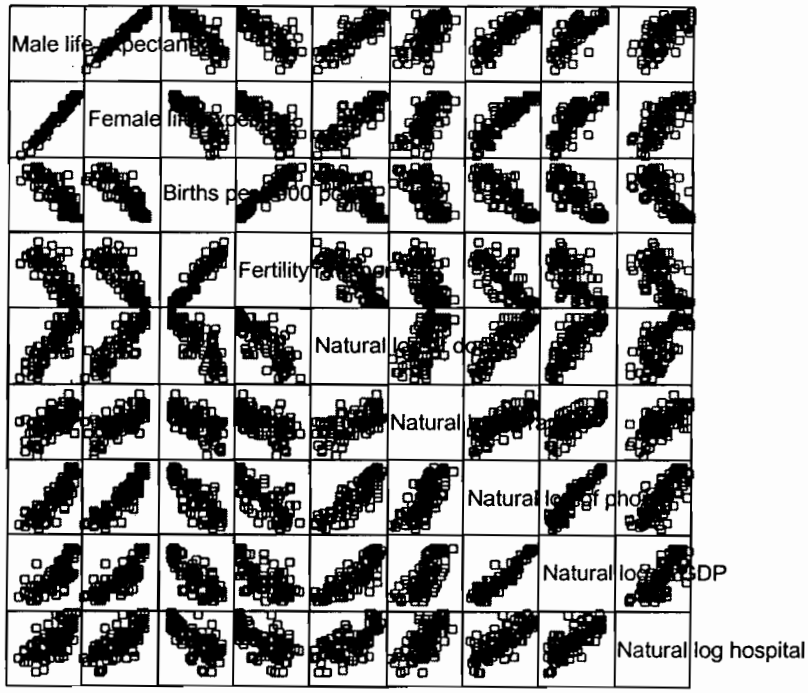


Figure 9.6 Table of Total Variance for One Component Solution (Example number 1).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608
2	.590	5.902	87.510			
3	.372	3.717	91.227			
4	.338	3.384	94.611			
5	.246	2.460	97.071			
6	.168	1.677	98.748			
7	5.491E-02	.549	99.297			
8	3.319E-02	.332	99.629			
9	2.978E-02	.298	99.927			
10	7.316E-03	7.316E-02	100.000			

Extraction Method: Principal Component Analysis.

Figure 9.7 Communalities for One-Component Solution (Example number 1).

Communalities

	Initial	Extraction
LIFEEXPM	1.000	.894
LIFEEXPF	1.000	.936
BIRTHRAT	1.000	.846
INFMR	1.000	.894
FERTRATE	1.000	.779
LNDOCS	1.000	.830
LNRADIO	1.000	.573
LNPHONE	1.000	.899
LNGDP	1.000	.839
LNBEDS	1.000	.670

Extraction Method: Principal Component Analysis.

Eigenvalue criteria is questionable since two communalities are less than .7.

A factor analysis was conducted again, this time eliminating the eigenvalue criteria and indicating that two factors should be retained. Varimax rotation was also applied. Since we have eliminated the eigenvalue criteria and found the scree plot and variance criteria both to indicate retaining only one component, we will immediately move to the fourth criteria—assessment of residuals (see Figure 9.9). This time only 11 residuals were greater than .05; consequently, the model has been improved. Because two components were retained, the model was rotated to improve model fit. The table of total variance displays the amount of variance for each component before and after rotation (see Figure 9.10). One should note that factor rotation does not effect the total variance accounted for by the model but does change how the variance is distributed among the retained components. Prior to rotation, the first component accounted for 81.61% and the second for only 5.9%. However, once rotated, the first component accounted for 53.54% and the second for 33.97%. Figure 9.11 displays how variables were loaded into the components after rotation. Assessment of component loadings is necessary to name each com-

ponent. Component number 1 was composed of both negative and positive loadings, which somewhat complicates matters. Positive loadings included the variables of male life expectancy, female life expectancy, and the number of doctors. Negative loadings included fertility rate, birth rate, and infant mortality. One should also note the variables with the highest loadings were fertility rate (-.879) followed by birth rate (-.859). Since these variables all seemed to relate to the health of one's life, this component will be named *Healthy Lifespan*. Component number 2 included all positive loadings and addressed the number of radios, number of hospital beds, gross domestic product, and the number of phones, respectively. This component will be labeled *Economic Stature*.

Figure 9.8 Reproduced Correlations and Residuals for One-Component Solution (Example number 1).

Reproduced Correlations

		LIFEEXPM	LIFEEXPF	BIRTHRAT	INFMR	FERTRATE	LNDOCS	LNRADIO	LNPHONE	LNGDP	LNBEDS
Reproduced Correlation	LIFEEXPM	.894[b]	.915	-.870	-.894	-.834	.861	.716	.896	.866	.774
	LIFEEXPF	.915	.936[b]	-.890	-.915	-.854	.881	.732	.917	.886	.792
	BIRTHRAT	-.870	-.890	.846[b]	.870	.812	-.838	-.696	-.872	-.843	-.753
	INFMR	-.894	-.915	.870	.894[b]	.835	-.862	-.716	-.897	-.866	-.774
	FERTRATE	-.834	-.854	.812	.835	.779[b]	-.804	-.668	-.837	-.809	-.723
	LNDOCS	.861	.881	-.838	-.862	-.804	.830[b]	.690	.864	.835	.746
	LNRADIO	.716	.732	-.696	-.716	-.668	.690	.573[b]	.718	.694	.620
	LNPHONE	.896	.917	-.872	-.897	-.837	.864	.718	.899[b]	.869	.776
	LNGDP	.866	.886	-.843	-.866	-.809	.835	.694	.869	.839[b]	.750
	LNBEDS	.774	.792	-.753	-.774	-.723	.746	.620	.776	.750	.670[b]
Residual[a]	LIFEEXPM		7.300E-02	3.658E-02	-6.64E-02	2.648E-02	1.548E-02	-7.40E-02	-1.949E-02	-4.82E-02	-8.83E-02
	LIFEEXPF	7.300E-02		3.122E-02	-5.68E-02	2.157E-02	-3.42E-03	-4.29E-02	-1.958E-02	-4.45E-02	-5.12E-02
	BIRTHRAT	3.658E-02	3.122E-02		-1.67E-02	.146	2.105E-02	7.335E-02	5.575E-02	3.509E-02	3.658E-02
	INFMR	-6.638E-02	-5.68E-02	-1.668E-02		-3.83E-03	2.955E-02	6.251E-02	3.408E-02	5.469E-02	5.166E-02
	FERTRATE	2.648E-02	2.157E-02	.146	-3.83E-03		3.657E-02	9.574E-02	6.335E-02	6.850E-02	6.203E-02
	LNDOCS	1.548E-02	-3.42E-03	2.105E-02	2.955E-02	3.657E-02		-6.38E-02	1.062E-03	-9.95E-03	-3.69E-02
	LNRADIO	-7.400E-02	-4.29E-02	7.335E-02	6.251E-02	9.574E-02	-6.38E-02		2.086E-02	1.551E-02	2.879E-02
	LNPHONE	-1.949E-02	-1.96E-02	5.575E-02	3.408E-02	6.335E-02	1.062E-03	2.086E-02		7.333E-02	-3.53E-03
	LNGDP	-4.818E-02	-4.45E-02	3.509E-02	5.469E-02	6.850E-02	-9.95E-03	1.551E-02	7.333E-02		1.684E-02
	LNBEDS	-8.831E-02	-5.12E-02	3.658E-02	5.166E-02	6.203E-02	-3.69E-02	2.879E-02	-3.529E-03	1.684E-02	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 18 (40.0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities

Figure 9.9 Reproduced Correlations and Residuals for Two-Component Solution (Example number 1).

Reproduced Correlations

		LIFEEXPM	LIFEEXPF	BIRTHRAT	INFMR	FERTRATE	LNDOCS	LNRADIO	LNPHONE	LNGDP	LNBEDS
Reproduced Correlation	LIFEEXPM	.919[b]	.931	-.903	-.917	-.881	.872	.632	.876	.840	.730
	LIFEEXPF	.931	.946[b]	-.911	-.930	-.884	.888	.679	.904	.870	.764
	BIRTHRAT	-.903	-.911	.890[b]	.900	.873	-.852	-.585	-.845	-.808	-.694
	INFMR	-.917	-.930	.900	.915[b]	.877	-.871	-.639	-.878	-.843	-.734
	FERTRATE	-.881	-.884	.873	.877	.865[b]	-.824	-.513	-.800	-.761	-.641
	LNDOCS	.872	.888	-.852	-.871	-.824	.834[b]	.654	.855	.824	.727
	LNRADIO	.632	.679	-.585	-.639	-.513	.654	.851[b]	.785	.780	.767
	LNPHONE	.876	.904	-.845	-.878	-.800	.855	.785	.915[b]	.890	.812
	LNGDP	.840	.870	-.808	-.843	-.761	.824	.780	.890	.866[b]	.796
	LNBEDS	.730	.764	-.694	-.734	-.641	.727	.767	.812	.796	.748[b]
Residual[a]	LIFEEXPM		5.692E-02	6.987E-02	-4.34E-02	7.294E-02	4.886E-03	9.570E-03	6.809E-04	-2.22E-02	-4.42E-02
	LIFEEXPF	5.692E-02		5.256E-02	-4.20E-02	5.134E-02	-1.02E-02	1.060E-02	-6.652E-03	-2.79E-02	-2.29E-02
	BIRTHRAT	6.987E-02	5.256E-02		-4.72E-02	8.397E-02	3.510E-02	-3.75E-02	2.898E-02	6.287E-04	-2.19E-02
	INFMR	-4.337E-02	-4.20E-02	-4.720E-02		-4.64E-02	3.926E-02	-1.41E-02	1.558E-02	3.088E-02	1.122E-02
	FERTRATE	7.294E-02	5.134E-02	8.397E-02	-4.64E-02		5.618E-02	-5.90E-02	2.600E-02	2.041E-02	-1.97E-02
	LNDOCS	4.886E-03	-1.02E-02	3.510E-02	3.926E-02	5.618E-02		-2.85E-02	9.578E-03	1.010E-03	-1.82E-02
	LNRADIO	9.570E-03	1.060E-02	-3.754E-02	-1.41E-02	-5.90E-02	-2.85E-02		-4.632E-02	-7.10E-02	-.118
	LNPHONE	6.809E-04	-6.65E-03	2.898E-02	1.558E-02	2.600E-02	9.578E-03	-4.63E-02		5.246E-02	-3.90E-02
	LNGDP	-2.221E-02	-2.79E-02	6.287E-04	3.088E-02	2.041E-02	1.010E-03	-7.10E-02	5.246E-02		-2.88E-02
	LNBEDS	-4.420E-02	-2.29E-02	-2.195E-02	1.122E-02	-1.97E-02	-1.82E-02	-.118	-3.899E-02	-2.88E-02	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 11 (24.0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities

Figure 9.10 Table of Total Variance for Two-Component Solution (Example number 1).

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608	5.354	53.542	53.542
2	.590	5.902	87.510	.590	5.902	87.510	3.397	33.968	87.510
3	.372	3.717	91.227						
4	.338	3.384	94.611						
5	.246	2.460	97.071						
6	.168	1.677	98.748						
7	5.491E-02	.549	99.297						
8	3.319E-02	.332	99.629						
9	2.978E-02	.298	99.927						
10	7.316E-03	7.316E-02	100.000						

Extraction Method: Principal Component Analysis.

Figure 9.11 Factor Loadings for Rotated Components (Example number 1).

Rotated Component Matrix[a]

	Component	
	1	2
LIFEEXPM	.846	.450
LIFEEXPF	.829	.509
BIRTHRAT	-.858	-.393
INFMR	-.839	-.461
FERTRATE	-.879	-.305
LNDOCS	.763	.502
LNRADIO	.279	.879
LNPHONE	.675	.678
LNGDP	.627	.688
LNBEDS	.480	.719

Loadings for Component #1.

Loadings for Component #2.

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Writing Up Results

Once again, the results narrative should always describe the elimination of subjects and transformation of variables. The summary should then describe the type of factor analysis conducted and indicate if any rotation method was utilized. The interpretation process is then presented in conjunction with the results. In general, it is only necessary to indicate the number of factors retained and the criteria that led to that decision. The results of the final solution are then presented. One should summarize each component by presenting the following: the percent of variance, the number and names of variables loaded into the component, and the component loadings. This is often displayed in table format, depending upon the number of components and variables. Finally, the researcher indicates the names of components. The following summary applies the output presented in Figures 9.5 – 9.11.

> Factor analysis was conducted to determine what, if any, underlying structure exists for measures on the following ten variables: male life expectancy (*lifexpm*), female life expectancy (*lifexpf*), births per 1,000 people (*birthrat*), infant mortality rate (*infmr*), fertility rate per woman (*fertrate*), doctors per 10,000 people (*docs*), radios per 100 people (*radio*), telephones per 100 people (*phone*), gross domestic product (*gdp*), and hospital beds per 10,000 people (*beds*). Prior to the analysis, evaluation of linearity and normality led to the natural log transformation of *docs*, *radios*, *phones*, *gdp*, and *beds*. Principal components analysis was conducted utilizing a varimax rotation. The initial analysis retained only one component. Four criteria were used to determine the appropriate number of components to retain: eigenvalue, variance, scree plot, and residuals. Criteria indicated that retaining two components should be investigated. Thus, principal components analysis was conducted to retain two components and apply the varimax rotation. Inclusion of two components increased the model fit as it decreased the number of residuals exceeding the .05 criteria.
>
> After rotation, the first component accounted for 53.54% and the second for 33.97%. Component number 1 included items with both negative and positive loadings. Positive loadings include the variables of male life expectancy, female life expectancy, and the number of

doctors. Negative loadings include fertility rate, birth rate, and infant mortality. Items with the highest loadings were fertility rate and birth rate. Component number 1 was named *Healthy Lifespan.* Component number 2 included the number of radios, number of hospital beds, gross domestic product, and the number of phones, respectively. This component was labeled *Economic Stature.* (See Table 1.)

Table 1 Component Loadings

	Loading
Component 1: Healthy Lifespan	
Fertility rate	–.879
Birth rate per 1,000 people	–.858
Male life expectancy	.846
Infant mortality	–.839
Female life expectancy	.829
Number of doctors per 10,000	.763
Component 2: Economic Stature	
Number of radios per 100 people	.879
Number of hospital beds per 10,000 people	.719
Gross domestic product	.688
Number of phones per 100 people	.678

SECTION 9.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of the process of factor analysis. This process includes the development of research questions, data screening, test methods, interpretation of output, and presentation of results. The example utilizes the data set *schools.sav* from the SPSS Web site.

Problem

We are interested in determining what, if any, underlying structures exist for measures on the following twelve variables: % graduating in 1993 (*grad93*); % graduating in 1994 (*grad94*); average ACT score in 1993 (*act93*); average ACT score in 1994 (*act94*); 10^{th} grade average score in 1993 for math (*math93*), reading (*read93*), science (*scienc93*); % meeting or exceeding state standards in 1994 for math (*math94me*), reading (*read94me*), science (*sci94me*); and % limited English proficiency in 1993 (*lep93*) and 1994 (*lep94*).

Method, Output, and Interpretation

Since factor analysis requires a great deal of interpretation throughout the process of analysis, we have combined the discussion of methods, output, and interpretation in this section.

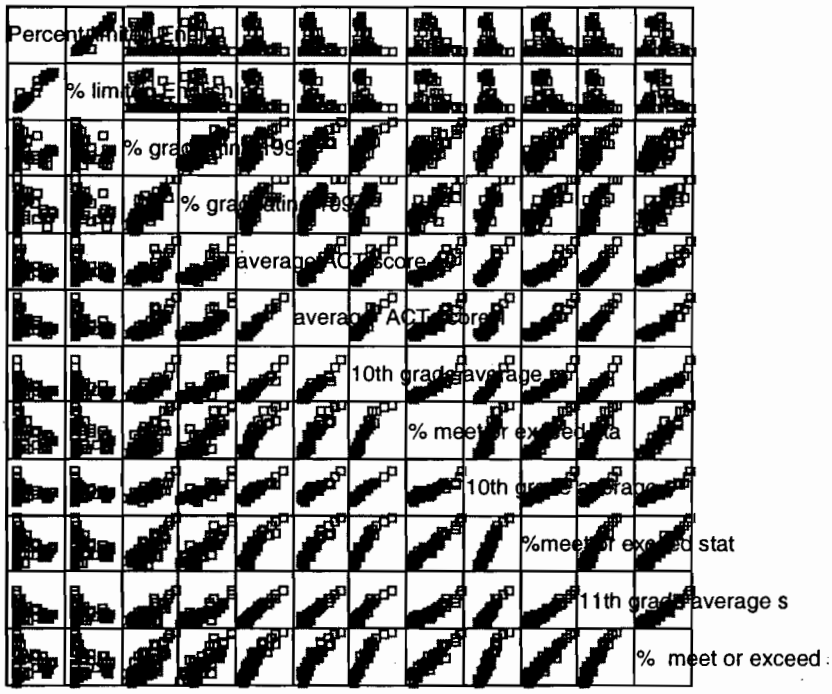
Figure 9.12 Outliers for Mahalanobis Distance (Example number 2).

Extreme Values

			Case Number	Value
MAH_1	Highest	1	37	40.30603
		2	64	38.07910
		3	46	23.28122
		4	6	23.09387
		5	39	22.68936
	Lowest	1	16	2.35371
		2	27	3.40542
		3	19	3.47370
		4	61	3.95236
		5	22	4.13483

Outliers exceeding the $\chi 2(12)=32.909$ at $\alpha=.001$.

Figure 9.13 Scatterplot Matrix (Example number 2).



Data were evaluated to screen for outliers and assess normality and linearity. Using Mahalanobis distance, two outliers (cases number 37 and number 64) were found and eliminated (see Figure 9.12). A scatterplot matrix reveals fairly normal distributions and linear relationship among variables (see Figure 9.13). A factor analysis was then conducted using **`Data Reduction`**, which utilized the eigenvalue criteria and varimax rotation. Applying the four methods of interpretation, we first examine the eigenvalues in the table of total variance (see Figure 9.14). Two components were retained since they have eigenvalues greater than 1. In this example, the application of the eigenvalue criteria seems appropriate since the number of variables is less than 30 and all communalities are greater than .7 (see Figure 9.15). Evaluation of variance is done by referring back to Figure 9.14. After rotation, the first

component accounts for 74.73% of the total variance in the original variables, while the second component accounts for 17.01%. The scree plot was then assessed and indicates that the eigenvalues after three components levels off (see Figure 9.16). Evaluation of residuals indicate that only three residuals are greater than .05 (see Figure 9.17). Although the scree plot suggests that the inclusion of the third component may improve the model, the residuals reveal that any model improvement would be minimal. Consequently, two components were retained.

Figure 9.14 Table of Total Variance for Two-Component Solution (Example number 2).

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.969	74.745	74.745	8.969	74.745	74.745	8.968	74.733	74.733
2	2.040	17.000	91.746	2.040	17.000	91.746	2.042	17.013	91.746
3	.407	3.390	95.136						
4	.205	1.705	96.841						
5	.136	1.132	97.973						
6	7.253E-02	.604	98.577						
7	5.635E-02	.470	99.047						
8	4.654E-02	.388	99.435						
9	2.780E-02	.232	99.667						
10	2.202E-02	.183	99.850						
11	1.197E-02	9.977E-02	99.950						
12	6.027E-03	5.022E-02	100.000						

Extraction Method: Principal Component Analysis.

Figure 9.15 Communalities (Example number 2).

Communalities

	Initial	Extraction
GRAD93	1.000	.701
GRAD94	1.000	.800
ACT94	1.000	.905
ACT93	1.000	.936
MATH93	1.000	.964
MATH94ME	1.000	.944
READ93	1.000	.932
READ94ME	1.000	.950
SCIENC93	1.000	.948
SCI94ME	1.000	.945
LEP93	1.000	.993
LEP94	1.000	.991

Extraction Method: Principal Component Analysis.

Eigenvalue criteria is acceptable since all communalities are greater than .7 and less than 30 factors are analyzed.

Figure 9.16 Scree Plot (Example number 2).

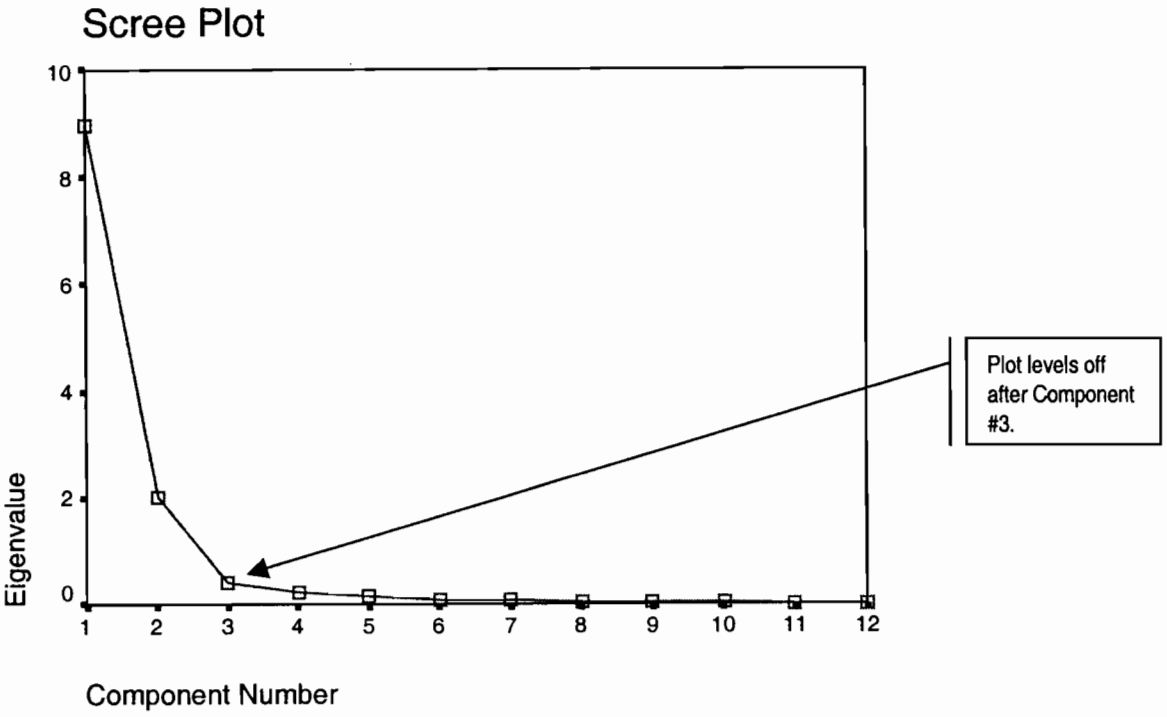


Figure 9.17 Reproduced Correlations and Residuals (Example number 2).

Reproduced Correlations

		GRAD93	GRAD94	ACT94	ACT93	MATH93	MATH94ME	READ93	READ94ME	SCIENC93	SCI94ME	LEP93	LEP94
Reproduced Correlation	GRAD93	.701[b]	.742	.783	.804	.806	.780	.796	.792	.794	.778	-.182	-.172
	GRAD94	.742	.800[b]	.850	.865	.876	.859	.863	.867	.867	.858	-7.54E-02	-6.45E-02
	ACT94	.783	.850	.905[b]	.919	.934	.919	.918	.926	.925	.919	-3.18E-02	-2.02E-02
	ACT93	.804	.865	.919	.936[b]	.947	.927	.933	.936	.937	.926	-9.69E-02	-8.51E-02
	MATH93	.806	.876	.934	.947	.964[b]	.950	.947	.956	.955	.950	-2.15E-02	-9.60E-03
	MATH94ME	.780	.859	.919	.927	.950	.944[b]	.932	.946	.944	.944	6.759E-02	7.930E-02
	READ93	.796	.863	.918	.933	.947	.932	.932[b]	.939	.939	.931	-4.20E-02	-3.03E-02
	READ94ME	.792	.867	.926	.936	.956	.946	.939	.950[b]	.949	.946	2.449E-02	3.630E-02
	SCIENC93	.794	.867	.925	.937	.955	.944	.939	.949	.948[b]	.944	6.128E-03	1.794E-02
	SCI94ME	.778	.858	.919	.926	.950	.944	.931	.946	.944	.945[b]	7.579E-02	3.750E-02
	LEP93	-.182	-7.54E-02	-3.18E-02	-9.69E-02	-2.15E-02	6.759E-02	-4.20E-02	2.449E-02	6.128E-03	7.579E-02	.993[b]	.992
	LEP94	-.172	-6.45E-02	-2.02E-02	-8.51E-02	-9.60E-03	7.930E-02	-3.03E-02	3.630E-02	1.794E-02	3.750E-02	.992	.991[b]
Residual[a]	GRAD93		6.407E-02	-5.45E-02	-2.35E-02	-3.25E-02	-4.69E-02	-1.52E-02	-2.87E-02	-6.046E-02	-5.00E-02	2.880E-02	3.362E-02
	GRAD94	6.407E-02		-6.92E-02	-3.97E-02	-2.27E-02	-2.88E-03	-2.41E-02	-9.59E-03	-4.452E-02	-2.70E-02	3.885E-03	1.825E-02
	ACT94	-5.45E-02	-6.92E-02		4.827E-02	-1.02E-02	-1.00E-02	-2.34E-02	-1.58E-02	3.731E-02	-9.30E-03	2.403E-03	-9.56E-03
	ACT93	-2.35E-02	-3.97E-02	4.827E-02		-3.47E-03	-1.09E-02	-2.18E-02	-2.20E-02	1.505E-02	-1.11E-02	3.867E-03	-2.40E-03
	MATH93	-3.25E-02	-2.27E-02	-1.02E-02	-3.47E-03		1.139E-02	2.071E-02	-3.84E-03	-7.250E-03	4.158E-03	-5.01E-03	-2.63E-03
	MATH94ME	-4.69E-02	-2.88E-03	-1.00E-02	-1.09E-02	1.139E-02		-1.83E-02	1.268E-02	-6.073E-03	7.046E-03	-6.21E-03	-9.15E-03
	READ93	-1.52E-02	-2.41E-02	-2.34E-02	-2.18E-02	2.071E-02	-1.83E-02		6.196E-03	-5.925E-03	3.737E-03	-5.08E-03	1.387E-03
	READ94ME	-2.87E-02	-9.59E-03	-1.58E-02	-2.20E-02	-3.84E-03	1.268E-02	6.196E-03		3.071E-04	5.325E-03	-2.94E-03	-8.03E-03
	SCIENC93	-6.05E-02	-4.45E-02	3.731E-02	1.505E-02	-7.25E-03	-6.07E-03	-5.92E-03	3.071E-04		8.884E-03	-4.41E-03	-8.25E-03
	SCI94ME	-5.00E-02	-2.70E-02	-9.30E-03	-1.11E-02	4.158E-03	7.046E-03	3.737E-03	5.325E-03	8.884E-03		-1.03E-02	-6.67E-03
	LEP93	2.880E-02	3.885E-03	2.403E-03	3.867E-03	-5.01E-03	-6.21E-03	-5.08E-03	-2.94E-03	-4.412E-03	-1.03E-02		-1.03E-04
	LEP94	3.362E-02	1.825E-02	-9.56E-03	-2.40E-03	-2.63E-03	-9.15E-03	1.387E-03	-8.03E-03	-8.250E-03	-6.67E-03	-1.03E-04	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 5 (7.0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities

The next step was to interpret each component. Figure 9.18 presents the factor loadings for the rotated components. Component number 1 consisted of ten of the twelve variables: *scienc93, read94me, math 93, sci94me, read93, math94me, act94, grad94,* and *grad93*. These variables had positive loadings and addressed *Academic Achievement*. The second component included the remaining two variables of percent of limited English proficiency in 1994 (*lep94*) and 1993 (*lep93*). Both variables had positive loadings. Component number 2 was named *Limited English Proficiency*.

Figure 9.18 Factor Loadings for Rotated Components (Example number 2).

Rotated Component Matrix[a]

	Component	
	1	2
MATH93	.982	-4.77E-03
READ94ME	.974	4.129E-02
SCIENC93	.973	2.284E-02
MATH94ME	.968	8.444E-02
SCI94ME	.968	9.268E-02
READ93	.965	-2.56E-02
ACT93	.964	-8.07E-02
ACT94	.951	-1.56E-02
GRAD94	.892	-6.04E-02
GRAD93	.820	-.169
LEP93	-1.71E-02	.996
LEP94	-4.94E-03	.996

Loadings for Component #1.

Loadings for Component #2.

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Presentation of Results

The following summary applies the output from Figures 9.12 – 9.18.

Factor analysis was conducted to determine what, if any, underlying structures exist for measures on the following twelve variables: % graduating in 1993 (*grad93*); % graduating in 1994 (*grad94*); average ACT score in 1993 (*act93*); average ACT score in 1994 (*act94*); 10th grade average score in 1993 for math (*math93*), reading (*read93*), science (*scienc93*); % meeting or exceeding state standards in 1994 for math (*math94me*), reading (*read94me*), science (*sci94me*); and % limited English proficiency in 1993 (*lep93*) and 1994 (*lep94*). Prior to analysis, two outliers were eliminated. Principal components analysis was conducted utilizing a varimax rotation. The analysis produced a two-component solution, which was evaluated with the following criteria: eigenvalue, variance, scree plot, and residuals. Criteria indicated a two-component solution was appropriate.

After rotation, the first component accounted for 74.73% of the total variance in the original variables, while the second component accounted for 17.01%. Table 1 presents the loadings for each component. Component number 1 consisted of ten of the twelve variables: *math93, read94me, scienc93, math94me, sci94me, read93, act93, act94, grad94,* and *grad93*. These variables had positive loadings and addressed *Academic Achievement*. The second component included the remaining two variables of % of limited English proficiency in 1994 (*lep94*)

and 1993 (*lep93*). Both variables had positive loadings. Component number 2 was labeled *Limited English Proficiency*. (See Table 1.)

Table 1 Component Loadings

	Loading
Component 1: Academic Achievement	
10th grade average math score (1993)	.982
% meeting/exceeding reading standards (1994)	.974
10th grade average science score	.973
% meeting/exceeding math standards (1994)	.968
% meeting/exceeding science standards (1994)	.968
10th grade average reading score (1993)	.965
Average ACT score (1993)	.964
Average ACT score (1994)	.951
% graduating (1994)	.892
% graduating (1993)	.820
Component 2: Limited English Proficiency	
% Limited English Proficiency (1994)	.996
% Limited English Proficiency (1993)	.996

SECTION 9.5 SPSS "HOW TO"

This section presents the steps for conducting factor analysis using **Data Reduction**. To open the Factor Analysis Dialogue Box, select the following:

Analyze
Data Reduction
Factor

Factor Analysis Dialogue Box (Figure 9.19)

Select each variable to be included in the analysis and move to the Variables box. Then click **Descriptives**.

Factor Analysis: Descriptive Dialogue Box (Figure 9.20)

Several descriptive statistics are provided in this Dialogue Box. Under Statistics, two options are provided: **Univariate Descriptives** and **Initial Solution**. **Univariate Descriptives** presents the means and standard deviations for each variable analyzed. **Initial Solution** is selected by default and will present the initial communalities, eigenvalues, and percent accounted for by each factor. For our example, we selected only **Initial Solution**. Under Correlation Matrix, the following options are frequently used:

Coefficients—Presents original correlation coefficients of variables.

Significance Levels—Indicates *p* values of each correlation coefficient.
KMO and Bartlett's Test of Sphericity—Tests both multivariate normality and sampling adequacy.
Reproduced—Presents reproduced correlation coefficients and residuals (difference between original and reproduced coefficients).

Our example only utilized the **Reproduced** option. After selecting the descriptive options, click **Continue**. Click **Extraction**.

Figure 9.19 Factor Analysis Dialogue Box.

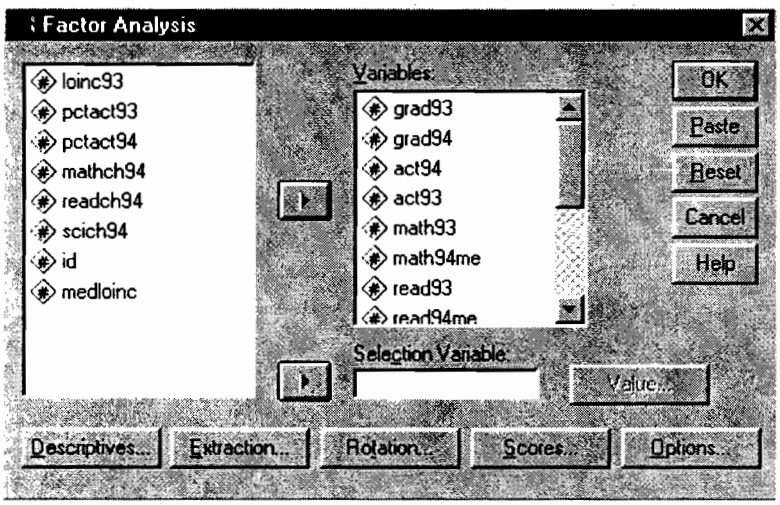
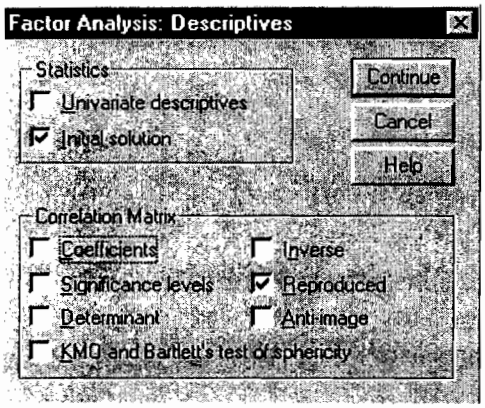


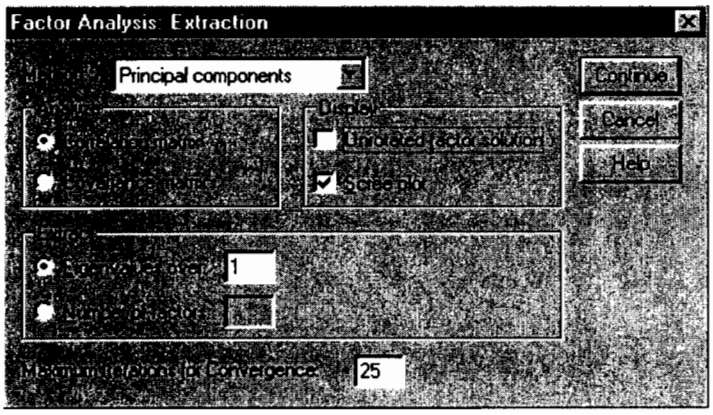
Figure 9.20 Factor Analysis: Descriptive Dialogue Box.



Factor Analysis: Extraction Dialogue Box (Figure 9.21)

For method, select **Principal Components**. Under Analyze, check **Correlation Matrix**. Under Display, select **Scree Plot**. Under Extract, the eigenvalue criteria of 1 is the default. Utilize the default, unless a previous analysis indicates that more components should be retained, at which time you would indicate the **Number of Factors**. Next, click **Continue**. Click **Rotation**.

Figure 9.21 Factor Analysis: Extraction Dialogue Box.



Factor Analysis: Rotation Dialogue Box (Figure 9.22)

Select the rotation method you prefer. Rotation methods available are described as follows:

Varimax—Orthogonal method that minimizes factor complexity by maximizing variance for each factor.

Direct Oblimin—Oblique method that simplifies factors by minimizing cross products of loadings.

Quartimax—Orthogonal method that minimizes factor complexity by maximizing variance loadings on each variable.

Equamax— Orthogonal method that combines both Varimax and Quartimax procedures.

Promax— Oblique method that rotates orthogonal factors to oblique positions.

Our example utilized **Varimax**. If a rotation method is indicated, check **Rotated Solution** under Display. Click **Continue**, then **Scores**.

Factor Analysis: Factor Scores Dialogue Box (Figure 9.23)

If you will be using the generated factors in future analyses, you will need to save factor scores. To do so, check **Save as Variables** and utilize the default method of **Regression**. Click **Continue**, then **Options**.

Factor Analysis: Options Dialogue Box (Figure 9.24)

Under Coefficient Display Format, check **Sorted by Size**. This will help in reading the Component Matrix. Click **Continue**. Click **OK**.

Figure 9.22 Factor Analysis: Rotation Dialogue Box.

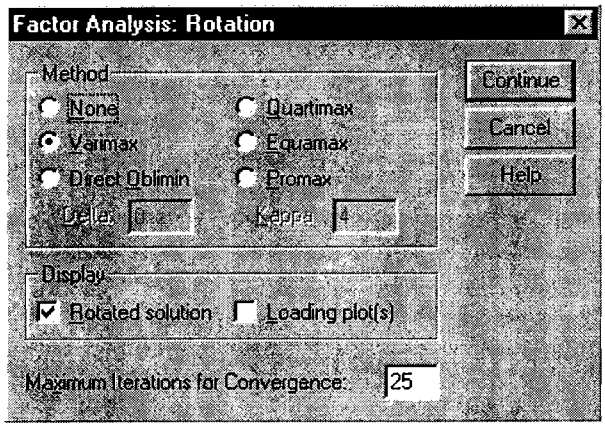


Figure 9.23 Factor Analysis: Factor Scores Dialogue Box.

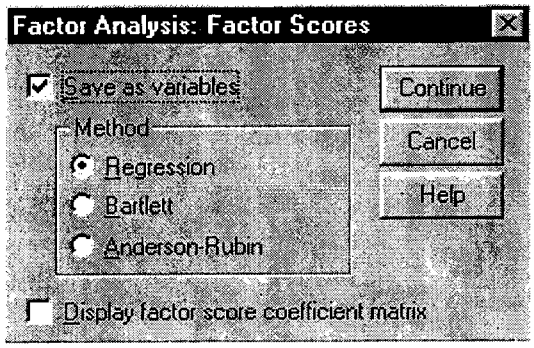
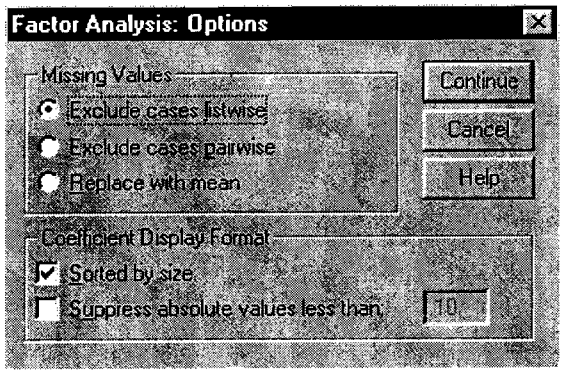


Figure 9.24 Factor Analysis: Options Dialogue Box.



Summary

Factor analysis is a technique used to identify factors that explain common variance among variables. This statistical method is often used to reduce data by grouping variables that measure a common construct. Principal components analysis is one of the most commonly used methods of extraction since this method will evaluate all sources of variability for each variable. Factors or components can also be rotated to make the components more interpretable. Orthogonal rotation methods (i.e., varimax, quartimax, equamax) result in uncorrelated factors and are the most frequently used methods. Oblique rotation methods (i.e., oblimin, promax, orthoblique) result in factors being correlated with each other.

Since principal components analysis is typically exploratory, the researcher must determine the appropriate number of components to retain. Four criteria are used in this decision-making process: 1) Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criteria is fairly reliable when: the number of variables is < 30 and communalities are $> .70$, or the number of individuals is > 250 and the mean communality is $\geq .60$; 2) Variance—Retain components that account for at least 70% total variability; 3) Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criterion is fairly reliable when the number of individuals is > 250 and communalities are $> .30$; 4) Residuals—Retain the components generated by the model if only a few residuals exceed .05. If several reproduced correlations differ, you may want to include more components.

Once the appropriate number of components to retain has been determined, the researcher must then interpret/name the components by evaluating the type of variables included in each factor, the strength of factor loadings, and the direction of factor loadings. Figure 9.25 provides a checklist for conducting factor analysis.

Figure 9.25 Checklist for Conducting Factor Analysis.

I. Screen Data

a. Missing Data?

b. Multivariate Outliers?

- ❑ Run preliminary Regression to calculate Mahalanobis' Distance.
  1. **Analyze...Regression...Linear.**
  2. Identify a variable that serves as a case number and move to Dependent Variable box.
  3. Identify all appropriate quantitative variables and move to Independent(s) box.
  4. **Save.**
  5. Check **Mahalanobis'**.
  6. **Continue.**
  7. **OK.**
  8. Determine chi square χ^2 critical value at $p<.001$.
- ❑ Conduct **Explore** to test outliers for Mahalanobis chi square χ^2.
  1. **Analyze...Descriptive Statistics...Explore**
  2. Move *mah_1* to Dependent Variable box.
  3. Leave Factor box empty.
  4. **Statistics.**
  5. Check **Outliers.**
  6. **Continue.**
  7. **OK.**
- ❑ Delete outliers for subjects when χ^2 exceeds critical χ^2 at $p<.001$.

c. Linearity and Normality?

- ❑ Create Scatterplot Matrix of all model variables.
  - ❑ Scatterplot shapes are not close to elliptical shapes→reevaluate univariate normality and consider transformations.

II. Conduct Factor Analysis

a. Run Factor Analysis using **Data Reduction**.

1. **Analyze...Data Reduction...Factor.**
2. Move each studied variable to the Variables box.
3. **Descriptives.**
4. Check: **Initial Solution** and **Reproduced**.
5. **Continue.**
6. **Extraction.**
7. Check **Correlation Matrix, Unrotated Factor Solution, Scree Plot,** and **Eigenvalue.**
8. **Continue.**
9. **Rotation.**
10. Check **Varimax** and **Rotated Solution.**
11. **Continue.**
12. **Scores.**
13. Check **Save as Variables** and **Regression.**
14. **Continue.**
15. **Options.**
16. Check **Sorted by Size.**
17. **Continue.**
18. **OK.**

Figure 9.25 Checklist for Conducting Factor Analysis. (*Continued*)

b. Determine appropriate number of components to retain.
   1. Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criteria is fairly reliable when: the number of variables is < 30 and communalities are > .70, or the number of individuals is > 250 and the mean communality is ≥ .60;
   2. Variance—Retain components that account for at least 70% total variability;
   3. Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criteria is fairly reliable when the number of individuals is > 250 and communalities are > .30;
   4. Residuals—Retain the components generated by the model if only a few residuals exceed .05. If several reproduced correlations differ, you may want to include more components.

c. Conduct factor analysis again if more components should be retained.

d. Interpret components.
   1. Evaluate the type of variables loaded into each component;
   2. Note the strength and direction of loadings;
   3. Label component accordingly.

III. Summarize Results

a. Describe any data elimination or transformation.

b. Describe the initial model.

c. Describe the criteria used to determine the number of components to retain.

d. Summarize the components generated by narrating: the variables loaded into each component, the strength and direction of loadings, the component labels, and the percent of variance.

e. Create a table that summarizes each component (report component loadings).

f. Draw conclusions.

Exercises for Chapter 9

The following exercises seek to determine what underlying structure exists among the following variables: highest degree earned (*degree*), hours worked per week (*hrs1*), job satisfaction (*satjob*), years of education (*educ*), hours per day watching TV (*tvhours*), general happiness (*happy*), degree to which life is exciting *(life*), and degree to which the lot of the average person is getting worse (*anomia5*).

1. The following output was generated for the initial analysis. Varimax rotation was utilized.