A Step-by-Step Approach to Using SAS® for

Factor Analysis and Structural Equation Modeling

Second Edition

Norm O'Rourke and Larry Hatcher

Contents

Chapter 1: Principal Component Analysis

Introduction: The Basics of Principal Component Analysis

Principal component analysis is used when you have obtained measures for a number of observed variables and wish to arrive at a smaller number of variables (called "principal components") that will account for, or capture, most of the variance in the observed variables. The principal components may then be used as predictors or criterion variables in subsequent analyses.

A Variable Reduction Procedure

Principal component analysis is a variable reduction procedure. It is useful when you have obtained data for a number of variables (possibly a large number of variables) and believe that there is redundancy among those variables. In this case, redundancy means that some of the variables are correlated with each other, often because they are measuring the same construct. Because of this redundancy, you believe that it should be possible to reduce the observed variables into a smaller number of principal components that will account for most of the variance in the observed variables.

Because it is a variable reduction procedure, principal component analysis is similar in many respects to exploratory factor analysis. In fact, the steps followed when conducting a principal component analysis are virtually identical to those followed when conducting an exploratory factor analysis. There are significant conceptual differences between the two, however, so it is important that you do not mistakenly claim that you are performing factor analysis when you are actually performing principal component analysis. The differences between these two procedures are described in greater detail in a later subsection titled "Principal Component Analysis Is *Not* Factor Analysis."

An Illustration of Variable Redundancy

We now present a fictitious example to illustrate the concept of variable redundancy. Imagine that you have developed a seven-item measure to gauge job satisfaction. The fictitious instrument is reproduced here:

Please respond to the following statements by placing your response to the left of each statement. In making your ratings, use a number from 1 to 7 in which 1 = "Strongly Disagree" and 7 = "Strongly Agree."

_____ 1. My supervisor(s) treats me with consideration.
_____ 2. My supervisor(s) consults me concerning important decisions that affect my work.
_____ 3. My supervisor(s) gives me recognition when I do a good job.
_____ 4. My supervisor(s) gives me the support I need to do my job well.
_____ 5. My pay is fair.
_____ 6. My pay is appropriate, given the amount of responsibility that comes with my job.
_____ 7. My pay is comparable to that of other employees whose jobs are similar to mine.

Perhaps you began your investigation with the intention of administering this questionnaire to 200 employees using their responses to the seven items as seven separate variables in subsequent analyses.

There are a number of problems with conducting the study in this manner, however. One of the more important problems involves the concept of redundancy as previously mentioned. Examine the content of the seven items in the questionnaire. Notice that items 1 to 4 each deal with employees' satisfaction with their supervisors. In this way, items 1 to 4 are somewhat redundant or overlapping in terms of what they are measuring. Similarly, notice also that items 5 to 7 each seem to deal with the same topic: employees' satisfaction with their pay.

Empirical findings may further support the likelihood of item redundancy. Assume that you administer the questionnaire to 200 employees and compute all possible correlations between responses to the seven items. Fictitious correlation coefficients are presented in Table 1.1:

Table 1.1: Correlations among Seven Job Satisfaction Items

Variable	1	2	3	4	5	6	7
1	1.00						
2	.75	1.00					
3	.83	.82	1.00				
4	.68	.92	.88	1.00			
5	.03	.01	.04	.01	1.00		
6	.05	.02	.05	.07	.89	1.00	
7	.02	.06	.00	.03	.92	.76	1.00

NOTE: *N* = 200.

When correlations among several variables are computed, they are typically summarized in the form of a **correlation matrix** such as the one presented in Table 1.1; this provides an opportunity to review how a correlation matrix is interpreted. (See Appendix A.5 for more information about correlation coefficients.)

The rows and columns of Table 1.1 correspond to the seven variables included in the analysis. Row 1 (and column 1) represents variable 1, row 2 (and column 2) represents variable 2, and so forth. Where a given row and column intersect, you will find the correlation coefficient between the two corresponding variables. For example, where the row for variable 2 intersects with the column for variable 1, you find a coefficient of .75; this means that the correlation between variables 1 and 2 is .75.

The correlation coefficients presented in Table 1.1 show that the seven items seem to *hang together* in two distinct groups. First, notice that items 1 to 4 show relatively strong correlations with each another. This could be because items 1 to 4 are measuring the same construct. In the same way, items 5 to 7 correlate strongly with one another, a possible indication that they also measure a single construct. Even more interesting, notice that items 1 to 4 are very weakly correlated with items 5 to 7. This is what you would expect to see if items 1 to 4 and items 5 to 7 were measuring two different constructs.

Given this apparent redundancy, it is likely that the seven questionnaire items are not really measuring seven different constructs. More likely, items 1 to 4 are measuring a single construct that could reasonably be labeled "satisfaction with supervision," whereas items 5 to 7 are measuring a different construct that could be labeled "satisfaction with pay."

If responses to the seven items actually display the redundancy suggested by the pattern of correlations in Table 1.1, it would be advantageous to reduce the number of variables in this dataset, so that (in a sense) items 1 to 4 are collapsed into a single new variable that reflects employees' satisfaction with supervision and items 5 to 7 are collapsed into a single new variable that reflects satisfaction with pay. You could then use these two new variables (rather than the seven original variables) as predictor variables in multiple regression, for instance, or another type of analysis.

In essence, this is what is accomplished by principal component analysis: it allows you to reduce a set of observed variables into a smaller set of variables called principal components. The resulting principal components may then be used in subsequent analyses.

What Is a Principal Component?

How Principal Components Are Computed

A **principal component** can be defined as a linear combination of optimally weighted observed variables. In order to understand the meaning of this definition, it is necessary to first describe how participants' scores on a principal component are computed.

In the course of performing a principal component analysis, it is possible to calculate a score for each participant for a given principal component. In the preceding study, for example, each participant would have scores on two components: one score on the "satisfaction with supervision" component; and one score on the "satisfaction with pay" component. Participants' actual scores on the seven questionnaire items would be optimally weighted and then summed to compute their scores for a given component.

Below is the general form of the formula to compute scores on the first component extracted (created) in a principal component analysis:

$$C_1 = b_{11}(X_1) + b_{12}(X_2) + \ldots b_{1p}(X_p)$$

where

C_1 = the participant's score on principal component 1 (the first component extracted)

b_{1p} = the coefficient (or weight) for observed variable p, as used in creating principal component 1

X_p = the participant's score on observed variable p

For example, assume that component 1 in the present study was "satisfaction with supervision." You could determine each participant's score on principal component 1 by using the following fictitious formula:

$$C_1 =.44\ (X_1) + .40\ (X_2) + .47\ (X_3) + .32\ (X_4)$$

$$+ .02\ (X_5) + .01\ (X_6) + .03\ (X_7)$$

In this case, the observed variables (the "X" variables) are participant responses to the seven job satisfaction questions: X_1 represents question 1; X_2 represents question 2; and so forth. Notice that different coefficients or weights were assigned to each of the questions when computing scores on component 1: questions 1 to 4 were assigned relatively large weights that range from .32 to .47, whereas questions 5 to 7 were assigned very small weights ranging from .01 to .03. This makes sense, because component 1 is the satisfaction with supervision component and satisfaction with supervision was measured by questions 1 to 4. It is therefore appropriate that items 1 to 4 would be given a good deal of weight in computing participant scores on this component, while items 5 to 7 would be given comparatively little weight.

Because component 2 measures a different construct, a different equation with different weights would be used to compute scores for this component (i.e., "satisfaction with pay"). Below is a fictitious illustration of this formula:

$$C_2 =.01\ (X_1) + .04\ (X_2) + .02\ (X_3) + .02\ (X_4)$$

$$+ .48\ (X_5) + .31\ (X_6) + .39\ (X_7)$$

The preceding example shows that, when computing scores for the second component, considerable weight would be given to items 5 to 7, whereas comparatively little would be given to items 1 to 4. As a result, component 2 should account for much of the variability in the three satisfaction with pay items (i.e., it should be strongly correlated with those three items).

But how are these weights for the preceding equations determined? PROC FACTOR in SAS generates these weights by using a special type of equation called an **eigenequation**. The weights produced by these eigenequations are optimal weights in the sense that, for a given set of data, no other set of weights could produce a set of components that are more effective in accounting for variance among observed variables. These weights are created to satisfy what is known as *the principle of least squares*. Later in this chapter we will show how PROC FACTOR can be used to extract (create) principal components.

It is now possible to understand the definition provided at the beginning of this section more fully. A principal component was defined as a linear combination of optimally weighted observed variables. The words "linear combination" refer to the fact that scores on a component are created by adding together scores for the observed variables being analyzed. "Optimally weighted" refers to the fact that the observed variables are weighted in such a way that the resulting components account for a maximal amount of observed variance in the dataset.

Number of Components Extracted

The preceding section may have created the impression that, if a principal component analysis were performed on data from our fictitious seven-item job satisfaction questionnaire, only two components would be created. Such an impression would not be entirely correct.

In reality, the number of components extracted in a principal component analysis is equal to the number of observed variables being analyzed. This means that an analysis of responses to the seven-item questionnaire would actually result in seven components, not two.

In most instances, however, only the first few components account for meaningful amounts of variance; only these first few components are retained, interpreted, and used in subsequent analyses. For example, in your analysis of the seven-item job satisfaction questionnaire, it is likely that only the first two components would account for, or capture, meaningful amounts of variance. Therefore, only these would be retained for interpretation. You could assume that the remaining five components capture only trivial amounts of variance. These latter components would therefore not be retained, interpreted, or further analyzed.

Characteristics of Principal Components

The first component extracted in a principal component analysis accounts for a maximal amount of total variance among the observed variables. Under typical conditions, this means that the first component will be correlated with at least some (often many) of the observed variables.

The second component extracted will have two important characteristics. First, this component will account for a maximal amount of variance in the dataset that was not accounted for or captured by the first component. Under typical conditions, this again means that the second component will be correlated with some of the observed variables that did not display strong correlations with component 1.

The second characteristic of the second component is that it will be uncorrelated with the first component. Literally, if you were to compute the correlation between components 1 and 2, that coefficient would be zero. (For the exception, see the following section regarding oblique solutions.)

The remaining components that are extracted exhibit the same two characteristics: each accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components; and each is uncorrelated with all of the preceding components. Principal component analysis proceeds in this manner with each new component accounting for progressively smaller amounts of variance. This is why only the first few components are retained and interpreted. When the analysis is complete, the resulting components will exhibit varying degrees of correlation with the observed variables, but will be completely uncorrelated with each another.

What is meant by "total variance" in the dataset? To understand the meaning of "total variance" as it is used in a principal component analysis, remember that the observed variables are standardized in the course of the analysis. This means that each variable is transformed so that it has a mean of zero and a standard deviation of one (and hence a variance of one). The "total variance" in the dataset is simply the sum of variances for these observed variables. Because they have been standardized to have a standard deviation of one, each observed variable contributes one unit of variance to the total variance in the dataset. Because of this, total variance in principal component analysis will always be equal to the number of observed variables analyzed. For example, if seven variables are being analyzed, the total variance will equal seven. The components that are extracted in the analysis will partition this variance. Perhaps the first component will account for 3.2 units of total variance; perhaps the second component will account for 2.1 units. The analysis continues in this way until all variance in the dataset has been accounted for.

Orthogonal versus Oblique Solutions

This chapter will discuss only principal component analyses that result in orthogonal solutions. An **orthogonal solution** is one in which the components are uncorrelated ("orthogonal" means uncorrelated).

It is possible to perform a principal component analysis that results in correlated components. Such a solution is referred to as an **oblique solution**. In some situations, oblique solutions are preferred to orthogonal solutions because they produce cleaner, more easily interpreted results.

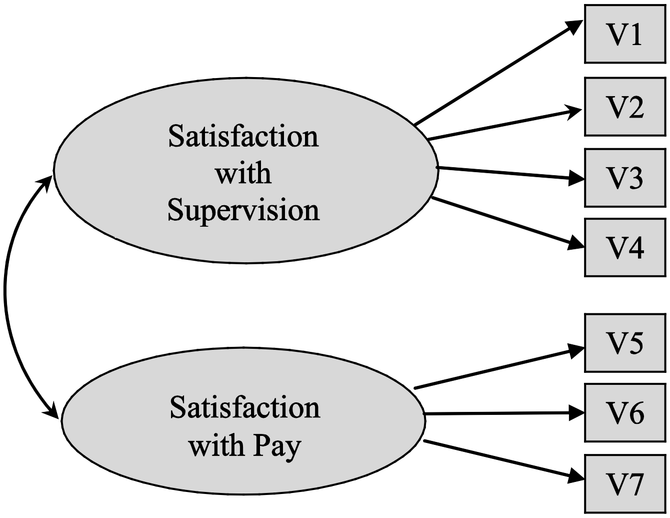
However, oblique solutions are often complicated to interpret. For this reason, this chapter will focus only on the interpretation of orthogonal solutions. The concepts discussed will provide a good foundation for the somewhat more complex concepts discussed later in this text.

Principal Component Analysis Is *Not* Factor Analysis

Principal component analysis is commonly confused with factor analysis. This is understandable because there are many important similarities between the two. Both are methods that can be used to identify groups of observed variables that tend to hang together empirically. Both procedures can also be performed with PROC FACTOR, and they generally provide similar results.

Nonetheless, there are some important conceptual differences between principal component analysis and factor analysis that should be understood at the outset. Perhaps the most important difference deals with the **assumption of an underlying causal structure**. Factor analysis assumes that covariation among the observed variables is due to the presence of one or more latent variables that exert directional influence on these observed variables. An example of such a structure is presented in Figure 1.1.

Figure 1.1: Example of the Underlying Causal Structure That Is Assumed in Factor Analysis



The ovals in Figure 1.1 represent the latent (unmeasured) factors of "satisfaction with supervision" and "satisfaction with pay." These factors are latent in the sense that it is assumed employees hold these beliefs but that these beliefs cannot be measured directly; however, they do influence employees' responses to the items that constitute the job satisfaction questionnaire described earlier. (These seven items are represented as the squares labeled V1 to V7 in the figure.) It can be seen that the "supervision" factor exerts influence on items V1 to V4 (the supervision questions), whereas the "pay" factor exerts influence on items V5 to V7 (the pay items).

Researchers use factor analysis when they believe that one or more unobserved or latent factors exert directional influence on participants' responses to observed variables. Exploratory factor analysis helps the researcher identify the number and nature of such latent factors. These procedures are described in the next chapter.

In contrast, principal component analysis makes no assumption about underlying causal structures; it is simply a variable reduction procedure that (typically) results in a relatively small number of components accounting for, or capturing, most variance in a set of observed variables (i.e., groupings of observed variables versus latent constructs).

Another important distinction between the two is that principal component analysis assumes no measurement error whereas factor analysis captures both true variance and measurement error. Acknowledgement and measurement of error is particularly germane to social science research because instruments are invariably incomplete measures of underlying constructs. Principal component analysis is sometimes used in instrument construction studies to overestimate precision of measurement (i.e., overestimate the effectiveness of the scale).

In summary, both factor analysis and principal component analysis are important in social science research, but their conceptual foundations are quite distinct.

Example: Analysis of the Prosocial Orientation Inventory

Assume that you have developed an instrument called the Prosocial Orientation Inventory (POI) that assesses the extent to which a person has engaged in helping behaviors over the preceding six months. This fictitious instrument contains six items and is presented here:

Instructions: Below are a number of activities in which people sometimes engage. For each item, please indicate how frequently you have engaged in this activity over the past six months. Provide your response by circling the appropriate number to the left of each item using the response key below:

7 = Very Frequently
6 = Frequently
5 = Somewhat Frequently
4 = Occasionally
3 = Seldom
2 = Almost Never
1 = Never

1 2 3 4 5 6 7 1. I went out of my way to do a favor for a coworker.
1 2 3 4 5 6 7 2. I went out of my way to do a favor for a relative.
1 2 3 4 5 6 7 3. I went out of my way to do a favor for a friend.
1 2 3 4 5 6 7 4. I gave money to a religious charity.
1 2 3 4 5 6 7 5. I gave money to a charity not affiliated with a religion.
1 2 3 4 5 6 7 6. I gave money to a panhandler.

When this instrument was developed, the intent was to administer it to a sample of participants and use their responses to the six items as separate predictor variables. As previously stated, however, you learned that this is a problematic practice and have decided, instead, to perform a principal component analysis on responses to see if a smaller number of components can successfully account for most variance in the dataset. If this is the case, you will use the resulting components as predictor variables in subsequent analyses.

At this point, it may be instructive to examine the content of the six items that constitute the POI to make an informed guess as to what is likely to result from the principal component analysis. Imagine that when you first constructed the instrument, you assumed that the six items were assessing six different types of prosocial behavior. Inspection of items 1 to 3, however, shows that these three items share something in common: they all deal with “going out of one’s way to do a favor for someone else.” It would not be surprising then to learn that these three items will hang together empirically in the principal component analysis to be performed. In the same way, a review of items 4 to 6 shows that each of these items involves the activity of “giving money to those in need.” Again, it is possible that these three items will also group together in the course of the analysis.

In summary, the nature of the items suggests that it may be possible to account for variance in the POI with just two components: a “helping others” component and a “financial giving” component. At this point, this is only speculation, of course; only a formal analysis can determine the number and nature of components measured by the inventory of items. (Remember that the preceding instrument is fictitious and used for purposes of illustration only and should not be regarded as an example of a good measure of prosocial orientation. Among other problems, this questionnaire obviously deals with very few forms of helping behavior.)

Preparing a Multiple-Item Instrument

The preceding section illustrates an important point about how *not* to prepare a multiple-item scale to measure a construct. Generally speaking, it is poor practice to throw together a questionnaire, administer it to a sample, and then perform a principal component analysis (or factor analysis) to determine what the questionnaire is measuring.

Better results are much more likely when you make *a priori* decisions about what you want the questionnaire to measure, and then take steps to ensure that it does. For example, you would have been more likely to obtain optimal results if you:

- began with a thorough review of theory and research on prosocial behavior
- used that review to determine how many types of prosocial behavior may exist
- wrote multiple questionnaire items to assess each type of prosocial behavior

Using this approach, you could have made statements such as “There are three types of prosocial behavior: acquaintance helping; stranger helping; and financial giving.” You could have then prepared a number of items to assess each of these three types, administered the questionnaire to a large sample, and performed a principal component analysis to see if three components did, in fact, emerge.

Number of Items per Component

When a variable (such as a questionnaire item) is given a weight in computing a principal component, we say that the variable **loads** on that component. For example, if the item “Went out of my way to do a favor for a coworker” is given a lot of weight on the “helping others” component, we say that this item “loads” on that component.

It is highly desirable to have a minimum of three (and preferably more) variables loading on each retained component when the principal component analysis is complete (see Clark and Watson 1995). Because some items may be dropped during the course of the analysis (for reasons to be discussed later), it is generally good practice to write at least five items for each construct that you wish to measure. This increases your chances that at least three items per component will survive the analysis. Note that we have violated this recommendation by writing only three items for each of the two *a priori* components constituting the POI.

Keep in mind that the recommendation of three items per scale should be viewed as an absolute minimum and certainly not as an optimal number. In practice, test and attitude scale developers normally desire that their scales contain many more than just three items to measure a given construct. It is not unusual to see individual scales that include 10, 20, or even more items to assess a single construct (e.g., Chou and O’Rourke 2012; O’Rourke and Cappeliez 2002). Up to a point, the greater the number of scale items, the more reliable it will be. The recommendation of three items per scale should therefore be viewed as a rock-bottom lower bound, appropriate only if practical concerns prevent you from including more items (e.g., total questionnaire length). For more information on scale construction, see DeVellis (2012) and, Saris and Gallhofer (2007).

Minimal Sample Size Requirements

Principal component analysis is a large-sample procedure. To obtain reliable results, the minimal number of participants providing usable data for the analysis should be the larger of 100 participants or 5 times the number of variables being analyzed (Streiner 1994).

To illustrate, assume that you wish to perform an analysis on responses to a 50-item questionnaire. (Remember that when responses to a questionnaire are analyzed, the number of variables is equal to the number of items on that questionnaire.) Five times the number of items on the questionnaire equals 250. Therefore, your final sample should provide usable (complete) data from at least 250 participants. Note, however, that any participant who fails to answer just one item will not provide usable data for the principal component analysis and will therefore be excluded from the final sample. A certain number of participants can always be expected to leave at least one question blank. To ensure that the final sample includes at least 250 usable responses, you would be wise to administer the questionnaire to perhaps 300 to 350 participants (see Little and Rubin 1987). A preferable alternative is to use an imputation procedure that assigns values for skipped items (van Buuren 2012). A number of such procedures are available in SAS but are not covered in this text.

These rules regarding the number of participants per variable again constitute a lower bound, and some have argued that they should be applied only under two optimal conditions for principal component analysis: when many variables are expected to load on each component, and when variable communalities are high. Under less optimal conditions, even larger samples may be required.

What is a communality? A **communality** refers to the percent of variance in an observed variable that is accounted for by the retained components (or factors). A given variable will display a large communality if it loads heavily on at least one of the study's retained components. Although communalities are computed in both procedures, the *concept* of variable communality is more relevant to factor analysis than principal component analysis.

SAS Program and Output

You may perform principal component analysis using the PRINCOMP, CALIS, or FACTOR procedures. This chapter will show how to perform the analysis using PROC FACTOR since this is a somewhat more flexible SAS procedure. (It is also possible to perform an exploratory factor analysis with PROC FACTOR or PROC CALIS.) Because the analysis is to be performed using PROC FACTOR, the output will at times make reference to factors rather than to principal components (e.g., component 1 will be referred to as FACTOR1 in the output). It is important to remember, however, that you are performing principal component analysis, not factor analysis.

This section will provide instructions on writing the SAS program and an overview of the SAS output. A subsequent section will provide a more detailed treatment of the steps followed in the analysis as well as the decisions to be made at each step.

Writing the SAS Program

The DATA Step

To perform a principal component analysis, data may be entered as raw data, a correlation matrix, a covariance matrix, or some other format. (See Appendix A.2 for further description of these data input options.) In this chapter's first example, raw data will be analyzed.

Assume that you administered the POI to 50 participants, and entered their responses according to the following guide:

Line	Column	Variable Name	Explanation
1	1–6	V1–V6	Participants' responses to survey questions 1 through 6. Responses were provided along a 7-point scale.

Here are the statements to enter these responses as raw data. The first three observations and the last three observations are reproduced here; for the entire dataset, see Appendix B.

```
data D1;
     input V1-V6 ;

datalines;
556754
567343
777222
.
.
.
767151
455323
455544
;
run;
```

The dataset in Appendix B includes only 50 cases so that it will be relatively easy to enter the data and replicate the analyses presented here. It should be restated, however, that 50 observations is an unacceptably small sample for principal component analysis. Earlier it was noted that a sample should provide usable data from the larger of either 100 cases or 5 times the number of observed variables. A small sample is being analyzed here for illustrative purposes only.

The PROC FACTOR Statement

The general form for the SAS program to perform a principal component analysis is presented here:

```
proc factor   data=dataset-name
              simple
              method=prin
              priors=one
              mineigen=p
              rotate=varimax
              round
              flag=desired-size-of-"significant"-factor-loadings ;
   var  variables-to-be-analyzed ;
run;
```

Options Used with PROC FACTOR

The PROC FACTOR statement begins the FACTOR procedure and a number of options may be requested in this statement before it ends with a semicolon. Some options that are especially useful in social science research are:

FLAG

causes the output to flag (with an asterisk) factor loadings with absolute values greater than some specified size. For example, if you specify

```
flag=.35
```

an asterisk will appear next to any loading whose absolute value exceeds .35. This option can make it much easier to interpret a factor pattern. Negative values are not allowed in the FLAG option, and the FLAG option can be used in conjunction with the ROUND option.

METHOD=factor-extraction-method

specifies the method to be used in extracting the factors or components. The current program specifies

```
method=prin
```

to request that the principal axis (principal factors) method be used for the initial extraction. This is the appropriate method for a principal component analysis.

MINEIGEN=p

specifies the critical eigenvalue a component must display if that component is to be retained (here, p = the critical eigenvalue). For example, the current program specifies

```
mineigen=1
```

This statement will cause PROC FACTOR to retain and rotate any component whose eigenvalue is 1.00 or larger. Negative values are not allowed.

NFACT=n

allows you to specify the number of components to be retained and rotated where n = the number of components.

OUT=name-of-new-dataset

creates a new dataset that includes all of the variables in the existing dataset, along with factor scores for the components retained in the present analysis. Component 1 is given the variable name FACTOR1, component 2 is given the name FACTOR2, and so forth. It must be used in conjunction with the NFACT option, and the analysis must be based on raw data.

PRIORS=prior-communality-estimates

specifies prior communality estimates. Users should always specify PRIORS=one to perform a principal component analysis.

ROTATE=rotation-method

specifies the rotation method to be used. The preceding program requests a varimax rotation that provides orthogonal (uncorrelated) components. Oblique rotations may also be requested (correlated components).

ROUND

factor loadings and correlation coefficients in the matrices printed by PROC FACTOR are normally carried out to several decimal places. Requesting the ROUND option, however, causes all coefficients to be limited to two decimal places, rounded to the nearest integer, and multiplied by 100 (thus eliminating the decimal point). This generally makes it easier to read the coefficients.

PLOTS=scree

creates a plot that graphically displays the size of the eigenvalues associated with each component. This can be used to perform a scree test to visually determine how many components should be retained.

SIMPLE

requests simple descriptive statistics: the number of usable cases on which the analysis was performed and the means and standard deviations of the observed variables.

The VAR Statement

The variables to be analyzed are listed on the VAR statement with each variable separated by at least one space. Remember that the VAR statement is a *separate* statement and not an option within the FACTOR statement, so don't forget to end the FACTOR statement with a semicolon before beginning the VAR statement.

Example of an Actual Program

The following is an actual program, including the DATA step, that could be used to analyze some fictitious data. Only a few sample lines of data appear here; the entire dataset may be found in Appendix B.

```
data D1;
     input  #1    @1   (V1-V6)    (1.)

datalines;
556754
567343
777222
.
.
.
767151
455323
455544
;
run;

proc factor   data=D1
              simple
              method=prin
              priors=one
              mineigen=1
              plots=scree
```

```
               rotate=varimax
               round
               flag=.40   ;
      var V1 V2 V3 V4 V5 V6;
run;
```

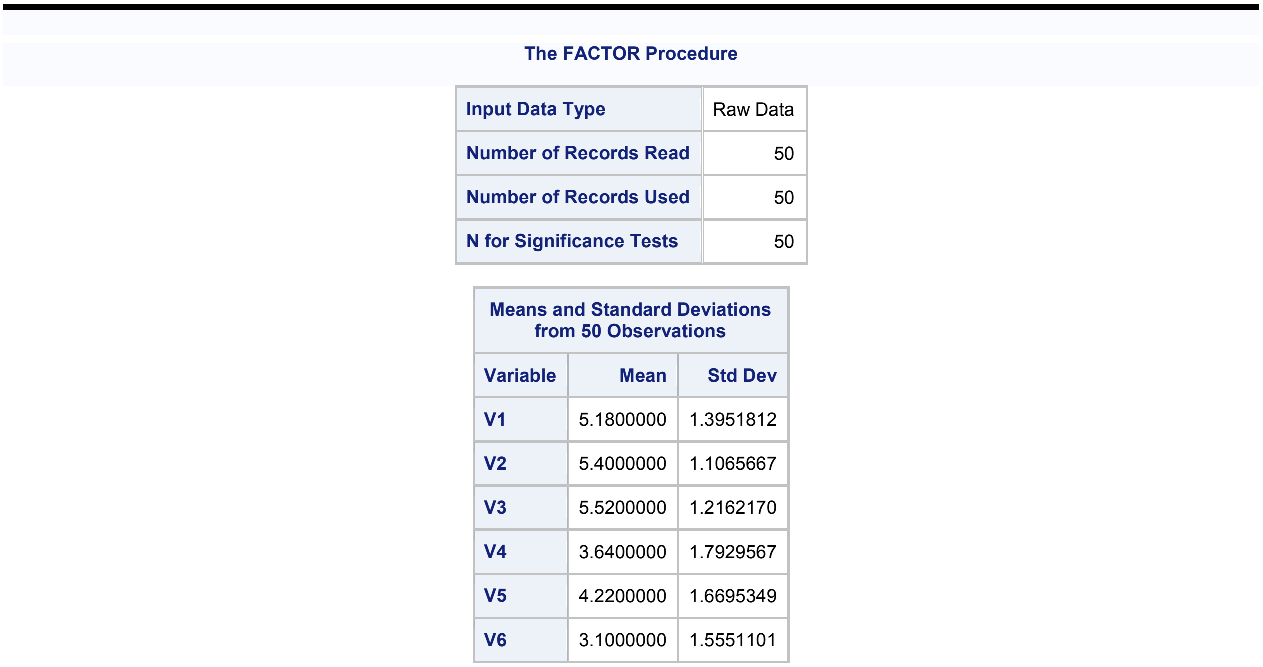
Results from the Output

The preceding program would produce three pages of output. Here is a list of some of the most important information provided by the output and the page on which it appears:

- page 1 includes simple statistics (mean values and standard deviations)
- page 2 includes scree plot of eigenvalues and cumulative variance explained
- page 3 includes the final communality estimates

The output created by the preceding program is presented here as Output 1.1.

Output 1.1: Results of the Initial Principal Component Analysis of the Prosocial Orientation Inventory (POI) Data (Page 1)



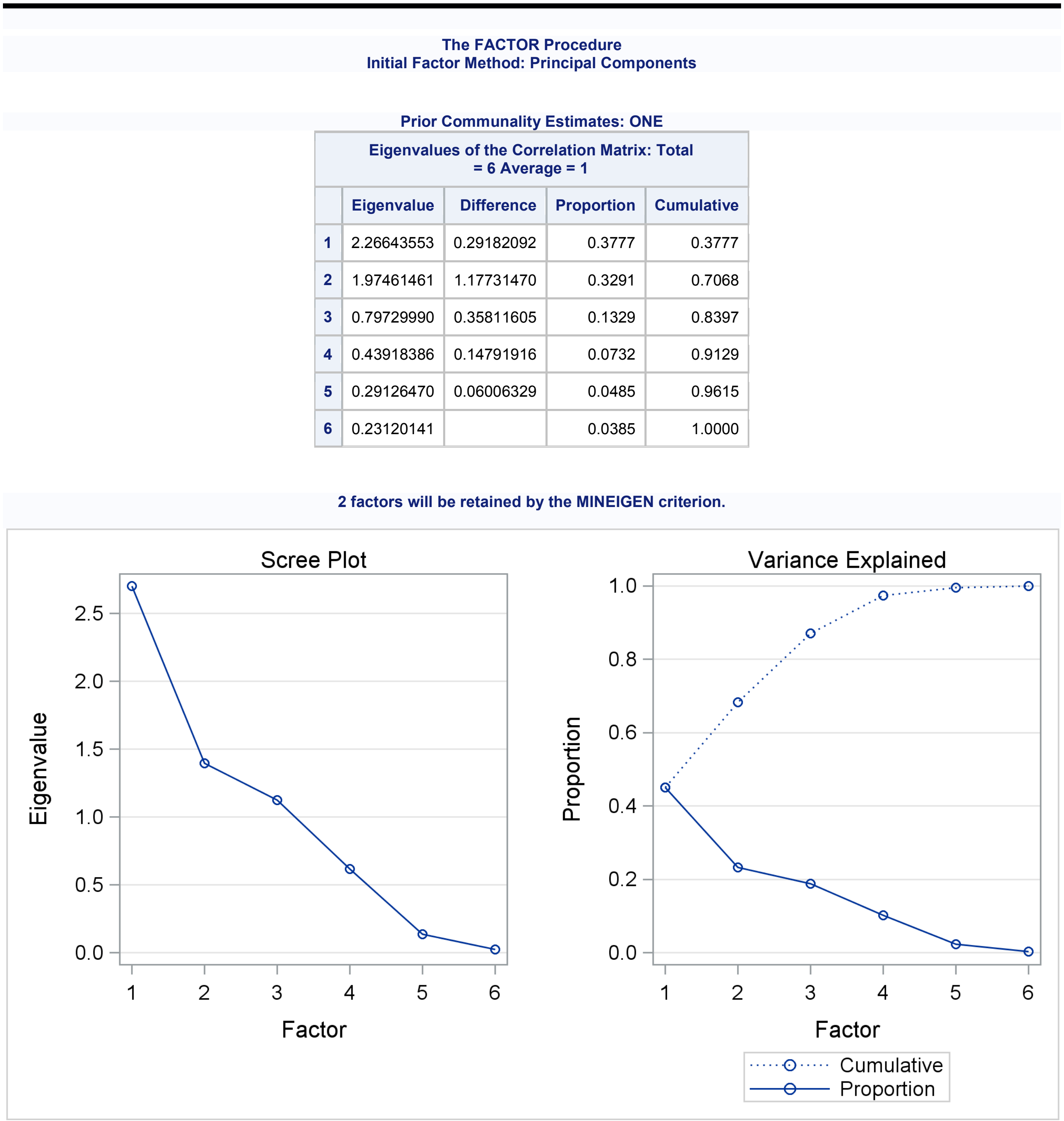
The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	50
Number of Records Used	50
N for Significance Tests	50

Means and Standard Deviations from 50 Observations

Variable	Mean	Std Dev
V1	5.1800000	1.3951812
V2	5.4000000	1.1065667
V3	5.5200000	1.2162170
V4	3.6400000	1.7929567
V5	4.2200000	1.6695349
V6	3.1000000	1.5551101

Output 1.1 (Page 2)



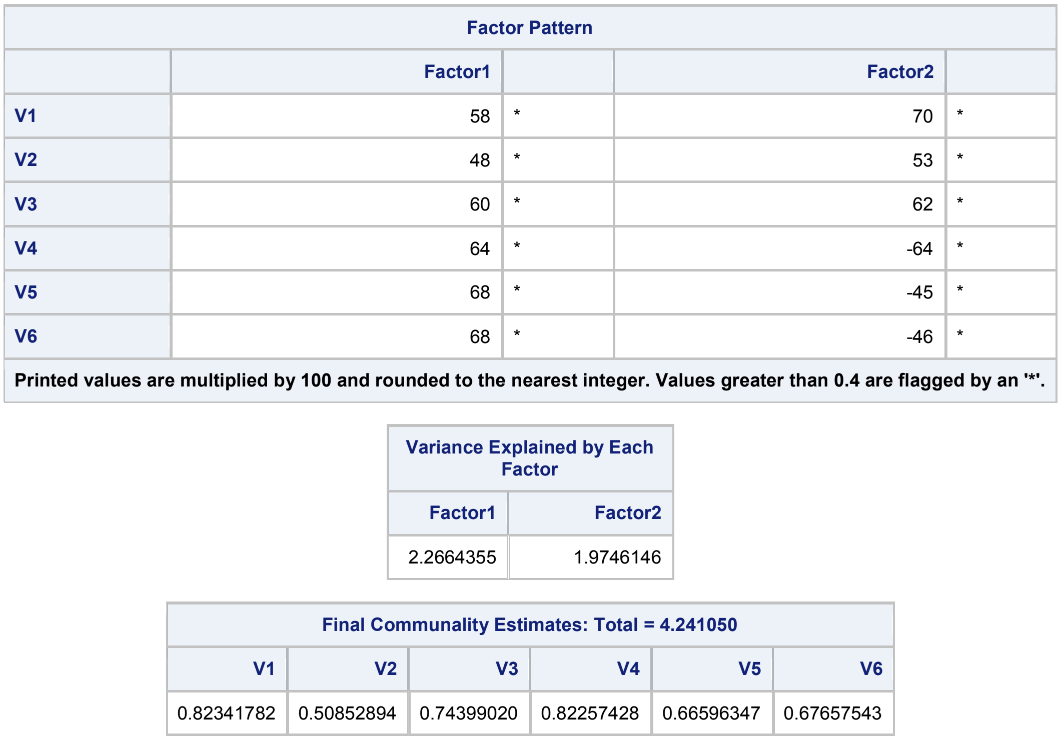
The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.26643553	0.29182092	0.3777	0.3777
2	1.97461461	1.17731470	0.3291	0.7068
3	0.79729990	0.35811605	0.1329	0.8397
4	0.43918386	0.14791916	0.0732	0.9129
5	0.29126470	0.06006329	0.0485	0.9615
6	0.23120141		0.0385	1.0000

2 factors will be retained by the MINEIGEN criterion.



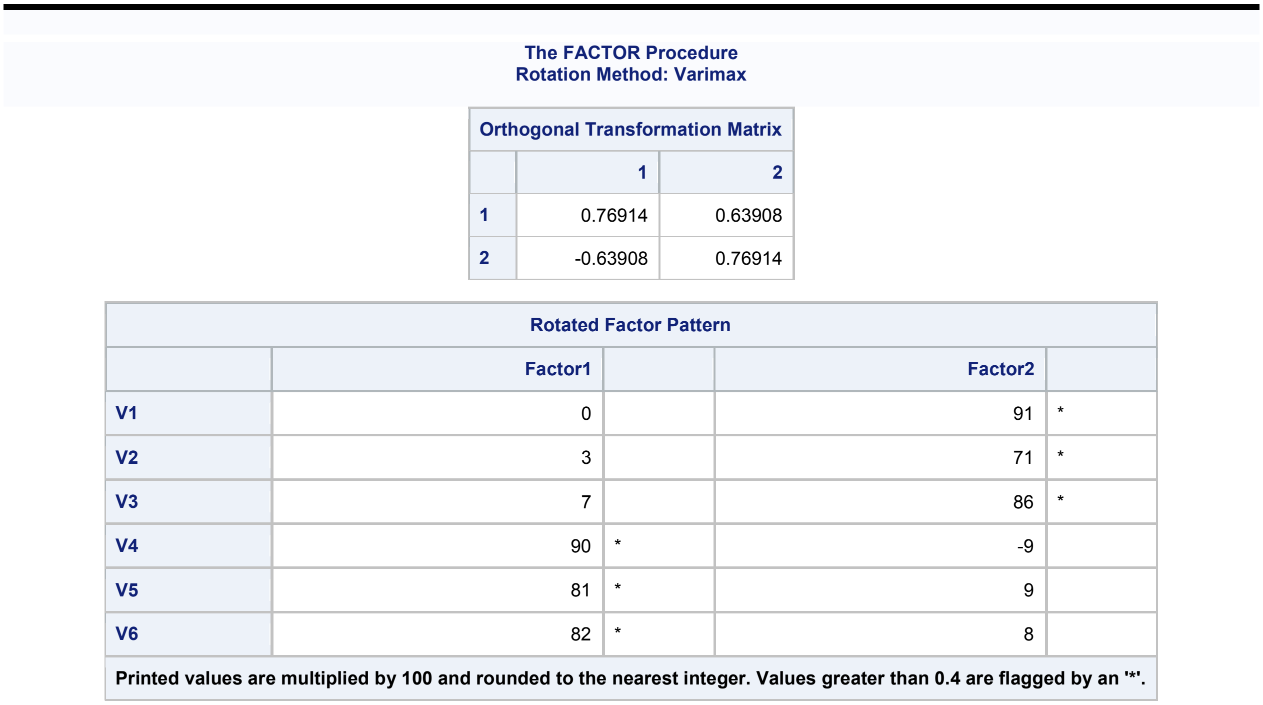
Factor Pattern				
	Factor1		Factor2	
V1	58	*	70	*
V2	48	*	53	*
V3	60	*	62	*
V4	64	*	-64	*
V5	68	*	-45	*
V6	68	*	-46	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Variance Explained by Each Factor	
Factor1	Factor2
2.2664355	1.9746146

Final Communality Estimates: Total = 4.241050					
V1	V2	V3	V4	V5	V6
0.82341782	0.50852894	0.74399020	0.82257428	0.66596347	0.67657543

Output 1.1 (Page 3)



The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix		
	1	2
1	0.76914	0.63908
2	-0.63908	0.76914

Rotated Factor Pattern				
	Factor1		Factor2	
V1	0		91	*
V2	3		71	*
V3	7		86	*
V4	90	*	-9	
V5	81	*	9	
V6	82	*	8	

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Variance Explained by Each Factor	
Factor1	Factor2
2.1472475	2.0938026

Final Community Estimates: Total = 4.241050					
V1	V2	V3	V4	V5	V6
0.82341782	0.50852894	0.74399020	0.82257428	0.66596347	0.67657543

Page 1 from Output 1.1 provides simple statistics for the observed variables included in the analysis. Once the SAS log has been checked to verify that no errors were made in the analysis, these simple statistics should be reviewed to determine how many usable observations were included in the analysis, and to verify that the means and standard deviations are in the expected range. On page 1, it says "Means and Standard Deviations from 50 Observations," meaning that data from 50 participants were included in the analysis.

Steps in Conducting Principal Component Analysis

Principal component analysis is normally conducted in a sequence of steps, with somewhat subjective decisions being made at various points. Because this chapter is intended as an introduction to the topic, this text will not provide a comprehensive discussion of all of the options available at each step; instead, specific recommendations will be made, consistent with common practice in applied research. For a more detailed treatment of principal component analysis and factor analysis, see Stevens (2002).

Step 1: Initial Extraction of the Components

In principal component analysis, the number of components extracted is equal to the number of variables being analyzed. Because six variables are analyzed in the present study, six components are extracted. The first can be expected to account for a fairly large amount of the total variance. Each succeeding component will account for progressively smaller amounts of variance. Although a large number of components may be extracted in this way, only the first few components will be sufficiently important to be retained for interpretation.

Page 2 from Output 1.1 provides the eigenvalue table from the analysis. (This table appears just below the heading "Eigenvalues of the Correlation Matrix: Total = 6 Average = 1".) An **eigenvalue** represents the amount of variance captured by a given component. In the column heading "Eigenvalue," the eigenvalue for each component is presented. Each row in the matrix presents information for each of the six components. Row 1 provides information about the first component extracted, row 2 provides information about the second component extracted, and so forth.

Where the column heading "Eigenvalue" intersects with rows 1 and 2, it can be seen that the eigenvalue for component 1 is approximately 2.27, while the eigenvalue for component 2 is 1.97. This pattern is consistent with our earlier statement that the first components tend to account for relatively large amounts of variance, whereas the later components account for comparatively smaller amounts.

Step 2: Determining the Number of "Meaningful" Components to Retain

Earlier it was stated that the number of components extracted is equal to the number of variables analyzed. This requires that you decide just how many of these components are truly meaningful and worthy of being retained for rotation and interpretation. In general, you expect that only the first few components will account for meaningful amounts of variance, and that the later components will tend to account for only trivial variance. The next step, therefore, is to determine how many meaningful components should be retained to interpret. This section will describe four criteria that may be used in making this decision: the eigenvalue-one criterion, the scree test, the proportion of variance accounted for, and the interpretability criterion.

The Eigenvalue-One Criterion

In principal component analysis, one of the most commonly used criterion for solving the number-of-components problem is the eigenvalue-one criterion, also known as the Kaiser-Guttman criterion (Kaiser 1960). With this method, you retain and interpret all components with eigenvalues greater than 1.00.

The rationale for this criterion is straightforward: each observed variable contributes one unit of variance to the total variance in the dataset. Any component with an eigenvalue greater than 1.00 accounts for a greater amount of variance than had been contributed by one variable. Such a component therefore accounts for a meaningful amount of variance and (in theory) is worthy of retention.

On the other hand, a component with an eigenvalue less than 1.00 accounts for less variance than contributed by one variable. The purpose of principal component analysis is to reduce a number of observed variables into a relatively smaller number of components. This cannot be effectively achieved if you retain components that account for less variance than had been contributed by individual variables. For this reason, components with eigenvalues less than 1.00 are viewed as trivial and are not retained.

The eigenvalue-one criterion has a number of positive features that contribute to its utility. Perhaps the most important reason for its use is its simplicity. It does not require subjective decisions; you merely retain components with eigenvalues greater than 1.00.

Yet this criterion often results in retaining the correct number of components, particularly when a small to moderate number of variables are analyzed and the variable communalities are high. Stevens (2002) reviews studies that have investigated the accuracy of the eigenvalue-one criterion and recommends its use when fewer than 30 variables are being analyzed and communalities are greater than .70, or when the analysis is based on more than 250 observations and the mean communality is greater than .59.

There are, however, various problems associated with the eigenvalue-one criterion. As suggested in the preceding paragraph, it can lead to retaining the wrong number of components under circumstances that are often encountered in research (e.g., when many variables are analyzed, when communalities are small). Also, the reflexive application of this criterion can lead to retaining a certain number of components when the actual difference in the eigenvalues of successive components is trivial. For example, if component 2 has an eigenvalue of 1.01 and component 3 has an eigenvalue of 0.99, then component 2 will be retained but component 3 will not. This may mistakenly lead you to believe that the third component was meaningless when, in fact, it accounted for almost the same amount of variance as the second component. In short, the eigenvalue-one criterion can be helpful when used judiciously, yet the reflexive application of this approach can lead to serious errors of interpretation. Almost always, the eigenvalue-one criterion should be considered in conjunction with other criteria (e.g., scree test, the proportion of variance accounted for, and the interpretability criterion) when deciding how many components to retain and interpret.

With SAS, the eigenvalue-one criterion can be applied by including the MINEIGEN=1 option in the PROC FACTOR statement and not including the NFACT option. The use of the MINEIGEN=1 will cause PROC FACTOR to retain any component with an eigenvalue greater than 1.00.

The eigenvalue table from the current analysis appears on page 2 of Output 1.1. The eigenvalues for components 1, 2, and 3 are 2.27, 1.97, and 0.80, respectively. Only components 1 and 2 have eigenvalues greater than 1.00, so the eigenvalue-one criterion would lead you to retain and interpret only these two components.

Fortunately, the application of the criterion is fairly unambiguous in this case. The last component retained (2) has an eigenvalue of 1.97, which is substantially greater than 1.00, and the next component (3) has an eigenvalue of 0.80, which is clearly lower than 1.00. In this instance, you are not faced with the difficult decision of whether to retain a component with an eigenvalue approaching 1.00 (e.g., an eigenvalue of .99). In situations such as this, the eigenvalue-one criterion may be used with greater confidence.

The Scree Test

With the scree test (Cattell 1966), you plot the eigenvalues associated with each component and look for a definitive “break” between the components with relatively large eigenvalues and those with relatively small eigenvalues. The components that appear *before* the break are assumed to be meaningful and are retained for rotation, whereas those appearing *after* the break are assumed to be unimportant and are not retained. Sometimes a scree plot will display several large breaks. When this is the case, you should look for the last big break before the eigenvalues begin to level off. Only the components that appear before this last large break should be retained.

Specifying the PLOTS=SCREE option in the PROC FACTOR statement tells SAS to print an eigenvalue plot as part of the output. This appears as page 2 of Output 1.1.

You can see that the component numbers are listed on the horizontal axis, while eigenvalues are listed on the vertical axis. With this plot, notice there is a relatively small break between components 1 and 2, and a relatively large break following component 2. The breaks between components 3, 4, 5, and 6 are all relatively small. It is often helpful to draw long lines with extended tails connecting successive pairs of eigenvalues so that these breaks are more apparent (e.g., measure degrees separating lines with a protractor).

Because the large break in this plot appears between components 2 and 3, the scree test would lead you to retain only components 1 and 2. The components appearing after the break (3 to 6) would be regarded as trivial.

The scree test can be expected to provide reasonably accurate results, provided that the sample is large (over 200) and most of the variable communalities are large (Stevens 2002). This criterion too has its weaknesses, most notably the ambiguity of scree plots under common research conditions. Very often, it is difficult to determine precisely where in the scree plot a break exists, or even if a break exists at all. In contrast to the eigenvalue-one criterion, the scree test is often more subjective.

The break in the scree plot on page 3 of Output 1.1 is unusually obvious. In contrast, consider the plot that appears in Figure 1.2.

Figure 1.2: A Scree Plot with No Obvious Break

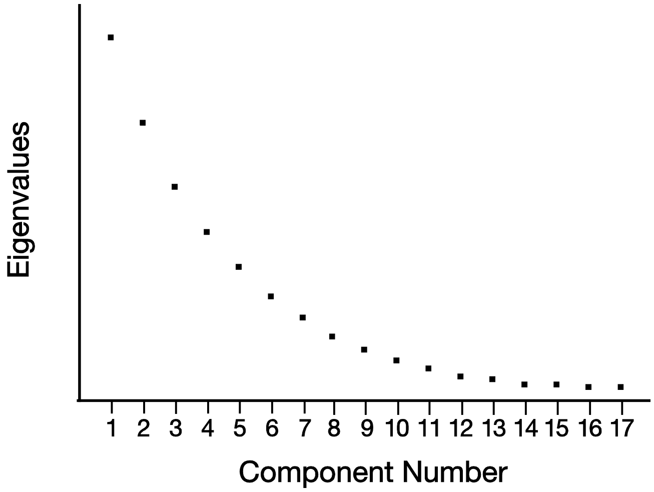


Figure 1.2 presents a fictitious scree plot from a principal component analysis of 17 variables. Notice that there is no obvious break in the plot that separates the meaningful components from the trivial components. Most researchers would agree that components 1 and 2 are probably meaningful whereas components 13 to 17 are probably trivial; but it is difficult to decide exactly where you should draw the line. This example underscores the qualitative nature of judgments based solely on the scree test.

Scree plots such as the one presented in Figure 1.2 are common in social science research. When encountered, the use of the scree test must be supplemented with additional criteria such as the “variance accounted for” criterion and the interpretability criterion, to be described later.

Why do they call it a “scree” test? The word “scree” refers to the loose rubble that lies at the base of a cliff or glacier. When performing a scree test, you normally hope that the scree plot will take the form of a cliff. At the top will be the eigenvalues for the few meaningful components, followed by a definitive break (the edge of the cliff). At the bottom of the cliff will lay the scree (i.e., eigenvalues for the trivial components).

Proportion of Variance Accounted For

A third criterion to address the number of factors problem involves retaining a component if it accounts for more than a specified proportion (or percentage) of variance in the dataset. For example, you may decide to retain any component that accounts for at least 5% or 10% of the total variance. This proportion can be calculated with a simple formula:

$$\text{Proportion} = \frac{\text{Eigenvalue for the component of interest}}{\text{Total eigenvalues of the correlation matrix}}$$

In principal component analysis, the “total eigenvalues of the correlation matrix” is equal to the total number of variables being analyzed (because each variable contributes one unit of variance to the analysis).

Fortunately, it is not necessary to actually compute these percentages by hand since they are provided in the results of PROC FACTOR. The proportion of variance captured by each component is printed in the eigenvalue table (page 2) and appears below the “Proportion” heading.

The eigenvalue table for the current analysis appears on page 2 of Output 1.1. From the “Proportion” column, you can see that the first component alone accounts for 38% of the total variance, the second component alone accounts for 33%, the third component accounts for 13%, and the fourth component accounts for 7%. Assume that you have decided to retain any component that accounts for at least 10% of the total variance in the dataset. With the present results, this criterion leads you to retain components 1, 2, and 3. (Notice that use of this criterion would result in retaining more components than would be retained using the two preceding criteria.)

An alternative criterion is to retain enough components so that the *cumulative* percent of variance is equal to some minimal value. For example, recall that components 1, 2, 3, and 4 accounted for approximately 38%, 33%, 13%, and 7% of the total variance, respectively. Adding these percentages together results in a sum of 91%. This means that the *cumulative* percent of variance accounted for by components 1, 2, 3, and 4 is 91%. When researchers use the “cumulative percent of variance accounted for” criterion for solving the number-of-components problem, they usually retain enough components so that the cumulative percent of variance is at least 70% (and sometimes 80%).

With respect to the results of PROC FACTOR, the cumulative percent of variance accounted for is presented in the eigenvalue table (from page 2), below the “Cumulative” heading. For the present analysis, this information appears in the eigenvalue table on page 2 of Output 1.1. Notice the values that appear below the heading “Cumulative.” Each value indicates the percent of variance accounted for by the present component as well as all preceding components. For example, the value for component 2 is approximately .71 (intersection of the column labeled “Cumulative” and the second row). This value of .71 indicates that approximately 71% of the total variance is accounted for by components 1 and 2. The corresponding entry for component 3 is approximately .84, indicating that 84% of the variance is accounted for by components 1, 2, and 3. If you were to use 70% as the “critical value” for determining the number of components to retain, you would retain only components 1 and 2 in the present analysis.

The primary advantage of the proportion of variance criterion is that it leads you to retain a group of components that combined account for a relatively large proportion of variance in the dataset. Nonetheless, the critical values discussed earlier (10% for individual components and 70% to 80% for the combined components) are quite arbitrary. Because of this and related problems, this approach has been criticized for its subjectivity.

The Interpretability Criterion

Perhaps the most important criterion for solving the number-of-components problem is the **interpretability criterion**: interpreting the substantive meaning of the retained components and verifying that this interpretation makes sense in terms of what is known about the constructs under investigation. The following list provides four rules to follow when applying this criterion. A later section (titled "Step 4: Interpreting the Rotated Solution") shows how to actually interpret the results of a principal component analysis. The following rules will be more meaningful after you have completed that section.

1. **Are there at least three variables (items) with significant loadings on each retained component?** A solution is less satisfactory if a given component is measured by fewer than three variables.
2. **Do the variables that load on a given component share the same conceptual meaning?** For example, if three questions on a survey all load on component 1, do all three of these questions appear to be measuring the same construct?
3. **Do the variables that load on different components seem to be measuring different constructs?** For example, if three questions load on component 1 and three other questions load on component 2, do the first three questions seem to be measuring a construct that is conceptually distinct from the construct measured by the other three questions?
4. **Does the rotated factor pattern demonstrate "simple structure"? Simple structure** means that the pattern possesses two characteristics: (a) most of the variables have relatively high factor loadings on only one component and near zero loadings on the other components; and (b) most components have relatively high loadings for some variables and near-zero loadings for the remaining variables. This concept of simple structure will be explained in more detail in "Step 4: Interpreting the Rotated Solution."

Recommendations

Given the preceding options, what procedures should you actually follow in solving the number-of-components problem? We recommend combining all four in a structured sequence. First, use the MINEIGEN=1 option to implement the eigenvalue-one criterion. Review this solution for interpretability but use caution if the break between the components with eigenvalues above 1.00 and those below 1.00 is not clear-cut (e.g., if component 1 has an eigenvalue of 1.01 and component 2 has an eigenvalue of 0.99).

Next, perform a scree test and look for obvious breaks in the eigenvalues. Because there will often be more than one break in the scree plot, it may be necessary to examine two or more possible solutions.

Next, review the amount of common variance accounted for by each individual component. You probably should not rigidly use some specific but arbitrary cutoff point such as 5% or 10%. Still, if you are retaining components that account for as little as 2% or 4% of the variance, it may be wise to take a second look at the solution and verify that these latter components are truly of substantive importance. In the same way, it is best if the combined components account for at least 70% of the cumulative variance. If less than 70% is captured, it may be prudent to consider alternate solutions that include a larger number of components.

Finally, apply the interpretability criteria to each solution. If more than one solution can be justified on the basis of the preceding criteria, which of these solutions is the most interpretable? By seeking a solution that is both interpretable and satisfies one or more of the other three criteria, you maximize chances of retaining the optimal number of components.

Step 3: Rotation to a Final Solution

Factor Patterns and Factor Loadings

After extracting the initial components, PROC FACTOR will create an unrotated **factor pattern matrix**. The rows of this matrix represent the variables being analyzed, and the columns represent the retained components. (Note that even though we are performing principal component analysis, components are labeled as FACTOR1, FACTOR2, and so forth in the output.)

The entries in the matrix are factor loadings. A **factor loading** (or, more correctly, a *component loading*) is a general term for a coefficient that appears in a factor pattern matrix or a factor structure matrix. In an analysis that results in oblique (correlated) components, the definition of a factor loading is different depending on whether it is in a factor *pattern* matrix or in a factor *structure* matrix. The situation is simpler, however, in an analysis that results in orthogonal components (as in the present chapter). In an orthogonal analysis, factor loadings are equivalent to bivariate correlations between the observed variables and the components.

For example, the factor pattern matrix from the current analysis appears on page 2 of Output 1.1. Where the rows for observed variables intersect with the column for FACTOR1, you can see that the correlation between V1 and the first component is .58, the correlation between V2 and the first component is .48, and so forth.

Rotations

Ideally, you would like to review the correlations between the variables and the components, and use this information to *interpret* the components. In other words, you want to determine what construct seems to be measured by component 1, what construct seems to be measured by component 2, and so forth. Unfortunately, when more than one component has been retained in an analysis, the interpretation of an unrotated factor pattern is generally quite difficult. To facilitate interpretation, you will normally perform an operation called a "rotation." A **rotation** is a linear transformation that is performed on the factor solution for the purpose of making the solution easier to interpret.

PROC FACTOR allows you to request several different types of rotations. The preceding program that analyzed data from the POI study included the statement

```
    rotate=varimax
```

A **varimax rotation** is an orthogonal rotation, meaning that it results in uncorrelated components. Compared to some other types of rotations, a varimax rotation tends to maximize the variance of a column of the factor pattern matrix (as opposed to a row of the matrix). This rotation is probably the most commonly used orthogonal rotation in the social sciences (e.g., Chou and O'Rourke 2012). The results of the varimax rotation for the current analysis appear on page 5 of Output 1.1.

Step 4: Interpreting the Rotated Solution

Interpreting a rotated solution means determining just what is measured by each of the retained components. Briefly, this involves identifying the variables with high loadings on a given component and determining what these variables share in common. Usually, a brief name is assigned to each retained component to describe its content.

The first decision to be made at this stage is how large a factor loading must be to be considered "large." Stevens (2002) discusses some of the issues relevant to this decision and even provides guidelines for testing the statistical significance of factor loadings. Given that this is an introductory treatment of principal component analysis, simply consider a loading to be "large" if its absolute value exceeds .40.

The rotated factor pattern for the POI study appears on page 3 of Output 1.1. The following text provides a structured approach for interpreting this factor pattern.

5. **Read across the row for the first variable.** All “meaningful loadings” (i.e., loadings greater than .40) have been flagged with an asterisk (“*”). This was accomplished by including the FLAG=.40 option in the preceding program. If a given variable has a meaningful loading on more than one component, cross out that variable and ignore it in your interpretation. In many situations, researchers drop variables that load on more than one component because the variables are not pure measures of any one construct. (These are sometimes referred to as *complex items*.) In the present case, this means looking at the row heading “V1” and reading to the right to see if it loads on more than one component. In this case it does not, so you may retain this variable.
6. **Repeat this process for the remaining variables, crossing out any variable that loads on more than one component.** In this analysis, none of the variables have high loadings on more than one component, so none will have to be deleted. In other words, there are no complex items.
7. **Review all of the surviving variables with high loadings on component 1 to determine the nature of this component.** From the rotated factor pattern, you can see that only items 4, 5, and 6 load on component 1 (note the asterisks). It is now necessary to turn to the questionnaire itself and review the content in order to decide what a given component should be named. What do questions 4, 5, and 6 have in common? What common construct do they appear to be measuring? For illustration, the questions being analyzed in the present case are reproduced here. Remember that question 4 was represented as V4 in the SAS program, question 5 was V5, and so forth. Read questions 4, 5, and 6 to see what they have in common.

1 2 3 4 5 6 7 1. Went out of my way to do a favor for a coworker.
1 2 3 4 5 6 7 2. Went out of my way to do a favor for a relative.
1 2 3 4 5 6 7 3. Went out of my way to do a favor for a friend.
1 2 3 4 5 6 7 4. Gave money to a religious charity.
1 2 3 4 5 6 7 5. Gave money to a charity not affiliated with a religion.
1 2 3 4 5 6 7 6. Gave money to a panhandler.

Questions 4, 5, and 6 all seem to deal with giving money to persons in need. It is therefore reasonable to label component 1 the “financial giving” component.

8. **Repeat this process to name the remaining retained components.** In the present case, there is only one remaining component to name: component 2. This component has high loadings for questions 1, 2, and 3. In reviewing these items, it is apparent that each seems to deal with helping friends, relatives, or other acquaintances. It is therefore appropriate to name this the “helping others” component.
9. **Determine whether this final solution satisfies the interpretability criteria.** An earlier section indicated that the overall results of a principal component analysis are satisfactory only if they meet a number of interpretability criteria. The adequacy of the rotated factor pattern presented on page 3 of Output 1.1 is assessed in terms of the following criteria:
   a. **Are there at least three variables (items) with significant loadings on each retained component?** In the present example, three variables loaded on component 1 and three also loaded on component 2, so this criterion was met.
   b. **Do the variables that load on a given component share similar conceptual meaning?** All three variables loading on component 1 measure giving to those in need, while all three loading on component 2 measure prosocial acts performed for others. Therefore, this criterion is met.
   c. **Do the variables that load on different components seem to be measuring different constructs?** The items loading on component 1 measure respondents’ financial contributions, while the items loading on component 2 measure helpfulness toward others. Because these seem to be conceptually distinct constructs, this criterion appears to be met as well.
   d. **Does the rotated factor pattern demonstrate “simple structure”?** Earlier, it was noted that a rotated factor pattern demonstrates simple structure when it has two characteristics. First, most of the variables should have high loadings on one component and near-zero loadings on other components. It can be seen that the pattern obtained here meets that requirement: items 1 to 3 have high loadings on component 2 and near-zero loadings on component 1. Similarly, items 4 to 6 have high loadings on component 1 and near-zero loadings on component 2. The second

characteristic of simple structure is that each component should have high loadings for some variables and near-zero loadings for the others. The pattern obtained here also meets this requirement: component 1 has high loadings for items 4 to 6 and near-zero loadings for the other items whereas component 2 has high loadings for items 1 to 3 and near-zero loadings on the remaining items. In short, the rotated component pattern obtained in this analysis does appear to demonstrate simple structure.

Step 5: Creating Factor Scores or Factor-Based Scores

Once the analysis is complete, it is often desirable to assign scores to participants to indicate where they stand on the retained components. For example, the two components retained in the present study were interpreted as “financial giving” and “helping others.” You may now want to assign one score to each participant to indicate that participant’s standing on the “financial giving” component and a second score to indicate that participant’s standing on the “helping others” component. Once assigned, these component scores could be used either as predictor variables or as criterion variables in subsequent analyses.

Before discussing the options for assigning these scores, it is important to first draw a distinction between factor scores and factor-based scores. In principal component analysis, a **factor score** (or **component score**) is a linear composite of the optimally weighted observed variables. If requested, PROC FACTOR will compute each participant’s factor scores for the two components by:

- determining the optimal weights
- multiplying participant responses to questionnaire items by these weights
- summing the products

The resulting sum will be a given participant’s score on the component of interest. Remember that a separate equation with different weights is computed for each retained component.

A **factor-based score**, on the other hand, is merely a linear composite of the variables that demonstrate meaningful loadings for the component in question. In the preceding analysis, for example, items 4, 5, and 6 demonstrated meaningful loadings for the “financial giving” component. Therefore, you could calculate the factor-based score on this component for a given participant by simply adding together her responses to items 4, 5, and 6. Notice that, with a factor-based score, the observed variables are not multiplied by optimal weights before they are summed.

Computing Factor Scores

Factor scores are requested by including the NFACT and OUT options in the PROC FACTOR statement. Here is the general form for a SAS program that uses the NFACT and OUT option to compute factor scores:

```
proc factor   data=dataset-name
              simple
              method=prin
              priors=one
              nfact=number-of-components-to-retain
              rotate=varimax
              round
              flag=desired-size-of-"significant"-factor-loadings
              out=name-of-new-SAS-dataset   ;
   var  variables-to-be-analyzed ;
run;
```

Here are the actual program statements (minus the DATA step) that could be used to perform a principal component analysis and compute factor scores for the POI study:

```
proc factor   data=D1
        simple
        method=prin
        priors=one
        nfact=2
        rotate=varimax
        round
        flag=.40
❶      out=D2   ;
    var V1 V2 V3 V4 V5 V6;
run;
```

Notice how this program differs from the original program presented earlier in the chapter (in the section titled “SAS Program and Output”). The MINEIGEN=1 option has been removed and replaced with the NFACT=2 option. The OUT=D2 option has also been added.

Line ❶ of the preceding program asks that an output dataset be created and given the name D2. This name is arbitrary; any name consistent with SAS requirements would be acceptable. The new dataset named D2 will contain all variables contained in the previous dataset (D1), as well as new variables named FACTOR1 and FACTOR2. FACTOR1 will contain factor scores for the first retained component, and FACTOR2 will contain scores for the second. The number of new “FACTOR” variables created will be equal to the number of components retained by the NFACT statement.

The OUT option may be used to create component scores only if the analysis has been performed on a raw data as opposed to a correlation or covariance matrix. The use of the NFACT statement is also required.

Having created the new variables named FACTOR1 and FACTOR2, you may be interested to see how they relate to the study’s original observed variables. This can be done by appending PROC CORR statements to the SAS program, following the last of the PROC FACTOR statements. The full program minus the DATA step is now presented:

```
  proc factor  data=D1
         simple
         method=prin
         priors=one
         nfact=2
         rotate=varimax
         round
         flag=.40
❶  out=D2   ;
     var V1 V2 V3 V4 V5 V6;
   run;

❷ proc corr   data=D2;
     var FACTOR1 FACTOR2;
     with V1 V2 V3 V4 V5 V6 FACTOR1 FACTOR2;
  run;
```

Notice that the PROC CORR statement on line ❷ specifies DATA=D2. This dataset (D2) is the name of the output dataset created on line ❶ the PROC FACTOR statement. The PROC CORR statement requests that the factor score variables (FACTOR1 and FACTOR2) be correlated with participants’ responses to questionnaire items 1 to 6 (V1 to V6).

The preceding program produces five pages of output. Pages 1 to 2 provide simple statistics, the eigenvalue table, and the unrotated factor pattern. Page 3 provides the rotated factor pattern and final communality estimates (same as before). Page 4 provides the standardized scoring coefficients used in creating factor scores.

Finally, page 5 provides the correlations requested by the corr procedure. Pages 3, 4, and 5 of the output created by the preceding program are presented here as Output 1.2.

Output 1.2: Output Pages 3, 4, and 5 from the Analysis of POI Data from Which Factor Scores Were Created (Page 3)

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix		
	1	2
1	-0.87835	0.47802
2	0.47802	0.87835

Rotated Factor Pattern				
	Factor1		Factor2	
V1	-86	*	7	
V2	-12		93	*
V3	85	*	-2	
V4	-40		-47	*
V5	79	*	-38	
V6	-37		67	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Variance Explained by Each Factor	
Factor1	Factor2
2.4042522	1.6940222

Final Community Estimates: Total = 4.098274					
V1	V2	V3	V4	V5	V6
0.75027648	0.88099977	0.73071122	0.38098475	0.76187043	0.59343168

Output 1.2 (Page 4)

The FACTOR Procedure
Rotation Method: Varimax

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor	
Factor1	Factor2
1.0000000	1.0000000

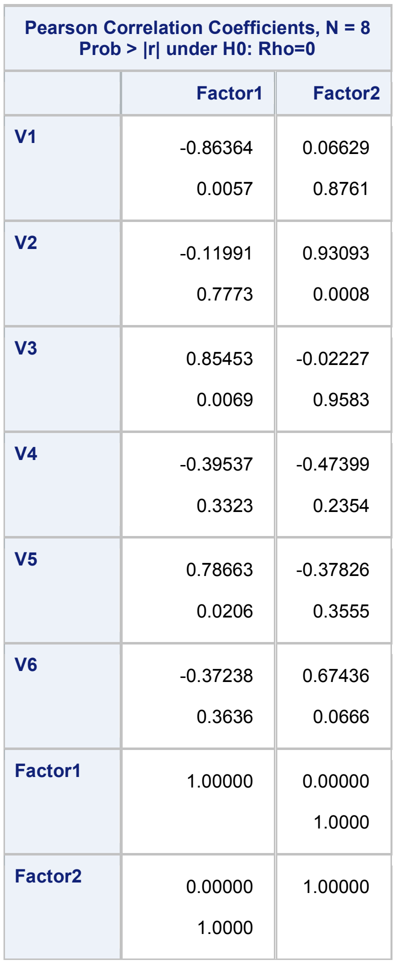
Standardized Scoring Coefficients		
	Factor1	Factor2
V1	-0.37829	-0.08350
V2	0.08170	0.57602
V3	0.38060	0.11024
V4	-0.24662	-0.35975
V5	0.29827	-0.12660
V6	-0.06907	0.37569

Output 1.2 (Page 5)

The CORR Procedure

8 With Variables:	V1 V2 V3 V4 V5 V6 Factor1 Factor2
2 Variables:	Factor1 Factor2

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
V1	8	560956	134602	4487647	353434	767153
V2	8	544528	182498	4356220	142441	676222
V3	8	574671	190693	4597367	265454	777222
V4	8	662603	80496	5300822	544444	777443
V5	8	621159	78894	4969272	445332	666665
V6	8	534284	175061	4274270	244342	767151
Factor1	8	0	1.00000	0	-1.38533	1.30018
Factor2	8	0	1.00000	0	-1.85806	1.32865



Pearson Correlation Coefficients, N = 8 Prob > |r| under H0: Rho=0		
	Factor1	**Factor2**
V1	-0.86364 0.0057	0.06629 0.8761
V2	-0.11991 0.7773	0.93093 0.0008
V3	0.85453 0.0069	-0.02227 0.9583
V4	-0.39537 0.3323	-0.47399 0.2354
V5	0.78663 0.0206	-0.37826 0.3555
V6	-0.37238 0.3636	0.67436 0.0666
Factor1	1.00000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000

The simple statistics for PROC CORR appear on page 5 in Output 1.2. Notice that the simple statistics for the observed variables (V1 to V6) are identical to those that appeared at the beginning of the factor output discussed earlier (at the top of Output 1.1, page 1). In contrast, note the simple statistics for FACTOR1 and FACTOR2 (the factor score variables for components 1 and 2, respectively). Both have means of 0 and standard deviations of 1; these variables were constructed to be standardized variables.

The correlations between FACTOR1 and FACTOR2 and the original observed variables appear at the bottom half of page 5. You can see that the correlations between FACTOR1 and V1 to V6 on page 4 of Output 1.2 are identical to the factor loadings of V1 to V6 on FACTOR1 on page 3 of Output 1.1, under “Rotated Factor Pattern.” This makes sense, as the elements of a factor pattern (in an orthogonal solution) are simply correlations between the observed variables and the components themselves. Similarly, you can see that the correlations between FACTOR2 and V1 to V6 from page 5 of Output 1.2 are also identical to the corresponding factor loadings from page 5 of Output 1.1.

Of particular interest is the correlation between FACTOR1 and FACTOR2, as computed by PROC CORR. This appears on page 5 of Output 1.2, where the row for FACTOR2 intersects with the column for FACTOR1. Notice that the observed correlation between these two components is zero. This is as expected; the rotation method used in the principal component analysis was the varimax method which produces orthogonal, or uncorrelated, components.

Computing Factor-Based Scores

A second (and less sophisticated) approach to scoring involves the creation of new variables that contain factor-based scores rather than true factor scores. A variable that contains factor-based scores is sometimes referred to as a **factor-based scale**.

Although factor-based scores can be created in a number of ways, the following method has the advantage of being relatively straightforward:

1. To calculate factor-based scores for component 1, first determine which questionnaire items had high loadings on that component.
2. For a given participant, add together that participant's responses to these items. The result is that participant's score on the factor-based scale for component 1.
3. Repeat these steps to calculate each participant's score on the remaining retained components.

Although this may sound like a cumbersome task, it is actually quite simple with the use of data manipulation statements contained in a SAS program. For example, assume that you have performed the principal component analysis on your questionnaire responses and have obtained the findings reported in this chapter. Specifically, you found that survey items 4, 5, and 6 loaded on component 1 (the "financial giving" component), while items 1, 2, and 3 loaded on component 2 (the "helping others" component).

You would now like to create two new SAS variables. The first variable, called GIVING, will include each participant's factor-based score for financial giving. The second variable, called HELPING, will include each participant's factor-based score for helping others. Once these variables are created, they can be used as criterion or predictor variables in subsequent analyses. To keep things simple, assume that you are simply interested in determining whether there is a significant correlation between GIVING and HELPING.

At this time, it may be useful to review Appendix A.3, "Working with Variables and Observations in SAS Datasets," particularly the section on creating new variables from existing variables. This review should make it easier to understand the data manipulation statements used here.

Assume that earlier statements in the SAS program have already entered responses to the six questionnaire items. These variables are included in a dataset called D1. The following are the subsequent lines that will then create a new dataset called D2. This dataset will include all of the variables in D1 as well as the newly created factor-based scales called GIVING and HELPING.

```
❶  data D2;
❷     set D1;

❸  GIVING   = (V4 + V5 + V6);
   HELPING  = (V1 + V2 + V3);

❹  proc corr   data=D2;
❺     var GIVING  HELPING;
❻  run;
```

Lines ❶ and ❷ request that a new dataset be created called D2, and that it be set up as a duplicate of existing dataset D1. On line ❸, the new variable called GIVING is created. For each participant, the responses to items 4, 5, and 6 are added together. The result is each participant's score on the factor-based scale for the first component. These scores are stored as a variable called GIVING. The component-based scale for the "helping others" component is created on line ❹, and these scores are stored as the variable called HELPING. Lines ❺ to ❻ request the correlations between GIVING and HELPING be computed. GIVING and HELPING can now be used as predictor or criterion variables in subsequent analyses. To save space, the results of this program will not be presented here. However, note that this output would probably display a nonzero correlation between GIVING and HELPING. This may come as a surprise because earlier it was shown that the factor scores contained in FACTOR1 and FACTOR2 (counterparts to GIVING and HELPING) were completely uncorrelated.

The reason for this apparent contradiction is simple: FACTOR1 and FACTOR2 are true principal components, and true principal components (created in an orthogonal solution) are always created with optimally weighted equations so that they will be mutually uncorrelated.

In contrast, GIVING and HELPING are not true principal components that consist of true factor scores; they are merely variables *based* on the results of a principal component analysis. Optimal weights (that would ensure orthogonality) were not used in the creation of GIVING and HELPING. This is why factor-based scales generally demonstrate nonzero correlations while true principal components (from an orthogonal solution) will not.

Recoding Reversed Items Prior to Analysis

It is almost always best to recode any reversed or negatively keyed items before conducting any of the analyses described here. In particular, it is essential that reversed items be recoded prior to the program statements that produce factor-based scales. For example, the three questionnaire items that assess financial giving appear again here:

1 2 3 4 5 6 7 4. Gave money to a religious charity.
1 2 3 4 5 6 7 5. Gave money to a charity not affiliated with a religion.
1 2 3 4 5 6 7 6. Gave money to a panhandler.

None of these items are reversed. With each item, a response of “7” indicates a high level of financial giving. In the following, however, item 4 is a reversed item; a response of “7” indicating a low level of giving:

1 2 3 4 5 6 7 4. **Refused** to give money to a religious charity.
1 2 3 4 5 6 7 5. Gave money to a charity not affiliated with a religion.
1 2 3 4 5 6 7 6. Gave money to a panhandler.

If you were to perform a principal component analysis on responses to these items, the factor loading for item 4 would most likely have a sign that is the opposite of the sign of the loadings for items 5 and 6 (e.g., if items 5 and 6 had positive loadings, then item 4 would have a negative loading). This would complicate the creation of a component-based scale: with items 5 and 6, higher scores indicate greater giving whereas with item 4, lower scores indicate greater giving. You would not want to sum these three items as they are presently coded. First, it will be necessary to reverse item 4. Notice how this is done in the following program (assume that the data have already been input in a SAS dataset named D1):

```
    data D2;
       set D1;

❶    V4 = 8 - V4;

      GIVING   = (V4 + V5 + V6);
      HELPING  = (V1 + V2 + V3);

    proc corr   DATA=D2;
       var GIVING    HELPING;
    run;
```

Line ❶ of the preceding program created a new, recoded version of variable V4. Values on this new version of V4 are equal to the quantity 8 minus the value of the old version of V4. For participants whose score on the old version of V4 was 1, their value on the new version of V4 is 7 (because 8 – 1 = 7) whereas for those whose score is 7, their value on the new version of V4 is 1 (because 8 – 7 = 1). Again, see Appendix A.3 for further description of this procedure.

The general form of the formula used to recode reversed items is

```
variable-name = constant - variable-name ;
```

In this formula, the “constant” is the following quantity:

the number of points on the response scale used with the questionnaire item plus 1

Therefore, if you are using the 4-point response format, the constant is 5. If using a 9-point scale, the constant is 10.

If you have prior knowledge about which items are going to appear as reversed (with reversed component loadings) in your results, it is best to place these recoding statements early in your SAS program, before the PROC FACTOR statements. This will make interpretation of the components more straightforward because it will eliminate significant loadings with opposite signs from appearing on the same component. In any case, it is essential that the statements used to recode reversed items appear before the statements that create any factor-based scales.

Step 6: Summarizing the Results in a Table

For reports that summarize the results of your analysis, it is generally desirable to prepare a table that presents the rotated factor pattern. When analyzed variables contain responses to questionnaire items, it can be helpful to reproduce the questionnaire items within this table. This is presented in Table 1.2:

Table 1.2: Rotated Factor Pattern and Final Communality Estimates from Principal Component Analysis of Prosocial Orientation Inventory

Component			
1	**2**	**h^2**	**Items**
.00	.91	.82	Went out of my way to do a favor for a coworker.
.03	.71	.51	Went out of my way to do a favor for a relative.
.07	.86	.74	Went out of my way to do a favor for a friend.
.90	-.09	.82	Gave money to a religious charity.
.81	.09	.67	Gave money to a charity not associated with a religion.
.82	.08	.68	Gave money to a panhandler.

Note: *N* = 50. Communality estimates appear in column headed h^2.

The final communality estimates from the analysis are presented under the heading “**h^2**” in the table. These estimates appear in the SAS output following “Variance Explained by Each Factor” (page 3 of Output 1.2).

Very often, the items that constitute the questionnaire are lengthy, or the number of retained components is large, so that it is not possible to present the factor pattern, the communalities, and the items themselves in the same table. In such situations, it may be preferable to present the factor pattern and communalities in one table and the items in a second. Shared item numbers (or single words or defining phrases) may then be used to associate each item with its corresponding factor loadings and communality.

Step 7: Preparing a Formal Description of the Results for a Paper

The preceding analysis could be summarized in the following way:

Principal component analysis was performed on responses to the 6-item questionnaire using ones as prior communality estimates. The principal axis method was used to extract the components, and this was followed by a varimax (orthogonal) rotation.

Only the first two components had eigenvalues greater than 1.00; results of a scree test also suggested that only the first two were meaningful. Therefore, only the first two components were retained for rotation. Combined, components 1 and 2 accounted for 71% of the total variance (38% plus 33%, respectively).

Questionnaire items and corresponding factor loadings are presented in Table 1.2. When interpreting the rotated factor pattern, an item was said to load on a given component if the factor loading was .40 or greater for that component and less than .40 for the other. Using these criteria, three items were found to load on the first component, which was subsequently labeled “financial giving.” Three items also loaded on the second component labeled “helping others.”

An Example with Three Retained Components

The Questionnaire

The next example involves fictitious research that examines Rusbult’s (1980) investment model (Le and Agnew 2003). This model identifies variables believed to affect a person’s commitment to a romantic relationship. In this context, **commitment** refers to the person’s intention to maintain the relationship and stay with a current romantic partner.

One version of the investment model predicts that commitment will be affected by three antecedent variables: satisfaction, investment size, and alternative value. **Satisfaction** refers to a person’s affective (emotional) response to the relationship. Among other things, people report high levels of satisfaction when their current relationship comes close to their perceived ideal relationship. **Investment size** refers to the amount of time, energy, and personal resources that an individual has put into the relationship. For example, people report high investments when they have spent a lot of time with their current partner and have developed mutual friends that may be lost if the relationship were to end. Finally, **alternative value** refers to the attractiveness of alternatives to one’s current partner. A person would score high on alternative value if, for example, it would be appealing to date someone else or perhaps just be alone for a while.

Assume that you wish to conduct research on the investment model and are in the process of preparing a 12-item questionnaire to assess levels of satisfaction, investment size, and alternative value in a group of participants involved in romantic relationships. Part of the instrument used to assess these constructs is presented here:

Indicate the extent to which you agree or disagree with each of the following statements by specifying the appropriate response in the space to the left of the statement. Please use the following response format to make these ratings:

7 = Strongly Agree
6 = Agree
5 = Slightly Agree
4 = Neither Agree Nor Disagree
3 = Slightly Disagree
2 = Disagree
1 = Strongly Disagree

_____ 1. I am satisfied with my current relationship.
_____ 2. My current relationship comes close to my ideal relationship.
_____ 3. I am more satisfied with my relationship than the average person.
_____ 4. I feel good about my current relationship.
_____ 5. I have invested a great deal of time in my current relationship.
_____ 6. I have invested a great deal of energy in my current relationship.
_____ 7. I have invested a lot of my personal resources (e.g., money) in developing my current relationship.
_____ 8. My partner and I have established mutual friends that I might lose if we were to break up.
_____ 9. There are plenty of other attractive people for me to date if I were to break up with my current partner.
_____ 10. It would be appealing to break up with my current partner and date someone else.
_____ 11. It would be appealing to break up with my partner to be alone for a while.
_____ 12. It would be appealing to break up with my partner and “play the field.”

In the preceding questionnaire, items 1 to 4 were written to assess satisfaction, items 5 to 8 were written to assess investment size, and items 9 to 12 were written to assess alternative value. Assume that you administer this questionnaire to 300 participants and now want to perform a principal component analysis on their responses.

Writing the Program

Earlier, it was noted that it is possible to perform a principal component analysis on a correlation matrix (or covariance matrix) as well as on raw data. This section shows how the former is done. The following program includes the correlation matrix that provides all possible correlation coefficients between responses to the 12 questionnaire items and performs a principal component analysis on these fictitious data:

```
data D1(type=corr)  ;
     input   _type_   $
             _name_   $
             V1-V12   ;
  datalines;
  n     .    300  300  300  300  300  300  300  300  300  300  300  300
  std   .   2.48 2.39 2.58 3.12 2.80 3.14 2.92 2.50 2.10 2.14 1.83 2.26
  corr V1   1.00  .    .    .    .    .    .    .    .    .    .    .
  corr V2    .69 1.00  .    .    .    .    .    .    .    .    .    .
  corr V3    .60  .79 1.00  .    .    .    .    .    .    .    .    .
  corr V4    .62  .47  .48 1.00  .    .    .    .    .    .    .    .
  corr V5    .03  .04  .16  .09 1.00  .    .    .    .    .    .    .
  corr V6    .05 -.04  .08  .05  .91 1.00  .    .    .    .    .    .
  corr V7    .14  .05  .06  .12  .82  .89 1.00  .    .    .    .    .
  corr V8    .23  .13  .16  .21  .70  .72  .82 1.00  .    .    .    .
  corr V9   -.17 -.07 -.04 -.05 -.33 -.26 -.38 -.45 1.00  .    .    .
  corr V10  -.10 -.08  .07  .15 -.16 -.20 -.27 -.34  .45 1.00  .    .
  corr V11  -.24 -.19 -.26 -.28 -.43 -.37 -.53 -.57  .60  .22 1.00  .
  corr V12  -.11 -.07  .07  .08 -.10 -.13 -.23 -.31  .44  .60  .26 1.00
  ;
run ;
    proc factor   data=D1
                  method=prin
                  priors=one
                  mineigen=1
                  plots=scree
                  rotate=varimax
                  round
                  flag=.40   ;
      var  V1-V12;
run;
```

The PROC FACTOR statement in the preceding program follows the general form recommended for the previous data analyses. Notice that the MINEIGEN=1 statement requests that all components with eigenvalues greater than 1.00 be retained and the PLOTS=SCREE option requests a scree plot of eigenvalues. These options are particularly helpful for the initial analysis of data as they can help determine the correct number of components to retain. If the scree test (or the other criteria) suggests retaining some number of components other than what would be retained using the MINEIGEN=1 option, that option may be dropped and replaced with the NFACT option.

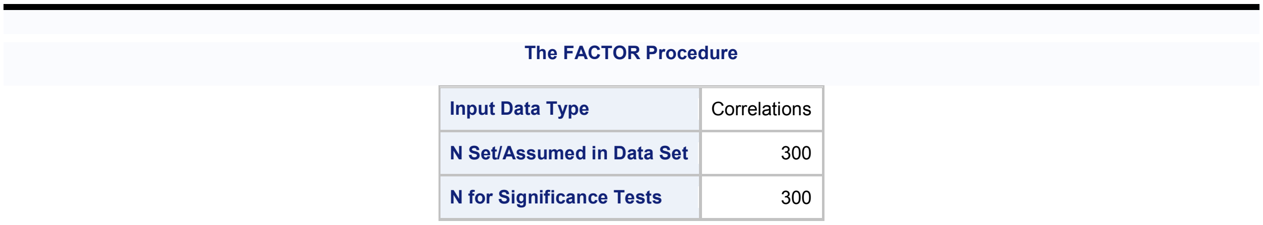
Results of the Initial Analysis

The preceding program produced three pages of output, with the following information appearing on each page:

- page 1 reports the data input procedure and sample size
- page 2 includes the eigenvalue table and scree plot of eigenvalues
- page 3 includes the rotated factor pattern and final communality estimates

The eigenvalue table from this analysis appears on page 1 of Output 1.3. The eigenvalues themselves appear in the left-hand column under the heading “Eigenvalue.” From these values, you can see that components 1, 2, and 3 have eigenvalues of 4.47, 2.73, and 1.70, respectively. Furthermore, you can see that only these first three components have eigenvalues greater than 1.00. This means that three components will be retained by the MINEIGEN criterion. Notice that the first nonretained component (component 4) has an eigenvalue of approximately 0.85 which, of course, is well below 1.00. This is encouraging, as you have more confidence in the eigenvalue-one criterion when the solution does not contain “near-miss” eigenvalues (e.g., .98 or .99).

Output 1.3: Results of the Initial Principal Component Analysis of the Investment Model Data (page 1)



The FACTOR Procedure

Input Data Type	Correlations
N Set/Assumed in Data Set	300
N for Significance Tests	300

Output 1.3 (page 2)

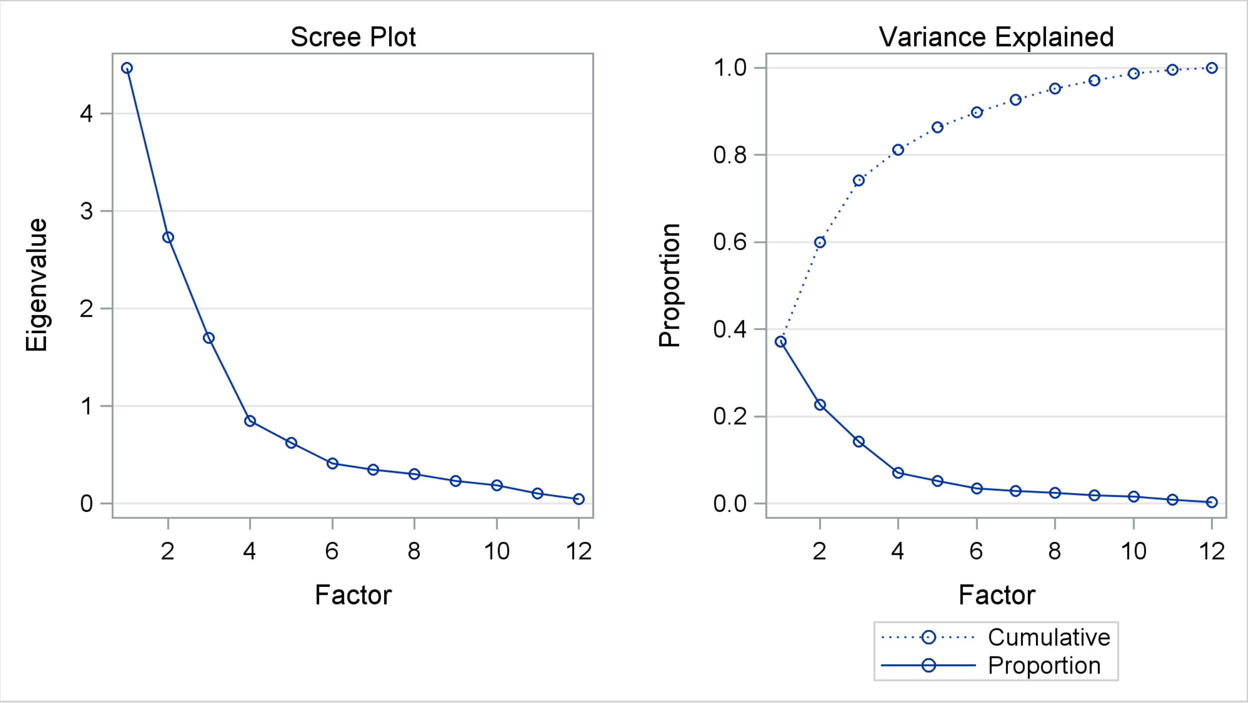
The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 12 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.47058134	1.73995858	0.3725	0.3725
2	2.73062277	1.02888853	0.2276	0.6001
3	1.70173424	0.85548155	0.1418	0.7419
4	0.84625269	0.22563029	0.0705	0.8124
5	0.62062240	0.20959929	0.0517	0.8642
6	0.41102311	0.06600575	0.0343	0.8984
7	0.34501736	0.04211948	0.0288	0.9272
8	0.30289788	0.07008042	0.0252	0.9524
9	0.23281745	0.04595812	0.0194	0.9718
10	0.18685934	0.08061799	0.0156	0.9874
11	0.10624135	0.06091129	0.0089	0.9962
12	0.04533006		0.0038	1.0000

3 factors will be retained by the MINEIGEN criterion.



Factor Pattern						
	Factor1		Factor2		Factor3	
V1	39		76	*	-14	
V2	31		82	*	-12	
V3	34		79	*	9	
V4	31		69	*	15	
V5	80	*	-26		41	*
V6	79	*	-32		41	*
V7	87	*	-27		26	
V8	88	*	-14		9	
V9	-61	*	14		47	*
V10	-43	*	23		68	*
V11	-72	*	-6		12	
V12	-40		19		72	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

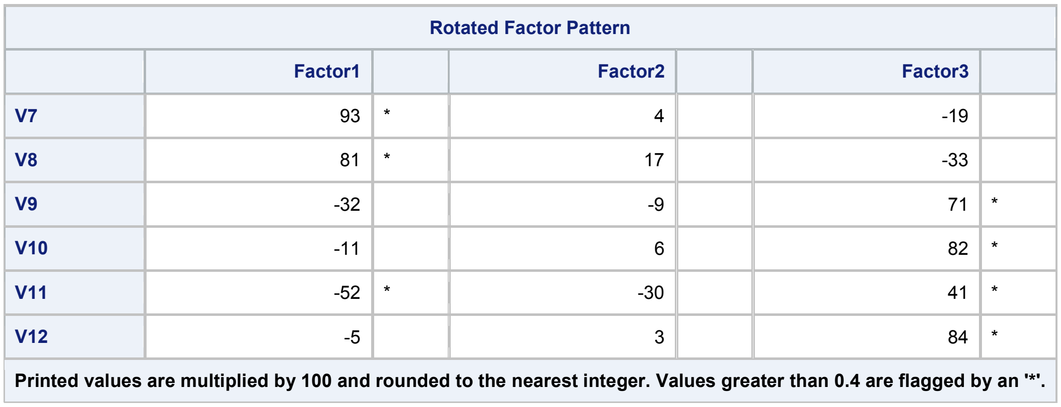
Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.4705813	2.7306228	1.7017342

Output 1.3 (page 3)

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix			
	1	2	3
1	0.83136	0.34431	-0.43623
2	-0.29481	0.93864	0.17902
3	0.47110	-0.02022	0.88185

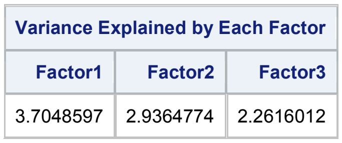
Rotated Factor Pattern						
	Factor1		Factor2		Factor3	
V1	3		85	*	-16	
V2	-4		88	*	-10	
V3	9		86	*	8	
V4	13		75	*	12	
V5	93	*	2		-3	
V6	95	*	-4		-4	



Rotated Factor Pattern

	Factor1		Factor2		Factor3	
V7	93	*	4		-19	
V8	81	*	17		-33	
V9	-32		-9		71	*
V10	-11		6		82	*
V11	-52	*	-30		41	*
V12	-5		3		84	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.



Variance Explained by Each Factor

Factor1	Factor2	Factor3
3.7048597	2.9364774	2.2616012

The eigenvalue table in Output 1.3 also shows that the first three components combined account for slightly more than 74% of the total variance. (This variance value can be observed at the intersection of the column labeled “Cumulative” and row “3”.) The “percentage of variance accounted for” criterion suggests that it may be appropriate to retain three components.

The scree plot from this solution appears on page 2 of Output 1.3. This scree plot shows that there are several large breaks in the data following components 1, 2, and 3, and then the line begins to flatten beginning with component 4. The last large break appears after component 3, suggesting that only components 1 to 3 account for meaningful variance. This suggests that only these first three components should be retained and interpreted. Notice how it is almost possible to draw a straight line through components 4 to 12. The components that lie along a semi-straight line such as this are typically assumed to be measuring only trivial variance (i.e., components 4 to 12 constitute the “scree” of your scree plot).

So far, the results from the eigenvalue-one criterion, the variance accounted for criterion, and the scree plot are in agreement, suggesting that a three-component solution may be most appropriate. It is now time to review the rotated factor pattern to see if such a solution is interpretable. This matrix is presented on page 3 of Output 1.3.

Following the guidelines provided earlier, you begin by looking for factorially complex items (i.e., items with meaningful loadings on more than one component). A review shows that item 11 (variable V11) is a complex item, loading on both components 1 and 3. Item 11 should therefore be discarded. Except for this item, the solution is otherwise fairly straightforward.

To interpret component 1, you read down the column for FACTOR1 and see that items 5 to 8 load significantly on this component. These items are:

_____ 5. I have invested a great deal of time in my current relationship.
_____ 6. I have invested a great deal of energy in my current relationship.
_____ 7. I have invested a lot of my personal resources (e.g., money) in developing my current relationship.
_____ 8. My partner and I have established mutual friends that I might lose if we were to break up.

All of these items deal with the investments that participants have made in their relationships, so it makes sense to label this the “investment size” component.

The rotated factor pattern shows that items 1 to 4 have meaningful loadings on component 2. These items are:

_____ 1. I am satisfied with my current relationship.

_____ 2. My current relationship comes close to my ideal relationship.

_____ 3. I am more satisfied with my relationship than the average person.

_____ 4. I feel good about my current relationship.

Given the content of the preceding items, it seems reasonable to label component 2 the “satisfaction” component.

Finally, items 9, 10, and 12 have meaningful loadings on component 3. (Again, remember that item 11 has been discarded.) These items are:

_____ 9. There are plenty of other attractive people around for me to date if I were to break up with my current partner.

_____ 10. It would be appealing to break up with my current partner and date someone else.

_____ 12. It would be appealing to break up with my partner and “play the field.”

These items all seem to deal with the attractiveness of alternatives to one’s current relationship, so it makes sense to label this the “alternative value” component.

You may now step back and determine whether this solution satisfies the interpretability criteria presented earlier.

1. Are there at least three variables with meaningful loadings on each retained component?
2. Do the variables that load on a given component share the same conceptual meaning?
3. Do the variables that load on different components seem to be measuring different constructs?
4. Does the rotated factor pattern demonstrate “simple structure”?

In general, the answer to each of these questions is “yes,” indicating that the current solution is, in most respects, satisfactory. There is, however, a problem with item 11, which loads on both components 1 and 3. This problem prevents the current solution from demonstrating a perfectly “simple structure” (criterion 4 from above). To eliminate this problem, it may be desirable to repeat the analysis, this time analyzing all of the items *except* for item 11. This will be done in the second analysis of the investment model data described below.

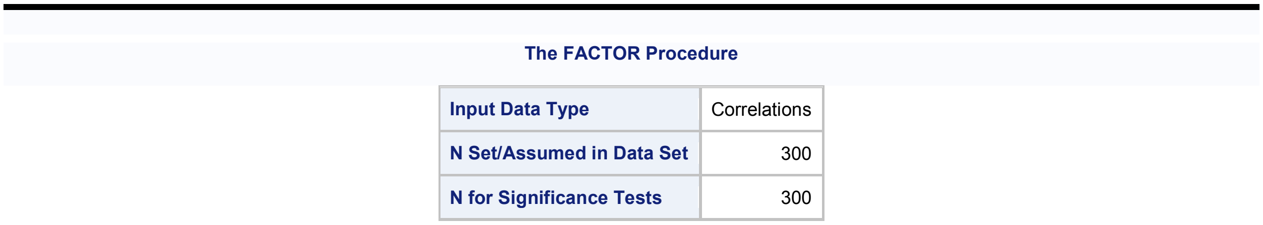
Results of the Second Analysis

To repeat the current analysis with item 11 deleted, it is necessary only to modify the VAR statement of the preceding program. This may be done by changing the VAR statement so that it appears as follows:

```
        var V1-V10 V12;
```

All other aspects of the program will remain as they were previously. The eigenvalue table, scree plot, the unrotated factor pattern, the rotated factor pattern, and final communality estimates obtained from this revised program appear in Output 1.4:

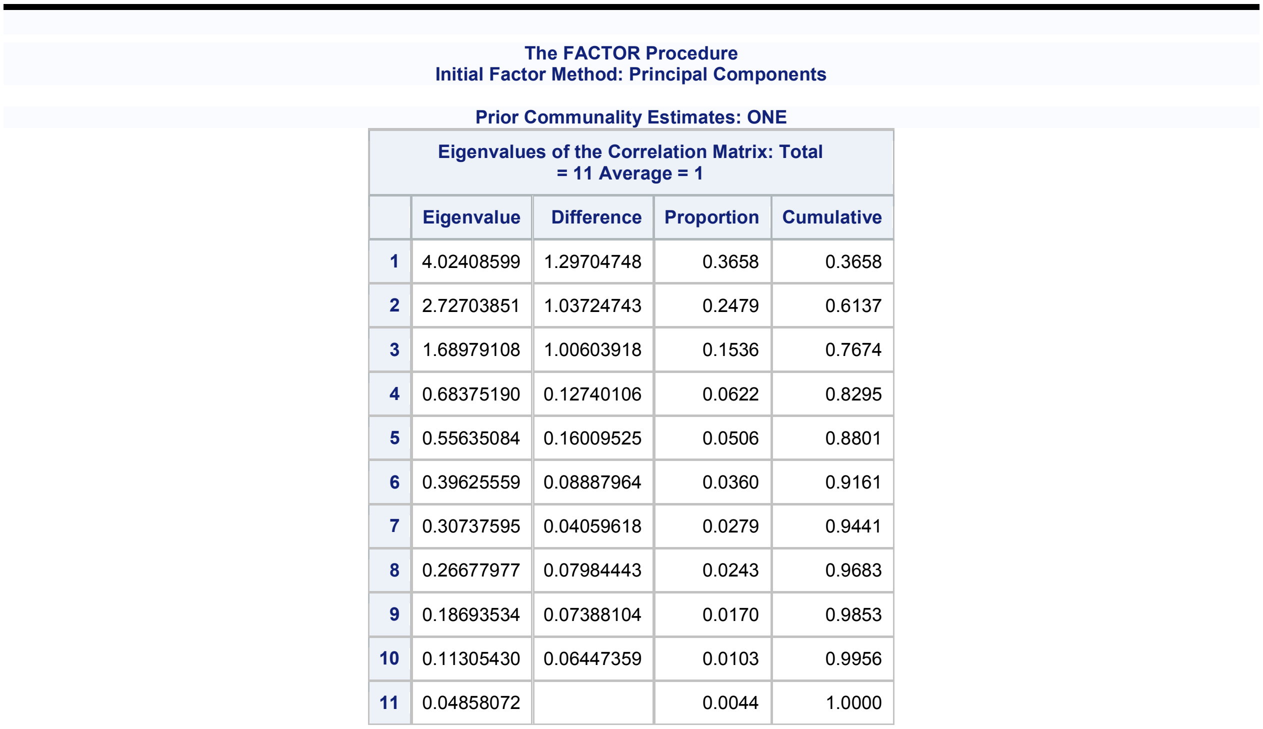
Output 1.4: Results of the Second Analysis of the Investment Model Data (Page 1)



The FACTOR Procedure

Input Data Type	Correlations
N Set/Assumed in Data Set	300
N for Significance Tests	300

Output 1.4 (page 2)



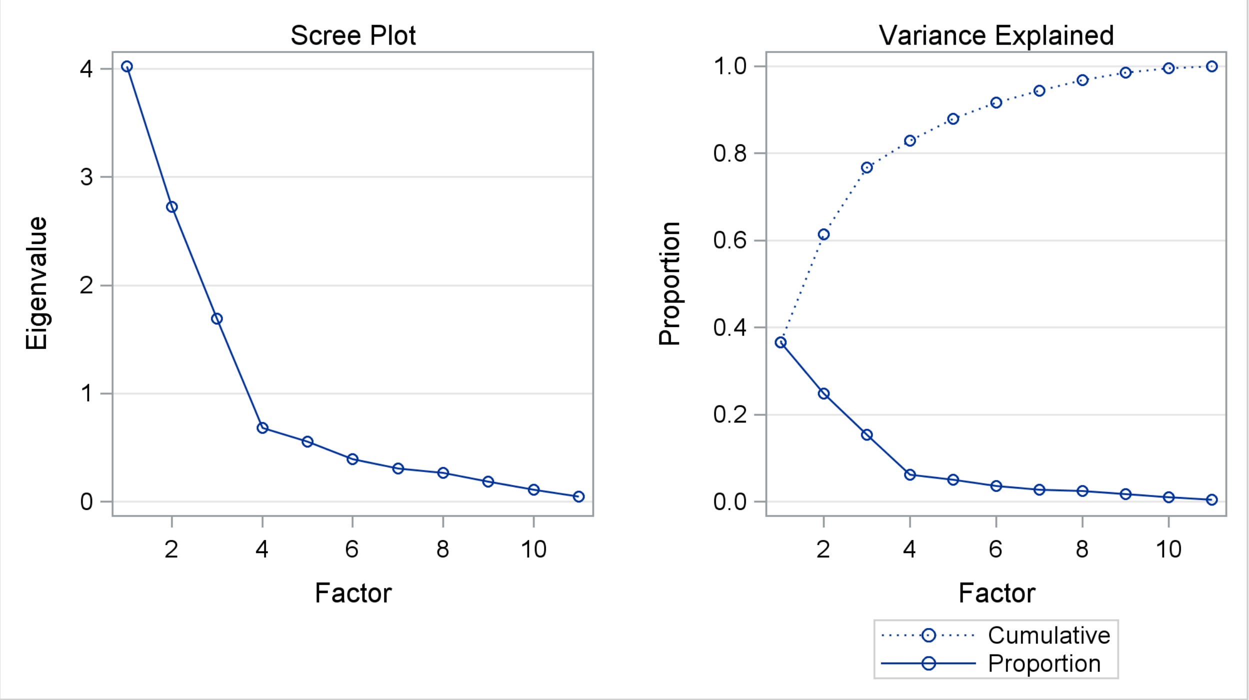
The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 11 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.02408599	1.29704748	0.3658	0.3658
2	2.72703851	1.03724743	0.2479	0.6137
3	1.68979108	1.00603918	0.1536	0.7674
4	0.68375190	0.12740106	0.0622	0.8295
5	0.55635084	0.16009525	0.0506	0.8801
6	0.39625559	0.08887964	0.0360	0.9161
7	0.30737595	0.04059618	0.0279	0.9441
8	0.26677977	0.07984443	0.0243	0.9683
9	0.18693534	0.07388104	0.0170	0.9853
10	0.11305430	0.06447359	0.0103	0.9956
11	0.04858072		0.0044	1.0000

3 factors will be retained by the MINEIGEN criterion.

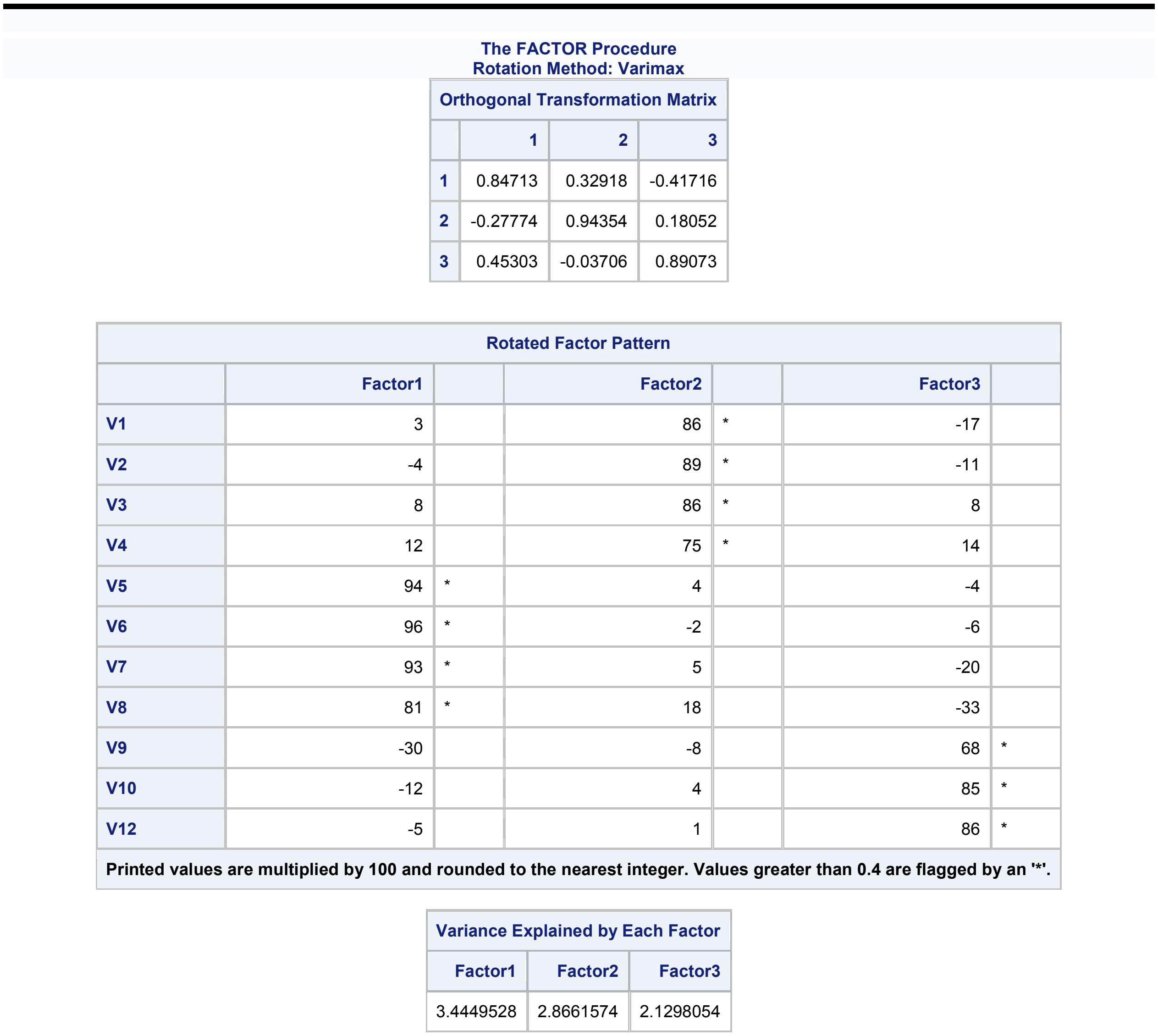


Factor Pattern						
	Factor1		Factor2		Factor3	
V1	38		77	*	-17	
V2	30		83	*	-15	
V3	32		80	*	8	
V4	29		70	*	15	
V5	83	*	-23		38	
V6	83	*	-30		38	
V7	89	*	-24		24	
V8	88	*	-12		7	
V9	-56	*	13		47	*
V10	-44	*	22		70	*
V12	-40		18		74	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.0240860	2.7270385	1.6897911

Output 1.4 (Page 3)



The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix			
	1	2	3
1	0.84713	0.32918	-0.41716
2	-0.27774	0.94354	0.18052
3	0.45303	-0.03706	0.89073

Rotated Factor Pattern

	Factor1		Factor2		Factor3	
V1	3		86	*	-17	
V2	-4		89	*	-11	
V3	8		86	*	8	
V4	12		75	*	14	
V5	94	*	4		-4	
V6	96	*	-2		-6	
V7	93	*	5		-20	
V8	81	*	18		-33	
V9	-30		-8		68	*
V10	-12		4		85	*
V12	-5		1		86	*

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
3.4449528	2.8661574	2.1298054

The results obtained when item 11 is deleted from the analysis are very similar to those obtained when it was included. The eigenvalue table of Output 1.4 shows that the eigenvalue-one criterion would again result in retaining three components. The first three components account for close to 77% of the total variance, which means that three components would also be retained if you used the variance-accounted-for criterion. Also, the scree plot from page 2 of Output 1.4 is cleaner than observed with the initial analysis; the break between components 3 and 4 is now more distinct and the eigenvalues again level off after this break. This means that three components would also likely be retained if the scree test were used to solve the number-of-components problem.

The biggest change can be seen in the rotated factor pattern that appears on page 4 of Output 1.4. The solution is now cleaner in the sense that no item loads on more than one component (i.e., no complex items). The current results now demonstrate a somewhat simpler structure than the initial analysis of the investment model data.

Conclusion

Principal component analysis is an effective procedure for reducing a number of observed variables into a smaller number that account for most of the variance in a dataset. This technique is particularly useful when you need a data reduction procedure that makes no assumptions concerning an underlying causal structure responsible for covariation in the data.

Appendix: Assumptions Underlying Principal Component Analysis

Because a principal component analysis is performed on a matrix of Pearson correlation coefficients, the data should satisfy the assumptions for this statistic. These assumptions are described in Appendix A.5, “Preparing Scattergrams and Computing Correlations,” and are briefly reviewed here:

- **Interval-level measurement.** All variables should be assessed on an interval or ratio level of measurement.
- **Random sampling.** Each participant will contribute one score on each observed variable. These sets of scores should represent a random sample drawn from the population of interest.
- **Linearity.** The relationship between all observed variables should be linear.
- **Bivariate normal distribution.** Each pair of observed variables should display a bivariate normal distribution (e.g., they should form an elliptical scattergram when plotted).

References

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

Chou, P. H. B., and O’Rourke, N. (2012). Development and initial validation of the Therapeutic Misunderstanding Scale for use with clinical trial research participants. *Aging and Mental Health, 16,* 45–15.

Clark, L. A., and Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

DeVellis, R. F. (2012). *Scale development theory and applications* (3rd Ed.). Thousand Oaks, CA: Sage.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151.

Le, B., and Agnew, C. R. (2003). Commitment and its theorized determinants: A meta-analysis of the investment model. *Personal Relationships, 10,* 37–57.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analyses with missing data.* New York: Wiley.

O’Rourke, N., and Cappeliez, P. (2002). Development and validation of a couples measure of biased responding: The Marital Aggrandizement Scale. *Journal of Personality Assessment, 78,* 301–320.

Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology, 16,* 172–186.

Saris, W. E., and Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research.* Hoboken, NJ: Wiley InterScience.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th Ed.). Mahwah, NJ: Lawrence Erlbaum.

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry, 39,* 135–140.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL. Chapman and Hall.

1
3
5
2
4
R