2

The Common Factor Model and Exploratory Factor Analysis

This chapter introduces the reader to the concepts, terminology, and basic equations of the common factor model. Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are based on the common factor model. In this chapter, the common factor model is discussed primarily in the context of EFA. Nonetheless, most of the concepts and terminology (e.g., common and unique variances, factor loadings, communalities) of EFA are also used in CFA. This chapter discusses some of the fundamental similarities and differences of EFA and CFA. In applied research, EFA and CFA are often conducted in conjunction with one another. For instance, CFA is frequently used in the later stages of scale development, after the factor structure of a testing instrument has been explored and refined by EFA. Thus, because the applied CFA researcher must have a working knowledge of EFA, the methods of conducting an EFA are reviewed in this chapter. This overview is also provided to allow more detailed comparisons of EFA and CFA in later chapters.

OVERVIEW OF THE COMMON FACTOR MODEL

Since its inception over a century ago (Spearman, 1904, 1927), factor analysis has become one of the most widely used multivariate statistical procedures in applied research endeavors across a multitude of domains (e.g., psychology, education, sociology, management, political science, public health). The fundamental intent of factor analysis is to determine the number and nature of latent variables or *factors* that account for the variation and covariation among a set of observed measures, commonly referred to as *indicators*. Specifically, a factor is an unobservable variable that influences more than one observed measure and that accounts for the correlations among these observed measures. In other words, the observed measures are intercorrelated because they share a common cause (i.e., they are influenced by the same underlying construct); if the latent

construct is partialed out, the intercorrelations among the observed measures will be zero.[1] Thus factor analysis attempts a more parsimonious understanding of the covariation among a set of indicators because the number of factors is less than the measured variables.

In applied research, factor analysis is most commonly used in psychometric evaluations of multiple-item testing instruments (e.g., questionnaires; cf. Floyd & Widaman, 1995). For example, a researcher may have generated 20 questionnaire items that he or she believes are indicators of the unidimensional construct of self-esteem. In the early stages of scale development, the researcher may use factor analysis to examine the plausibility of this assumption (i.e., the ability of a single factor to account for the intercorrelations among the 20 indicators) and to determine if all 20 items are reasonable indicators of the underlying construct of self-esteem (i.e., how strongly is each item related to the factor?). In addition to psychometric evaluation, other common uses for factor analysis include construct validation (e.g., obtaining evidence of convergent and discriminant validity by demonstrating that indicators of selected constructs load onto separate factors in the expected manner; e.g., Brown, Chorpita, & Barlow, 1998) and data reduction (e.g., reducing a larger set of intercorrelated indicators to a smaller set of composite variables, and using these composites—i.e., factor scores—as the units of analysis in subsequent statistical tests; e.g., Cox, Walker, Enns, & Karpinski, 2002).

These concepts emanate from the *common factor model* (Thurstone, 1947), which postulates that each indicator in a set of observed measures is a linear function of one or more common factors and one unique factor. Thus factor analysis partitions the variance of each indicator (derived from the sample correlation/covariance matrix which is used as input for the analysis) into two parts: (1) *common variance*, or the variance accounted for by the factor, which is estimated on the basis of variance shared with other indicators in the analysis; and (2) *unique variance*, which is a combination of reliable variance that is specific to the indicator (i.e., systematic factors that influence only one indicator) and random error variance (i.e., measurement error or unreliability in the indicator). There are two main types of analyses based on the common factor model: *exploratory factor analysis* (EFA) and *confirmatory factor analysis* (CFA; Jöreskog, 1969, 1971a). Both EFA and CFA aim to reproduce the observed relationships among a group of indicators with a smaller set of latent variables, but they differ fundamentally by the number and nature of a priori specifications and restrictions made on the factor model. EFA is a data-driven approach, such that no specifications are made in regard to the number of factors (initially) or the pattern of relationships between the common factors and the indicators (i.e., the *factor loadings*). Rather, a researcher employs EFA as an exploratory or descriptive technique to determine the appropriate number of common factors, and to uncover which measured variables are reasonable indicators of the various latent dimensions (e.g., by the size and differential magnitude of factor loadings). In CFA, the researcher specifies the number of factors and the pattern of indicator–factor loadings in advance, as well as other parameters such as those bearing on the independence or covariance of the factors and indicator unique variances.[2] The prespecified factor solution is evaluated in terms of how well it reproduces the sample correlation (covariance)

matrix of the measured variables. Thus, unlike EFA, CFA requires a strong empirical or conceptual foundation to guide the specification and evaluation of the factor model. Accordingly, EFA is typically used earlier in the process of scale development and construct validation, whereas CFA is used in later phases after the underlying structure has been established on prior empirical (EFA) and theoretical grounds. Other important differences between EFA and CFA are discussed in Chapter 3.

A brief, applied example is used to illustrate some of the key concepts of the common factor model. In this basic example, four behavioral observation ratings (O1–O4) have been collected on 300 individuals admitted to an inpatient psychiatric facility. The four ratings are hopelessness (O1), feelings of worthlessness/guilt (O2), psychomotor retardation (O3), and sleep disturbance (O4). As shown in Table 2.1, these four clinical ratings (indicators) are moderately intercorrelated. It is conjectured that each of these ratings is a manifest indicator of the latent construct of Depression; that is, each of the observed symptoms (e.g., hopelessness, worthlessness) has the shared influence of Depression, the single latent variable (factor) that accounts for the intercorrelations among these observed measures. The only reason the indicators are correlated is that they share the common cause of Depression; if this latent variable is partialed out, no relationship among these indicators will be seen.

With the sample correlations presented in Table 2.1 as input, a factor analysis is conducted by using the EFA routines provided in SPSS (FACTOR), SAS (PROC FACTOR), and Mplus (see Table 2.2). For reasons noted later in this chapter, only a one-factor solution can be pursued. Because EFA typically uses correlations as the units of analysis, it can be run in SPSS and SAS by embedding the sample correlation matrix in the body of the syntax (as shown in Table 2.2), although both programs can generate this matrix by reading raw input data files. In Mplus, the input data must be read in from an external file (in text format). The procedures of EFA are discussed later in this chapter (e.g., methods of factor extraction and selection), but for purposes of this illustration, consider the selected results of the analysis presented in Table 2.2. In the selected output from SPSS, of particular interest is the output under the heading "Factor Matrix," which provides the factor loadings for the four clinical ratings. In EFA, the factor loadings are completely standardized estimates of the regression slopes for predicting the indicators from the factor, and thus are interpreted along the lines of standardized regression (β)

TABLE 2.1. Intercorrelations among Four Behavioral Observation Ratings of Depression

	O1	O2	O3	O4
O1	1.00			
O2	0.70	1.00		
O3	0.65	0.66	1.00	
O4	0.62	0.63	0.60	1.00

Note. (N = 300) O1, hopelessness; O2, feelings of worthlessness/guilt; O3, psychomotor retardation; O4, sleep disturbance.

TABLE 2.2. SPSS, SAS, and Mplus Syntax and Selected Output for a Basic One-Factor Model

SPSS syntax

```
MATRIX DATA VARIABLES=ROWTYPE_ O1 O2 O3 O4.
BEGIN DATA.
N 300 300 300 300
COR 1.0
COR .70 1.0
COR .65 .66 1.0
COR .62 .63 .60 1.0
END DATA.
FACTOR
  /MATRIX=IN(COR=*)
  /MISSING LISTWISE
  /ANALYSIS O1 O2 O3 O4
  /PRINT INITIAL EXTRACTION
  /CRITERIA FACTORS(1) ITERATE(25)
  /EXTRACTION ML
  /ROTATION NOROTATE.
```

SAS syntax

```
data sm (type=CORR);
   input _type_ $ _name_ $ O1 O2 O3 O4;
cards;
n     .  300    .     .      .
corr  O1 1.00   .     .      .
corr  O2 0.70  1.00   .      .
corr  O3 0.65  0.66   1.00   .
corr  O4 0.62  0.63   0.60  1.00
;
proc factor data=sm method=ml nfactors=1 scree;
run;
```

Mplus syntax

```
TITLE: MPLUS EXAMPLE OF EFA
DATA:     FILE IS tab2.2.dat;
          NOBSERVATIONS = 300;
          TYPE=CORRELATION;
VARIABLE: NAMES ARE O1 O2 O3 O4;
ANALYSIS: TYPE = EFA 1 1;
          ESTIMATOR=ML;
OUTPUT:   FSDETERMINACY;
```

Selected output (SPSS)

```
Initial Statistics:

Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
                          *
O1               .57811   *     1       2.93118       73.3         73.3
O2               .59175   *     2        .41039       10.3         83.5
O3               .53077   *     3        .35924        9.0         92.5
O4               .48795   *     4        .29919        7.5        100.0
```

(continued)

TABLE 2.2. *(continued)*

```
Test of fit of the    1-factor model:

Chi-square statistic:      .2031, D.F.:    2, Significance:      .9035



Factor Matrix:

              Factor  1

O1               .82822
O2               .84090
O3               .78766
O4               .75228


Final Statistics:

Variable     Communality  *  Factor   SS Loadings  Pct of Var   Cum Pct
                          *
O1               .68595   *     1       2.57939       64.5         64.5
O2               .70712   *
O3               .62040   *
O4               .56592   *
```

Selected output (Mplus)

```
RESULTS FOR EXPLORATORY FACTOR ANALYSIS

           EIGENVALUES FOR SAMPLE CORRELATION MATRIX
                1             2             3             4
           ________      ________      ________      ________
      1         2.931         0.410         0.359         0.299

MODEL FIT INFORMATION

Number of Free Parameters                        8

Chi-Square Test of Model Fit
          Value                             0.206
          Degrees of Freedom                    2
          P-Value                           0.9023

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.000
          90 Percent C.I.                   0.000  0.048
          Probability RMSEA <= .05          0.952

CFI/TLI
          CFI                               1.000
          TLI                               1.009
```

(continued)

TABLE 2.2. *(continued)*

```
SRMR (Standardized Root Mean Square Residual)
          Value                              0.003

          GEOMIN ROTATED LOADINGS (* significant at 5% level)
                 1
                 ________
O1               0.828*
O2               0.841*
O3               0.788*
O4               0.752*

          ESTIMATED RESIDUAL VARIANCES
             O1             O2             O3             O4
             ________       ________       ________       ________
      1        0.314          0.293          0.380          0.434

          S.E. GEOMIN ROTATED LOADINGS
                 1
                 ________
O1               0.024
O2               0.023
O3               0.027
O4               0.030

          Est./S.E. GEOMIN ROTATED LOADINGS
                 1
                 ________
O1              34.141
O2              35.963
O3              28.926
O4              25.112

          FACTOR DETERMINACIES
                 1
                 ________
      1          0.940
```

or correlation (*r*) coefficients as in multiple regression/correlational analysis (cf. Cohen, Cohen, West, & Aiken, 2003).[3] For instance, the factor loading estimate for O1 (hopelessness) is .828, which is interpreted as indicating that a standardized score increase in the factor (Depression) is associated with an .828 standardized score increase in tearfulness. Squaring the factor loadings provides the estimate of the amount of variance in the indicator accounted for by the latent variable (e.g., $.828^2$ = 68.5% variance explained). In factor analysis, the amount of variance in the indicator explained by the common factors is often referred to as the *communality* (shown at the bottom of the SPSS output in Table 2.2). Thus, for the O1 (hopelessness) indicator, the factor model estimates that 68.5% of its total variance is *common variance* (variance explained by the latent variable of Depression), whereas the remaining 31.5% (i.e., 1 – .685 = .315) is *unique variance*. As

stated earlier, unique variance is some combination of specific factor and measurement error variance. It is important to note that EFA and CFA do not provide separate estimates of specific variance and error variance.

In addition, Table 2.2 provides selected output from the Mplus program. As would be expected, many of the results are identical to those generated by SPSS (e.g., eigenvalues, factor loadings). However, Mplus also provides other useful output, including an expanded set of goodness-of-fit statistics; standard errors and significance tests for the factor loadings (as well as for the residual variances, not shown in Table 2.2); and an estimate of factor determinacy (if requested by the user on the OUTPUT line; see Table 2.2). Each of these additional aspects of the Mplus output is discussed later in this book.

Path diagrams of the one-factor measurement model are provided in Figure 2.1. The first diagram presents the solution, using common symbols for the various elements of

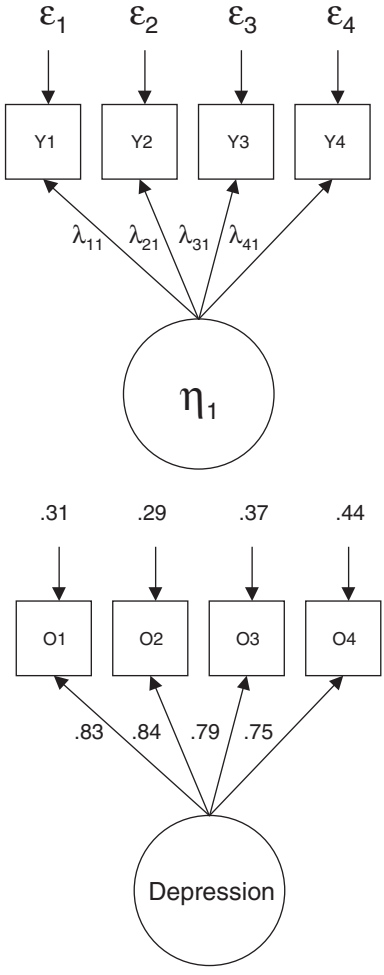


FIGURE 2.1. Path diagram of the one-factor model.

factor models (and LISREL latent Y variable notation); the second diagram replaces these elements with the sample estimates obtained from the EFA presented in Table 2.1. Following the conventions of factor analysis and structural equation modeling (SEM), the latent variable (factor) of Depression is depicted by a circle or an oval, whereas the four clinical ratings (indicators) are represented by squares or rectangles. The unidirectional arrows (→) represent the *factor loadings* (λ, or *lambda*), which are the regression slopes (direct effects) for predicting the indicators from the factor (η, or *eta*). These arrows are also used to relate the unique variances (ε, or *epsilon*) to the indicators.[4]

A fundamental equation of the common factor model is

$$y_j = \lambda_j 1\eta_1 + \lambda_j 2\eta_2 + \ldots + \lambda_{jm}\eta_m + \varepsilon_j \tag{2.1}$$

where y_j represents the jth of p indicators (in the case $p = 4$; O1, O2, O3, O4) obtained from a sample of n independent participants (in this case, $n = 300$); λ_{jm} represents the factor loading relating variable j to the mth factor η (in the case $m = 1$; the single factor of Depression); and ε_j represents the variance that is unique to indicator y_j and is independent of all ηs and all other εs. As will be seen in subsequent chapters, similar notation is used to represent some of the equations of CFA. In this simple factor solution entailing a single factor (η_1) and four indicators, the regression functions depicted in Figure 2.1 can be summarized by four separate equations:

$$\begin{aligned} O1 &= \lambda_{11}\eta_1 + \varepsilon_1 \\ O2 &= \lambda_{21}\eta_1 + \varepsilon_2 \\ O3 &= \lambda_{31}\eta_1 + \varepsilon_3 \\ O4 &= \lambda_{41}\eta_1 + \varepsilon_4 \end{aligned} \tag{2.2}$$

This set of equations can be summarized in a single equation that expresses the relationships among observed variables (y), factors (η), and unique variances (ε):

$$y = \Lambda_y \eta + \varepsilon \tag{2.3}$$

or in expanded matrix form:

$$\Sigma = \Lambda_y \Psi \Lambda'_y + \Theta\varepsilon \tag{2.4}$$

where Σ is the $p \times p$ symmetric correlation matrix of p indicators; Λ_y is the $p \times m$ matrix of factor loadings λ (in this case, a 4×1 vector); Ψ is the $m \times m$ symmetric correlation matrix of the factor correlations (1×1); and Θε is the $p \times p$ diagonal matrix of unique variances ε ($p = 4$). In accord with matrix algebra, matrices are represented in factor analysis and SEM by uppercase Greek letters (e.g., Λ, Ψ, and Θ), and specific elements of these matrices are denoted by lowercase Greek letters (e.g., λ, ψ, and ε). With minor variations, these fundamental equations can be used to calculate various aspects of the sample data from the factor analysis parameter estimates, such as the variances, covari-

ances, and means of the input indicators (the latter can be conducted in context of CFA with mean and covariance structures; see Chapter 7). For example, the following equation reproduces the variance in the O1 indicator:

$$\begin{aligned} \text{VAR(O1)} = \sigma_{11} &= \lambda_{11}^2\psi_{11} + \varepsilon_1 \\ &= .828^2(1) + .315 \\ &= 1.00 \end{aligned} \tag{2.5}$$

where ψ_{11} is the variance of the factor η_1, and ε_1 is the unique variance of O1. Note that both ψ_{11} and σ_{11} equal 1.00 because the EFA model is completely standardized; that is, when variables are standardized, their variances equal 1.00. Similarly, the model estimate of the covariance (correlation) of O1 and O2 can be obtained from the following equation:

$$\begin{aligned} \text{COV(O1, O2)} = \sigma_{21} &= \lambda_{11}\psi_{11}\lambda_{21} \\ &= (.828)(1)(.841) \\ &= .696 \end{aligned} \tag{2.6}$$

Because the solution is completely standardized, this covariance is interpreted as the factor model estimate of the sample correlation of O1 and O2. In other words, the model-implied correlation of the indicators is the product of their completely standardized factor loadings. Note that the sample correlation of O1 and O2 is .70, which is very close to the factor-model-implied correlation of .696. As discussed in further detail in Chapter 3, the acceptability of factor analysis models is determined in large part by how well the parameter estimates of the factor solution (e.g., the factor loadings) are able to reproduce the observed relationships among the input variables. The current illustration should exemplify the point made earlier that common variance (i.e., variance explained by the factors as reflected by factor loadings and communalities) is estimated on the basis of the shared variance among the indicators used in the analysis. EFA generates a matrix of factor loadings (Λ) that best explain the correlations among the input indicators.

PROCEDURES OF EFA

Although a full description of EFA is beyond the scope of this book, an overview of its concepts and procedures is helpful to make later comparisons to CFA. The reader is referred to papers by Fabrigar, Wegener, MacCallum, and Strahan (1999); Floyd and Widaman (1995); and Preacher and MacCallum (2003) for detailed guidelines on conducting EFA in applied data sets.

As stated earlier, the overriding objective of EFA is to evaluate the dimensionality of a set of multiple indicators (e.g., items from a questionnaire) by uncovering the smallest number of interpretable factors needed to explain the correlations among them. Whereas the researcher must ultimately specify the number of factors, EFA is an "explor-

atory" analysis because no a priori restrictions are placed on the pattern of relationships between the observed measures and the latent variables. This is a key difference between EFA and CFA. In CFA, a researcher must specify in advance several key aspects of the factor model (e.g., number of factors, patterns of indicator–factor loadings).

After determining that EFA is the most appropriate analytic technique for the empirical question at hand, the researcher must decide which indicators to include in the analysis, and determine if the size and the nature of the sample are suitable for the analysis (for more details on these issues, see Chapters 9 and 10). Other procedural aspects of EFA include (1) selection of a specific method to estimate the factor model; (2) selection of the appropriate number of factors; (3) in the case of models that have more than one factor, selection of a technique to rotate the initial factor matrix to foster the interpretability of the solution; and (4) if desired, selection of a method to compute factor scores.

Factor Extraction

There are many methods that can be used to estimate the common factor model, such as maximum likelihood (ML), principal factors (PL), weighted least squares, unweighted least squares, generalized least squares, imaging analysis, minimum residual analysis, and alpha factoring, to name just some. For EFA with continuous indicators (i.e., observed measures that approximate an interval-level measurement scale), the most frequently used factor extraction methods are ML and PF. ML is also the most commonly used estimation method in CFA, and its fundamental properties are discussed in Chapter 3. A key advantage of the ML estimation method is that it allows for a statistical evaluation of how well the factor solution is able to reproduce the relationships among the indicators in the input data. That is, how closely do the correlations among the indicators predicted by the factor analysis parameters approximate the relationships seen in the input correlation matrix (see Eq. 2.6)? This feature is very helpful for determining the appropriate number of factors. However, as discussed in Chapter 9, ML estimation requires the assumption of multivariate normal distribution of the variables. If the input data depart substantially from a multivariate normal distribution, important aspects of the results of ML-estimated EFA model can be distorted and not trustworthy (e.g., goodness of model fit, significance tests of model parameters).

Another potential disadvantage of ML estimation is its occasional tendency to produce *improper solutions*. An improper solution exists when a factor model does not converge on a final set of parameter estimates, or produces an *out-of-range* estimate such as an indicator with a communality above 1.0. On the other hand, PF has the strong advantages of being free of distributional assumptions and of being less prone to improper solutions than ML (Fabrigar et al., 1999). Unlike ML, PF does not provide goodness-of-fit indices useful in determining the suitability of the factor model and the number of latent variables. Thus PF might be preferred in instances where marked non-normality is evident in the observed measures, or perhaps when ML estimation produces an improper solution. However, as discussed later in this book, the presence of improper solutions may be a sign of more serious problems, such as a poorly specified

factor model or a poorly behaved input data matrix. If distributional assumptions hold, ML may be favored because of its ability to produce a wide range of fit indices that guide other important aspects of the factor analytic procedure. As noted in Chapter 3, ML is a full information estimator that provides standard errors that can be used for statistical significance testing and confidence intervals of key parameters such as factor loadings and factor correlations. Strategies for dealing with non-normal, continuous outcomes and categorical indicators are discussed in Chapter 9.

Although related to EFA, principal components analysis (PCA) is frequently mistaken to be an estimation method of common factor analysis. Unlike the estimators discussed in the preceding paragraphs (ML, PF), PCA relies on a different set of quantitative methods that are not based on the common factor model. PCA does not differentiate common and unique variance. Instead, PCA aims to account for the *variance* in the observed measures rather than explain the *correlations* among them. Thus PCA is more appropriately used as a data reduction technique to reduce a larger set of measures to a smaller, more manageable number of composite variables for use in subsequent analyses. However, some methodologists have argued that PCA is a reasonable or perhaps superior alternative to EFA, in view of the fact that PCA possesses several desirable statistical properties. For instance, it is computationally simpler; it is not susceptible to improper solutions; it often produces results similar to EFA; and PCA is able to calculate a participant's score on a principal component, whereas the indeterminate nature of EFA complicates such computations. Although debate on this issue continues, Fabrigar et al. (1999) provide several reasons in opposition to the argument for the place of PCA in factor analysis. These authors underscore the situations where EFA and PCA produce dissimilar results (e.g., when communalities are low or when there are only a few indicators of a given factor; cf. Widaman, 1993). Regardless, if the overriding rationale and empirical objectives of an analysis are in accord with the common factor model, then it is conceptually and mathematically inconsistent to conduct PCA; that is, EFA is more appropriate if the stated objective is to reproduce the intercorrelations of a set of indicators with a smaller number of latent dimensions, recognizing the existence of measurement error in the observed measures. Floyd and Widaman (1995) make the related point that estimates based on EFA are more likely to generalize to CFA than are those obtained from PCA (because, unlike PCA, EFA and CFA are based on the common factor model). This is a noteworthy consideration in light of the fact that EFA is often used as a precursor to CFA in scale development and construct validation. A detailed demonstration of the computational differences between PCA and EFA can be found in multivariate and factor analytic textbooks (e.g., Tabachnick & Fidell, 2013).

Factor Selection

Next, the factor analysis is run with the selected estimation method (e.g., ML, PF). The results of the initial analysis are used to determine the appropriate number of factors to be extracted in subsequent analyses. This is often considered to be the most crucial decision in EFA because *underfactoring* (selecting too few factors) and *overfactoring* (select-

ing too many factors) can severely compromise the validity of the factor model and its resulting estimates (e.g., can introduce considerable error in the factor loading estimates), although some research suggests that the consequences of overfactoring are less severe than those of underfactoring (cf. Fabrigar et al., 1999). Despite the fact that EFA is an exploratory or descriptive technique by nature, the decision about the appropriate number of factors should be guided by substantive considerations, in addition to the statistical guidelines discussed below. For instance, the validity of a given factor should be evaluated in part by its interpretability (e.g., does a factor revealed by the EFA have substantive importance?). A firm theoretical background and previous experience with the variables will strongly foster the interpretability of factors and the evaluation of the overall factor model. Moreover, factors in the solution should be well defined (i.e., composed of several indicators that strongly relate to it). Factors that are represented by two or three indicators may be underdetermined (have poor determinacy; see below) and highly unstable across replications. The solution should also be evaluated with regard to whether "trivial" factors exist in the data—for instance, factors based on differential relationships among indicators that stem from extraneous or methodological artifacts (e.g., method effects arising from subsets of very similarly worded or reverse-worded items; see Chapter 5).

It is also important to note that the number of factors (m) that can be extracted by EFA is limited by the number of observed measures (p) that are submitted to the analysis. The upper limit on the number of factors varies across estimation techniques. For instance, in EFA using PF, the maximum number of factors that can be extracted is $p - 1$.[5] In ML EFA, the number of elements in the input correlation or covariance matrix (a) must be equal to or greater than the number of parameters that are estimated in the factor solution (b) (i.e., $a \geq b$). As the number of factors (m) increases, so does the number of estimated parameters (b) in the solution. The fact that the maximum number of factors is mathematically limited by the input data can be problematic for ML analyses that use a small set of indicators; that is, the data may not support extraction of the number of factors that are posited to exist on conceptual grounds. For example, because only four observed measures ($p = 4$) are involved, it is possible to extract only one factor ($m = 1$) in the EFA presented in Table 2.2. Although a two-factor solution may be conceptually viable (e.g., Cognitive Depression: O1, O2; Somatic Depression: O3, O4), the number of parameters associated with a two-factor model (b) would exceed the number of pieces of information in the input correlation matrix (a). a and b can be readily calculated by the following equations:

$$a = [p * (p + 1)] / 2 \tag{2.7}$$

$$b = (p * m) + [(m * (m + 1)] / 2) + p - m^2 \tag{2.8}$$

where p is the number of observed variables (indicators), and m is the number of factors.

Solving for a indicates that input matrix contains 10 pieces of information (see Table 2.1), corresponding to the 6 correlations in the off-diagonal and the 4 standard-

ized variances on the diagonal; that is, $a = (4 * 5) / 2 = 10$. Solving for b (when $m = 1$) indicates that there are 8 parameters estimated in a one-factor solution; that is, $b = (4 * 1) + [(1 * 2) / 2] + 4 - 1 = 4 + 1 + 4 - 1 = 8$. Because the number of elements of the input matrix ($a = 10$) is greater than the number of parameter estimates ($b = 8$), a single factor can be extracted from the data. (As seen in Table 2.2, the degrees of freedom associated with the χ^2 fit statistic is 2, corresponding to the difference $a - b$, $10 - 8 = 2$; see Chapter 3.) However, two factors cannot be extracted because the number of parameters to be estimated in this model exceeds the number of elements of the input matrix by one; that is, $b = (4 * 2) + [(2 * 3) / 2] + 4 - 4 = 8 + 3 + 4 - 4 = 11$.

Each aspect of the equation used to solve for b corresponds to specific parameters and mathematical restrictions in the EFA model (cf. Eq. 2.4). The first aspect, ($p * m$), indicates the number of factor loadings (Λ_y). The second aspect, ($[m * (m + 1)] / 2$), indicates the number of factor variances and covariances (Ψ). The third aspect, p, corresponds to the number of residual variances (θ_ε). The final aspect, m^2, reflects the number of restrictions that are required to identify the EFA model (e.g., mathematically convenient restrictions, which include fixing factor variances to unity). For example, as depicted in Figure 2.1, in the one-factor model there are 4 factor loadings ($p * m$), 1 factor variance ($[m * (m + 1)] / 2$), and 4 indicator residuals (p); however, for identification purposes, the factor variance is fixed to 1.0 ($m^2 = 1^2 = 1$), and thus the model contains 8 estimated parameters. A two-factor solution would entail 8 factor loadings ($4 * 2$), 2 factor variances and 1 factor covariance [$(2 * 3) / 2$], and 4 residual variances (total number of parameters = 15). After subtracting the identifying restrictions ($m^2 = 2^2 = 4$; $15 - 4 = 11$), the number of parameters to be estimated in the two-factor model ($b = 11$) still exceeds the pieces in the input matrix ($a = 10$). Thus two factors cannot be extracted from the data by ML when $p = 4$.

Especially when an estimation procedure other than ML is used (e.g., PF), factor selection is often guided by the *eigenvalues* generated from either the *unreduced correlation matrix* (**R**; i.e., the input correlation matrix with unities—1.0s—in the diagonal) or the *reduced correlation matrix* ($\mathbf{R}_r$; i.e., the correlation matrix with communality estimates in the diagonal). For example, the selected SPSS output in Table 2.2 provides eigenvalues from the unreduced correlation matrix under the heading "Initial Statistics."[6] Most multivariate procedures such as EFA rely on eigenvalues and their corresponding *eigenvectors* because they summarize variance in a given correlation or variance–covariance matrix. The calculation of eigenvalues and eigenvectors is beyond the scope of this chapter (for an informative illustration, see Tabachnick & Fidell, 2013), but for practical purposes, it is useful to view eigenvalues as representing the variance in the indicators explained by the successive factors. This is illustrated in the final two sections of Table 2.2—specifically, the eigenvalue corresponding to the single factor that was extracted to account for the interrelationships of the four ratings of clinical depression. In the SPSS printout, this eigenvalue is listed under the heading "SS Loadings" and equals 2.579. Calculating the sum of squares of the four factor loadings (i.e., $.82822^2 + \ldots + .75228^2 = 2.579$) provides the eigenvalue for this factor. Dividing this eigenvalue by the total variance of the input matrix (because indicators are standardized, total

variance is equal to the number of input measures, *p*) yields the proportion of variance in the indicators that is accounted for by the factor model (i.e., 2.579 / 4 = .645), as also denoted under the heading "Pct of Var" (64.5%) in the "Final Statistics" section of the SPSS printout in Table 2.2.

The previous paragraph has discussed eigenvalues (e.g., 2.579) that are derived from the reduced correlation matrix ($\mathbf{R}_r$) produced by the EFA solution. The SPSS printout (Table 2.2) also presents eigenvalues for **R**, listed under the "Initial Statistics" heading (i.e., 2.93, .410, .359, .299). In line with the notion that eigenvalues communicate variance, note that the sum of the eigenvalues for **R** is 4 (i.e., total variance = number of input indicators, *p*). As was the case for eigenvalues associated with $\mathbf{R}_r$, dividing the eigenvalue by 4 yields an estimate of explained variance (e.g., 2.93 / 4 = .733; see Table 2.2). Thus eigenvalues guide the factor selection process by conveying whether a given factor explains a considerable portion of the total variance of the observed measures.

Three commonly used factor selection procedures are based on eigenvalues. They are (1) the Kaiser–Guttman rule, (2) the scree test, and (3) parallel analysis. The *Kaiser–Guttman rule* (also referred to as the *Kaiser criterion* or the *eigenvalues > 1.0 rule*) is very straightforward: (1) Obtain the eigenvalues derived from the input correlation matrix, **R** (as noted by Fabrigar et al., 1999, researchers occasionally make the mistake of using eigenvalues of the reduced correlation matrix, $\mathbf{R}_r$); (2) determine how many eigenvalues are greater than 1.0; and (3) use that number to determine the number of nontrivial latent dimensions that exist in the input data. As seen in the "Initial Statistics" section of the selected SPSS output provided in Table 2.2, a single eigenvalue from the input correlation matrix (**R**) is above 1.0 (i.e., 2.93); thus the Kaiser–Guttman rule suggests a unidimensional latent structure.

The logic of the Kaiser–Guttman rule is that when an eigenvalue is less than 1.0, the variance explained by a factor is less than the variance of a single indicator. Recall that eigenvalues represent variance, and that EFA standardizes both the latent and observed variables (e.g., the variance that each standardized input variable contributes to the factor extraction is 1.0). Thus, because a goal of EFA is to reduce a set of input indicators (the number of factors should be smaller than the number of input indicators), if an eigenvalue is less than 1.0, then the corresponding factor accounts for less variance than the indicator (whose variance equals 1.0). The Kaiser–Guttman rule has wide appeal (and in fact is the default in popular statistical software such as SPSS) because of its simplicity and objectivity. Nevertheless, many methodologists have criticized this procedure because it can result in either overfactoring or underfactoring, and because of its somewhat arbitrary nature. For example, sampling error in the input correlation matrix may result in eigenvalues of .99 and 1.01, but nonetheless the Kaiser–Guttman rule indicates that the latter is an important factor, whereas the former is not.

Another popular approach, called the *scree test* (Cattell, 1966), also uses the eigenvalues that can be taken from either the input or reduced correlation matrix (although Fabrigar et al., 1999, note reasons why scree tests based on $\mathbf{R}_r$ might be preferred). To provide a more realistic illustration of this procedure, a larger data set is used ($p = 20$). As shown in Figure 2.2, the scree test employs a graph in which the eigenvalues form

the vertical axis and the factors form the horizontal axis. The graph is inspected to determine the last substantial decline in the magnitude of the eigenvalues—or the point where lines drawn through the plotted eigenvalues change slope. A limitation of this approach is that the results of the scree test may be ambiguous (e.g., there is no clear shift in the slope) and open to subjective interpretation. This is evident in Figure 2.2, where the results can be interpreted as indicating either a four- or five-factor solution. However, as noted by Gorsuch (1983), the scree test performs reasonably well under conditions when the sample size is large and well-defined factors are present in the data (i.e., factors defined by multiple items with high communalities).

Another eigenvalue-based procedure for guiding factor selection is *parallel analysis* (Horn, 1965; Humphreys & Montanelli, 1975). The approach is based on a scree plot of the eigenvalues obtained from the sample data against eigenvalues that are estimated from a data set of random numbers (i.e., the means of eigenvalues produced by multiple sets of completely random data).[7] Both the observed sample and random data eigenvalues are plotted, and the appropriate number of factors is indicated by the point where the two lines cross. Thus factor selection is guided by the number of real eigenvalues greater than the eigenvalues generated from the random data; that is, if the "real" factor explains less variance than the corresponding factor obtained from random numbers, it should not be included in the factor analysis. The term *parallel analysis* refers to the fact that the random data set(s) should parallel aspects of the actual research data (e.g., sam-

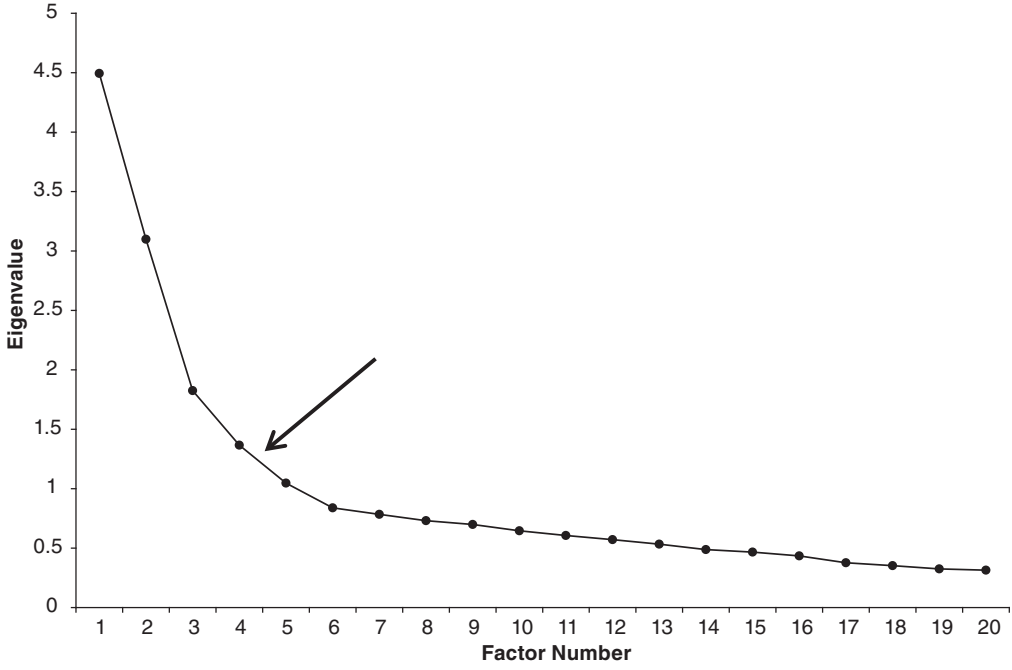


FIGURE 2.2. Scree test of eigenvalues from the unreduced correlation matrix. Arrow indicates region of curve where slope changes.

ple size, number of indicators). The rationale of parallel analysis is that the factor should account for more variance than is expected by chance (as opposed to more variance than is associated with a given indicator, according to the logic of the Kaiser–Guttman rule). Using the 20-item data set, parallel analysis suggests four factors (see Figure 2.3). After the eigenvalue for the fourth factor, the eigenvalues from the randomly generated data (averages of 50 replications) exceed the eigenvalues of the research data. Although parallel analysis frequently performs well, like the scree test it is sometimes associated with somewhat arbitrary outcomes (e.g., chance variation in the input correlation matrix may result in eigenvalues falling just above or below the parallel analysis criterion). A practical drawback of the procedure is that it is not available in major statistical software packages such as SAS and SPSS, although parallel analysis is an option in the Mplus and Stata software programs, and in various shareware programs found on the Internet (e.g., O'Connor, 2001). In addition, Hayton, Allen, and Scarpello (2004) have provided syntax for conducting parallel analysis in SPSS, although the user must save and summarize the eigenvalues generated from random data outside of SPSS.

As noted above, when a factor estimation procedure other than ML is employed, eigenvalue-based procedures such as the Kaiser–Guttman rule, the scree test, and par-

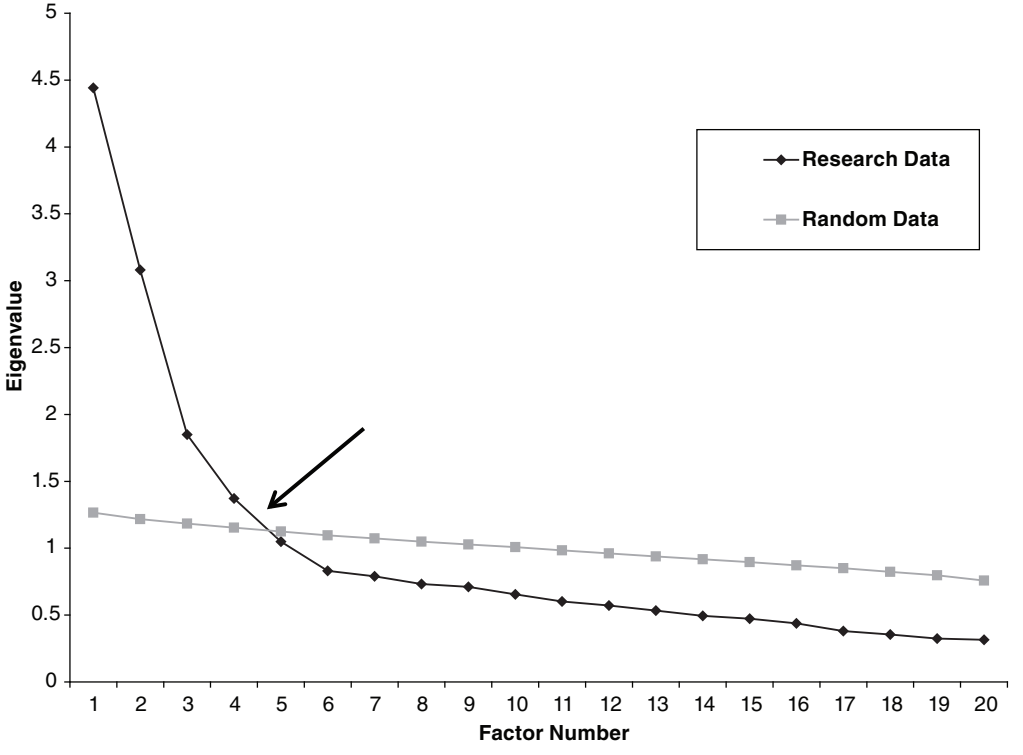


FIGURE 2.3. Parallel analysis using eigenvalues from research and random data (average of 50 replications). Arrow indicates that eigenvalues from random data exceed the eigenvalues from research data after the fourth factor.

allel analysis can be used to assist in factor selection. Although these methods can also assist in determining the appropriate number of factors in ML factor analysis, ML has the advantage of being a full information estimator that allows for goodness-of-fit evaluation and statistical inference such as significance testing and confidence interval estimation. ML is covered extensively in later chapters, so only a brief overview relevant to EFA is provided here. It is helpful to consider ML EFA as a special case of SEM. For example, like CFA and SEM, ML EFA provides goodness-of-fit information that can be used to determine the appropriate number of factors. Various goodness-of-fit statistics (such as χ^2 and the root mean square error of approximation, or RMSEA; Steiger & Lind, 1980) provide different pieces of information about how well the parameter estimates of the factor model are able to reproduce the sample correlations. As seen earlier in this chapter, the factor loadings of O1 and O2 yield a predicted correlation of .696 (i.e., Eq. 2.6), which is very similar to the correlation of these indicators in the sample data (i.e., .70; see correlation between O1 and O2 in Table 2.1). If the remaining observed relationships in the input matrix are reproduced as well by the factor loading estimates in this solution, descriptive fit statistics such as χ^2 and RMSEA will indicate that the one-factor model provides a good fit to the data. As shown in Table 2.2, the SPSS and Mplus output provides a χ^2 test of the fit of the one-factor solution (as seen in Table 2.2, Mplus also provides several other goodness-of-fit statistics that are discussed in Chapter 3). Because the χ^2 is statistically nonsignificant, $\chi^2(2) = .20$, $p = .90$, it can be concluded that the one-factor model provides a reasonable fit to the data. The nonsignificant χ^2 test suggests that the correlation matrix predicted by the factor model parameter estimates does not differ from the sample correlation matrix. However, it will be seen in Chapter 3 that χ^2 has serious limitations, and thus it should not be used as the sole index of overall model fit.

The goal of goodness-of-fit approaches is to identify the solution that reproduces the observed correlations considerably better than more parsimonious models (i.e., models involving fewer factors), but that is able to reproduce these observed relationships equally or nearly as well as more complex solutions (i.e., models with more factors). Accordingly, a researcher conducting ML EFA is apt to estimate the factor model several times (specifying different numbers of factors) to compare the fit of the solutions. As in other approaches (e.g., eigenvalue-based methods), factor selection should not be determined by goodness of fit alone, but should be strongly assisted by substantive considerations (e.g., prior theory and research evidence) and other aspects of the resulting solution. Although a factor solution may provide a reasonable fit to the data, it may be unacceptable for other reasons—such as the presence of factors that have no strong conceptual basis or utility (e.g., factors arising from methodological artifacts—see Chapter 5); poorly defined factors (e.g., factors in which only one or two indicators have strong primary loadings); indicators that do not have salient loadings on any factor; or indicators that have high loadings on multiple factors. Again, EFA is a largely exploratory procedure, but substantive and practical considerations should strongly guide the factor analytic process. Because of this and other issues (e.g., the role of sampling error), the results of an initial EFA should be interpreted cautiously and should be cross-validated (additional EFAs or CFAs should be conducted with independent data sets).

Factor Rotation

Once the appropriate number of factors has been determined, the extracted factors are rotated, to foster their interpretability. In instances when two or more factors are involved (rotation does not apply to one-factor solutions), rotation is relevant because of the indeterminate nature of the common factor model. That is, for any given multiple-factor model, there exist an infinite number of equally good-fitting solutions (each represented by a different factor loading matrix). The term *simple structure* was coined by Thurstone (1947) to refer to the most readily interpretable solutions, in which (1) each factor is defined by a subset of indicators that load highly on the factor; and (2) each indicator (ideally) has a high loading on one factor (often referred to as a *primary loading*) and has a trivial or close to zero loading on the remaining factors (referred to as a *cross-loading* or *secondary loading*). In applied research, factor loadings greater than or equal to .30 or .40 are often interpreted as *salient* (i.e., the indicator is meaningfully related to a primary or secondary factor), although explicit or widely accepted guidelines do not exist, and the criteria for salient and nonsalient loadings often depend on the empirical context. Thus, for models that contain two or more factors (where an infinite number of equally fitting solutions is possible), rotation is conducted to produce a solution with the best simple structure. It is important to emphasize that rotation does not alter the fit of the solution; for example, in ML EFA, model χ^2 is the same before and after factor rotation. Rather, factor rotation is a mathematical transformation (i.e., rotation in multidimensional space) that is undertaken to foster interpretability by maximizing larger factor loadings closer to one and minimizing smaller factor loadings closer to zero. For a mathematical demonstration of this procedure, the reader is referred to Comrey and Lee (1992).

There are two major types of rotation: *orthogonal* and *oblique*. In orthogonal rotation, the factors are constrained to be uncorrelated (i.e., factors are oriented at 90° angles in multidimensional space); in oblique rotation, the factors are allowed to intercorrelate (i.e., to permit factor axis orientations of less than 90°). The correlation between two factors is equal to the cosine of the angle between the rotational axes. Because cos(90) = 0, the factors are uncorrelated in orthogonal rotation. In oblique rotations, the angle of the axis is allowed to be greater or less than 90°, and thus the cosine of the angle may yield a factor correlation between zero and one.

In applied social sciences research, orthogonal rotation is used most often, perhaps because it has historically been the default in major statistical programs such as SPSS (*varimax rotation*), and researchers have traditionally perceived that orthogonally rotated solutions are more easily interpreted because the factor loadings represent correlations between the indicators and the factors (e.g., squaring the factor loadings provides the proportion of variance in the indicator that the factor solution explains). In oblique solutions, factor loadings usually do not reflect simple correlations between the indicators and the factors unless the factors themselves have no overlap. Because oblique rotations allow the factors to intercorrelate, the correlations between indicators and factors may be inflated by the covariation of the factors; that is, an indicator may correlate with one factor in part through its correlation with another factor. However, orthogonal rota-

tion may produce misleading solutions in situations where the factors are expected to be intercorrelated (e.g., a questionnaire whose latent structure entails several interrelated dimensions of a broader construct). In other words, although substantial correlations may exist among factors, orthogonal rotation constrains the solution to yield uncorrelated latent variables.

Thus, in most cases, oblique rotation is preferred because it provides a more realistic representation of how factors are interrelated. If the factors are in fact uncorrelated, oblique rotation will produce a solution that is virtually the same as one produced by orthogonal rotation. On the other hand, if the factors are interrelated, oblique rotation will yield a more accurate representation of the magnitude of these relationships. In addition, estimation of factor correlations provides important information, such as the existence of redundant factors or a potential higher-order structure. Factor intercorrelations above .80 or .85 may imply poor discriminant validity, and suggest that a more parsimonious solution could be obtained (see Chapter 5). If all factors in the solution are moderately intercorrelated at roughly the same magnitude, a single higher-order factor may account for these relationships (see Chapter 8). Moreover, when EFA is used as a precursor to CFA (see Chapter 5), oblique solutions are more likely to generalize to CFA than orthogonal solutions (i.e., constraining factors to be uncorrelated in CFA will typically result in poor model fit).

Several forms of oblique rotation have been developed (e.g., promax, geomin, quartamin, orthooblique). When oblique rotation is requested, most software programs (such as SPSS) output both a *pattern matrix* and a *structure matrix*. The loadings in the pattern matrix convey the *unique* relationship between a factor and an indicator. They are interpreted in the same fashion as partial regression coefficients in standard multiple regression; that is, the coefficient represents the relationship between the predictor (factor) and outcome (indicator), while controlling for the influence of all other predictors (other factors). Thus indicator variance that is explained by more than one factor is omitted from the loadings in the pattern matrix. The structure matrix is calculated by multiplying the pattern matrix by the factor correlation matrix (oblique rotation produces a factor correlation matrix, but orthogonal rotation does not). Hence loadings in the structure matrix reflect both the unique relationship between the indicator and factor (as in the pattern matrix) *and* the relationship between the indicator and the shared variance among the factors. In other words, the loadings in the structure matrix reflect a zero-order relationship between the indicator and a given factor, without holding the other factors in the solution constant. Unless the correlations among factors are minimal, loadings in the structure matrix will typically be larger than those in the pattern matrix, because they are inflated by the overlap in the factors (akin to zero-order correlations vs. partial regression coefficients in standard multiple regression). Although there is some debate about whether the pattern matrix or structure matrix should be used, the pattern matrix is by far more often interpreted and reported in applied research. In fact, some popular latent variable software programs only provide the pattern matrix. As noted above, the mathematical operations for generating the structure matrix are quite straightforward when the pattern matrix and factor intercorrelation are available. Thus,

with the aid of software (e.g., SAS PROC IML), either matrix can be readily computed on the basis of the other (i.e., structure matrix = pattern matrix multiplied by the factor correlation matrix; pattern matrix = structure matrix multiplied by the inverse of factor correlation matrix).

The factor intercorrelations produced by an oblique rotation of a PCA solution are often lower than those of obliquely rotated solutions based on the common factor model (e.g., PF, ML; Fabrigar et al., 1999; Widaman, 1993). This is because in common factor analysis, random error is removed from the factors. In PCA, random error is included in the components (i.e., PCA does not differentiate common and unique variance). Thus another potential adverse consequence of PCA is a mistaken conclusion that components share modest variance when in fact the intercorrelations have been attenuated by random error (or a conclusion that components are distinct when in fact the error-disattenuated correlations would be above .80). Because factor correlations arising from common factor analysis are more likely to be closer to population values, this is another reason why methodologists usually advocate EFA over PCA.

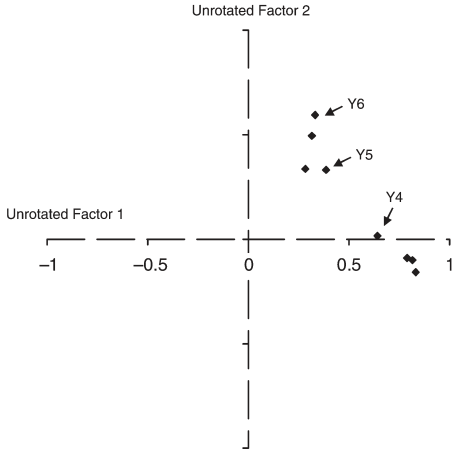
Factor rotation is illustrated in Figure 2.4, using a real data set of eight indicators collected from a sample of 500 participants. A scree test and parallel analysis suggest a two-factor solution. Results indicate that the first four indicators (Y1–Y4) load on Factor 1 and the remaining four indicators (Y5–Y8) load on Factor 2. Figure 2.4 displays a geometric representation of unrotated, orthogonally rotated (varimax), and obliquely rotated (promax) factor matrices. ML estimation produces the unrotated factor loadings presented in Figure 2.4A. Figure 2.4B shows the results of the varimax rotation. The factor axes remain at 90° angles, but are rotated in the most optimal fashion to maximize high factor loadings and minimize low loadings. Rotation produces a transformation matrix. With matrix algebra, the unrotated factor loading matrix is multiplied by the transformation matrix to produce the rotated factor loading matrix. In this data set, the varimax transformation matrix is as follows:

	Factor 1	Factor 2
Factor 1	.93347	.35867
Factor 2	–.35867	.93347

These values convey how much the axes are rotated to foster simple structure. Specifically, the values on the diagonal (.93347) are cosines, and the values on the off-diagonal (.35867, –.35867) are sines and –sines. As shown in Figure 2.4B, the axes are rotated 21° to better transect the clusters of indicators. Within rounding error, the cos(21) equals .933 and the sin(19) equals .359, the same as the transformation coefficients shown above. Because orthogonal rotation is used, the axes of Factor 1 and Factor 2 remain at right angles, and thus the factors are constrained to be uncorrelated; that is, cos(90) = 0.

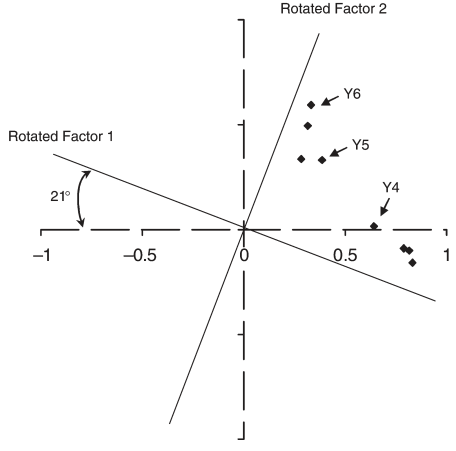
To witness the effects of rotation on maximizing and minimizing factor loadings, consider the fifth indicator, Y5. Before rotation, the loadings of Y5 on Factor 1 and Factor 2 are very similar (.386 and .329, respectively; Figure 2.4A). A 21° rotation of the

A. Unrotated Factor Matrix



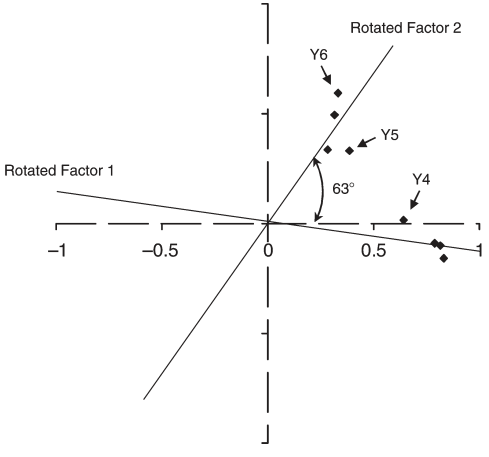
	Factor	
	1	2
Y1	.834	–.160
Y2	.813	–.099
Y3	.788	–.088
Y4	.642	.015
Y5	.386	.329
Y6	.333	.593
Y7	.313	.497
Y8	.284	.336

B. Orthogonally Rotated Factor Matrix (Varimax)



	Factor	
	1	2
Y1	.836	.150
Y2	.794	.199
Y3	.767	.201
Y4	.594	.244
Y5	.242	.445
Y6	.098	.673
Y7	.114	.576
Y8	.145	.416

C. Obliquely Rotated Factor Matrix (Promax)



	Factor	
	1	2
Y1	.875	–.062
Y2	.817	.003
Y3	.788	.012
Y4	.588	.106
Y5	.154	.418
Y6	–.059	.704
Y7	–.018	.595
Y8	.055	.413

FIGURE 2.4. Geometric representations of unrotated, orthogonally rotated, and obliquely rotated factor matrices.

factor axes raises Y5's position on the Factor 2 axis (.445), and decreases this indicator's position on the Factor 1 axis (.242) (Figure 2.4B). Although this transformation fosters the interpretability of the solution, it does not alter the communality of Y5 or any other indicator. In a solution entailing more than one latent variable, communalities in an orthogonal EFA are calculated by taking the sum of squared loadings for a given indicator across all factors.[8] Before and after rotation, the proportion of variance explained in Y5 is .257; unrotated solution: $.386^2 + .329^2 = .257$, rotated solution: $.242^2 + .445^2 = .257$. Thus rotation does not alter the fit of the factor solution.

Figure 2.4B also suggests that oblique rotation may be more appropriate. A factor solution is best defined when the indicators are clustered around the upper end of their respective factor axes. The higher up the axis, the higher the factor loading; if an indicator is in close proximity to one factor axis, it does not load highly on another factor. As shown in Figure 2.4B, orthogonal rotation moves the Factor 2 axis closer to the Y5–Y8 indicators. As a result, the rotation has an overall effect of increasing the primary loadings of Y5–Y8 on Factor 2, and decreasing their cross-loadings on Factor 1 (compared to the unrotated solution, Figure 2.4A). However, orthogonal rotation moves the Factor 1 axis away from the Y1–Y4 indicators, which has the general effect of increasing the magnitude of the cross-loadings of Y1–Y4 on Factor 2 (e.g., Y4: .015 vs. .244 for the unrotated and rotated solutions, respectively). Indeed, in instances where all the indicators fall in between the factor axes after orthogonal rotation (as seen in Figure 2.4B), the restriction of maintaining a 90° orientation of the factor axes may not be tenable. Oblique rotations such as promax begin with an orthogonal rotation, but then "break" the 90° angle to allow the factor axes to pass through the clusters of indicators better. The angle of the factor axes reflects the factor correlation. If factors are uncorrelated, the angle of factor axes will remain close to 90°. If the factors are correlated, the angle of factor axes will deviate from 90°.

Figure 2.4C provides a geometric depiction of oblique rotation (promax) of the two-factor solution. Note that the axes of Factor 1 and Factor 2 are both turned inward somewhat to better transect the two clusters of indicators. Compared to orthogonal rotation, the oblique rotation increases the values of most primary loadings further. A more notable impact of oblique rotation is its success at moving the cross-loadings closer to zero (but, as before, the overall fit of the solution is the same). To accomplish this, the angle of the factor axes is shifted from 90° to 63° (see Figure 2.4C). The results of the analysis indicate that the correlation of Factor 1 and Factor 2 is .45. This corresponds to the cosine of the factor angle; that is, cos(63) = .45.

Factor Scores

After an appropriate factor solution has been established, a researcher may wish to calculate factor scores by using the factor loadings and factor correlations. Factor scores are used for various purposes (e.g., to serve as proxies for latent variables, to determine a participant's relative standing on the latent dimension). Conceptually, a factor score is the score that would have been observed for a person if it had been possible to measure

the factor directly. In applied research, factor scores are often computed by creating *coarse factor scores*, which are simple unweighted composites of the raw scores of indicators (e.g., averaging or summing) found to have salient loadings on the factor. However, there are many reasons why coarse factor scores may poorly represent factors (e.g., they may be highly intercorrelated even when the factors are truly orthogonal; Grice, 2001). Alternatively, factor scores can be estimated by multivariate methods that use various aspects of the reduced or unreduced correlation matrix and factor analysis coefficients. The resulting values are called *refined factor scores*. A frequently used method of estimating refined factor scores is Thurstone's (1935) least squares regression approach, although several other strategies have been developed (e.g., Bartlett, 1937; Harman, 1976; McDonald, 1981). Most statistical software packages provide options to compute refined factor scores by one or more of these methods. In the majority of instances, refined factor scores have less bias than coarse factor scores and thus are favored over coarse factor scores as proxies for the factors (Grice, 2001). However, a complicating issue in factor score estimation is the indeterminate nature of the common factor model. With respect to factor scores, this indeterminacy means that an infinite number of sets of factor scores can be computed from any given factor analysis that will be equally consistent with the same factor loadings (Grice, 2001). The degree of indeterminacy depends on several aspects, such as the ratio of items to factors and the size of the item communalities (e.g., factors defined by several items with strong communalities have better determinacy). If a high degree of indeterminacy is present, the sets of factor scores can vary so widely that an individual ranked high on the dimension in one set may receive a low ranking on the basis of another set. In such scenarios, the researcher has no way of discerning which set of scores or rankings is most accurate.

Thus, although typically neglected in applied factor analytic research, the degree of factor score indeterminacy should be examined as part of EFA, especially in instances when factor scores are to be computed for use in subsequent statistical analyses. Grice (2001) has specified three criteria for evaluating the quality of factor scores: (1) *validity coefficients*, or correlations between the factor score estimates and their respective factor scores; (2) *univocality*, or the extent to which the factor scores are excessively or insufficiently correlated with other factors in the same analysis; and (3) *correlational accuracy*, or how closely the correlations among factor scores correspond to the correlations among the factors. For instance, Gorsuch (1983) has recommended that validity coefficients should be at least .80, although higher values (e.g., >.90) may be required in some situations (e.g., when factor scores are used as dependent variables). Unfortunately, procedures for evaluating factor scores are not standard options in most software packages. As shown in Table 2.2, an exception is the Mplus program, where validity coefficients can be requested as part of EFA and CFA by using the FSDETERMINACY option of the OUTPUT command (the validity coefficient is .94, which indicates an acceptable level of factor score determinacy; see also Chapter 5). In addition, Grice (2001) has developed SAS PROC IML computer code for assessing the degree of factor score indeterminacy (validity coefficients, univocality, correlational accuracy) in the context of EFA (this syntax can be downloaded from *http://psychology.okstate.edu/faculty/jgrice/factorscores*).

SUMMARY

Procedural recommendations for conducting applied EFA are summarized in Table 2.3. In addition to providing a practical overview of EFA (e.g., procedural considerations for factor estimation, selection, rotation, and interpretation), the goal of this chapter has been to introduce key concepts that are carried forward in the subsequent chapters on CFA (e.g., observed vs. latent variables, factor loadings, factor correlations, common and unique variance, basic equations and notation). Some fundamental differences of EFA and CFA have been described. Unlike EFA, in CFA the number of factors and the pattern of indicator–factor loadings are specified in advance on the basis of strong empirical knowledge or theory. The acceptability of the CFA model is evaluated in part by descriptive fit statistics that convey the ability of the solution to reproduce the observed relationships among the input indicators (although similar testing can be applied in EFA when the ML estimator is used). As will be seen in Chapter 3, EFA and CFA differ in several other important manners.

NOTES

1. Exceptions to this rule are discussed in Chapters 3 and 4 (e.g., when indicator measurement errors are correlated).
2. As discussed in Chapter 5, hybrid models that combine the features of EFA and CFA have been developed recently (e.g., exploratory structural equation modeling).
3. For instance, in the current example, which entails a single factor, the factor loadings can be interpreted as zero-order correlation coefficients between the factor and the observed measures (i.e., factor loading = the standardized regression slope = zero-order correlation). In solutions involving multiple, correlated factors (oblique rotation), factor loadings from the factor pattern matrix are interpreted as partial regression coefficients.
4. The reader will encounter many variations in this notational system across factor analysis and SEM texts. For instance, because indicator unique variances ($\boldsymbol{\varepsilon}$) are not observed, it is common to see these parameters depicted as circles in path diagrams. In Chapter 3, this notation is expanded by differentiating latent X (exogenous) and latent Y (endogenous) solutions.
5. In PCA, the limit on the number of components is equal to p.
6. Because eigenvalues are drawn from the unreduced correlation matrix (**R**), PCA is always conducted initially, regardless of the type of factor analysis requested (e.g., PF).
7. Some researchers (e.g., Glorfeld, 1995) have recommended that the 95th percentile of eigenvalues from random data be used in place of average eigenvalues, in part to adjust for parallel analysis's slight tendency to overfactor (regardless of the method used, research has shown that parallel analysis is accurate in the vast majority of cases; e.g., Humphreys & Montanelli, 1975; Zwick & Velicer, 1986).
8. Although communalities can also be hand-calculated from the estimates of an obliquely rotated EFA solution, this computation is less straightforward, because the factors are permitted to be intercorrelated and thus the factor loadings are partial regression coefficients. Later chapters (e.g., Chapter 3) discuss the tracing rules necessary to compute these estimates.

TABLE 2.3. Fundamental Steps and Procedural Recommendations for EFA

Factor extraction

- Use an estimator based on the common factor model, such as:
  - *Principal factors*: No distributional assumptions; less prone to improper solutions than maximum likelihood
  - *Maximum likelihood*: Assumes multivariate normality, but provides goodness-of-fit evaluation and, in some cases, significance tests and confidence intervals of parameter estimates

Factor selection

- Determine the appropriate number of factors by:
  - *Scree plot* of eigenvalues from the reduced correlation matrix,
  - *Parallel analysis*, and/or
  - *Goodness of model fit* (e.g., χ^2, RMSEA; see Chapter 3)

Factor rotation

- In multifactorial models, rotate the solution to obtain simple structure by:
  - Using an *oblique rotation* method (e.g., promax, geomin)

Interpret the factors and evaluate the quality of the solution

- Consider the meaningfulness and interpretability of the factors:
  - Factors should have substantive meaning and conceptual/empirical relevance
  - Rule out nonsubstantive explanations such as method effects (e.g., factors composed of reverse- and non-reverse-worded items; see Chapters 3 and 5)
- Eliminate poorly defined factors, such as:
  - Factors on which only two or three items have salient loadings
  - Factors defined by items that have small loadings (i.e., *low communalities*)
  - Factors with low *factor determinacy* (poor correspondence between the factors and their factor scores; see Grice, 2001)
- Eliminate poorly behaved items (indicators), such as:
  - Items with high loadings on more than one factor (i.e., *cross-loadings*)
  - Items with small loadings on all factors (i.e., *low communalities*)

Rerun and (ideally) replicate the factor analysis

- If items or factors are dropped in preceding step, rerun the EFA in the same sample
- Replicate the final EFA solution in an independent sample
- Consider further replications/extensions of the factor solution by:
  - Developing tentative CFA models (e.g., exploratory SEM; see Chapter 5)
  - Larger-scale CFA investigations
  - Measurement invariance evaluation in population subgroups (e.g., equivalence of solution between sexes; see Chapter 7)

Note. EFA, exploratory factor analysis; RMSEA, root mean square error of approximation; CFA, confirmatory factor analysis; SEM, structural equation modeling.