tion of the computational differences between PCA and EFA can be found in multivariate and factor analytic textbooks (e.g., Tabachnick & Fidell, 2001).

Factor Selection

Next, the factor analysis is run using 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" (selecting too many factors) can severely compromise the validity of the factor model and its resulting estimates (e.g., 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; for example, 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—that is, comprised 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 $p - 1$.[4] In ML EFA, the number of parameters that are estimated in the factor solution (a) must be equal to or less than the number of elements (b) in the input correlation or covariance matrix (i.e., $a \leq b$). As the number of factors (m)

increases, so does the number of estimated parameters (a) 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 ($p = 4$) were involved, it was 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: D1, D2; Somatic Depression: D3, D4), 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:

$$a = (p * m) + [(m * (m + 1)] / 2) + p - m^2 \quad (2.7)$$

$$b = [(p * (p + 1)] / 2 \quad (2.8)$$

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

Solving for b indicates that the input matrix contains 10 pieces of information (see Table 2.1), corresponding to the 6 correlations in the off-diagonal and the 4 standardized variances on the diagonal; that is, $b = (4 * 5) / 2 = 10$. Solving for a (when $m = 1$) indicates that there are 8 parameters that are estimated in a one-factor solution; that is, $a = (4 * 1) + [(1 * 2) / 2)] + 4 - 1 = 4 + 1 + 4 - 1 = 8$. Because the number of elements of the input matrix ($a = 10$) is greater than the number of parameters ($b = 8$), 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, $a = (4 * 2) + [(2 * 3) / 2)] + 4 - 4 = 8 + 3 + 4 - 4 = 11$.

Each aspect of the equation used to solve for a 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 four factor loadings ($p * m$), one factor variance ($[m * (m + 1)] / 2$), and four 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 eight estimated parameters. A two-factor solution would entail eight factor loadings ($4 * 2$), two factor variances and one factor covariance [$(2 * 3) / 2$], and four 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."[5] 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, 2001), 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. On 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 discussed eigenvalues (e.g., 2.579) that were 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**) was above 1.0 (i.e., 2.93); thus, the Kaiser–Guttman rule would suggest 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 latent 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 because of its simplicity and objectivity; in fact, it is the default in popular statistical software packages such as SPSS. 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 would indicate 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 $\mathbf{R}_r$ 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 whereby 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., no clear shift in the slope) and open to subjective interpretation. This is evident in Figure 2.2 where the results could 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 such as when the sample size is large and when well-defined factors are present in the data.

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

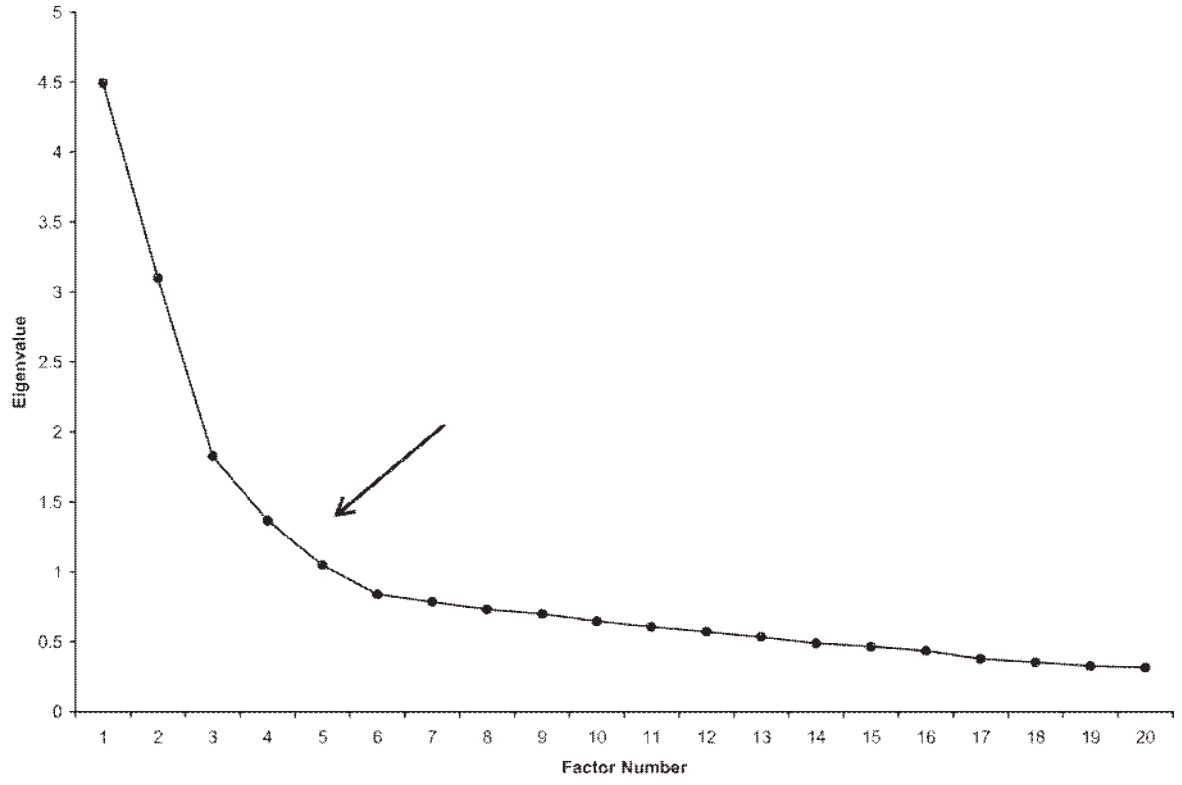

FIGURE 2.2. Scree test of eigenvalues from the unreduced correlation matrix. Arrow indicates region of curve where slope changes.

completely random data).[6] 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., sample 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, per 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

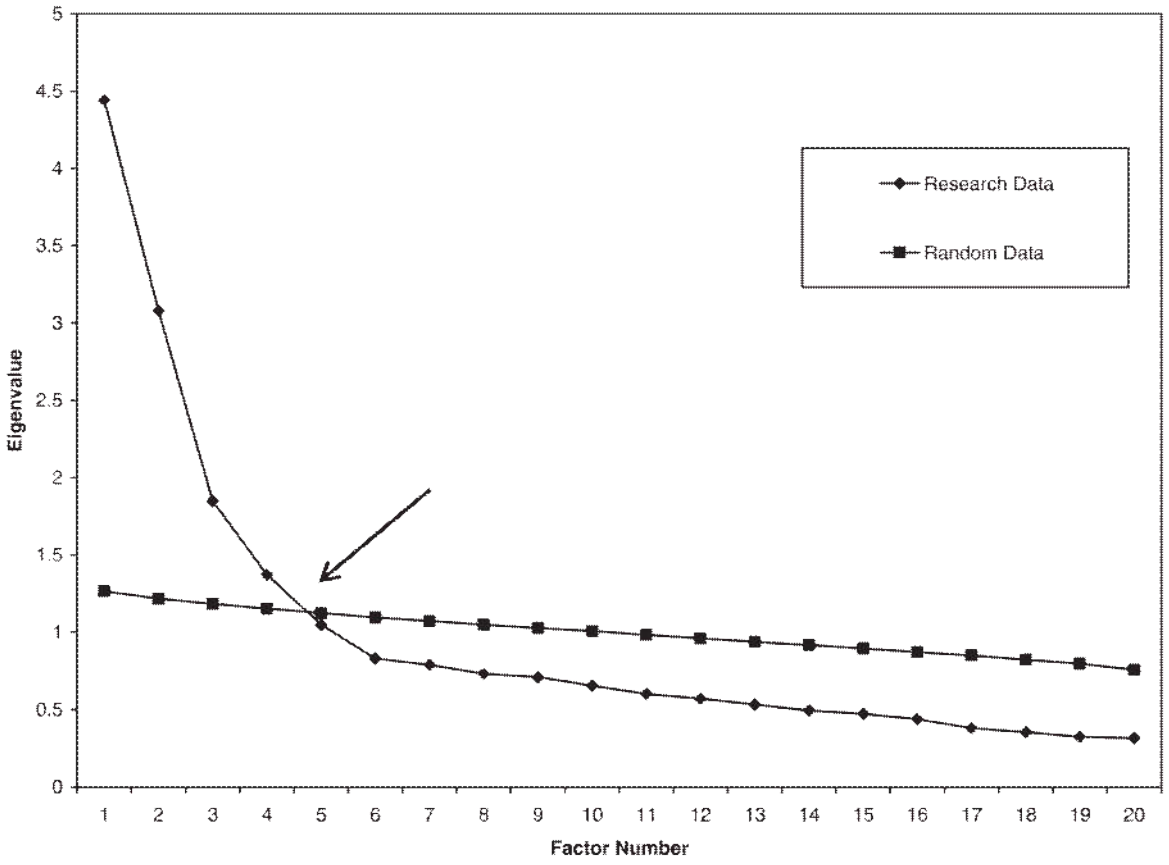

FIGURE 2.3. Parallel analysis using eigenvalues from research and random data (average of 50 replications). Arrow indicates that eigenvalues from random data exceed the eigenvalues from research data after the fourth factor.

scree test it is sometimes associated with somewhat arbitrary outcomes; for instance, 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 *Stata* software and 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 application of the Kaiser–Guttman rule, the scree test, and parallel 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 of approximation, RMSEA; Steiger & Lind, 1980) provide different pieces of information about how well the parameters of the factor model are able to reproduce the sample correlations. As seen earlier in this chapter, the factor loadings of D1 and D2 yielded 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 D1 and D2 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 the χ^2 and RMSEA will indicate that the one-factor model provided a good fit to the data. As shown in Table 2.2, the SPSS output provides a χ^2 test of the fit of the one-factor solution. Because the χ^2 was statistically nonsignificant, $\chi^2(2) = .20$, $p = .90$, it could be concluded that the one-factor model provides a reasonable fit to the data. The nonsignificant χ^2 test suggests the correlation matrix predicted by the factor model parameters 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 is able to reproduce these observed relationships equally or nearly as well as more complex solutions (i.e., models with more factors). Accordingly, the 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 might 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 largely an 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 using 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 possible 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*; that is, the indicator is mean-