including variable transformations, may markedly enhance [illegible]
confirmatory purposes. However, the limitations apply with greater force to [illegible]

13.3.2.1 Sample Size and Missing Data

Correlation coefficients tend to be less reliable when estimated from small samples. Therefore, it is important that sample size be large enough that correlations are reliably estimated. The required sample size also depends on magnitude of population correlations and number of factors: if there are strong, reliable correlations and a few, distinct factors, a smaller sample size is adequate.

Comrey and Lee (1992) give as a guide sample sizes of 50 as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 as excellent. *As a general rule of thumb, it is comforting to have at least 300 cases for factor analysis.* Solutions that have several high loading marker variables (>.80) do not require such large sample sizes (about 150 cases should be sufficient) as solutions with lower loadings and/or fewer marker variables (Guadagnoli and Velicer, 1988).

If cases have missing data, either the missing values are estimated, the cases deleted, or a missing data (pairwise) correlation matrix is analyzed. Consult Chapter 4 for methods of finding and estimating missing values and cautions about pairwise deletion of cases. Consider the distribution of missing values (is it random?) and remaining sample size when deciding between estimation and deletion. If cases are missing values in a nonrandom pattern or if sample size becomes too small, estimation is in order. However, beware of using estimation procedures (such as regression) that are likely to overfit the data and cause correlations to be too high. These procedures may "create" factors.

13.3.2.2 Normality

As long as PCA and FA are used descriptively as convenient ways to summarize the relationships in a large set of observed variables, assumptions regarding the distributions of variables are not in force. If variables are normally distributed, the solution is enhanced. To the extent that normality fails, the solution is degraded but may still be worthwhile.

However, multivariate normality is assumed when statistical inference is used to determine the number of factors. Multivariate normality is the assumption that all variables, and all linear combinations of variables, are normally distributed. Although tests of multivariate normality are overly sensitive, *normality among single variables is assessed by skewness and kurtosis* (see Chapter 4 and Section 13.7.1.2). If a variable has substantial skewness and kurtosis, variable transformation is considered.

13.3.2.3 Linearity

Multivariate normality also implies that relationships among pairs of variables are linear. The analysis is degraded when linearity fails, because correlation measures linear relationship and does not reflect nonlinear relationship. *Linearity among pairs of variables is assessed through inspection of scatterplots.* Consult Chapter 4 and Section 13.7.1.3 for methods of screening for linearity. If nonlinearity is found, transformation of variables is considered.

13.3.2.4 Absence of Outliers among Cases

As in all multivariate techniques, cases may be outliers either on individual variables (univariate) or on combinations of variables (multivariate). Such cases have more influence on the factor solution

than other cases. Consult Chapter 4 and Section 13.7.1.4 for methods of detecting and reducing the influence of both univariate and multivariate outliers.

13.3.2.5 Absence of Multicollinearity and Singularity

In PCA, multicollinearity is not a problem because there is no need to invert a matrix. For most forms of FA and for estimation of factor scores in any form of FA, singularity or extreme multicollinearity is a problem. For FA, if the determinant of **R** and eigenvalues associated with some factors approach 0, multicollinearity or singularity may be present.

To investigate further, look at the SMCs for each variable where it serves as DV with all other variables as IVs. If any of the SMCs is one, singularity is present; if any of the SMCs is very large (near one), multicollinearity is present. Delete the variable with multicollinearity or singularity. Chapter 4 and Section 13.7.1.5 provide examples of screening for and dealing with multicollinearity and singularity.

13.3.2.6 Factorability of R

A matrix that is factorable should include several sizable correlations. The expected size depends, to some extent, on *N* (larger sample sizes tend to produce smaller correlations), but if no correlation exceeds .30, use of FA is questionable because there is probably nothing to factor analyze. *Inspect* **R** *for correlations in excess of .30 and, if none is found, reconsider use of FA.*

High bivariate correlations, however, are not ironclad proof that the correlation matrix contains factors. It is possible that the correlations are between only two variables and do not reflect underlying processes that are simultaneously affecting several variables. For this reason, it is helpful to examine matrices of partial correlations where pairwise correlations are adjusted for effects of all other variables. If there are factors present, then high bivariate correlations become very low partial correlations. SPSS and SAS produce partial correlation matrices.

Bartlett's (1954) test of sphericity is a notoriously sensitive test of the hypothesis that the correlations in a correlation matrix are zero. The test is available in SPSS FACTOR but because of its sensitivity and its dependence on *N*, the test is likely to be significant with samples of substantial size even if correlations are very low. Therefore, use of the test is recommended only if there are fewer than, say, five cases per variable.

Several more sophisticated tests of the factorability of **R** are available through SPSS and SAS. Both programs give significance tests of correlations, the anti-image correlation matrix, and Kaiser's (1970, 1974) measure of sampling adequacy. Significance tests of correlations in the correlation matrix provide an indication of the reliability of the relationships between pairs of variables. If **R** is factorable, numerous pairs are significant. The anti-image correlation matrix contains the negatives of partial correlations between pairs of variables with effects of other variables removed. If **R** is factorable, there are mostly small values among the off-diagonal elements of the anti-image matrix. Finally, Kaiser's measure of sampling adequacy is a ratio of the sum of squared correlations to the sum of squared correlations plus sum of squared partial correlations. The value approaches 1 if partial correlations are small. Values of .6 and above are required for good FA.

13.3.2.7 Absence of Outliers among Variables

After FA, in both exploratory and confirmatory FA, variables that are unrelated to others in the set are identified. These variables are usually not correlated with the first few factors although they often

correlate with factors extracted later. These factors are usually unreliable, both because they account for very little variance and because factors that are defined by just one or two variables are not stable. Therefore, one never knows whether these factors are "real." Suggestions for determining reliability of factors defined by one or two variables are in Section 13.6.2.

If the variance accounted for by a factor defined by only one or two variables is high enough, the factor is interpreted with great caution or ignored, as pragmatic considerations dictate. In confirmatory FA, the factor represents either a promising lead for future work or (probably) error variance, but its interpretation awaits clarification by more research.

A variable with a low squared multiple correlation with all other variables and low correlations with all important factors is an outlier among the variables. The variable is usually ignored in the current FA and either deleted or given friends in future research. Screening for outliers among variables is illustrated in Section 13.7.1.7.

13.4 Fundamental Equations for Factor Analysis

Because of the variety and complexity of the calculations involved in preparing the correlation matrix, extracting factors, and rotating them, and because, in our judgment, little insight is produced by demonstrations of some of these procedures, this section does not show them all. Instead, the relationships between some of the more important matrices are shown, with an assist from SPSS FACTOR for underlying calculations.

Table 13.1 lists many of the important matrices in FA and PCA. Although the list is lengthy, it is composed mostly of *matrices of correlations* (between variables, between factors, and between variables and factors), *matrices of standard scores* (on variables and on factors), *matrices of regression weights* (for producing scores on factors from scores on variables), and the *pattern matrix* of unique relationships between factors and variables after oblique rotation.

Also in the table are the matrix of eigenvalues and the matrix of their corresponding eigenvectors. Eigenvalues and eigenvectors are discussed here and in Appendix A, albeit scantily, because of their importance in factor extraction, the frequency with which one encounters the terminology, and the close association between eigenvalues and variance in statistical applications.

A data set appropriate for FA consists of numerous subjects each measured on several variables. A grossly inadequate data set appropriate for FA is in Table 13.2. Five subjects who were trying on ski boots late on a Friday night in January were asked about the importance of each of four variables to their selection of a ski resort. The variables were cost of ski ticket (COST), speed of ski lift (LIFT), depth of snow (DEPTH), and moisture of snow (POWDER). Larger numbers indicate greater importance. The researcher wanted to investigate the pattern of relationships among the variables in an effort to understand better the dimensions underlying choice of ski area.

Notice the pattern of correlations in the correlation matrix as set off by the vertical and horizontal lines. The strong correlations in the upper left and lower right quadrants show that scores on COST and LIFT are related, as are scores on DEPTH and POWDER. The other two quadrants show that scores on DEPTH and LIFT are unrelated, as are scores on POWDER and LIFT, and so on. With