4

Data Preparation and Psychometrics Review

Just as in other types of statistical analyses, data preparation is critical in SEM for three reasons. First, it is easy to make a mistake entering data into computer files. Second, the most widely used estimation methods in SEM make specific distributional assumptions about the data. These assumptions must be taken seriously because violation of them could result in bias. Third, data-related problems can make SEM computer tools fail to yield a logical solution. A researcher who has not carefully prepared and screened the data could mistakenly believe that the model is at fault, and confusion ensues. How to select good measures is also considered along with review of basic psychometric issues, including the evaluation of score reliability and validity. It is not possible to cover all aspects of data screening or psychometrics in a single chapter, but more advanced works are cited throughout, and they should be consulted for more information. This adage attributed to Abraham Lincoln sets the tone for this chapter: If I had eight hours to chop a tree, I'd spend six sharpening my axe.

FORMS OF INPUT DATA

Most primary researchers—those who conduct original studies—input raw data files for analysis with SEM computer programs. Just as in multiple regression, raw data themselves are not necessary for many—and perhaps most—types of SEM. For example, when analyzing continuous outcomes with default maximum likelihood estimation, a matrix of summary statistics instead can be the input to an SEM computer tool instead of a raw data file. In fact, you can replicate most of the analyses described in this book using the data matrix summaries that accompany them—see the website for this book. This is a great way to learn because you can make mistakes using someone else's data before analyzing your own. Many journal articles about the results of SEM contain

enough information, such as correlations and standard deviations, to create a matrix summary of the data, which can then be submitted to a computer program for analysis. Thus, readers of these works can, with no access to the raw data, replicate the original analyses or estimate alternative models not considered in the original work.

Basically, all SEM computer tools accept either a raw data file or a matrix summary of the data. If a raw data file is submitted, the program will create its own matrix, which is then analyzed. You should consider the following issues when choosing between a raw data file and a matrix summary as program input:

1. *Some special types of analyses require raw data files.* There are three basic kinds. One is when continuous outcomes have severely non-normal distributions and the data are analyzed with a method that assumes normality, but test statistics and standard errors are calculated that adjust for non-normality. A second situation concerns missing data. You should know that default maximum likelihood estimation does not handle incomplete raw data files. But special versions of the maximum likelihood method are available in many SEM computer tools that analyze incomplete data sets. The third case is when outcome variables are not continuous, that is, they are ordinal or nominal variables. Such outcomes can be analyzed in SEM, but raw data files are needed. For analyses that do not involve any of these applications, either the raw data or a matrix summary of them can be analyzed.

2. *Matrix input offers a potential economy over raw data files.* Suppose that 1,000 cases are measured on 10 continuous variables. The data file may be 1,000 lines (or more) in length, but a matrix summary for the same data might be only 10 lines long.

3. *Sometimes a researcher might "make up" a data matrix using theory or results from a meta-analysis, so there are no raw data, only a matrix summary.* A made-up data matrix can be submitted to an SEM computer tool for analysis. This is also a way to diagnose certain kinds of technical problems that can crop up in SEM. This point is elaborated in later chapters.

If means are not analyzed, there are two basic summaries of raw data—correlation matrices with standard deviations and covariance matrices. For example, presented in the top part of Table 4.1 are the correlation matrix with standard deviations (left) and the covariance matrix (right) for the raw data in Table 2.1 on three continuous variables. Whenever possible, at least four-decimal accuracy is recommended for matrix input. Precision at this level helps to minimize rounding error in the analysis. All summary matrices in Table 4.1 are in **lower diagonal form** where only the unique values of correlations or covariances are reported in the lower-left-hand side of the matrix. Most SEM computer tools accept lower diagonal matrices as alternatives to full ones, with redundant entries above and below the diagonal, and can "assemble" a covariance matrix given the correlations and standard deviations. Exercise 1 asks you to reproduce the covariance matrix in the upper right part of Table 4.1 from the correlations and standard deviations in the upper left part of the table.

TABLE 4.1. Matrix Summaries of the Data in Table 2.1

X	*W*	*Y*	*X*	*W*	*Y*
Summaries without means					
	r, *SD*			cov	
1.0000			9.0421		
.2721	1.0000		2.3053	7.9368	
.6858	.4991	1.0000	22.4158	15.2842	118.1553
3.0070	2.8172	10.8699			
Summaries with means					
	r, *SD*, *M*			cov, *M*	
1.0000			9.0421		
.2721	1.0000		2.3053	7.9368	
.6858	.4991	1.0000	22.4158	15.2842	118.1553
3.0070	2.8172	10.8699	16.9000	49.4000	102.9500
16.9000	49.4000	102.9500			

It may be problematic to submit for analysis just a correlation matrix without standard deviations, specify that all standard deviations equal 1.0 (which standardizes everything), or convert the raw scores to normal deviates (*z* scores) and then submit for analysis the data file of standardized scores. This is because most estimation methods in SEM, including default maximum likelihood estimation, assume that the variables are unstandardized. This implies that if a correlation matrix without the original standard deviations is analyzed, the results may not be correct. Potential problems include the derivation of incorrect standard errors for standardized estimates if special methods for standardized variables are not used. Some SEM computer programs give warning messages or terminate the run if the researcher requests the analysis of a correlation matrix only with default maximum likelihood estimation. Thus, it is generally safer to analyze a covariance matrix (or a correlation matrix with standard deviations). Accordingly, covariances are analyzed for most examples in this book. When a correlation matrix is analyzed, I use a special estimation method for standardized variables described in Chapter 11. The issues just discussed about the pitfalls of analyzing correlation matrices without standard deviations explain why you must clearly state in written reports the specific kind of data matrix analyzed and the estimation method used.

Matrix summaries of raw data must consist of the covariances and means whenever means are analyzed in SEM. Presented in the lower part of Table 4.1 are matrix summaries of the data in Table 2.1 that include the correlations, standard deviations, and means (left) and the covariances and means (right). Both matrices convey the same information. Even if your analyses do not concern means, you should nevertheless report the means of all variables. You may not be interested in analyzing means, but someone else

may be. Always report sufficient descriptive statistics (including the means) so that others can reproduce your results.

POSITIVE DEFINITENESS

The data matrix that you submit—or the one calculated by the computer from your raw data—to an SEM computer program should be **positive definite**, which is required for most estimation methods. A matrix that lacks this characteristic is **nonpositive definite**; therefore, attempts to analyze such a data matrix would probably fail. A positive definite data matrix has the properties summarized next and then discussed:

1. The matrix is **nonsingular** or has an inverse. A matrix with no inverse is **singular**.
2. All eigenvalues of the matrix are positive (> 0), which also says that the matrix determinant is positive.
3. There are no out-of-bounds correlations or covariances.

In most kinds of multivariate analyses (SEM included), the computer needs to derive the inverse of the data matrix as part of its linear algebra operations. If the matrix has no inverse, these operations fail. An **eigenvalue** is the variance of an **eigenvector**, and both are from a principal components analysis of the data matrix, or **eigendecomposition**, that creates a total of v orthogonal linear combinations, or eigenvectors, of the observed variables, where v is the total number of those variables. The maximum number of eigenvectors for a data matrix equals v, and the set of all possible eigenvectors explains all the variance of the original variables.

If any eigenvalue equals zero, then (1) the matrix is singular, and (2) there is some pattern of perfect collinearity that involves at least two variables (e.g., $r_{XY} = 1.0$) or three or more variables in a more complex configuration (e.g., $R_{Y.X,W} = 1.0$). Perfect collinearity means that some denominators in matrix calculations will be zero, which results in illegal (undefined) fractions (estimation fails). Near-perfect collinearity, such as $r_{XY} = .95$, manifested as near-zero eigenvalues, can also cause this problem.

Negative eigenvalues (< 0) may indicate a data matrix element—a correlation or covariance—that is **out of bounds**. Such an element would be mathematically impossible to derive if all elements were calculated from the same cases with no missing data. For example, the value of the Pearson correlation between two variables X and Y is limited by the correlations between these variables and a third variable W. Specifically, the value of r_{XY} must fall within the range defined next:

$$(r_{XW} \times r_{WY}) \pm \sqrt{(1 - r_{XW}^2)(1 - r_{WY}^2)} \tag{4.1}$$

Given $r_{XW} = .60$ and $r_{WY} = .40$, for example, the value of r_{XY} must fall within the range

.24 ± .73, or from –.49 to .97

Any other value of r_{XY} would be out of bounds. (You should verify this result using Equation 4.1.) Another way to view Equation 4.1 is that it specifies a **triangle inequality** for values of correlations among three variables measured in the same sample.[1]

In a positive definite data matrix, the maximum absolute value of cov_{XY}, the covariance between *X* and *Y*, must respect the limit defined next:

$$\max \left| \text{cov}_{XY} \right| \leq \sqrt{s_X^2 \times s_Y^2} \tag{4.2}$$

where s_X^2 and s_Y^2 are, respectively, the sample variances of *X* and *Y*. In words, the maximum absolute value for the covariance between two variables is less than or equal to the square root of the product of their variances; otherwise, the value of cov_{XY} is out of bounds. For example, given

$$\text{cov}_{XY} = 13.00,\ s_X^2 = 12.00, \text{ and } s_Y^2 = 10.00$$

the covariance between *X* and *Y* is out of bounds because

$$13.00 > \sqrt{12.00 \times 10.00} = 10.95$$

which violates Equation 4.2. The value of r_{XY} for this example is also out of bounds because it equals 1.19 (an impossible result), given these variances and covariance. Exercise 2 asks you to verify this fact.

The **determinant** of the data matrix is the serial product (the first times the second times the third, and so on) of the eigenvalues. Assuming that all eigenvalues are positive, the determinant is a kind of matrix variance. Specifically, it is the volume of the multivariate space "mapped" by the set of observed variables.[2] If any eigenvalue equals zero, then the determinant is zero; in this case, the matrix has no inverse (it is singular). A close-to-zero eigenvalue will probably make the determinant be close to zero, which signals the potential inability of the computer to derive the inverse. If some odd number of the eigenvalues (1 or 3 or 5, etc.) is negative, then the determinant will be negative, too. A data matrix with a negative determinant may have an inverse, but the whole matrix is still nonpositive definite, perhaps due to out-of-bounds correlations or covariances. See Topic Box 4.1 for more information about causes of nonpositive definiteness in the data matrix and possible solutions.

Before analyzing in SEM either a raw data file or a matrix summary, the original data file should be screened for the problems considered next. Some of these potential

[1] In a geometric triangle, the length of a given side must be less than the sum of the lengths of the other two sides but greater than the difference between the lengths of the two sides.

[2] For diagrams, see *http://en.wikipedia.org/wiki/Determinant*

TOPIC BOX 4.1

Causes of Nonpositive Definiteness and Solutions

Many points summarized here are from Wothke (1993) and Rigdon (1997). Some causes of nonpositive definite data matrices are listed next. Most can be detected through data screening.

1. Extreme bivariate or multivariate collinearity among the observed variables.
2. The presence of outliers that force the values of correlations to be extremely high.
3. Pairwise deletion of cases with missing data.
4. Making a typing mistake when transcribing a data matrix from one source, such as a table in a journal article, to another, such as a command file for computer analysis, can result in a nonpositive definite data matrix. For example, if the value of a covariance in the original matrix is 15.00, then mistakenly typing 150.00 in the transcribed matrix could generate a nonpositive definite matrix.
5. Plain old sampling error can generate nonpositive definite data matrices, especially in small or unrepresentative samples.
6. Sometimes matrices of estimated Pearson correlations, such as polyserial or polychoric correlations derived for noncontinuous observed variables, can be nonpositive definite.

Here are some tips about diagnosing whether a data matrix is positive definite before submitting it for analysis to an SEM computer tool: Copy the full matrix (with redundant entries above and below the diagonal) into a text (ASCII) file, such as Microsoft Windows Notepad. Next, point your Internet browser to a free, online matrix calculator and then copy the data matrix into the proper window on the calculating webpage. Finally, select options on the calculating webpage to derive the determinant and eigenvalues with the corresponding eigenvectors. Look for outcomes that indicate nonpositive definiteness, such as near-zero, zero, or negative eigenvalues. A handy matrix calculator is available at *www.bluebit.gr/matrix-calculator.*

Suppose that the covariances among continuous variables *X*, *W*, and *Y*, respectively, are

$$\begin{bmatrix} 1.00 & & \\ .30 & 2.00 & \\ .65 & 1.15 & .90 \end{bmatrix} \quad \textbf{(I)}$$

The eigenvalues for this matrix (**I**) derived using the online matrix calculator just mentioned are

(2.918, .982, 0)

The third eigenvalue is zero, so let us inspect the weights for the third eigenvector, which for *X*, *W*, and *Y*, respectively, are

(–.408, –.408, .816)

Some other online matrix calculators report the eigenvector weights as –1, –1, 2, but these values are proportional to the weights just reported. None of these weights equals zero, so all three variables are involved in perfect collinearity. The pattern for these data is

$$R_{Y \cdot X,\ W} = R_{W \cdot X,\ Y} = R_{X \cdot Y,\ W} = 1.0$$

To verify this pattern, I used the SPPS syntax listed next to automatically convert the covariance matrix for this example to a correlation matrix:

```
comment convert covariance matrix to correlation matrix.
matrix data variables=rowtype_ x w y/format=full.
begin data
cov 1.00 .30 .65
cov .30 2.00 1.15
cov .65 1.15 .90
end data.
mconvert.
```

The correlation matrix (**II**) for *X*, *W*, and *Y*, respectively, in lower diagonal form is

$$\begin{bmatrix} 1.0 & & \\ .2121 & 1.0 & \\ .6852 & .8572 & 1.0 \end{bmatrix} \quad \textbf{(II)}$$

Given these correlations, you should verify that $R_{Y \cdot X,\ W} = R_{W \cdot X,\ Y} = R_{X \cdot Y,\ W} = 1.0$.

The LISREL program offers an option for **ridge adjustment**, which multiplies the diagonal entries by a constant > 1.0 until negative eigenvalues disappear (the matrix becomes positive definite). These adjustments increase the variances until they are large enough to exceed any out-of-bounds covariance entry in the off-

diagonal part of the matrix. This technique "fixes up" a data matrix so that necessary algebraic operations can be performed (Wothke, 1993), but parameter estimates, standard errors, and fit statistics are biased after ridge adjustment. A better solution is to try to solve the problem of nonpositive definiteness through data screening.

There are other contexts where you may encounter nonpositive definite matrices in SEM, but these generally concern (1) matrices of parameter estimates for your model or (2) matrices of correlations or covariances predicted from your model. A problem is indicated if any of these matrices is nonpositive definite. We will deal with these contexts in later chapters.

difficulties are causes of nonpositive definite data matrices, but others concern distributional assumptions for continuous outcomes.

EXTREME COLLINEARITY

Extreme collinearity can occur because what appear to be separate variables actually measure the same thing. Suppose that X measures accuracy and Y measures speed. If $r_{XY} = .95$, for instance, then X and Y are redundant notwithstanding their different labels (speed is accuracy, and vice versa). Either one or the other could be included in the same analysis, but not both. Researchers can inadvertently cause extreme collinearity when composites and their constituents are analyzed together. Suppose that a questionnaire has 10 items and the total score is summed across the items. Although the bivariate correlation between the total score and each of the individual items may not be high, the multiple correlation between the total score and the items must equal 1.0, which is multivariate collinearity in its most extreme.

Methods to detect collinearity among three or more continuous variables are summarized next. Most of these methods are available in regression diagnostics procedures of computer programs for general statistical analyses:

1. Calculate R^2 between each variable and all the rest. The observation that $R^2 > .90$ for a particular variable analyzed as the criterion suggests extreme multivariate collinearity.
2. A related statistic is **tolerance**, or $1 - R^2$, which indicates the proportion of total standardized variance that is unique. Tolerance values < .10 may indicate extreme multivariate collinearity.
3. Another statistic is the **variance inflation factor** (VIF), or $1/(1 - R^2)$, the ratio of the total standardized variance over the proportion of unique variance (tolerance). The variable in question may be redundant, if VIF > 10.0.

There are two basic options for dealing with extreme collinearity: eliminate variables or combine redundant ones into a composite. For example, if X and Y are highly correlated, one could be dropped or their scores could be averaged or summed to form a single new variable, but note that this new variable must replace both X and Y in the analysis. Extreme collinearity can also happen between latent variables when their estimated correlation is so high that they are not distinct. Later we will consider this type of extreme collinearity in the technique of confirmatory factor analysis (CFA).

OUTLIERS

Outliers are scores that are very different from the rest. A **univariate outlier** is a score that is extreme on a single variable. There is no single definition of "extreme," but one heuristic is that scores more than three standard deviations beyond the mean may be outliers. Univariate outliers are easy to find by inspecting frequency distributions of z scores (e.g., $|z| > 3.0$ indicates an outlier). But this method is susceptible to distortion by the very outliers that it is supposed to detect; that is, it is not robust. Suppose that scores for five cases are

19, 25, 28, 32, and 10,000

The last score (10,000) is obviously an outlier, but it so distorts the mean and standard deviation for all scores that the $|z| > 3.0$ rule fails, also called **masking**:

$$M = 2{,}020.80 \quad SD = 4{,}460.51 \quad \text{and} \quad z = \frac{10{,}000 - 2{,}020.80}{4{,}460.51} = 1.79$$

A more robust decision rule for detecting univariate outliers is

$$\frac{|X - Mdn|}{1.483\ (\text{MAD})} > 2.24 \tag{4.3}$$

where Mdn designates the sample median—which is more robust against outliers than the mean—and MAD is the **median absolute deviation** (MAD) of all scores from the sample median. The quantity MAD does not estimate the population standard deviation σ, but the product of MAD and the scale factor 1.43 is an unbiased estimator of σ in a normal distribution. The value of the ratio in Equation 4.3 is the distance between a score and the median expressed in robust standard deviation units. The constant 2.24 in Equation 4.3 is the square root of the approximate 97.5th percentile in a central χ^2 distribution with a single degree of freedom. A potential outlier thus has a score on the ratio in Equation 4.3 that exceeds 2.24. For the five scores in the example, Mdn = 28.00, and the absolute values of median deviations are, respectively,

9.00, 3.00, 0, 4.00, and 9,972.00

The median of the deviations just listed is MAD = 4.00, and so for X = 10,000 we calculate

$$\frac{9{,}972.00}{1.483\ (4.00)} = 1{,}681.05 > 2.24$$

which obviously detects the score of 10,000 as an outlier. Wilcox (2012) describes additional robust outlier detection methods.

A **multivariate outlier** has extreme scores on two or more variables, or a pattern of scores that is atypical. For example, a case may have scores between two and three standard deviations above the mean on all variables. Although no individual score might be considered extreme, the case could be a multivariate outlier if this pattern is unusual. Here are some options for detecting multivariate outliers without extreme individual scores:

1. Some SEM computer programs, such as Amos and EQS, identify cases that contribute the most to multivariate non-normality, and such cases may be multivariate outliers.
2. Calculate for each case its squared **Mahalanobis distance**, D_M^2, which indicates the distance in variance units between the profile of scores for that case and the vector of sample means, or **centroid**, correcting for intercorrelations.

In large samples with normal distributions, D_M^2 is distributed as central χ^2 with degrees of freedom equal to the number of variables, or v. A relatively high D_M^2 with a low p value in the corresponding $\chi^2(v)$ distribution may lead to the rejection of the null hypothesis that the case comes from the same population as the rest. A conservative level of statistical significance is usually recommended for this test, such as .001. Some standard computer procedures for multiple regression can automatically calculate and save D_M^2 values to the raw data file.

Let us assume that an outlier is not due to a data entry error (e.g., 99 was entered instead of 9) or the failure to specify a missing data code (e.g., –9) in the data editor of a statistics computer program; that is, the outlier is a valid score. Now, what to do with the outlier? One possibility is that the case does not belong to the population from which the researcher intended to sample. Suppose that a senior graduate student audits an undergraduate class in which a questionnaire is administered. The auditing student is from a different population, and his or her questionnaire responses may be extreme compared with those of classmates. If it is determined that a case with outliers is not from the same population, then it is best to delete that case; otherwise, there are ways to reduce the influence of extreme-but-valid scores if they are retained. One option is to convert extreme scores to a value that equals the next most extreme score that is within three standard deviations of the mean. Another is to apply a mathematical transformation to a variable with outliers. Transformations are considered later in this chapter.

NORMALITY

The default estimation method in SEM, maximum likelihood, assumes **multivariate normality** (**multinormality**) for continuous outcome variables. This means that

1. all the individual univariate distributions are normal;
2. all joint distributions of any pair of variables is bivariate normal; that is, each variable is normally distributed for each value of every other variable; and
3. all bivariate scatterplots are linear with homoscedastic residuals.

Because it is often impractical to examine all joint frequency distributions, it can be difficult to assess all aspects of multivariate normality. There are significance tests intended to detect violation of multivariate normality, including Mardia's (1985) test, but all such tests have limited usefulness. One reason is that slight departures from multivariate normality could be significant in large samples, and power in small samples may be low, so larger departures could be missed. Fortunately, many instances of multivariate non-normality are detectable through inspection of univariate frequency distributions.

Skew and kurtosis are two ways that a univariate distribution can be non-normal, and they can occur either separately or together in the same variable. Skew implies that the shape of a unimodal distribution is asymmetrical about its mean. **Positive skew** indicates that most of the scores are below the mean, and **negative skew** indicates just the opposite. Presented in Figure 4.1(a) are examples of distributions with either positive skew or negative skew compared with a normal curve. For a unimodal, symmetric distribution, **positive kurtosis** indicates heavier tails and a higher peak and **negative kurtosis** indicates just the opposite, both relative to a normal curve with the same variance. A distribution with positive kurtosis is described as **leptokurtic**, and a distribution with negative kurtosis is described as **platykurtic**. Presented in Figure 4.1(b) are examples of distributions with either positive kurtosis or negative kurtosis compared with a normal curve. Skewed distributions are generally leptokurtic, which means that remedies for skew also may fix kurtosis. Blest (2003) describes a kurtosis measure that adjusts for skewness.

Extreme skew is easy to detect through inspection of frequency distributions or histograms. Two other types of displays helpful for spotting skew are **stem-and-leaf plots** and **box plots** (**box-and-whisker plots**). For example, presented in the left side of Figure 4.2 is a stem-and-leaf plot for $N = 64$ scores. The lowest and highest scores are, respectively, 10 and 27. The latter score is an outlier ($z > 5.0$). In the stem-and-leaf plot, the numbers to the left side of the vertical line ("stems") represent the "tens" digit of each score, and each number to the right ("leaf") represents the "ones" digit. The shape of the stem-and-leaf plot indicates positive skew.

Presented in the right side of Figure 4.2 is a box plot for the same scores. The bottom and top borders of the rectangle in a box plot correspond to, respectively, the 25th percentile (1st quartile) and the 75th percentile (3rd quartile). The line inside the rect-

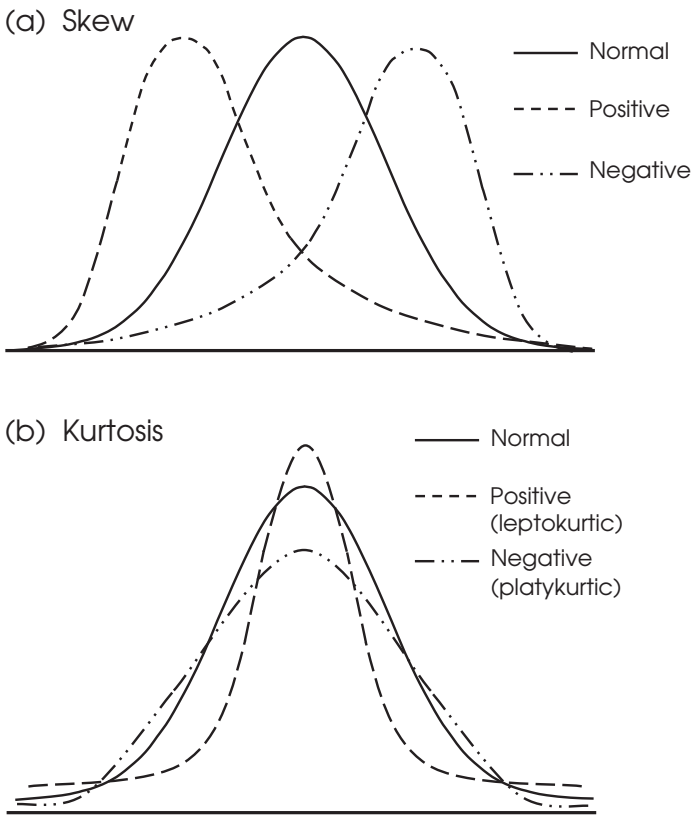

FIGURE 4.1. Distributions with (a) positive skew or negative skew and with (b) positive kurtosis or negative kurtosis relative to a normal curve.

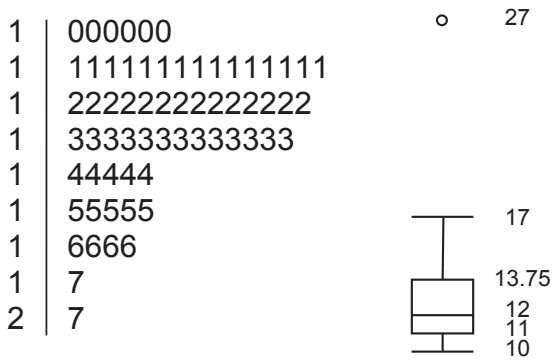

FIGURE 4.2. A stem-and-leaf plot (left) and a box plot (right) for the same distribution (N = 64).

angle of a box plot represents the median (50th percentile, or 2nd quartile). The "whiskers" are the vertical lines that connect the first and third quartiles with, respectively, the lowest and highest scores that are not extremes, or outliers. The length of the whiskers shows how far nonextreme scores spread away from the median. Skew is indicated in a box plot if the median line does not fall within the center of the rectangle or if the "whiskers" have unequal lengths. In the box plot of Figure 4.2, the high score of 27 is extreme and thus is represented in the box plot as a single open circle above the upper "whisker." The box plot in the figure indicates positive skew because there is greater spread of scores above the median than below the median.

Kurtosis is harder to spot by eye when inspecting frequency distributions, stem-and-leaf plots, or box plots, especially in distributions that are more or less symmetrical. Departures from normality due to skew or kurtosis may be apparent in **normal probability plots**, in which data are plotted against a theoretical normal distribution in such a way that points should approximate a straight line. The distribution is otherwise non-normal, but it is hard to discern the degree of non-normality due to skew or kurtosis apparent in normal probability plots. An example of a normal probability plot is presented later in this chapter.

Fortunately, there are more precise measures of skew and kurtosis. Perhaps the best known standardized measures of these characteristics that permit comparison of different distributions to the normal curve are the **skew index** ($\hat{\gamma}_1$) and **kurtosis index** ($\hat{\gamma}_2$), which are calculated as follows:

$$\hat{\gamma}_1 = \frac{S^3}{(S^2)^{3/2}} \quad \text{and} \quad \hat{\gamma}_2 = \frac{S^4}{(S^2)^2} - 3.0 \tag{4.4}$$

where S^2, S^3, and S^4 are, respectively, the second through fourth **moments about the mean**:

$$S^2 = \frac{\sum (X - M)^2}{N}, \quad S^3 = \frac{\sum (X - M)^3}{N}, \quad \text{and} \quad S^4 = \frac{\sum (X - M)^4}{N} \tag{4.5}$$

The sign of $\hat{\gamma}_1$ indicates the direction of the skew, positive or negative, and a value of zero indicates a symmetrical distribution. The value of $\hat{\gamma}_2$ in a normal distribution equals zero, and its sign indicates the type of kurtosis, positive or negative.

The ratio of either $\hat{\gamma}_1$ or $\hat{\gamma}_2$ over its standard error is interpreted in large samples as a z test of the null hypothesis that there is no population skew or kurtosis. These tests may not be helpful in large samples because even slight departures from normality could be statistically significant, and low power in small samples means that appreciable skew or kurtosis can go undetected. Significance testing with $\hat{\gamma}_1$ or $\hat{\gamma}_2$ is not generally helpful in data screening. An alternative is to interpret absolute values of $\hat{\gamma}_1$ or $\hat{\gamma}_2$, but there are few clear-cut standards for doing so. Some guidelines can be offered based on computation simulation studies of estimation methods used in SEM (e.g., Nevitt & Hancock, 2000). Variables where $|\hat{\gamma}_1| > 3.0$ are described as "severely" skewed by some authors of these studies. There is less consensus about $\hat{\gamma}_2$, for which absolute values from about 8.0 to 20.0 have been described as indicating "severe" kurtosis. A conservative rule of

thumb, then, seems to be that $|\hat{\gamma}_2| > 10.0$ suggests a problem and $|\hat{\gamma}_2| > 20.0$ indicates a more serious one. For the data in Figure 4.2, $\hat{\gamma}_1 = 3.10$ and $\hat{\gamma}_2 = 15.73$, so the distribution is severely non-normal by the standards just suggested. Do not conclude that a distribution is normal, if $|\hat{\gamma}_1| \leq 3.0$ and $|\hat{\gamma}_2| \leq 10.0$. This is because $\hat{\gamma}_1 = \hat{\gamma}_2 = 0$ in a true normal distribution; otherwise, the only thing that can be reasonably said is that the shape of the distribution may not be severely non-normal.

TRANSFORMATIONS

In a **normalizing transformation**, the original scores are converted with a mathematical operation to new ones that may be more normally distributed. The effect of applying such a transformation is to compress one part of a distribution more than another, thereby changing its shape but not the rank of the scores. This describes a **monotonic transformation**. There are basically two situations in SEM when normalizing transformations might be considered:

1. The researcher plans to use a **normal theory method**, such as default maximum likelihood, that requires normal distributions, but distributions of continuous outcomes are severely non-normal.
2. There are multiple observed measures, or indicators, of the same theoretical construct, but some of their relations with each other are curvilinear. Transformations that normalize distributions also tend to linearize relations among multiple indicators.

Before applying a normalizing transformation, you should think about the variables of interest and whether the expectation of normality is reasonable. Some variables are expected to have non-normal distributions, such as reports of alcohol or drug use and certain personality characteristics (Bentler, 1987). If so, then transforming an inherently non-normal variable to force a normal distribution may fundamentally alter it (the target variable is not actually studied). In this case, it would be better to use a different estimation method for continuous outcomes in SEM, one that does not assume normality, such as robust maximum likelihood. Another consideration is whether the metric of outcome variables is meaningful, such as athletic performance in seconds or postoperative survival time in years. Applying a transformation means that the original meaningful metric is lost, which could be a sacrifice.

Normalizing transformations may be more useful when there is no expectation of normality or metrics of outcome variables are arbitrary. An example is the total score for a set of true–false items. Because responses can be coded using any two different numbers, the total score is arbitrary. Standard scores such as percentiles and normal deviates are arbitrary, too, because one standardized metric can be substituted for another. Described in Topic Box 4.2 are types of normalizing transformations that may work in

different situations with practical suggestions for using them—see Osborne (2002) for more information. Exercise 3 asks you to find a normalizing transformation for the data in Figure 4.2.

Sometimes normalizing transformations can linearize relations between indicators of the same construct. For example, Budtz-Jørgensen, Keiding, Grandjean, and Weihe (2002) studied the effect of prenatal methylmercury exposure, through maternal consumption of contaminated pilot whale meat, on child neurobehavioral status among

TOPIC BOX 4.2

Normalizing Transformations

Three kinds of normalizing transformation are described next with suggestions for their use:

1. *Positive skew.* Before applying these transformations, add a constant to the scores so that the lowest value is 1.0. A basic transformation is the square root transformation, or $X^{1/2}$. It works by compressing the differences between scores in the upper end of the distribution more than the differences between lower scores. Logarithmic transformations are another option. A logarithm is the power (exponent) to which a base number must be raised in order to get the original number, such as $10^2 = 100$, so the logarithm of 100 in base 10 is 2.0. Distributions with extremely high scores may require a transformation with a higher base, such as $\log_{10} X$, but a lower base may suffice for less extreme cases, such as the natural base e (approximately 2.7183) for the transformation $\log_e X = \ln X$. The inverse function $1/X$ is an option for even more severe positive skew. Because inverting the scores reverses their order, (1) reflect (reverse) the original scores (multiply them by –1.0) and (2) add a constant to the reflected scores so that the maximum score is at least 1.0 before taking the inverse.

2. *Negative skew.* All the transformations just mentioned also work for negative skew when they are applied as follows: First, reflect the scores, and then add a constant so that the lowest score equals 1.0. Next, apply the transformation, and reflect the scores again to restore their original ordering.

3. *Other types of non-normality.* Odd-root functions (e.g., $X^{1/3}$) and sine functions tend to bring in outliers from both tails of the distribution toward the mean. Odd-powered polynomial transformations, such as X^3, may help for negative kurtosis. If the scores are proportions, the arcsine square root transformation function, or arcsin $X^{1/2}$, may normalize the distribution.

There are other kinds of normalizing transformations, and this is one of their potential problems: It can be difficult to find a transformation that works with a

particular set of scores. The **Box–Cox transformations** (Box & Cox, 1964) may require less trial and error. The most basic form is defined next only for positive scores:

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log X, & \text{if } \lambda = 0. \end{cases}$$

where the exponent λ is a constant that normalizes the scores. There are computer algorithms for finding the value of λ that maximizes the correlation between original and transformed scores (Friendly, 2006). There are other variations of the Box–Cox transformation (Osborne, 2010), some of which can be applied in regression analyses to deal with heteroscedasticity.

residents in the Farose Islands. Two biological markers were mercury concentration in cord blood and maternal hair, and the third measure was the amount of self-reported monthly consumption of whale meat. Blood or hair concentration scores can be so high that they have curvilinear relations with questionnaire data, so Budtz-Jørgensen et al. (2002) applied logarithmic transformations to the blood and hair concentrations before analyzing them.

Some distributions can be so severely non-normal that no transformation will work. Count variables are an example. A **count variable** is the number of times a discrete event happens over a period of time such as the number of serious automobile accidents over the past 5 years. Distributions of such variables tend to be positively skewed, and many cases may have scores of zero. Count variables generally follow non-normal distributions known as **Poisson distributions**, where the mean and variance are approximately equal. Some SEM computer tools, such as Mplus, offer special methods for analyzing count variables. These methods are related to the technique of **Poisson regression**, which also analyzes log linear models for count data (Agresti, 2007).

Little (2013) described **percentage or proportion of maximum scoring** (POMS) transformations for rescaling questionnaire items that measure a common domain but where responses are recorded on different Likert scales. After transformation, all items will have the same metric. Suppose that items with a 5-point Likert scale are administered to participants at time 1 but at time 2 the same items have a 7-point Likert scale. To compare responses over time, a transformation is needed. One option is to convert the narrower scale in this example (1-5) to the wider scale (1-7), as follows:

$$R7 = \left(\frac{O5 - 1}{4} \right) \times 6 + 1 \tag{4.6}$$

where $R7$ is the rescaled item with a 1-7 response format and $O5$ is the original item with a 1–5 scale. In Equation 4.6, the term $O5 - 1$ transforms the original 1–5 scale to a 0–4

scale; dividing by 4 then converts the scale to 0–1; then multiplying by 6 yields a 0–6 scale; and finally adding 1 generates the final 0–7 Likert scale.

Linearity and Homoscedasticity

Linear relations between continuous outcomes and homoscedastic residuals are part of multivariate normality. Curvilinear relations are easy to detect by looking at bivariate scatterplots. Heteroscedastic residuals may be caused by non-normality in either variable, greater measurement error at some levels of either variable than others, or outli-

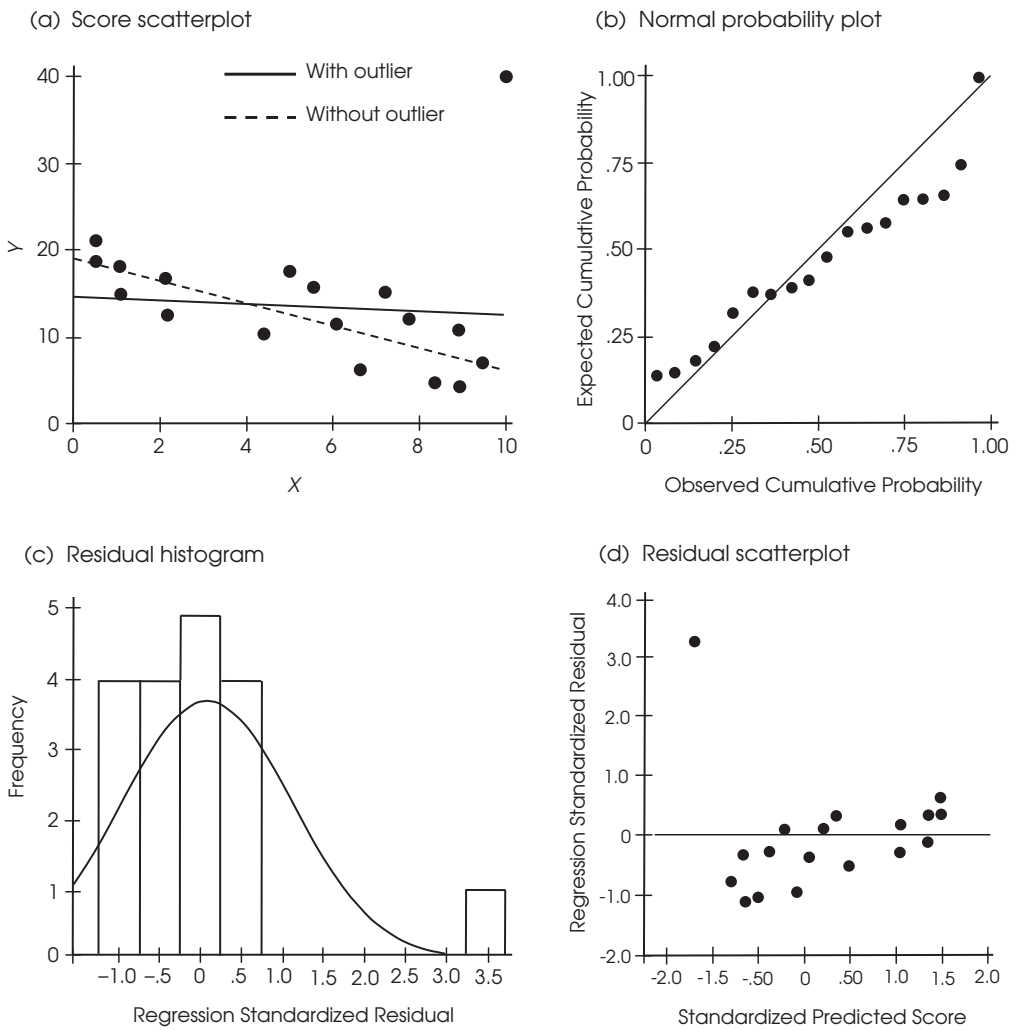

FIGURE 4.3. (a) Score scatterplot with outlier (N = 18) and the linear regression lines with and without the outlier (N = 17). (b) A normal probability plot of the regression standardized residuals. (c) A histogram of the regression standardized residuals. (d) Residual scatterplot.

ers. For example, presented in Figure 4.3(a) is a scatterplot for $N = 18$ scores. One score (40) on Y is > 3 standard deviations above the mean. For these data, $r_{XY} = -.074$, and the bivariate regression line is nearly horizontal, but these results are affected by the outlier. After removing the outlier ($N = 17$), then $r_{XY} = -.772$, and the new regression line better fits the remaining data—see Figure 4.3(a).

The regression standardized residuals for all the data ($N = 18$) in Figure 4.3(a) are plotted in various ways over Figures 4.3(b)–4.3(d). Figure 4.3(b) is a normal probability plot of the expected versus observed cumulative probabilities, which do not fall among a diagonal line. The histogram of the residuals with a superimposed normal curve is presented in Figure 4.3(c). The residuals are not normally distributed. A scatterplot of the residuals and standardized predicted scores is presented in Figure 4.3(d). The residuals are heteroscedastic because they are not evenly distributed about zero throughout the entire length of this scatterplot.

RELATIVE VARIANCES

In an **ill-scaled covariance matrix**, the ratio of the largest to smallest variance is greater than say, 100.0. Analysis of such a matrix in SEM can cause problems. Most estimation methods in SEM are iterative, which means that initial estimates are derived by the computer and then modified through subsequent cycles of calculation. The goal is to derive better estimates at each stage, estimates that improve the overall fit of the model to the data. When improvements from step to step become sufficiently small—that is, they fall below the **convergence criterion**—iterative estimation stops because the solution is stable. But if the estimates do not settle down to stable values, the process may fail. One cause is variances of observed variables that are very different in magnitude. When the computer adjusts the estimates from one step to the next in an iterative method for an ill-scaled matrix, the sizes of these changes may be huge for variables with small variances but trivial for others with large variances. Consequently, the entire set of estimates may head toward worse rather than better fit.

To prevent this problem, variables with extremely low or high variances can be rescaled by multiplying their scores by a constant, which changes the variance by a factor that equals the squared constant. For example,

$$s_X^2 = 12.0 \quad \text{and} \quad s_Y^2 = .12$$

so their variances differ by a factor of 100.0. Using the constant .10, we can rescale X as follows:

$$s_{X\times.10}^2 = .10^2 \times 12.0 = .12$$

so now variables $X \times .10$ and Y have the same variance, or .12. Now we rescale Y so that it has the same variance as X, or 12.0, by applying the constant 10.0, or

$$s^2_{Y\times 10.0} = 10^2 \times .12 = 12.0$$

Rescaling a variable in this way changes its average and variance but not its correlation with other variables. This is because multiplying a variable by a constant is just a linear transformation that does not affect relative differences among the scores. An example with real data follows.

Roth, Wiebe, Fillingham, and Shay (1989) administered measures of exercise, hardiness (resiliency, tough mindedness), fitness, stress, and level of illness in a sample of university students. Reported in Table 4.2 is a summary matrix of these data (correlations, means, and variances). The largest and smallest variances in this matrix (see the table) differ by a factor of more than 27,000, so the covariance matrix is ill-scaled. I have seen some SEM computer programs fail to analyze this matrix due to this characteristic. To prevent this problem, I multiplied the original variables by the constants listed in the table in order to make their variances more homogeneous (the constant 1.0 means no change). Among the rescaled variables, the largest variance is only about 13 times greater than the smallest variance. The rescaled matrix is not ill-scaled.

MISSING DATA

The topic of how to analyze data sets with missing observations is complicated. Entire books are devoted to it (Enders, 2010; McKnight, McKnight, Sidani, & Figueredo, 2007); there are also articles or chapters about methods for dealing with missing data in SEM (Allison, 2003; Graham & Coffman, 2012; Peters & Enders, 2002). This is fortunate

TABLE 4.2. Example of an Ill-Scaled Data Matrix

Variable	1	2	3	4	5
1. Exercise	—				
2. Hardiness	–.03	—			
3. Fitness	.39	.07	—		
4. Stress	–.05	–.23	–.13	—	
5. Illness	–.08	–.16	–.29	.34	—
M	40.90	0.0	67.10	4.80	716.70
Original s^2	4,422.25	14.44	338.56	44.89	390,375.04
Constant	1.00	10.00	1.00	5.00	.10
Rescaled s^2	4,422.25	1,440.00	338.56	1,122.25	3,903.75
Rescaled *SD*	66.50	38.00	18.40	33.50	62.48

Note. These data (correlations, means, and variances) are from Roth et al. (1989); $N = 373$. Note that low scores on the hardiness measure used by these authors indicate greater hardiness. In order to avoid confusion due to negative correlations, the signs of the correlations that involve the hardiness measure were reversed before they were recorded in this table.

because it is not possible here to give a comprehensive account of the topic. The goal instead is to give you a sense of basic analysis options and to explain the relevance of these options for SEM.

Ideally, researchers would always work with complete data sets, ones with no missing values; otherwise, prevention is the best strategy. For example, questionnaire items that are clear and unambiguous may prevent missing responses, and completed forms should be reviewed for missing responses before participants submit a computer-administered survey or leave the laboratory. In the real world, missing values occur in many, if not most, data sets, despite the best efforts at prevention. Missing data occur for many reasons, including hardware failure, missed appointments, and item nonresponse. A few missing values, such as < 5% in the total data set, may be of little concern. This is because selection among methods to deal with missing data is arbitrary in that the method used tends not to make much difference. Higher rates of data loss present more challenges, especially if the **data loss mechanism** is not truly random (or at least predictable). In this case, the choice of method can appreciably affect the results. This is why researchers should always explain how missing data were handled in the analysis.

Data Loss Mechanisms

There are basically three data loss mechanisms. All can operate within the same data set because each can affect different subsets of variables. Also, it is not always clear which pattern holds for a particular variable with missing values. The most optimistic case—and probably the most unrealistic in actual data—is when data are **missing completely at random** (MCAR). For variable *Y*, this means that (1) missing observations differ from the observed scores only by chance; that is, whether scores on *Y* are missing or not missing is unrelated to *Y* itself. (2) The presence versus absence of data on *Y* is unrelated to all other variables in the data set. In this case, the observed (nonmissing) data are just a random sample of scores that the researcher would have analyzed had the data been complete (Enders, 2010). Results based on the complete cases only should not be biased, although power may be reduced due to a smaller effective sample size. An example of haphazard missing data is when questionnaire responses to items about mental health are lost due to sporadic computer problems that have nothing to do with either respondents' true mental health status or their responses to questions about other topics.

A second data loss mechanism is indicated when the property of missingness on *Y* is unrelated to *Y* itself but is correlated with other variables in the data set; that is, missing data arise from a process that is both measured and predictable in a particular sample (Little, 2013). This process is called **missing at random** (MAR), which is an odd term because the data loss mechanism depends on other variables, and thus is not random. An example of an MAR process would be when men are less likely to respond to questions about mental health than women, but among men the probability of responding is unrelated to their true mental health status.

Information lost due to an MAR process is potentially recoverable through imputation, where missing scores are replaced by predicted scores. The predicted scores

are generated from other variables in the data set that predict missingness on *Y*. If the strength of that prediction is reasonably strong, then results on *Y* after imputation may be relatively unbiased. In this sense, the MAR pattern is described as **ignorable** concerning potential bias. Note that both the MAR and MCAR patterns of data loss can affect the same variable.

A strategy that anticipates the MAR pattern is to measure **auxiliary variables**. Such variables may not be of substantive interest, but they predict missingness on other variables in the data set. For example, gender, socioeconomic status, and parental involvement are potential auxiliary variables in longitudinal studies of children, and inclusion of these predictors when imputing scores on other variables may decrease bias (Little, 2013). Auxiliary variables require care in their selection. This is because the inclusion of too many auxiliary variables in smaller samples can increase imprecision by so much that more sophisticated methods for imputation can fail. This is especially true if less than about 10% or so of the variance in missingness on *Y* is explained by auxiliary variables ($R^2 < .10$) (Hardt, Herke, & Leonhart, 2012).

When data are **missing not at random** (MNAR), the data loss mechanism is **nonignorable**, which means that the presence versus absence of scores on *Y* depends on *Y* itself. An example from medicine occurs when patients drop out of a study when a particular treatment causes unpleasant side effects. Because that discomfort is not measured, however, the data are missing due to a process that is unknown in a particular data set. Results based on the complete cases only can be severely biased when the data loss pattern is MNAR. For example, a treatment may look more beneficial than it really is if data from patients who were unable to tolerate the treatment are lost. Some bias may be reduced if other measured variables happen to covary with unmeasured causes of data loss, but whether this is true in a particular sample is usually unknown. The choice of method to deal with the incomplete records can make a difference in the results when the MNAR pattern holds.

Diagnosing Missing Data

It is not easy in practice to determine whether the data loss mechanism is random or systematic, especially when each variable is measured only once. Specifically, there are ways to determine whether the assumption of MCAR is reasonable, but there is no definitive test that provides direct evidence of either MAR or MNAR if the former hypothesis is rejected. Little and Rubin (2002) describe a multivariate statistical test of the MCAR assumption that simultaneously compares complete versus incomplete cases on *Y* across all other variables. If this comparison is significant, then the MCAR hypothesis is rejected. Plausibility of the MCAR assumption can also be examined through a series of univariate comparisons of the *t* test of cases that have missing scores on *Y* with cases that have complete records on other variables. The problems with these significance tests are that they can have low power in smaller samples and they can flag trivial differences as significant in larger samples.

A related tactic involves creating a dummy-coded variable that indicates whether a score is missing and then examining cross tabulations with other categorical variables, such as gender or treatment condition. Some computer programs for general statistical analysis have special commands or procedures for analyzing missing data patterns. An example is the Missing Values procedure of SPSS, which can conduct all the diagnostic tests just mentioned. The PRELIS program of LISREL also has extensive capabilities for analyzing missing data patterns.

If the assumption of MCAR is rejected, then we cannot ever really be sure whether the data loss mechanism is MAR or MNAR. This is because variables may be omitted that account for data loss on *Y* that are related to *Y* itself. Because these variables are unmeasured, the true extent of nonrandom, systematic data loss will not be known. It helps if other, measured variables in the data set predict missingness on *Y*, but only some of the information on *Y* may be recovered in an imputation process based on these predictors. For this reason it is prudent to ascertain potential auxiliary variables when planning a study.

There is no magical statistical "fix" that will eliminate bias due to systematic data loss. About the best that can be done is to understand the nature of the underlying data loss pattern and accordingly modify your interpretation of the results. If the selection of one option for dealing with missing data instead of another makes a difference in the results and it is unclear which option is best, then you should report both sets of findings. This approach makes it plain that the results depend on how missing observations were handled. This tactic is a kind of **sensitivity analysis** in which data are reanalyzed under different assumptions—here, using alternative missing data techniques—and the results are compared with the original findings.

Classical Methods

Classical techniques for handling incomplete cases have been around for a long time and are available as options in many computer programs for general statistical analysis, but they are increasingly considered obsolete. One reason is that such methods generally assume that the missing value mechanism is MCAR, which is often unrealistic. Such methods tend to yield biased estimates under the less strict assumption of MAR, and even more so when the data loss mechanism is MNAR. They also take relatively little advantage of the structure in the data. Classical methods are briefly reviewed next, but there are better, more modern methods, which will also be discussed in this chapter.

There are two general kinds of classical methods: **available case methods,** which analyze data available through deletion of incomplete cases, and **single-imputation methods**, which replace each missing score with a single calculated (imputed) score. Available case methods include **listwise deletion** in which cases with missing scores on any variable are excluded from all analyses. The effective sample size with listwise deletion includes only cases with complete records. This number can be much smaller than the original sample size if missing observations are scattered across many records.

In regression analyses, listwise deletion of incomplete cases generates reasonably good estimates when the data loss mechanism depends on the predictors but not on the criterion (Little & Rubin, 2002).

An advantage of listwise deletion is that all analyses are conducted with the same cases. This is not so with **pairwise deletion**, in which cases are excluded only if they have missing data on variables involved in a particular analysis. Suppose that $N = 300$ for an incomplete data set. If 250 cases have no missing scores on variables X and Y, then the effective sample size for cov_{XY} is this number. If fewer or more cases have valid scores on X and W, however, the effective sample size for cov_{XW} will not be 250. It is this property of the method that can give rise to out-of-bounds correlations or covariances. Presented in Table 4.3 is a small data set with missing scores on all three variables. The covariance matrix generated by pairwise deletion for these data is nonpositive definite. Exercise 4 asks you to verify this statement.

The most basic single-imputation method is **mean substitution**, which involves replacing a missing score with the overall sample mean. A variation is **group-mean substitution**, in which a missing score in a particular group (e.g., women) is replaced by the group mean. This variation may be preferred when group membership is a predictor in the analysis or when a model in SEM is analyzed over multiple groups. Both methods are simple, but they can distort the distribution of the data by reducing variability. Suppose in a data set where $N = 75$ that 15 cases have missing values on some variable. Substituting the mean of the 60 valid cases does not change the mean for $N = 75$ after imputation compared with the mean for $N = 60$ before imputation. But the variance for the $N = 60$ scores before substitution will be greater than the variance for the $N = 75$ scores after substitution. Mean substitution also tends to make distributions more peaked at the mean, too, which further distorts the underlying distribution of the data (Vriens & Melton, 2002).

Regression substitution is somewhat more sophisticated. In this method, each missing score is replaced by a predicted score using multiple regression based on nonmissing scores on other variables. Regression substitution uses more information than mean substitution, but it assumes that variables with missing observations can be predicted reasonably well from other variables in the same data set; otherwise, there is little

TABLE 4.3. Example of an Incomplete Data Set

Case	X	W	Y
A	42	13	8
B	34	12	10
C	22	—	12
D	—	8	14
E	24	7	16
F	16	10	—
G	30	10	—

point in imputing missing scores with predicted scores. A variation is **stochastic regression imputation**, in which the computer adds a randomly sampled error term from the normal distribution or other user-specified distribution to each predicted score, which reflects uncertainty in the score. This capability is implemented in the Missing Values procedure in SPSS.

A more sophisticated single-imputation method is **pattern matching**, in which the computer replaces a missing observation with a score from a case with the most similar profile on other variables. Pattern matching is available in the PRELIS program of LISREL. Another option is **random hot-deck imputation**, which separates complete from incomplete cases; sorts both sets of records so that cases with similar profiles on background variables are grouped together; randomly interleaves the incomplete cases and complete ones; and replaces missing scores with those on the same variable from the nearest complete record. Myers (2011) describes a macro that performs random hot-deck imputation in SPSS. All single-imputation methods tend to underestimate error variance, especially if the proportion of missing observations is relatively high (Vriens & Melton, 2002).

Modern Methods

Contemporary methods generally assume a data loss pattern that is MAR, not MCAR. When the pattern is not random (MNAR), these more sophisticated techniques will also yield biased estimates, but probably less so compared with classical techniques (Peters & Enders, 2002).

There are two basic kinds of modern methods for analyses with missing data. A **model-based method** takes the researcher's model as the starting point. Next, the procedure partitions the cases in a raw data file into subsets, each with the same pattern of missing observations, including none (complete cases). Relevant statistical information, including the means and the variances, is extracted from each subset, so all cases are retained in the analysis. The parameters of the researcher's model are then estimated after combining all available information over the subsets of cases. Thus, parameter estimates and their standard errors are calculated directly from the available data without deletion or imputation of missing values. Some SEM computer tools, including Amos, LISREL, and Mplus, offer a special version of the maximum likelihood method—sometimes called **full information maximum likelihood** (FIML)—for incomplete data files that works in the way just described. Some FIML procedures for incomplete data allow the specification of auxiliary variables (see Graham & Coffman, 2012).

Multiple imputation is a **data-based method** that typically works with the whole raw data file, not just with the observed variables that comprise the researcher's model. As the name suggests, **multiple imputation** can generally replace a missing score with multiple estimated (imputed) values from a predictive distribution that models the data loss mechanism. In nontechnical terms, a model for both the complete and incomplete data is defined under these methods. The computer then estimates means and variances in the whole sample that satisfy a statistical criterion. The process of imputation

is repeated so that the analysis is actually conducted with multiple versions of imputed data sets. In large data sets, a relatively high number of imputed data sets may need to be generated (e.g., 100) in order for the results to have reasonable precision (Little, 2013). The final set of estimates comes after the computer synthesizes the results from all replications.

Some methods for multiple imputation are based on the **expectation-maximization** (EM) **algorithm**, which has two steps. In the E (expectation) step, missing observations are imputed by predicted scores in a series of regressions in which each incomplete variable is regressed on the remaining variables for a particular case. In the M (maximization) step, the whole imputed data set is submitted for maximum likelihood estimation. These two steps are repeated until a stable solution is reached across the M steps. The EM algorithm for multiple imputation is available in EQS and LISREL, among other SEM computer tools. In addition, some methods are based on the **Markov Chain Monte Carlo** (MCMC) approach, which is a class of methods for random sampling from a theoretical probability distribution. The MCMC method is used to draw from a predictive distribution for the missing data, and these draws become the imputed scores. Multiple imputation in Mplus is based on the MCMC method.

There may be times in SEM when multiple imputation is favored over the FIML method (Graham & Coffman, 2012). Not all SEM computer programs feature FIML estimation for incomplete data files. In this case, the researcher could use procedures for multiple imputation in computer tools for general statistical analyses. For example, the MI procedure in SAS/STAT could be used to impute the data, and later the MIANALYZE procedure could combine results from the imputed data sets after they have been analyzed with a computer tool for SEM. It is also generally easier to incorporate auxiliary variables in multiple imputation than in the FIML method. But if the FIML method is available in your SEM computer tool, it is a reasonable option for conducting the analysis with an incomplete data set.

SELECTING GOOD MEASURES AND REPORTING ABOUT THEM

It is just as critical in SEM as in other types of analyses to (1) select measures with strong psychometric properties and (2) report these characteristics in written summaries. This is because the product of measures, or scores, is what is analyzed. If the scores do not have good psychometrics, then the results can be meaningless.

Presented in Table 4.4 is a checklist of descriptive, practical, and technical information that should be considered before selecting a measure. Not all of these points may be relevant in a particular study, and some types of research have special measurement needs that may not be represented in the table. If so, just modify the checklist to better reflect a particular situation. The *Mental Measurements Yearbook* (Carlson, Geisinger, & Jonson, 2014) is a good source of information about commercial tests. It is also available as a searchable electronic database in many university libraries. Maddox (2008) describes measures in psychology, education, and business. A directory of noncommer-

TABLE 4.4. Checklist for Evaluating Potential Measures

General
Stated purpose of the measure
Attribute(s) claimed to be measured
Characteristics of samples in which measure was developed (e.g., normative sample)
Language of test materials
Costs (manuals, forms, software, etc.)
Limitations of the measure
Academic or professional affiliation(s) of author(s) consistent with test development
Publication date and publisher
Administration
Test length and testing time
Measurement method (e.g., self-report, interview, unobtrusive)
Response format (e.g., multiple choice, free response)
Availability of alternative forms (versions)
Individual or group administration
Paper-and-pencil or computer administration
Scoring method, requirements, and options
Materials or testing facilities needed (e.g., computer, quiet testing room)
Training requirements for test administrators or scorers (e.g., test user qualifications)
Accommodations for test takers with physical or sensory disabilities
Test documentation
Test manual available
Manual's description of how to correctly derive and interpret scores
Evidence for score reliability and characteristics of samples (e.g., reliability induction)
Evidence for score validity and characteristics of samples
Evidence for test fairness (e.g., lack of gender, race, or age bias)
Results of independent reviews of the measure

cial measures from articles in psychology, sociology, or education journals is available in Goldman and Mitchell (2007). These measures are not protected by copyright, but as a professional courtesy you should ask the author's permission before using or adapting a particular test. There is also the freely accessible Measurement Instrument Database for the Social Sciences, an online test database.[3]

Readers who have already taken a measurement course are at some advantage when it comes to selecting a test because they can critically evaluate candidate measures. They should also know how to evaluate whether those scores in their own samples are reliable and valid. Readers without this background are encouraged to fill in this gap. Formal coursework is not the only way to learn more about measurement. Just like learning about SEM, more informal ways to learn measurement theory include partici-

[3] *www.midss.org*

pation in seminars or workshops and self-study. A good undergraduate-level book that emphasizes classical measurement theory in psychology and education is Thorndike and Thorndike-Christ (2010), and the graduate-level work by Raykov (2011) deals with modern measurement theory.

Unfortunately, the state of practice about reporting on the psychometric characteristics of scores analyzed is too often poor. For example, Vacha-Haase and Thompson (2011) found that 55% of authors did not even mention score reliability in over 13,000 primary studies from a total of 47 meta-analyses of reliability generalization in the behavioral sciences. Authors mentioned reliability in about 16% of the studies, but they merely inducted values reported in other sources, such as test manuals. Such **reliability induction**, or inferring from particular coefficients calculated in other samples to a different population, requires explicit justification. But researchers rarely compare characteristics of their sample with those from cited studies of score reliability. For example, scores from a computer-based task of reaction time developed in samples of young adults may not be as precise for elderly adults. A better practice is for researchers to report estimates of score reliability from their own samples. They should also cite reliability coefficients reported in published sources (reliability induction) but with comment on similarities between samples described in those other sources and the researcher's sample.

Thompson and Vacha-Haase (2000) speculated that another cause of poor reporting practices is the apparently widespread but false belief that it is *tests* that are reliable or unreliable, not *scores* in a particular sample. In other words, if researchers believe that reliability, once established, is an immutable property of tests, then they may put little effort into estimating reliability in their own samples. They may also adopt a "black box" mentality that assumes that reliability can be established by others, such as a select few academics who conduct measurement-related research. The truth is that reliability and validity are attributes of scores in particular samples where the intended uses of those scores must also be considered.

Measurement is a broad topic, so it is impossible to succinctly cover all its aspects, but familiarity with the issues considered next should help you to select good measures and report necessary information about scores generated from them. This presentation will also help you to better understand certain analysis options in CFA, the factor-analytic technique in SEM.

SCORE RELIABILITY

Score reliability is the degree to which scores in a particular sample are precise. It is estimated as one minus the proportion of total observed variance due to random error. These estimates are reliability coefficients, which for measure X are often designated with the symbol r_{XX}. Because r_{XX} is a proportion of variance, its theoretical range is 0–1.0. For example, if r_{XX} = .80, then 1 − .80 = .20, or 20% of total variance is unsystematic. But the remaining standardized variance, or 80%, may not all be systematic.

This is because a particular type of reliability coefficient may estimate a single source of random error, and scores can be affected by multiple sources of error. As r_{XX} approaches zero, the scores are increasingly more like random numbers, and random numbers measure nothing. It can happen that an empirical reliability coefficient is less than zero. A negative reliability coefficient is interpreted as though its value were zero, but such a result ($r_{XX} < 0$) indicates a serious problem with the scores.

The type of reliability coefficient reported most often in the literature is **coefficient alpha**, also called **Cronbach's alpha**. It measures **internal consistency reliability**, or the degree to which responses are consistent across the items of a measure. If internal consistency is low, then the content of the items may be so heterogeneous that the total score is not the best possible unit of analysis. A conceptual equation is

$$\alpha_C = \frac{n_i \, \overline{r}_{ij}}{1 + (n_i - 1)\overline{r}_{ij}} \tag{4.7}$$

where n_i is the number of items, not cases, and $\overline{r}_{ij}$ is the average Pearson correlation between all pairs of items. For example, given $n_i = 20$ with a mean interitem correlation of .30, then

$$\alpha_C = \frac{20\,(.30)}{1 + (20 - 1)\,(.30)} = .90$$

Internal consistency reliability is higher as there are more items or the average interitem correlation increases. In observed variable analyses, it is best to analyze scores from measures that are internally consistent. This is also generally good advice for latent variable analyses, including SEM, but see Little, Lindenberger, and Nesselroade (1999) about exceptions. Exercise 5 asks you to calculate and interpret α_C for a small set of items.

An older method of estimating internal consistency is that of **split-half reliability**, where a single test is split into two parts, such as an odd–even item split, and scores from two halves are correlated. The observed correlation is then corrected for test length, and the corrected result is the split-half reliability coefficient. For a particular set of items, the value of α_C is the average of all possible split-half coefficients (e.g., odd vs. even items, first-half vs. second-half items, etc.), so in this sense α_C is a more general estimate of internal consistency than any split-half coefficient.

A drawback of α_C is that it is actually not a very good indicator of whether a set of items measures a single factor. This is because lower values of $\overline{r}_{ij}$ can be offset by greater numbers of items, n_i. Suppose that $\overline{r}_{ij} = .01$ for 1,000 items. The average correlation across the items is practically zero, so they clearly do not measure a common domain. But with so many items in this example, $\alpha_C = .91$. (You should verify this statement.) In this example, the large number of items overwhelms the near-zero value of $\overline{r}_{ij}$. This means that a high value of α_C does not guarantee internal consistency because long, multidimensional scales will also have high values of α_C (Streiner, 2003). At the other extreme, very high values of α_C can suggest redundancy in a small item set. For example, given $\alpha_C = .95$ for $n_i = 2$ items, then $\overline{r}_{ij} = .90$, which indicates that the two items

are not distinct (they are extremely collinear). Better ways to estimate the reliability of construct measurement in SEM are described in Chapter 13.

Estimation of other kinds of score reliability may require multiple occasions, test forms, or examiners. **Test–retest reliability** involves the readministration of a measure to the same group on a second occasion. If the two sets of scores are highly correlated, error due to temporal factors may be minimal. **Alternate- (parallel-) forms reliability** involves the evaluation of score precision across different versions of a test. This method estimates whether variation in items drawn from the same domain leads to changes in rank order between the two forms. If so, then scores are unstable across different versions, which raises doubts that a common domain is measured. **Interrater reliability** is relevant for subjectively scored tests: If independent raters do not agree in scoring, then examiner-specific factors may contribute unduly to score variability.

In observed variable analyses, there is no gold standard as to how high coefficients should be in order to conclude that score reliability is satisfactory, but here are some guidelines: Generally, coefficients around .90 are considered "excellent," values around .80 as "very good," and values about .70 as "adequate." Note that somewhat lower levels of score reliability can be tolerated in latent variable methods compared with observed variable methods, if the sample size is sufficiently large (Little et al., 1999).

Low score reliability has detrimental effects in observed variable analyses. Poor reliability reduces statistical power; it also generally reduces effect sizes below their true (population) values. Unreliability in scores of two different variables, X or Y, attenuates their observed correlation. This formula from classical measurement theory shows the exact relation:

$$\max\left|\hat{r}_{XY}\right| = \sqrt{r_{XX} \times r_{XX}} \tag{4.8}$$

where $\max\left|\hat{r}_{XY}\right|$ is the theoretical (estimated) maximum absolute value of the correlation. In other words, the absolute correlation between X and Y can equal 1.0 only if scores on both variables are perfectly reliable. Suppose that $r_{XX} = .10$ and $r_{YY} = .90$. Given this information, the theoretical maximum absolute value of r_{XY} can be no higher than

$$\max\left|\hat{r}_{XY}\right| = \sqrt{.10 \times .90} = .30$$

A variation of Equation 4.8 is the **correction for attenuation**:

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX} \times r_{XX}}} \tag{4.9}$$

where $\hat{r}_{XY}$ is the *estimated* validity coefficient if scores on both measures were perfectly reliable. In general, $\hat{r}_{XY}$ is greater in absolute value than r_{XY}, the observed correlation. For example, given

$$r_{XY} = .30,\ r_{XX} = .90, \text{ and } r_{YY} = .40$$

then $\hat{r}_{XY}$ = .50; that is, we expect that the "true" correlation between *X* and *Y* would be .50, controlling for measurement error. Because disattenuated correlations are only estimates, it can happen that their absolute values exceed 1.0. A better way to control for measurement error is to use SEM where constructs are specified as latent variables, each measured by multiple indicators. In fact, SEM is more accurate at estimating correlations between factors or between indicators and factors than observed-variable methods (Little et al., 1999).

SCORE VALIDITY

Score validity concerns the soundness of inferences based on the scores, and information about score validity conveys to the researcher whether applying a test is capable of achieving certain aims. Kane (2013) elaborated on this theme by describing **interpretation–use arguments**, an approach to validity that concerns the plausibility and appropriateness of both the interpretation and the proposed uses of scores. In this view, validity is not a fixed property of tests; rather, it involves the proposed interpretation and intended uses of the scores. As the range of potential generalizations from test scores increases, such as from an observed sample of performances (test data) to predicted performances in other settings, more evidence is needed. Messick (1995) emphasized the qualities of relevance, utility, value implications, and social consequences of test use and interpretation in validation. An example of the social consequences of testing includes the fair and accurate assessment of cognitive abilities among minority children.

Construct validity involves whether scores measure a target hypothetical construct, which is latent and thus can be measured only indirectly through its indicators. There is no single, definitive test of construct validity, nor is it established in a single study. Instead, measurement-based research usually concerns a particular aspect of construct validity. For instance, **criterion-related validity** concerns whether test scores (*X*) relate to a criterion (*Y*) against which the scores can be evaluated. Specifically, are sample values of r_{XY} large enough to support the claim that a test explains an appreciable amount of the variability in the criterion? Whether an admissions test for graduate school predicts eventual program completion is a question of criterion-related validity.

Convergent validity and discriminant validity involve the evaluation of measures against each other instead of against an external standard. Variables presumed to measure the *same* construct show **convergent validity** if their intercorrelations are appreciable in magnitude. But if measures that supposedly reflect the same construct also share the same measurement method, their intercorrelations could be inflated by **common method variance**. Thus, the best case for convergent validity occurs when measures of the same presumed trait are each based on a different measurement method (Campbell & Fiske, 1959). Likewise, **discriminant validity** is supported if the intercorrelations among a set of variables presumed to measure *different* constructs are not too high, but this evidence is stronger when the measures are not based on the same method. If r_{XY}

= .90 and these two variables are each based on a different measurement method, one cannot claim that *X* and *Y* assess distinct constructs. Hypotheses about convergent and discriminant validity are routinely tested in CFA.

Content validity deals with whether test items are representative of the domain(s) they are supposed to measure. Content validity is often critical for scholastic achievement measures, such as tests that should assess specific skills at a particular grade level (e.g., Grade 4 math). It is important for other kinds of tests, too, such as symptom rating scales. The items of a depression rating scale, for example, should represent the symptom areas thought to reflect clinical depression. Expert opinion is the basis for establishing content validity, not statistical analysis.

As in other kinds of statistical methods, SEM requires the analysis of scores with good evidence for validity. Because score reliability is generally required for score validity—but does not guarantee it—this requirement includes good score reliability, too (see Little et al., 1999, for exceptions). Otherwise, the accuracy of the interpretation of the results is doubtful. So using SEM does not free researchers from having to think about measurement (just the opposite is true).

ITEM RESPONSE THEORY AND ITEM CHARACTERISTIC CURVES

For two reasons, it is worthwhile to know about **item response theory** (IRT), also known as **latent trait theory**. First, techniques in IRT permit more sophisticated estimation of item psychometrics than is possible in classical measurement theory. Methods in IRT can be used to equate scores from one test to another, evaluate the extent of item bias over different populations, and construct individualized tests for examinees of different ability levels, or **tailored testing**, among other possibilities. Second, it is an alternative to CFA for analyzing ordinal data. In the past, researchers who analyzed IRT models used specialized software, but now some SEM computer programs such as LISREL and Mplus can analyze at least basic kinds of IRT models. How to analyze ordinal data in CFA is considered later in the book, but part of the logic for doing so is related to that of IRT.

The body of IRT consists of mathematical models that relate responses on individual items to a continuous latent variable θ. Assume for this discussion that items are dichotomously scored (0 = incorrect, 1 = correct) and that θ is an ability dimension with a normal deviate (z) metric. Presented in Figure 4.4 is an **item characteristic curve** (ICC), or a sigmoid function that relates θ to the probability of a correct answer. This ICC depicts a **two-parameter IRT model**, where the parameters are item difficulty and item discrimination. Difficulty is the level of ability that corresponds to a 50% chance of getting the item correct, and discrimination is the slope of the tangent line to the ICC at that point. In the figure, difficulty is $\theta = 0$ (i.e., the mean) because this level of ability predicts that 50% of examinees will pass the item, and discrimination is the slope of the tangent line at this point. The steeper the slope, the more discriminating the item, and the stronger its relation with θ. **Three-parameter IRT models** also include a guessing

parameter, and it indicates the probability that an examinee of low ability would correctly guess the answer. A **Rasch model** has a single parameter, item difficulty. Uniform discrimination for all items implies a constant construct, one that can be measured in the same way for all examinees regardless of ability level. In this way, evaluation of Rasch models can be viewed as more confirmatory than fitting more complex IRT models to the data.

Figure 4.4 might look familiar. This is because the shape of an ICC and the sigmoid functions analyzed in logistic regression and probit regression are similar (see Figure 2.4). Shared among all these techniques is the analysis of a continuous latent variable that underlies responses to categorical observed variables. Parameter estimates in IRT can be scaled in either logistic units or probit units, and we will see later in the book that estimates in CFA can be mathematically transformed to estimates of the type generated in IRT. Baylor et al. (2011) gives a clear introduction to IRT.

SUMMARY

The most widely used methods for continuous outcomes in SEM require screening the data for multivariate normality. It is critical to select appropriate methods for handling missing data. Such methods generally assume that the data loss mechanism is random or at least predictable. Modern options, such as multiple imputation or a special maximum likelihood method for incomplete data files, are generally better choices than classical methods, such as case deletion or single imputation. Computer tools for SEM can analyze either raw data files or matrix summaries. Most estimation methods in SEM assume unstandardized variables, so a covariance matrix is preferred over a correlation

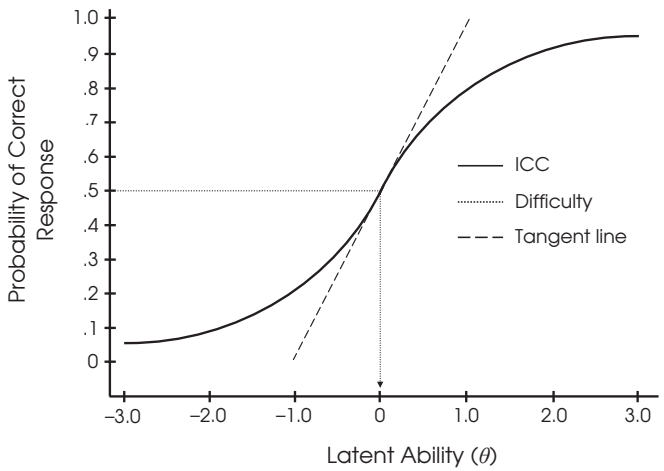

FIGURE 4.4. Item characteristic curve (ICC) for the predicted probability of a correct response for a dichotomously scored item in a two-parameter item response theory model. Item difficulty is $\theta = 0$, and item discrimination is the slope of the tangent line at $\theta = 0$.

matrix without standard deviations when a matrix summary is the input and means are not analyzed. In written reports, researchers should provide information about the psychometrics of their scores. Analysis of scores with poor reliability or validity can jeopardize the results. Computer tools for SEM are described in the next chapter.

LEARN MORE

The description of modern methods for analyses with missing values by Enders (2010) is exceptionally clear, and Graham and Coffman (2012) discuss specific options in SEM. Malone and Lubansky (2012) consider data screening in SEM with several examples.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277–295). New York: Guilford Press.

Malone, P. S., & Lubansky, J. B. (2012). Preparing data for structural equation modeling: Doing your homework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 263–276). New York: Guilford Press.

EXERCISES

1. Reproduce the covariance matrix in the upper right part of Table 4.1 from the correlations and standard deviations in the upper left part of the table.
2. Given $\text{cov}_{XY} = 13.00$, $s^2_X = 12.00$, and $s^2_Y = 10.00$, show that the corresponding correlation is out of bounds.
3. Find a normalizing transformation for the data in Figure 4.2.
4. Calculate the covariance matrix for the incomplete data in Table 4.3 using pairwise deletion. Show that the corresponding correlation matrix has an element that is out of bounds.
5. Presented next are scores on five dichotomously-scored items (0 = wrong, 1 = correct) for eight cases (A–H). Calculate the internal consistency reliability α_C for these data using Equation 4.7. If a reliability procedure is available in your computer program for general statistical analyses, use it to verify your calculations:

```
A: 1, 1, 0, 1, 1    B: 0, 0, 0, 0, 0
C: 1, 1, 1, 1, 0    D: 1, 1, 1, 0, 1
E: 1, 0, 1, 1, 1    F: 0, 1, 1, 1, 1
G: 1, 1, 1, 1, 1    H: 1, 1, 0, 1, 1
```