Structural Equation Modeling

Applications in ecological and evolutionary biology

Edited by

BRUCE H. PUGESEK

US Geological Survey – Biological Resources Division,
Northern Rocky Mountain Science Center, Bozeman

ADRIAN TOMER

Department of Psychology, Shippensburg University, Pennsylvania

and

ALEXANDER VON EYE

Department of Psychology, Michigan State University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Bembo 11/13 pt and Univers *System* LaTeX 2ε [TB]

A catalogue record for this book is available from the British Library

ISBN 0 521 78133 7 hardback

Contents

1 Structural equation modeling: an introduction

Scott L. Hershberger, George A. Marcoulides, and Makeba M. Parramore

Abstract

This chapter provides an introduction to structural equation modeling (SEM), a statistical technique that allows scientists and researchers to quantify and test scientific theories. As an example, a model from behavioral genetics is examined, in which genetic and environmental influences on a trait are determined. The many procedures and considerations involved in SEM are outlined and described, including defining and specifying a model diagrammatically and algebraically, determining the identification status of the model, estimating the model parameters, assessing the fit of the model to the data, and respecifying the model to achieve a better fit to the data. Since behavioral genetic models typically require family members of differing genetic relatedness, multisample SEM is introduced. All of the steps involved in evaluating the behavioral genetic model are accomplished with the assistance of LISREL, a popular software program used in SEM.

Introduction

Structural equation modeling (SEM) techniques are considered today to be a major component of applied multivariate statistical analyses and are used by biologists, economists, educational researchers, marketing researchers, medical researchers, and a variety of other social and behavioral scientists. Although the statistical theory that underlies the techniques appeared decades ago, a considerable number of years passed before SEM received the widespread attention it holds today. One reason for the recent attention is the availability of specialized SEM programs (e.g., AMOS, EQS, LISREL, Mplus, Mx, RAMONA, SEPATH). Another reason has been the publication of several introductory and advanced texts on SEM (e.g., Hayduk, 1987, 1996: Bollen, 1989; Byrne, 1989, 1994, 2000; Bollen & Long, 1993; Hoyle, 1995; Marcoulides & Schumacker, 1996; Schumacker & Lomax, 1996; Schumacker & Marcoulides, 1998; Raykov & Marcoulides, 2000), and a

journal, devoted exclusively to SEM, entitled *Structural Equation Modeling: A Multidisciplinary Journal.*

In its broadest sense, SEM models represent translations of a series of hypothesized cause–effect relationships between variables into a composite hypothesis concerning patterns of statistical dependencies (Shipley, 2000). The relationships are described by parameters that indicate the magnitude of the effect (direct or indirect) that independent variables (either observed or latent) have on dependent variables (either observed or latent). By enabling the translation of hypothesized relationships into testable mathematical models, SEM offers researchers a comprehensive method for the quantification and testing of theoretical models. Once a theory has been proposed, it can then be tested against empirical data. The process of testing a proposed theoretical model is commonly referred to as the "confirmatory" aspect of SEM (Raykov & Marcoulides, 2000). Another aspect of SEM is the so-called "exploratory" mode. This aspect allows for theory development and often involves repeated applications of the same data in order to explore potential relationships between variables of interest (either observed or latent).

Latent variables are hypothetical or theoretical variables (constructs) that cannot be observed directly. Latent variables are of major importance to most disciplines but generally lack an explicit or precise way of measuring their existence or influence. For example, many behavioral and social scientists study the constructs of aggression and dominance. Because these constructs cannot be measured explicitly, they are are inferred through observing or measuring specific features that operationally define them (e.g., tests, scales, self-reports, inventories, or questionnaires). SEM can also be used to test the plausibility of hypothetical assertions about potential interrelationships between constructs and their observed measures or indicators. Latent variables are hypothesized to be responsible for the outcome of observed measures (e.g., aggression is the underlying factor influencing one's score on a questionnaire that attempts to assess offensive driving behavior). In other words, the score on the explicit questionnaire would be an indicator of the construct or latent variable – aggression. Researchers often use a number of indicators or observed variables to examine the influences of a theoretical factor or latent variable. It is generally recommended that researchers use *multiple indicators* (preferably more than two) for each latent variable considered in order to obtain a more complete and reliable "picture" than that provided by a single indicator (Raykov & Marcoulides, 2000). Because both observed and latent variables can be independent or dependent in a proposed model, a more detailed description of this issue will be provided later in this chapter.

Definition and specification of a structural equation model

The definition of a SEM model begins with a simple statement of the verbal theory that makes explicit the hypothesized relationships among a set of studied variables (Marcoulides, 1989). Typically, researchers communicate a SEM model by drawing a picture of it (Marcoulides & Hershberger, 1997). These pictures, or so-called *path diagrams*, are simple mathematical representations (but in graphical form) of the proposed theorical model. Figure 1.1 presents the most commonly used graphical notation for the representation of SEM models. As will become clear later, path diagrams not only aid in the conceptualization and communication of theoretical models, but also substantially contribute to the creation of the appropriate input file that is necessary to test and fit the model to collected data using particular software packages (Raykov & Marcoulides, 2000).

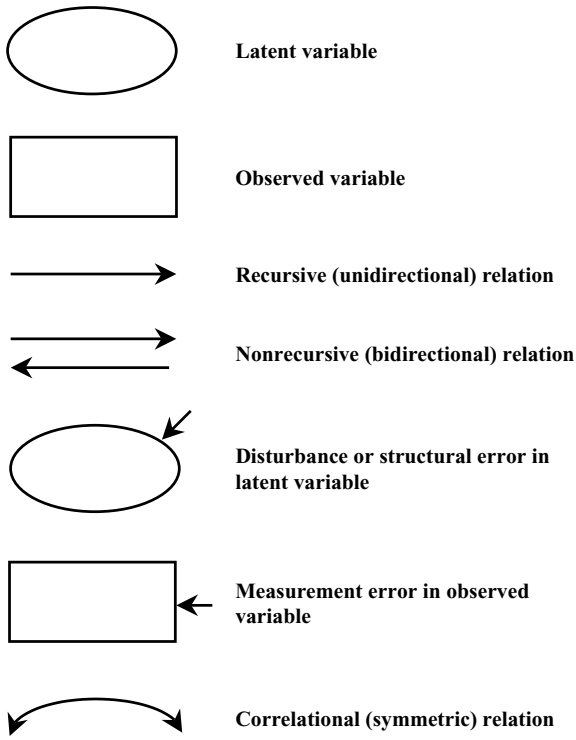

Figure 1.1. Commonly used graphical notation for the representation of SEM models.

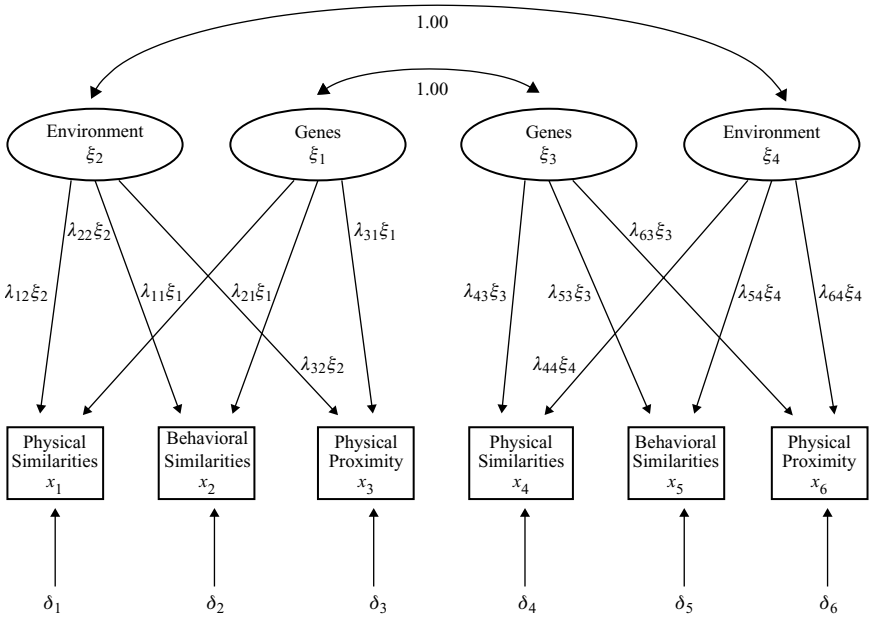

Figure 1.2. A model of sibling relatedness, in which the squares denote observed variables; the circles denote latent variables; the $\lambda_{ij}\xi_j$ are paths connecting latent with observed variables; and the δ_i are errors in the observed variables.

Figure 1.2 presents a simple example of a proposed theoretical model about sibling relatedness from the field of behavioral genetics. For years researchers have tried to understand the the "nature–nurture" phenomena by studying monozygotic twins, dizygotic twins, and nontwin siblings. To assess the amount of "relatedness" between siblings, researchers often use different types of questionnaire, standardized scales and tests, and independent observations. Two possible sources of relatedness between siblings are each sibling's genotype and environment. One may therefore define two different latent variables (i.e., genotype and environment) for each sibling, and denote each latent variable in the model by using the Greek letter ξ (ksi). Three possible observable variables (measures) of genotype and environment might be physical similarities, behavioral similarities, and physical proximity (Segal *et al.*, 1997). As it turns out, the scores or results observed for individuals on these variables will make up the correlation or covariance matrix that is analyzed to test a proposed model. The x values, which represent the observed variables or so-called *indicators*, are representative of the latent variables and make up the LAMBDA ($\mathbf{\Lambda}_x$) matrix. The error terms

(error of measurement in each indicator) are denoted by the Greek letter δ (delta) and are assumed to be associated with each indicator.

As indicated previously, the hypothesized relationships among the various observed and/or latent variables in a model are typically the primary focus of most SEM investigations. These relationships are represented graphically by one-way and two-way arrows in the path diagram. These arrows or *paths* are often interpreted as symbolizing a functional relationship. In other words, the variable at the end of the arrow is assumed to be affected by the variable at the beginning of the path. Two-way arrows are representative of a covariance or association between the connected variables. These paths are not directional in nature, but are interpreted as correlational. Note that in Figure 1.2 the two-way arrow between the latent genotype variables has been set to 1, based upon known genetic relatedness between monozygotic twins, and that the two-way arrow between the latent environment variables has been set to 1 as well. This setting of the two-way arrow between the environments of the monozygotic twins forces the environment latent variables to be interpreted as the twins' shared environment, or those environmental influences completely common to the twins.

The path coefficients from the proposed model are subsequently derived from the following *model definition equations*:

$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \delta_2$$
$$x_3 = \lambda_{31}\xi_1 + \lambda_{32}\xi_2 + \delta_3$$
$$x_4 = \lambda_{43}\xi_3 + \lambda_{44}\xi_4 + \delta_4$$
$$x_5 = \lambda_{53}\xi_3 + \lambda_{54}\xi_4 + \delta_5$$
$$x_6 = \lambda_{63}\xi_3 + \lambda_{64}\xi_4 + \delta_6$$

where x_1 is the observed physical similarities for twin 1; x_2 is the observed behavioral similarities for twin 1; x_3 is the observed physical proximity for twin 1; x_4 is the observed physical similarities for twin 2; x_5 is the observed behavioral similarities for twin 2; x_6 is the observed physical proximity for twin 2; $\lambda_{11}\xi_1$ to $\lambda_{64}\xi_4$ are the factor loadings that will be estimated based on the observed data; ξ_1 and ξ_3 are the genetic latent variables for twins 1 and 2, respectively; ξ_2 and ξ_4 are the environmental latent variables for twins 1 and 2, respectively; and δ_1 through δ_6 are the measurement errors attributed to a particular variable.[1]

[1] In the context of behavioral genetic modeling, the errors-in-variables (δ) not only represent measurement error but environmental influences *unique* to each twin.

These coefficients or parameters can be *free*, i.e., to be estimated from the collected data; *fixed*, i.e., set to some selected constant value; or *constrained*, i.e., set equal to one or more other parameters. In this model, both the correlations between the monozygotic twins' genotypes and their environments have been fixed to 1 on the basis of quantitative genetic theory (Plomin *et al.*, 1997). Further, note that the variances of ξ_1 to ξ_4 have been fixed to 1 as well. This is done to establish a metric for the latent variables. Since latent variables cannot be measured directly, it is difficult to work numerically with them without first assigning them some scale of measurement. A natural choice is to standardize these variances to a value of 1. In addition, comparable paths between the two twins should be constrained to be equal, since there is no reason to believe genetic or environmental effects will be stronger for one twin or the other:

$$\lambda_{11}\xi_1 = \lambda_{43}\xi_3$$
$$\lambda_{21}\xi_1 = \lambda_{53}\xi_3$$
$$\lambda_{31}\xi_1 = \lambda_{63}\xi_3$$

for the genetic paths, and

$$\lambda_{12}\xi_2 = \lambda_{44}\xi_4$$
$$\lambda_{22}\xi_2 = \lambda_{54}\xi_4$$
$$\lambda_{32}\xi_2 = \lambda_{64}\xi_4$$

for the environmental paths.

Comparable measurement errors should similarly be constrained as equal between the two twins; i.e.,

$$\delta_1 = \delta_4$$
$$\delta_2 = \delta_5$$
$$\delta_3 = \delta_6.$$

Model identification

With the definition and specification of the model complete, the next important consideration is the identification of the model. It is important to note that once the model and the parameters to be estimated are specified, the parameters are combined to form a model-implied variance–covariance matrix that will be tested against the observed variance–covariance matrix (i.e., the variance–covariance matrix obtained from the empirical data). In

a general way, the amount of unique information in the observed variance–covariance matrix is what will determine whether the model will be identified, and this verification procedure must be performed before any model can be appropriately tested. As it turns out, there are three levels of identification in SEM. The first and most problematic is that of an *under-identified* model. An *under-identified* model exists if one or more parameters cannot be estimated from the observed variance–covariance matrix. This type of model should be looked at with skepticism because the parameter estimates are most likely quite unstable. A *just-identified* model is a model that utilizes all of the uniquely estimable parameters. This type of model will always result in a "perfect fit" to the empirical data. Since there is no way one can really test or confirm the plausibility of a *just-identified* model (also referred to as a *saturated* model), this type of model is also problematic. As it turns out, the most desirable type of identification is the *over-identified* model. This type of model occurs when the number of available variance–covariances (units of information) is greater in number than the number of parameters to be estimated (Marcoulides & Hershberger, 1997). In other words, there is more than one way to estimate the specified parameters. The difference between the number of nonredundant elements of the variance–covariance matrix and the number of model parameters to be estimated is known as the *degrees of freedom* (df) of the model. For example, if the number of nonredundant elements of a variance–covariance matrix was 20 and 10 parameters were required to estimate the model, the degrees of freedom would be 20 minus 10, i.e., 10.

Model identification is an extremely complicated topic and requires several procedures to verify the status of a proposed model (for further discussion, see Marcoulides & Hershberger, 1997, or Raykov & Marcoulides, 2000). The *t*-rule, $\frac{1}{2}\, p(p+1)$ as cited by Marcoulides & Hersberger (1997), is one of the most frequently used necessary identification rules. Basically, the *t*-rule for identification is that the number of nonredundant elements in the variance–covariance (or correlation) matrix of the observed variables (p) must be greater than or equal to the number of unknown parameters in the proposed model (Marcoulides & Hersberger, 1997, p. 225). For example, Figure 1.2 has six observed variables or ($p = 6$), so there are $6(7)/2 = 21$ nonredundant elements in the variance–covariance matrix. If we attempt to estimate each path from an observed variable (x_1 to x_6) to each latent variable (ξ) and each error term associated with each observed variable (δ), we are estimating a total of 12 parameters. However, the six paths for twin 1 have been constrained to equal the six paths of twin 2 (e.g., $\lambda_{12}\xi_2 = \lambda_{44}\xi_4$), and the three indicator errors of twin 1 have been

constrained to equal the three indicator errors of twin 2 (e.g., $\delta_1 = \delta_4$), resulting in a reduction of six parameters to be estimated, or, altogether, only six parameters are to be estimated. Therefore, we have an *over-identified* model with 15 degrees of freedom (i.e., 21 unique elements of the variance–covariance matrix minus six parameters to be estimated = 15 df). Of course, it is important to note that having positive degrees of freedom in a proposed model is only a necessary condition for identification; it is not a sufficient condition. There can be cases in which the degrees of freedom for a proposed model are positive and yet some parameters remain under-identified (Raykov & Marcoulides, 2000).

Suppose we wanted to expand our proposed model and incorporate another latent variable. In behavior genetic modeling, each of the twins' observed variables is corrected for age, since twins within a pair are necessarily of the same age – age creating a spurious source of twin similarity. If age is incorporated as a latent variable, the new model appears as Figure 1.3. Note that the genetic and environmental latent variables are now symbolized by

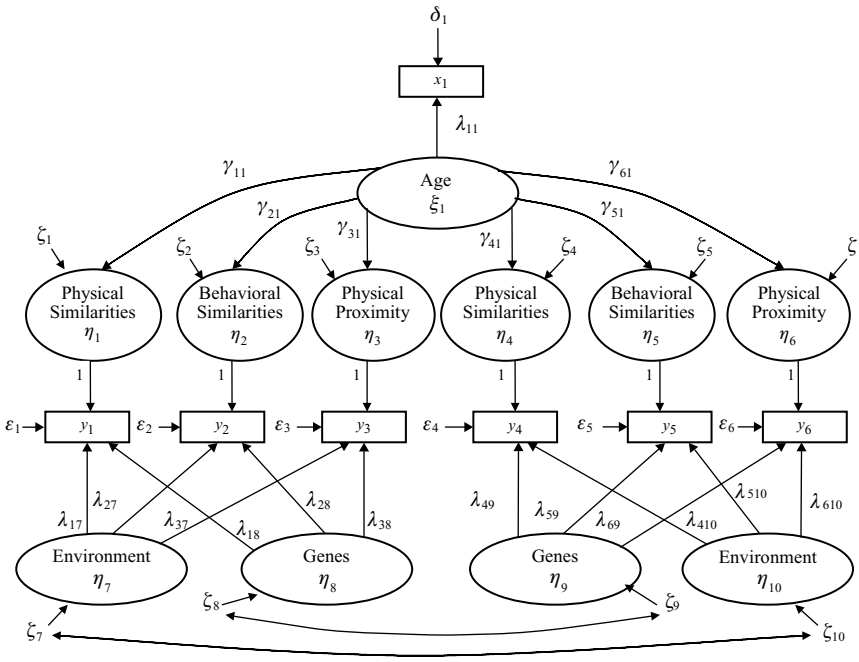

Figure 1.3. The model of sibling relatedness extended, with age as a covariate. Symbols are explained in the text.

the Greek letter η (eta). The indicators of these two latent variables, previously designated as x variables, are now designated as y variables. Further, these y values of the latent variables make up the so-called LAMBDA-y ($\mathbf{\Lambda}_y$) matrix. The error terms for the y variables are now denoted by the Greek letter ε (epsilon) and are associated (and correspondingly numbered) with each indicator. Age is now the single x variable in the model, with its latent variable denoted by ξ (ksi) and its error as δ (delta). The path connecting the observed variable age with the latent variable age is now the only entry in the LAMBDA-x ($\mathbf{\Lambda}_x$) matrix.

Let us now consider the questions "Why have we changed the symbolism of the x variables to y, and why has age been incorporated into the model as an x variable?" The answer to these two questions lies in the distinction between *dependent* and *independent variables*. *Dependent* (or "endogenous") *variables* are those variables that receive at least one path (one-way arrow) from another variable in the model. *Independent* (or "exogenous") *variables* are those variables from which paths only emanate but to which none is directed. Independent variables can be correlated among each other (i.e., connected in the path diagram by two-way arrows). It is important to note that a dependent variable may act as an independent variable with respect to one variable, but this does not change its dependent variable status. As long as there is at least one path ending into the variable, it is considered to be a dependent variable, no matter how many other dependent variables in the model are explained by that variable (Raykov & Marcoulides, 2000).

It was necessary to incorporate age as an independent variable based on our desire to have age act as a covariate of the original six observed indicators. Note the path that connects the latent age variable to each of the six original indicators. That parameter is expressed with the one-way arrow from one latent factor to the other and is represented by the Greek letter γ (gamma)[2]. There are other features to note about the new model in Figure 1.3. The loading of the age indicator variable on the latent age factor in the $\mathbf{\Lambda}_x$ matrix has been fixed to 1, with the indicator age error term (δ) set

[2] The reader will also note another change made to the original six indicators. Now each indicator has been the sole indicator of a latent η variable. For theoretical reasons, this change was unnecessary, but for practical reasons, this change was required. The LISREL program used to solve the model parameters only defines a parameter (γ) connecting latent independent with latent independent variables and not a parameter that connects latent independent variables with observed dependent variables. Again, this restriction requires a symbolic reformulation of the model but not one that is either theoretical or substantive.

to be zero and its variance freed. Due to identification difficulties, whenever a single indicator exists for a latent variable, a choice must be made between solving for either the value of the loading or the variance of the variable. Measurement errors are not generally identifiable with a single indicator. Another important feature to note is the two-way arrows connecting the dependent latent variables (η). This would seem to be in contradiction to the statement made above that only independent, and not dependent, latent variables may be correlated in a path model. Examining the two-way arrows in Figure 1.3, it is apparent that they do *not* directly connect the latent dependent variables, but rather connect another parameter denoted by the Greek letter ζ (zeta). Each dependent latent variable (η) has one ψ which represents the *residual error* in the variable. In other words, ψ represents all of the influences on the latent dependent variables not explicitly accounted for in the model. Some authors refer to these residual errors as *structural errors*. The two-way arrows between the ζ values of the twins' latent dependent variables is algebraically equivalent to the original formulation in Figure 1.2 of having the two-way arrows directly connect the twins' latent independent variables. Further, now the variance of the ζ values has been fixed to 1 in order to establish a metric for the latent dependent variables.

Our original model in Figure 1.3 was over-identified with 15 df. The addition of an observed age variable requires that we recalculate the degrees of freedom of the model. As before, comparable paths connecting the indicators to the latent variables should also be equated between the two twins (e.g., $\lambda_1\eta_1 = \lambda_4\eta_2$), as well as comparable indicator errors (e.g., $\varepsilon_1 = \varepsilon_4$) and comparable gammas (e.g., $\gamma_1 = \gamma_4$). In addition, we will be estimating the variance of age, and not its error or loading on ξ. In total, the model requires that 10 parameters be estimated, with the proper constraints imposed on the model (i.e., $3\,\mathbf{\Lambda}_y$, $3\,\varepsilon$, and $3\,\gamma$ values, and 1 variance). Using the *t*-rule to determine our model identification, one finds that there are $7(8)/2 = 28$ nonredundant elements in the variance–covariance matrix. Thus, our degrees of freedom are $28 - 10$ or 18 df, which results in an over-identified model suitable for model estimation.

Model estimation

In any SEM model, paths or *parameters* are estimated in such a way that the model becomes capable of "*emulating*" the observed sample variance–covariance (or correlation) matrix. The proposed theoretical model represented by the path diagram and equations makes certain assumptions about

the relationships between the involved variables, and hence has specific implications for their variances and covariances. It turns out that these implications can be worked out using a few simple relations that govern the variances and covariances of linear combinations of variables. These relations are illustrated below (for further details, see Raykov & Marcoulides, 2000, p. 19).

Let us denote variance by the letters Var and covariance by the letters Cov. For variable y (e.g., physical similarities) the first relation is stated as follows:

- Relation 1: $\text{Cov}(y, y) = \text{Var}(y)$.

This relation simply states that the covariance of a variable with itself is equal to that variable's variance.

- Relation 2: $\text{Cov}(a\,x + b\,y, c\,z + d\,u) = a\,c\,\text{Cov}(x, z) + a\,d\,\text{Cov}(x, u) + b\,c\,\text{Cov}(y, z) + b\,d\,\text{Cov}(y, u)$.

The second relation allows one to find out the covariance of two linear combinations of variables. Suppose that a, b, c, and d are four constants and assume that x, y, z, and u are four variables, e.g., those denoting the scores on tests of physical similarities, behavioral similarities, physical proximity, and age. The relation is obtained according to the product of the constants with the attached covariance of each combination of variables.

- Relation 3: $\text{Var}(a\,x + b\,y) = \text{Cov}(a\,x + b\,y, a\,x + b\,y) = a^2\text{Cov}(x, x) + b^2\text{Cov}(y, y) + ab\text{Cov}\,(x, y) + ab\text{Cov}\,(x, y)$,

which, on the basis of Relation 1, leads to $a^2\text{Var}(x) + b^2\text{Var}(y) + 2a\,b\,\text{Cov}(x, y)$.

This relation simply states that the variance of a linear combination of variables is equal to their covariance (e.g., see Relation 1). And in the case that variables x and y are uncorrelated (i.e., $\text{Cov}(x, y) = 0$), leads to $\text{Var}(a\,x + b\,y) = a^2\text{Var}(x) + b^2\text{Var}(y)$.

Any proposed theoretical model has certain implications for the variances and covariances (and the means if considered) of the involved observed variables. In order to see these implications, the above three relations are generally used. For example, consider the first two manifest variables y_1 and y_2 presented in Figure 1.3. Because both variables load on the same latent factors η_1 and η_2 we obtain directly from Relations 1 and 2 the

following equality:

$$\begin{aligned}
\mathrm{Cov}(y_1, y_2) &= ((\lambda_{11} \times \eta_1) + (\lambda_{12} \times \eta_2) + (1 \times \eta_5) + \varepsilon_1, (\lambda_{21} \times \eta_1) \\
&\quad + (\lambda_{22} \times \eta_2) + (1 \times \eta_6) + \varepsilon_2) \\
&= (\lambda_{11} \times \eta_1)(\lambda_{21} \times \eta_1) + (\lambda_{12} \times \eta_2)(\lambda_{21} \times \eta_1) \\
&\quad + (\eta_5)(\lambda_{21} \times \eta_1) + (\lambda_{11} \times \eta_1)(\lambda_{22} \times \eta_2) \\
&\quad + (\lambda_{12} \times \eta_2)(\lambda_{22} \times \eta_2) + (\eta_5)(\lambda_{22} \times \eta_2) \\
&\quad + (\lambda_{11} \times \eta_1)(\eta_6) + (\lambda_{12} \times \eta_2)(\eta_6) + (\eta_5)(\eta_6) \\
&\quad + \varepsilon_1((\lambda_{21} \times \eta_1) + (\lambda_{22} \times \eta_2) + \eta_6) \\
&\quad + \varepsilon_2((\lambda_{11} \times \eta_1) + (\lambda_{12} \times \eta_2) + \eta_5) + (\varepsilon_1, \varepsilon_2) \\
&= (\lambda_{11}\lambda_{21} \times \mathrm{Var}(\eta_1)) + (\lambda_{12}\lambda_{22} \times \mathrm{Cov}(\eta_1, \eta_2)) \\
&\quad + (\lambda_{21} \times \mathrm{Cov}(\eta_1, \eta_5)) + (\lambda_{11}\lambda_{22} \times \mathrm{Cov}(\eta_2, \eta_1)) \\
&\quad + (\lambda_{12}\lambda_{22} \times \mathrm{Var}(\eta_2)) + (\lambda_{22} \times \mathrm{Cov}(\eta_2, \eta_5)) \\
&\quad + (\lambda_{11} \times \mathrm{Cov}(\eta_1, \eta_6)) + (\lambda_{22} \times \mathrm{Cov}(\eta_2, \eta_6)) \\
&\quad + \mathrm{Cov}(\eta_5, \eta_6) + \mathrm{Cov}(\varepsilon_1, \lambda_{21}) + \mathrm{Cov}(\varepsilon_1, \lambda_{22}) \\
&\quad + \mathrm{Cov}(\varepsilon_1, \eta_6) + \mathrm{Cov}(\varepsilon_2, \lambda_{11}) + \mathrm{Cov}(\varepsilon_2, \lambda_{12}) \\
&\quad + \mathrm{Cov}(\varepsilon_2, \eta_5) + \mathrm{Cov}(\varepsilon_1, \varepsilon_2).
\end{aligned}$$

However, considerable simplification of the above expression is possible, since $\mathrm{Var}(\eta_1) = \mathrm{Var}(\eta_2) = 1$,

$$\begin{aligned}
\mathrm{Cov}(\eta_1, \eta_2) &= \mathrm{Cov}(\eta_1, \eta_5) = \mathrm{Cov}(\eta_2, \eta_5) = \mathrm{Cov}(\eta_1, \eta_6) \\
&= \mathrm{Cov}(\eta_2, \eta_6) = \mathrm{Cov}(\eta_5, \eta_6) = 0,
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Cov}(\varepsilon_1, \lambda_{21}) + \mathrm{Cov}(\varepsilon_1, \lambda_{22}) + \mathrm{Cov}(\varepsilon_1, \eta_6) + \mathrm{Cov}(\varepsilon_2, \lambda_{11}) \\
&\qquad + \mathrm{Cov}(\varepsilon_2, \lambda_{12}) + \mathrm{Cov}(\varepsilon_2, \eta_5) + \mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0.
\end{aligned}$$

Therefore:

$$\mathrm{Cov}(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22}.$$

If this process were continued for every combination of p observed variables (i.e., y_1 to y_6 and x_1), the result would be the determination of every element of a model-implied variance–covariance matrix. This matrix can be denoted by $\boldsymbol{\Sigma}$ (the capital Greek letter sigma) and is generally referred to as the *reproduced* (or *model implied*) *covariance matrix*. For the proposed model in Figure 1.3, the reproduced covariance matrix in Table 1.1 is determined.

Table 1.1. *Reproduced covariance matrix of y_1 through y_6 and x_1 for MZ twins*

$\lambda_{11}^2 + \lambda_{12}^2 + \mathrm{Var}(\eta_5) + \varepsilon_1$							
$\lambda_{11} \times \lambda_{21} + \lambda_{12} \times \lambda_{22} + \gamma_1\gamma_2 \times \mathrm{Var}(\xi_1)$	$\lambda_{21}^2 + \lambda_{22}^2 + \mathrm{Var}(\eta_6) + \varepsilon_2$						
$\lambda_{11} \times \lambda_{31} + \lambda_{12} \times \lambda_{32} + \gamma_1\gamma_3 \times \mathrm{Var}(\xi_1)$	$\lambda_{21} \times \lambda_{31} + \lambda_{22} \times \lambda_{32} + \gamma_2\gamma_3 \times \mathrm{Var}(\xi_1)$	$\lambda_{31}^2 + \lambda_{32}^2 + \mathrm{Var}(\eta_7) + \varepsilon_3$					
$\lambda_{11} \times \lambda_{43} + \lambda_{12} \times \lambda_{44} + \gamma_1\gamma_4 \times \mathrm{Var}(\xi_1)$	$\lambda_{21} \times \lambda_{43} + \lambda_{22} \times \lambda_{44} + \gamma_2\gamma_4 \times \mathrm{Var}(\xi_1)$	$\lambda_{31} \times \lambda_{43} + \lambda_{32} \times \lambda_{44} + \gamma_3\gamma_4 \times \mathrm{Var}(\xi_1)$	$\lambda_{43}^2 + \lambda_{44}^2 + \mathrm{Var}(\eta_8) + \varepsilon_4$				
$\lambda_{11} \times \lambda_{53} + \lambda_{12} \times \lambda_{54} + \gamma_1\gamma_5 \times \mathrm{Var}(\xi_1)$	$\lambda_{21} \times \lambda_{53} + \lambda_{22} \times \lambda_{54} + \gamma_2\gamma_5 \times \mathrm{Var}(\xi_1)$	$\lambda_{31} \times \lambda_{53} + \lambda_{32} \times \lambda_{54} + \gamma_3\gamma_5 \times \mathrm{Var}(\xi_1)$					
$\lambda_{43} \times \lambda_{53} + \lambda_{44} \times \lambda_{54} + \gamma_4\gamma_5 \times \mathrm{Var}(\xi_1)$	$\lambda_{53}^2 + \lambda_{54}^2 + \mathrm{Var}(\eta_9) + \varepsilon_5$						
$\lambda_{11} \times \lambda_{63} + \lambda_{12} \times \lambda_{64} + \gamma_1\gamma_6 \times \mathrm{Var}(\xi_1)$	$\lambda_{21} \times \lambda_{63} + \lambda_{22} \times \lambda_{64} + \gamma_2\gamma_6 \times \mathrm{Var}(\xi_1)$	$\lambda_{31} \times \lambda_{63} + \lambda_{32} \times \lambda_{64} + \gamma_3\gamma_6 \times \mathrm{Var}(\xi_1)$					
$\lambda_{43} \times \lambda_{63} + \lambda_{44} \times \lambda_{64} + \gamma_4\gamma_6 \times \mathrm{Var}(\xi_1)$	$\lambda_{53} \times \lambda_{63} + \lambda_{54} \times \lambda_{64} + \gamma_5\gamma_6 \times \mathrm{Var}(\xi_1)$	$\lambda_{63}^2 + \lambda_{64}^2 + \mathrm{Var}(\eta_{10}) + \varepsilon_6$					
$\gamma_1 \times \mathrm{Var}(\xi_1)$	$\gamma_4 \times \mathrm{Var}(\xi_1)$	ξ_1	$\gamma_2 \times \mathrm{Var}(\xi_1)$	$\gamma_5 \times \mathrm{Var}(\xi_1)$	$\gamma_3 \times \mathrm{Var}(\xi_1)$	$\gamma_6 \times \mathrm{Var}(\xi_1)$	

Note: Rows 6 and 8 are continuations of rows 5 and 7 respectively. Row 9 contains all the elements but because of spatial restrictions does not align with the elements in rows 5–9.

It is important to note that the elements of $\boldsymbol{\Sigma}$ are all *functions of model parameters.* In addition, each element of $\boldsymbol{\Sigma}$ has as a counterpart a corresponding numerical element (entry) in the observed sample covariance matrix obtained for the seven observed variables considered (i.e., y_1 to y_6 and x_1). Assume that the observed covariance matrix (denoted by **S**) was as follows:

$$\begin{bmatrix} 16.51 & & & & & & \\ 4.26 & 28.12 & & & & & \\ -2.10 & 3.38 & 3.35 & & & & \\ 10.04 & 3.40 & -2.50 & 16.51 & & & \\ 3.40 & 10.35 & 1.74 & 4.36 & 28.12 & & \\ -2.50 & 1.74 & 1.12 & -2.10 & 3.38 & 3.35 & \\ 2.44 & 3.18 & 1.10 & 2.44 & 3.18 & 1.10 & 4.00 \end{bmatrix}.$$

For example, the top element of **S** (i.e., 16.51) corresponds to $\lambda_{11}^2 + \lambda_{12}^2 + \text{Var}(\eta_5) + \varepsilon_1$ in the reproduced matrix $\boldsymbol{\Sigma}$. Now imagine setting the counterpart elements of **S** and $\boldsymbol{\Sigma}$ equal to one another. That is, according to the proposed model displayed in Figure 1.3, set $16.51 = \lambda_{11}^2 + \lambda_{12}^2 + \text{Var}(\eta_5) + \varepsilon_1$, then $4.26 = \lambda_{11} \times \lambda_{21} + \lambda_{12} \times \lambda_{22} + \gamma_1\gamma_2 \times \text{Var}(\xi_1)$, and so on until, for the last element of **S**, $4.00 = \xi_1$ is set. As a result of this equality setting, a system of 28 equations (i.e., the number of nonredundant elements, with 21 covariances and 7 variances) is generated. Thus one can conceive of the process of fitting a structural equation model as a way of solving a system of equations. For each equation, its left-hand side is a subsequent numerical entry of the sample observed variance–covariance matrix **S** while its right-hand side is the corresponding expression of model parameters defined in the $\boldsymbol{\Sigma}$ matrix. Hence, fitting a structural equation model is conceptually "equivalent" to solving in an optimal way (discussed in the next section) this system of equations obtained according to the proposed model. This discussion also demonstrates that the model presented in Figure 1.3, like any structural equation model, *implies* a specific structuring of the elements of the covariance matrix reproduced by the model in terms of specific expressions (functions) of unknown model parameters. Therefore, if certain values for the parameters were entered into these functions, one would obtain a covariance matrix that has numbers as elements. In fact, the process of fitting a model to data with SEM programs can be thought of as repeated "insertion" of appropriate values for the parameters in the matrix $\boldsymbol{\Sigma}$ until a certain optimality criterion (discussed in the next section) in terms of its proximity to the matrix **S** is satisfied. Every available SEM program has built into its "memory" the exact way in which these functions of model parameters in $\boldsymbol{\Sigma}$

can be obtained. Although for ease of computation most programs make use of matrix algebra, the programs in effect determine each of the expressions presented in the above-mentioned 28 equations (for further discussion, see Marcoulides & Hershberger, 1997). Fortunately, this occurs quite automatically once the user has communicated to the program the model parameters.

Model assessment and fit

The previous section illustrated how a proposed SEM model leads to the reproduction of a variance–covariance matrix $\boldsymbol{\Sigma}$ that is then fit to the observed sample variance–covariance matrix **S**. Now it would seem that the next logical question is "How can one measure or evaluate the extent to which the matrices **S** and $\boldsymbol{\Sigma}$ differ?". As it turns out, this question is particularly important in SEM because it actually permits one to evaluate the goodness of fit of the model. Indeed, if the difference between **S** and $\boldsymbol{\Sigma}$ is negligible, then one can conclude that the model represents the observed data reasonably well. On the other hand, if the difference is large, one can conclude that the proposed model is not consistent with the observed data. There are at least two reasons for such inconsistencies: (1) the proposed model may be deficient, in the sense that it is not capable of "emulating" the analyzed matrix even with most favorable parameter values; and/or (2) the data may not be good. Thus, in order to proceed with assessing model fit, we need a method for evaluating the degree to which the reproduced matrix $\boldsymbol{\Sigma}$ differs from the sample covariance matrix **S**.

In order to clarify this method, a new concept is introduced, that of distance between matrices. Obviously, if the values to be compared were scalars (single numbers) a simple subtraction of one from the other (and possibly taking the absolute value of the resulting difference) would suffice to evaluate the distance between them. However, this cannot be done directly with the two matrices **S** and $\boldsymbol{\Sigma}$. Subtracting the matrix **S** from the matrix $\boldsymbol{\Sigma}$ does not result in a single number. Rather, a matrix of differences is obtained.

Fortunately, there are some meaningful ways to assess the distance between two matrices and, interestingly, the resulting distance measure is a single number that is easier to interpret. Perhaps the simplest way to obtain this single number involves taking the sum of squares of the differences between the corresponding elements of the two matrices. Other more complicated ways involve a multiplication of these squares with some appropriately chosen weights and then taking their sum. Perhaps the most commonly used weight is based on maximum likelihood estimation. In either case, the single number represents a sort of generalized distance measure between the

two matrices considered. The bigger the number, the more different are the matrices, while the smaller the number, the more similar are the matrices.

Because in SEM this number results after comparison of the elements of **S** with those of the model-implied covariance matrix $\boldsymbol{\Sigma}$, the generalized distance is a function of the model parameters as well as the elements of the observed variances and covariances. Therefore it is customary to refer to the relationship between the matrix distance, on the one hand, and the model parameters and **S** on the other, as a *fit function* that is typically denoted by *F*. Since it equals the distance between two matrices, *F* is always equal to a positive value or 0. Whenever the value of *F* is 0, then the two matrices considered are identical.

Before particular measures of model fit are discussed, a word of warning is in order. Even if all possible fit indices point to an acceptable model, one can *never* claim to have found the *true* model that has generated the analyzed data (of course, we exclude from consideration the cases where data are simulated according to a preset known model). SEM is most concerned with finding a model that does not contradict the data. That is to say, in an empirical session of SEM, one is typically interested in retaining the proposed model whose validity is the essence of the null hypothesis. Statistically speaking, when using SEM methodology, one is usually interested in not rejecting the null hypothesis (Raykov & Marcoulides, 2000, p. 34).

When testing a model for fit, the complete fit of the model as well as the individual parameters should be examined. Typically, choosing the appropriate fit statistic is difficult for many researchers. One of the most widely used statistics for assessing the fit of a model is the χ^2 (chi-square) goodness-of-fit statistic. This statistic is an assessment of the magnitude of difference between the initial observed covariance matrix and the reproduced matrix. The probability level that is associated with this statistic indicates whether the difference between the reproduced matrix and the original data is significant or not. A significant χ^2 test states that the difference between the two matrices is due to sampling error or variation. Typically, researchers are most interested in a nonsignificant χ^2 test. This indicates that the observed matrix and the reproduced matrix are not statistically different, therefore indicating a good fit of the model to the data. However, the χ^2 test suffers from several weaknesses, including a dependence on sample size, and vulnerability to departures of the data from multivariate normality. Thus it is suggested that a researcher should examine a number of fit criteria in addition to the χ^2 value to assess the fit of the proposed model (Raykov & Marcoulides, 2000).

To assist in the process of assessing model fit, there are many other descriptive fit statistics that are typically formulated in values that range from

1 (perfect fit) to zero (no fit). One of the more popular fit indices is the goodness-of-fit index (GFI), which can loosely be considered as a measure of the proportion of variance and covariance that the proposed model is able to explain. If the number of parameters is also taken into account then the resulting index is the adjusted goodness of fit (AGFI) (Raykov & Marcoulides, 2000, p. 38). Unfortunately, there is not a strict norm for these indices. As a rough guide, it is currently viewed that a model with a GFI or AGFI of 0.95 or above may well represent a reasonably good approximation of the data (Hu & Bentler, 1999). Quite a few other indices of model fit have been developed, each with its own strengths and weaknesses. For more comprehensive discussions of evaluating model fit, see Bollen & Long (1993) or Marsh *et al.* (1996).

The fit indices proposed above were concerned with evaluating the fit of the entire model. Although this is certainly useful to have, one should also be interested in how well various parts of the model fit. It is entirely possible for the model as a whole to fit well, but for individual sections not to fit well. Aside from this, if a model does not fit well, it is of considerable value to determine which parts of the model are contributing to model misfit. Perhaps the most useful way to determine the fit of specific sections of the model is to examine the residual matrix (Bollen, 1989). The residual matrix results from the difference between the $\mathbf{S}$ and $\mathbf{\Sigma}$ matrices. The individual residual covariances (or correlations) are $(s_{ij} - \sigma_{ij})$ where s_{ij} is the ij-th element of $\mathbf{S}$ and σ_{ij} is the corresponding element in $\mathbf{\Sigma}$. A positive residual means that the model underpredicts the covariance between two variables, whereas a negative one means that the model overpredicts the covariance. Of course it can be difficult to interpret the absolute magnitude of the residuals, since the magnitude of a residual is in part a function of the scaling of the two variables. Thus, examining the *correlation residuals* or the *normalized residuals* can frequently better convey a sense of the fit of a specific part of a model (Jöreskog & Sörbom, 1996).

Model modification

The requirement for SEM is that the details of the proposed model be known before the model is fit and tested with data (Marcoulides & Drezner, 2001). Often, however, theories are poorly developed and require changes or adjustments throughout the testing process. Jöreskog & Sörbom (1996) have addressed three types of situation that concern model fitting and testing. The first situation is the *strictly confirmatory* notion in which the initial model is tested against empirical data and is either accepted or rejected. The second

type is the *competing or alternative model* situation. This procedure entails several proposed models that are then assessed and selected on the basis of which model more appropriately fits the observed data. The final situation is the *model generating* technique in which the scientist repeatedly modifies the proposed model until some level of fit is acquired. The decision as to which procedure will be utilized is based on the initial theory. A researcher who is firmly entrenched in his or her theory or hypotheses will conduct SEM differently from a scientist who is unsure of the interrelationships between the observed and latent variables. No matter how SEM is conducted, however, once a researcher attempts to respecify an initial model after it has been rejected by the data, the process of confirmation is over. Now SEM enters into an *exploratory* mode, in which the researcher searches for revisions to the model that will most significantly increase its fit to the data. These revisions usually entail freeing a previously fixed parameter and/or fixing a previously free parameter. Such a process of exploration is generally referred to as a *specification search* (Leamer, 1978).

All SEM computer programs come equipped with various statistics to assist in the specification search. Two of the most popular statistics are the *modification index* (MI) and the *t-ratio* (Jöreskog & Sörbom, 1996). The MI is used to determine which parameter, if freed, would contribute most to an increase in model fit and indicates the amount the χ^2 goodness-of-fit statistic would decrease if in fact the parameter were specified in the model. (Recall that with 1 df, a single parameter would significantly improve the fit of a model if it decreased the goodness of fit χ^2 by at least 8.841 points, $p < 0.05$.) On the other hand, the t-ratio assesses the significance of the individual parameters in a specified model; t-ratios of less than 2 are generally considered nonsignificant, $p = 0.05$. Presumably, those parameters which are not significant may be removed from the model without causing the model to fit significantly more badly (i.e., without causing the χ^2 goodness-of-fit statistic to increase significantly). Generally, the best strategy is first to determine which parameters should be added to the model by examining their individual MIs; then, once the list of significant MIs has been exhausted, the t-ratios should be examined to decide which parameters should be deleted from the model (Marcoulides & Hershberger, 1997). Marcoulides & Drezner (2001) have also proposed automated specification search procedures based on genetic algorithms and Tabu search procedures.

On the surface, the availability of MIs, t-ratios, similar indices, and automated specification searches may appear to be of tremendous benefit to the process of model respecification. However, certain cautionary remarks are in order. First, parameters should be added (or deleted) to the model one

at a time, each time the model is re-evaluated and the indices recalculated because changes in the model result (sometimes) in dramatic changes in the values of the indices. In other words, with one parameter not in the model, another parameter may appear to be potentially significant on the basis of its MI, but, with the addition of the first parameter, the significance of the second parameter's MI disappears (fortunately, this issue is addressed in the automated specification searches proposed by Marcoulides and his colleagues). Second, as can well be imagined, even covariance matrices of moderate size (for instance, our example of a 7×7 covariance matrix) may make possible the specification of hundreds of free parameters in a model. Leaving aside the desirability of any one of these parameters, the possibility of Type I errors looms (Green *et al.*, 1998). Green *et al.* (1999) have proposed methods for controlling Type I errors during SEM specification searches. Third, even though adding a parameter may cause the model to finally fit, if the parameter is theoretically meaningless or statistically suspect, it should be avoided. Similarly even though a parameter may appear to be nonsignificant as indicated by its small *t*-value, it should *not* be removed from a model if it is considered theoretically or logically important.

Multisample models

Before we introduce the LISREL program and its approach to the evaluation of the model in Figure 1.3, a final, critical issue must be addressed. Although we stated earlier that according to the *t*-rule, the model was identified, this is in fact not so. The *t*-rule is a necessary but not sufficient criterion for model identification. Rather than delve into a complex discussion as to why the model is not identified, or introduce alternative sufficient criteria for model identification, a simple demonstration will suffice to show why this model in under-identified. Remember that the primary reason for solving this behavioral genetic model is to estimate genetic and environmental influences on the observed variables. Recall also that we used one type of family relation to do this, monozygotic or MZ twins. Since MZs share *all* of their genes and *all* of their common environments, we can express the covariance between MZs for an observed variables as

$$\text{Cov(MZ)} = \text{Var(G)} + \text{Var(E)} + 2\text{Cov(G, E)},$$

where G denotes genotype and E environment. However, our model stipulates no covariance between G and E, so the MZ covariance simplifies to

$$\text{Cov(MZ)} = \text{Var(G)} + \text{Var(E)}.$$

Astute readers will no doubt question how we are to solve for two unknowns, the variance of G and the variance of E, with only one observed statistic – the covariance between MZ twins. The answer is that we cannot. Including the variance of the observed variables is of no help, for its expression is identical to that for Cov(MZ):

$$\text{Var(MZ)} = \text{Var(G)} + \text{Var(E)}.$$

How then are we to identify this model?

The solution is actually quite simple, and pre-dates the existence of SEM methodology. If we also include family members of genetic relatedness differing from that of MZs, we are now able to solve for the genetic and environmental variances of the observed variables. Traditionally, dizygotic, or DZ, twins have been used in conjunction with MZ twins, to solve for the values of these variances. This method is referred to as the "classical twin method", first used by Galton in the 1870s (Eaves *et al.*, 1989). DZ twins are a useful group to compare with MZ twins, since DZs share on average only half their genes but all of their common environments. Thus

$$\text{Cov(DZ)} = 0.5 \times \text{Var(G)} + \text{Var(E)}.$$

In fact, if we had only a single observed variable, the genetic variance of that variable[3] would be expressed as

$$\text{Var(G)} = 2 \times (\text{Cov(MZ)} - \text{Cov(DZ)}).$$

Therefore, in order to solve for the genetic and environmental variances of our model, we include a sample of MZ twins as well as one of DZ twins. For our model, the **S** (observed covariance matrix) for the DZ twins is

$$\begin{bmatrix} 16.51 & & & & & & \\ 4.26 & 28.12 & & & & & \\ -2.10 & 3.38 & 3.35 & & & & \\ 5.02 & 1.70 & -1.25 & 16.51 & & & \\ 1.70 & 5.18 & 0.87 & 4.36 & 28.12 & & \\ -1.25 & 0.87 & 0.56 & -2.10 & 3.38 & 3.35 & \\ 2.44 & 3.18 & 1.10 & 2.44 & 3.18 & 1.10 & 4.00 \end{bmatrix}$$

[3] Both the SEM model and the simple solution of the classical twin method (i.e., $2 \times (\text{Cov(MZ)} - \text{Cov(DZ)})$) rely on the validity of certain assumptions. If these assumptions are incorrect, then the estimate of genetic variance will be inaccurate. Among these assumptions are: (1) all genetic effects are additive (i.e., linear), (2) the covariance between genetic and environmental effects is zero, and (3) there is no assortative mating for the observed variable. For an extended discussion of the meaning and likelihood of these assumptions being met, see Eaves *et al.* (1989).