JOURNAL OF PERSONALITY ASSESSMENT, *87*(1), 35–50

Structural Equation Modeling: Reviewing the Basics and Moving Forward

Jodie B. Ullman
Department of Psychology
California State University, San Bernardino

This tutorial begins with an overview of structural equation modeling (SEM) that includes the purpose and goals of the statistical analysis as well as terminology unique to this technique. I will focus on confirmatory factor analysis (CFA), a special type of SEM. After a general introduction, CFA is differentiated from exploratory factor analysis (EFA), and the advantages of CFA techniques are discussed. Following a brief overview, the process of modeling will be discussed and illustrated with an example using data from a HIV risk behavior evaluation of homeless adults (Stein & Nyamathi, 2000). Techniques for analysis of nonnormally distributed data as well as strategies for model modification are shown. The empirical example examines the structure of drug and alcohol use problem scales. Although these scales are not specific personality constructs, the concepts illustrated in this article directly correspond to those found when analyzing personality scales and inventories. Computer program syntax and output for the empirical example from a popular SEM program (EQS 6.1; Bentler, 2001) are included.

In this article, I present an overview of the basics of structural equation modeling (SEM) and then present a tutorial on a few of the more complex issues surrounding the use of a special type of SEM, confirmatory factor analysis (CFA) in personality research. In recent years SEM has grown enormously in popularity. Fundamentally, *SEM* is a term for a large set of techniques based on the general linear model. After reviewing the statistics used in the *Journal of Personality Assessment* over the past 5 years, it appears that relatively few studies employ structural equation modeling techniques such as CFA, although many more studies use exploratory factor analysis techniques (EFA). SEM is a potentially valuable technique for personality assessment researchers to add to their analysis toolkit. It may be particularly helpful to those already employing EFA. Many different types of research questions can be addressed through SEM. A full tutorial on SEM is outside the scope of this article, therefore the focus of this is on one type of SEM, CFA, which might be of particular interest to researchers who analyze personality assessment data.

OVERVIEW OF SEM

SEM is a collection of statistical techniques that allow a set of relations between one or more independent variables (IVs), either continuous or discrete, and one or more dependent variables (DVs), either continuous or discrete, to be examined. Both IVs and DVs can be either measured variables (directly observed), or latent variables (unobserved, not directly observed). SEM is also referred to as causal modeling, causal analysis, simultaneous equation modeling, analysis of covariance structures, path analysis, or CFA. The latter two are actually special types of SEM.

A model of substance use problems appears in Figure 1. In this model, Alcohol Use Problems and Drug Use Problems are latent variables (factors) that are not directly measured but rather assessed indirectly, by eight measured variables that represent targeted areas of interest. Notice that the factors (often called latent variables or constructs) are signified with circles. The observed variables (measured variables) are signified with rectangles. These measured variables could be items on a scale. Instead of simply combining the items into a scale by taking the sum or average of the items, creating a composite containing measurement error, the scale items are employed as indicators of a latent construct. Using these items as indicators of a latent variable rather than components of a scale allows for estimation and removal of the measurement error associated with the observed variables. This model is a type of SEM analysis called a CFA. Often in

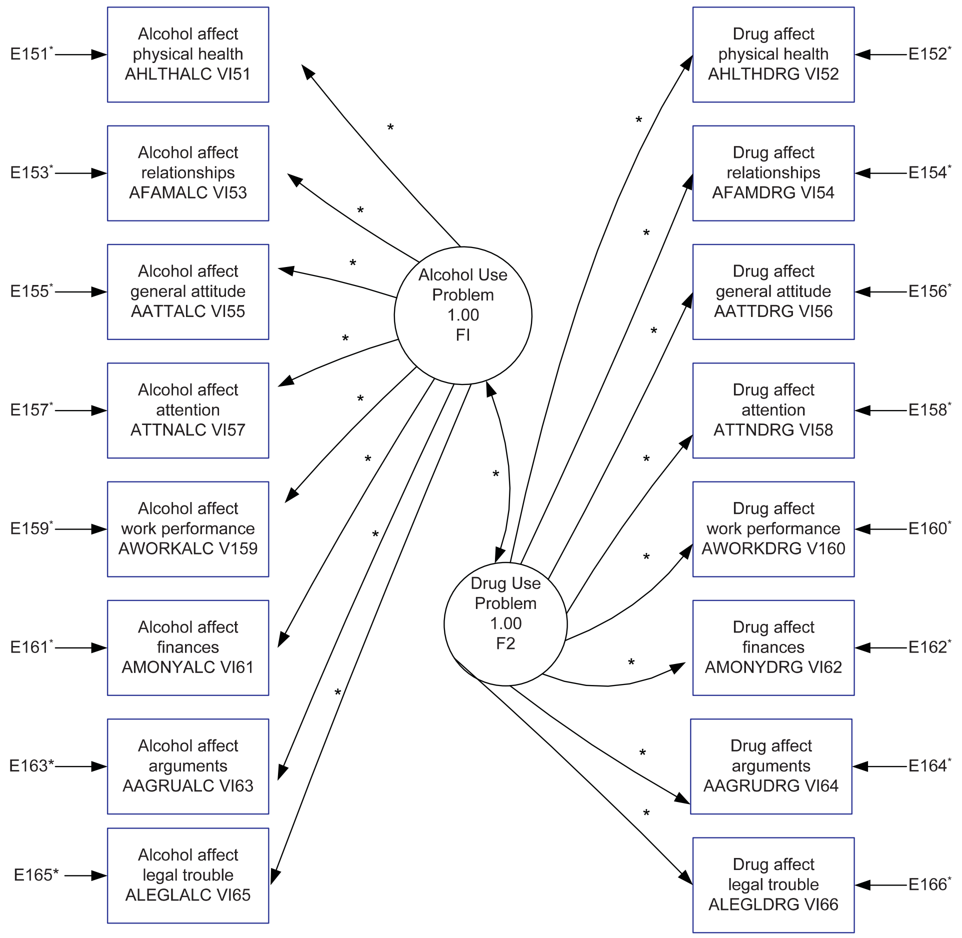


FIGURE 1 Hypothesized substance use problem confirmatory factor analysis model. Alcohol Use Problems: Physical health = AHLTHALC; relationships = AFAMALC; general attitude = AATTALC; attention = AATTNALC; work = AWORKALC; money = AMONEYALC; arguments = AARGUALC; legal trouble = ALEGLALC; Drug Use Problems: physical health = AHLTHDRG; relationships = AFAMDRG; general attitude = AATTDRG; attention = AATTNDRG; work = AWORKDRG; money = AMONEYDRG; arguments = AARGUDRG; legal trouble = ALEGLDRG.

later stages of research, after exploratory factor analyses (EFA), it is helpful to confirm the factor structure with new data using CFA techniques.

Path Diagrams and Terminology

Figure 1 is an example of a path diagram. Diagrams are fundamental to SEM because they allow the researcher to diagram the hypothesized set of relations—the model. The diagrams are helpful in clarifying a researcher's ideas about the relations among variables. There is a one to one correspondence between the diagrams and the equations needed for the analysis. For clarity in the text, initial capitals are used for names of factors and lowercase letters for names of measured variables.

Several conventions are used in developing SEM diagrams. Measured variables, also called *observed variables, indicators,* or *manifest variables* are represented by squares or rectangles. Factors have two or more indicators and are also called *latent variables*, *constructs*, or *unobserved variables.* Circles or ovals in path diagrams represent factors. Lines indicate relations between variables; lack of a line connecting variables implies that no direct relationship has been hypothesized. Lines have either one or two arrows. A line with one arrow represents a hypothesized direct relationship between two variables. The variable with the arrow pointing to it is the DV. A line with an arrow at both ends indicates a covariance between the two variables with no implied direction of effect.

In the model of Figure 1, both Alcohol Use Problems and Drug Use Problems are latent variables. Latent variables are unobservable and require two or more measured indicators. In this model the indicators, in rectangles, are predicted by the latent variable. There are eight measured variable indicators (problems with health, family, attitudes, attention, work, money, arguments, and legal issues) for each latent variable. The line with an arrow at both ends, connecting Alcohol Use Problems and Drug Use Problems, implies that there is a covariance between the latent variables. Notice the direction of the arrows connecting each construct (factor) to its' indicators: The construct predicts the measured variables. The theoretical implication is that the underlying constructs, Alcohol Use Problems and Drug Use Problems, drive the degree of agreement with the statements such as "How often has your use of alcohol affected your medical or physical health?" and "How often has your use of drugs affected your attention and concentration?" We are unable to measure these constructs directly, so we do the next best thing and measure indicators of Alcohol and Drug Use Problems.

Now look at the other lines in Figure 1. Notice that there is another arrow pointing to each measured variable. This implies that the factor does not predict the measured variable perfectly. There is variance (residual) in the measured variable that is not accounted for by the factor. There are no lines with single arrows pointing to Alcohol Problems and Drug Problems, these are independent variables in the model. Notice all of the measured variables have lines with single headed arrows pointing to them, these variables are dependent variables in the model. Notice that all the DVs have arrows labeled "E" pointing toward them. This is because nothing is predicted perfectly; there is always residual or error. In SEM, the residual, the variance not predicted by the IV(s), is included in the diagram with these paths.

The part of the model that relates the measured variables to the factors is called the *measurement model*. In this example, the two constructs (factors), Alcohol Use Problems and Drug Use Problems and the indicators of these constructs (factors) form the *measurement model.* The goal of an analysis is often to simply estimate the measurement model. This type of analysis is called CFA and is common after researchers have established hypothesized relations between measured variables and underlying constructs. This type of analysis addresses important practical issues such as the validity of the structure of a scale. In the example illustrated in Figure 1, theoretically we hope that we are able to tap into homeless adults' Alcohol and Drug Use problems by measuring several observable indicators. However, be aware that although we are interested in the theoretical constructs of Alcohol Use Problems and Drug Use Problems we are essentially defining the construct by the indicators we have chosen to use. Other researchers also interested in Alcohol and Drug Use Problems could define these constructs with completely different indicators and thus define a somewhat different construct. A common error in SEM CFA analyses is to forget that we have defined the construct by virtue of the measured variables we have chosen to use in the model.

The hypothesized relations among the constructs, in this example, the single covariance between the two constructs, could be considered the *structural model.* Note, the model presented in Figure 1 includes hypotheses about relations among variables (covariances) but not about means or mean differences. Mean differences can also be tested within the SEM framework but are not demonstrated in this article.

How Does CFA Differ From EFA?

In EFA the researcher has a large set of variables and hypothesizes that the observed variables may be linked together by virtue of an underlying structure; however, the researcher does not know the exact nature of the structure. The goal of an EFA is to uncover this structure. For example the EFA might determine how many factors exist, the relationship between factors, and how the variables are associated with the factors. In EFA various solutions are estimated with varying numbers of factors and various types of rotation. The researcher chooses among the solutions and selects the best solution based on theory and various descriptive statistics. EFA, as the name suggests, is an exploratory technique. After a solution is selected, the reproduced correlation matrix, calculated from the factor model, can be empirically compared to the sample correlation matrix.

CFA, is as the name implies a confirmatory technique. In a CFA the researcher has a strong idea about the number of factors, the relations among the factors, and the relationship between the factors and measured variables. The goal of the analysis is to test the hypothesized structure and perhaps test competing theoretical models about the structure. Factor extraction and rotation are not part of confirmatory factor analyses.

CFA is typically performed using sample covariances rather than the correlations used in EFA. A covariance could be thought of as an unstandardized correlation. Correlations indicate degree of linear relationships in scale-free units whereas covariances indicate degree of linear relationships in terms of the scale of measurement for the specific variables. Covariances can be converted to correlations by dividing the covariance by the product of the standard deviations of each variable.

Perhaps one of the most important differences between EFA and CFA is that CFA offers a statistical test of the comparison between the estimated unstructured population covariance matrix and the estimated structured population covariance matrix. The sample covariance matrix is used as an estimate of the unstructured population covariance matrix and the parameter estimates in the model combine to form the estimated structured population covariance matrix. The hypothesized CFA model provides the underlying structure for the estimated population covariance matrix. Said another way, the idea is that the observed sample covariance matrix is an estimate of the unstructured population covariance matrix. In this unstructured matrix there are $(p(p + 1))/2$, where p = number of measured variables, separate parameters (variances and covariances). However, we hypothesize that this covariance matrix is a really function of fewer parameters, that is, has an underlying simpler structure. This underlying structure is the given by the hypothesized CFA model. If the CFA model is justified, then we conclude that the relationships observed in the covariance matrix can be explained with fewer parameters than the $(p(p + 1))/2$ nonredundant elements of the sample covariance matrix.

There are a number of advantages to the use of SEM. When relations among factors are examined, the relations are theoretically free of measurement error because the error has been estimated and removed, leaving only common variance. Reliability of measurement can be accounted for explicitly within the analysis by estimating and removing the measurement error. In addition, as seen in Figure 1, complex relations can be examined. When the phenomena of interest are complex and multidimensional, SEM is the only analysis that allows complete and simultaneous tests of all the relations. In the social sciences we often pose hypotheses at the level of the construct. With other statistical methods these construct level hypotheses are tested at the level of a measured variable (an observed variable with measurement error). Mismatching the level of hypothesis and level of analysis although problematic, and often overlooked, may lead to faulty conclusions. A distinct advantage of SEM is the ability to test construct level hypotheses at the appropriate level.

THREE GENERAL TYPES OF RESEARCH QUESTIONS THAT CAN BE ADDRESSED WITH SEM

The focus of this article is on techniques and issues especially relevant to a type of SEM analysis called CFA. At least three questions may be answered with this type of analysis

1. Do the parameters of the model combine to estimate a population covariance matrix (estimated structured covariance matrix) that is highly similar to the sample covariance matrix (estimated unstructured covariance matrix)?
2. What are the significant relationships among variables within the model?
3. Which nested model provides the best fit to the data?

In the following section these three general types of research questions will be discussed and examples of types of hypotheses and models will be presented.

Adequacy of Model

The fundamental question that is addressed through the use of CFA techniques involves a comparison between a data set, an empirical covariance matrix (technically this is the estimated unstructured population covariance matrix), and an estimated structured population covariance matrix that is produced as a function of the model parameter estimates. The major question asked by SEM is, “Does the model produce an estimated population covariance matrix that is consistent with the sample (observed) covariance matrix?” If the model is good the parameter estimates will produce an estimated structured population covariance matrix that is close to the sample covariance matrix. “Closeness” is evaluated primarily with the chi-square test statistic and fit indexes. Appropriate test statistics and fit indexes will be discussed later.

It is possible to estimate a model, with a factor structure, at one time point and then test if the factor structure, that is, the measurement model, remains the same across time points. For example, we could assess the strength of the indicators of Drug and Alcohol Use Problems when young adults are 18 years of age and then assess the same factor structure when the adults are 20, 22, and 24. Using this longitudinal approach we could assess if the factor structure, the construct itself, remains the same across this time period or if the relative weights of the indicators change as young adults develop.

Using multiple-group modeling techniques it is possible to test complex hypotheses about moderators. Instead of using young adults as a single group we could divide the sample into men and women, develop single models of Drug and Alcohol Use Problems for men and women separately and then compare the models to determine if the measurement structure was the same or different for men and women, that is, does gender moderate the structure of substance use problems.

Significance of Parameter Estimates

Model estimates for path coefficients and their standard errors are generated under the implicit assumption that the model fit is very good. If the model fit is very close, then the estimates and standard errors may be taken seriously, and individual significance tests on parameters (path coefficients, variances, and covariances) may be performed. Using the example illustrated in Figure 1, the hypothesis that Drug Use problems are related to Alcohol Use problems can be tested. This would be a test of the null hypothesis that there is no covariance between the two latent variables, Alcohol Use Problems and Drug Use Problems. This parameter estimate (covariance) is then evaluated with a z test (the parameter estimate divided by the estimated standard error). The null hypothesis is the same as in regression, the path coefficient equals zero. If the path coefficient is significantly larger than zero then there is statistical support for the hypothesized predictive relationship.

Comparison of Nested Models

In addition to evaluating the overall model fit and specific parameter estimates, it is also possible to statistically compare nested models to one another. Nested models are models that are subsets of one another. When theories can be specified as nested hypotheses, each model might represent a different theory. These nested models are statistically compared, thus providing a strong test for competing theories (models). Notice in Figure 1 the items are identical for both drug use problems and alcohol use problems. Some of the common variance among the items may be due to wording and the general domain area (e.g., health problems, relationships problems), not solely due to the underlying substance use constructs. We could compare the model given in Figure 1 to a model that also includes paths that account for the variance explained by the general domain or wording of the item. The model with the added paths to account for this variability would be considered the full model. The model in Figure 1 would be thought of as nested within this full model. To test this hypothesis, the chi-square from the model with paths added to account for domain and wording would be subtracted from the chi-square for the nested model in Figure 1 that does not account for common domains and wording among the items. The corresponding degrees of freedom for these two models would also be subtracted. Given nested models and normally distributed data, the difference between two chi-squares is a chi-square with degrees of freedom equal to the difference in degrees of freedom between the two models. The significance of the chi-square difference test can then be assessed in the usual manner. If the difference is significant, the fuller model that includes the extra paths is needed to explain the data. If the difference is not significant, the nested model, which is more parsimonious than the fuller model, would be accepted as the preferred model. This hypothesis is examined in the empirical section of this article.

AN EMPIRICAL EXAMPLE—THE STRUCTURE OF SUBSTANCE USE PROBLEMS IN HOMELESS ADULTS

The process of modeling could be thought of as a four-stage process: model specification, model estimation, model evaluation, and model modification. These stages will be discussed and illustrated with data collected as part of a study that examines risky sex behavior in homeless adults (for a compete discussion of the study, see Nyamathi, Stein, Dixon, Longshore, & Galaif, 2003; Stein & Nyamathi, 2000). The primary goal of this analysis is to determine if a set of items that query both alcohol and drug problems are adequate indicators of two underlying constructs: Alcohol Use Problems and Drug Use Problems.

Model Specification/Hypotheses

The first step in the process of estimating a CFA model is model specification. This stage consists of: (a) stating the hypotheses to be tested in both diagram and equation form, (b) statistically identifying the model, and (c) evaluating the statistical assumptions that underlie the model. This section contains discussion of each of these components using the problems with drugs and alcohol use model (Figure 1) as an example.

Model hypotheses and diagrams. In this phase of the process, the model is specified, that is, the specific set of hypotheses to be tested is given. This is done most frequently through a diagram. The problems with substance use diagram given in Figure 1 is an example of hypothesis specification. This example contains 16 measured variables. Descriptive statistics for these Likert scaled (0 to 4) items are presented in Table 1.

In these path diagrams the asterisks indicate parameters to be estimated. The variances and covariances of IVs are parameters of the model and are estimated or fixed to a particular value. The number 1.00 indicates that a parameter, either a

TABLE 1
Descriptive Statistics for Measured Variables

			Skewness			*Kurtosis*		
Construct Variable	*M*	*SD*		*SE*	*Z*		*SE*	*Z*
Alcohol Use Problems								
Physical Health (AHLTHALC)	0.80	1.28	1.39	.09	15.41*	0.57	.02	3.19*
Relationships (AFAMALC)	1.18	1.52	0.82	.09	9.11*	–0.92	.02	–5.09*
Attitude (AATTALC)	1.23	1.52	0.74	.09	8.17*	–1.03	.02	–5.71*
Attention (AATTNALC)	1.21	1.53	0.80	.09	8.92*	–0.93	.02	–5.16*
Work (AWORKALC)	1.10	1.54	0.96	.09	10.68*	–0.72	.02	–4.00*
Finances (AMONYALC)	1.24	1.61	0.79	.09	8.72*	–1.08	.02	–6.03*
Arguments(AARGUALC)	1.19	1.54	0.82	.09	9.12*	–0.94	.02	–5.21*
Legal (ALEGLALC)	0.84	1.39	1.38	.09	15.29*	0.34	.02	1.90
Drug Use Problems								
Physical Health (AHLTHDRG)	1.21	1.57	0.81	.09	9.04*	–0.98	.02	–5.42*
Relationships (AFAMDRG)	1.59	0.71	0.38	.09	4.21*	–1.59	.02	–8.82*
Attitude (AATTDRG)	1.54	1.67	0.43	.09	4.73*	–1.51	.02	–8.41*
Attention (AATTNDRG)	1.54	1.69	0.43	.09	4.75*	–1.54	.02	–8.58*
Work (AWORKDRG)	1.47	1.74	0.53	.09	5.90*	–1.52	.02	–8.43*
Finances (AMONYDRG)	1.72	1.81	0.26	.09	2.84	–1.77	.02	–9.82*
Arguments (AARGUDRG)	1.46	1.68	0.54	.09	5.94*	–1.43	.02	–7.95*
Legal (ALEGLDRG)	1.15	1.61	0.94	.09	10.39*	–0.86	.02	–4.77*

Note. $N = 736$.
$^{*}p < .001$.

path coefficient or a variance, has been set (fixed) to the value of 1.00. In this figure the variance of both factors (F1 and F2) have been fixed to 1.00. The regression coefficients are also parameters of the model. (The rationale behind "fixing" paths will be discussed in the section about identification).

We hypothesize that the factor, Alcohol Use Problems, predicts the observed problems with physical health (AHLTHALC), relationships (AFAMALC), general attitude (AATTALC), attention (AATTNALC), work (AWORKALC), money (AMONEYALC), arguments (AARGUALC), and legal trouble (ALEGLALC) and the factor, Drug Use Problems, predicts problems with physical health (AHLTHDRG), relationships (AFAMDRG), general attitude (AATTDRG), attention (AATTNDRG), work (AWORKDRG), money (AMONEYDRG), arguments (AARGUDRG), and legal trouble (ALEGLDRG).

It is also reasonable to hypothesize that alcohol problems may be correlated to drug problems. The double-headed arrow connecting the two latent variables indicates this hypothesis. Carefully examine the diagram and notice the other major hypotheses indicated. Notice that each measured variable is an indicator for just one factor, this is sometimes called simple structure in EFA.

Model statistical specification. The relations in the diagram are directly translated into equations and the model is then estimated. One method of model specification is the Bentler–Weeks method (Bentler & Weeks, 1980). In this method every variable in the model, latent or measured, is either an IV or a DV. The parameters to be estimated are the (a) regression coefficients, and (b) the variances and the covariances of the independent variables in the model (Bentler, 2001). In Figure 1 the regression coefficients and covariances to be estimated are indicated with an asterisk. Initially, it may seem odd that a residual variable is considered an IV but remember the familiar regression equation, the residual is on the right hand side of the equation and therefore is considered an IV:

$$Y = X\beta + e, \tag{1}$$

where Y is the DV and X and e are both IVs.

In fact the Bentler–Weeks model is a regression model, expressed in matrix algebra:

$$\boldsymbol{\eta} = \boldsymbol{\beta\eta} + \boldsymbol{\gamma\xi} \tag{2}$$

where if q is the number of DVs and r is the number of IVs, then $\boldsymbol{\eta}$ (eta) is a $q \times 1$ vector of DVs, $\boldsymbol{\beta}$ (beta) is a $q \times q$ matrix of regression coefficients between DVs, $\boldsymbol{\gamma}$ (gamma) is a $q \times r$ matrix of regression coefficients between DVs and IVs, and $\boldsymbol{\xi}$ (xi) is a $r \times 1$ vector of IVs.

This model is different from ordinary regression models because of the possibility of having latent variables as DVs and predictors as well as the possibility of DVs predicting other DVs. In the Bentler–Weeks model only independent variables have variances and covariances as parameters of the model. These variances and covariances are in $\boldsymbol{\phi}$ (phi), an $r \times r$ matrix. Therefore, the parameter matrices of the model are $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\phi}$.

The model in the diagram can be directly translated into a series of equations. For example the equation predicting problems with health due to alcohol, (AHLTHALC, V151) is, V151 = *F1 + E151, or in Bentler–Weeks notation, $\eta_1 = \hat{\gamma}_{1,17}\xi_{17} + \xi_1$, , where the symbols are defined as above and we estimate, $\hat{\gamma}_{1,17}$, the regression coefficient predicting the measured variable AHLTHALC from Factor 1, Alcohol Use Problems.

There is one equation for each dependent variable in the model. The set of equations forms a syntax file in EQS (Bentler, 2001; a popular SEM computer package). The syntax for this model is presented in the Appendix. An asterisk indicates a parameter to be estimated. Variables included in the equation without asterisks are considered parameters fixed to the value 1.

Model identification. A particularly difficult and often confusing topic in SEM is identification. A complete discussion is outside the scope of this article. Therefore only the fundamental issues relevant to the empirical example will be discussed. The interested reader may want to read Bollen (1989) for an in-depth discussion of identification. In SEM a model is specified, parameters (variances and covariances of IVs and regression coefficients) for the model are estimated using sample data, and the parameters are used to produce the estimated population covariance matrix. However only models that are identified can be estimated. A model is said to be identified if there is a unique numerical solution for each of the parameters in the model. The following guidelines are rough, but may suffice for many models.

The first step is to count the number of data points and the number of parameters that are to be estimated. The data in SEM are the variances and covariances in the sample covariance matrix. The number of data points is the number of nonredundant sample variances and covariances,

$$\text{Number of data points} = \frac{p(p+1)}{2}, \tag{3}$$

where p equals the number of measured variables. The number of parameters is found by adding together the number of regression coefficients, variances, and covariances that are to be estimated (i.e., the number of asterisks in a diagram).

A required condition for a model to be estimated is that there are more data points than parameters to be estimated. Hypothesized models with more data than parameters to be estimated are said to be *over identified.* If there are the same number of data points as parameters to be estimated, the model is said to be *just identified.* In this case, the estimated parameters perfectly reproduce the sample covariance matrix, and the chi-square test statistic and degrees of freedom are equal to zero, hypotheses about the adequacy of the model cannot be tested. However, hypotheses about specific paths in the model can be tested. If there are fewer data points than parameters to be estimated, the model is said to be *under identified* and parameters cannot be estimated. The number of parameters needs to be reduced by fixing, constraining, or deleting some of them. A parameter may be fixed by setting it to a specific value or constrained by setting the parameter equal to another parameter.

In the substance use problem example of Figure 1, there are 16 measured variables so there are 136 data points: 16(16 +1)/2 = 136 (16 variances and 120 covariances). There are 33 parameters to be estimated in the hypothesized model: 16 regression coefficients, 16 variances, and 1 covariance. The hypothesized model has 103 fewer parameters than data points, so the model may be identified.

The second step in determining model identifiability is to examine the measurement portion of the model. The measurement part of the model deals with the relationship between the measured indicators and the factors. In this example the entire model is the measurement model. It is necessary both to establish the scale of each factor and to assess the identifiability of this portion of the model.

Factors, in contrast to measured variables, are hypothetical and consist, essentially of common variance, as such they have no intrinsic scale and therefore need to be scaled. Measured variables have scales, for example, income may be measured in dollars or weight in pounds. To establish the scale of a factor, the variance for the factor is fixed to 1.00, or the regression coefficient from the factor to one of the measured variables is fixed to 1.00. Fixing the regression coefficient to 1 gives the factor the same variance as the common variance portion of the measured variable. If the factor is an IV, either alternative is acceptable. If the factor is a DV the regression coefficient is set to 1.00. In the example, the variances of both latent variables are set to 1.00 (normalized). Forgetting to set the scale of a latent variable is easily the most common error made when first identifying a model.

Next, to establish the identifiability of the measurement portion of the model the number of factors and measured variables is examined. If there is only one factor, the model may be identified if the factor has at least three indicators with nonzero loading and the errors (residuals) are uncorrelated with one another. If there are two or more factors, again consider the number of indicators for each factor. If each factor has three or more indicators, the model may be identified if errors associated with the indicators are not correlated, each indicator loads on only one factor and the factors are allowed to covary. If there are only two indicators for a factor, the model may be identified if there are no correlated errors, each indicator loads on only one factor, and none of the covariances among factors is equal to zero.

In the example, there are eight indicators for each factor. The errors are uncorrelated and each indicator loads on only one factor. In addition, the covariance between the factors is not zero. Therefore, this hypothesized CFA model may be identified. Please note that identification may still be possible if errors are correlated or variables load on more than one factor, but it is more complicated.

Sample size and power. SEM is based on covariances and covariances are less stable when estimated from small samples. So generally, large sample sizes are needed for SEM analyses. Parameter estimates and chi-square tests of fit are also very sensitive to sample size; therefore, SEM is a large sample technique. However, if variables are highly reliable it may be possible to estimate small models with fewer participants. MacCallum, Browne, and Sugawara (1996) presented tables of minimum sample size needed for tests of goodness-of-fit based on model degrees of freedom and effect size. In addition, although SEM is a large sample technique and test statistics are effected by small samples, promising work has been done by Bentler and Yuan (1999) who developed test statistics for small samples sizes.

Missing data. Problems of missing data are often magnified in SEM due to the large number of measured variables employed. The researcher who relies on using complete cases only is often left with an inadequate number of complete cases to estimate a model. Therefore missing data imputation is particularly important in many SEM models. When there is evidence that the data are missing at random (MAR; missingness on a variable may depend on other variables in the dataset but the missingness does not depend on the variable itself) or missing completely at random (MCAR; missingness is unrelated to the variable missing data or the variables in the dataset), a preferred method of imputing missing data, the EM (expectation maximization) algorithm to obtain maximum likelihood (ML) estimates, is appropriate (Little & Rubin, 1987). Some of the software packages now include procedures for estimating missing data, including the EM algorithm. EQS 6.1 (Bentler, 2004) produces the EM-based maximum likelihood solution automatically, based on the Jamshidian–Bentler (Jamshidian & Bentler, 1999) computations. LISREL and AMOS also produce EM-based maximum likelihood estimates. It should be noted that, if the data are not normally distributed, maximum likelihood test statistics—including those based on the EM algorithm—may be quite inaccurate. Although not explicitly included in SEM software, another option for treatment of missing data is multiple imputation see Schafer and Olsen (1998) for a nontechnical discussion of multiple imputation.

Multivariate normality and outliers. In SEM the most commonly employed techniques for estimating models assume multivariate normality. To assess normality it is often helpful to examine both univariate and multivariate normality indexes. Univariate distributions can be examined for outliers and skewness and kurtosis. Multivariate distributions are examined for normality and multivariate outliers. Multivariate normality can be evaluated through the use of Mardia's (1970) coefficient and multivariate outliers can be evaluated through evaluation of Mahalanobis distance.

Mardia's (1970) coefficient evaluates multivariate normality through evaluation of multivariate kurtosis. Mardia's coefficient can be converted to a normalized score (equivalent to a *z* score), often normalized coefficients greater than 3.00 are indicative of nonnormality (Bentler, 2001; Ullman, 2006). Mahalanobis distance is the distance between a case and the centroid (the multivariate mean with that data point removed). Mahalanobis distance is distributed as a chi-square with degrees of freedom equal to the number of measured variables used to calculate the centroid. Therefore, a multivariate outlier can be defined as a case that is associated with a Mahalanobis distance greater than a critical distance specified typically by a $p < .001$ (Tabachnick & Fidell, 2006).

In other multivariate analyses if variable distributions are nonnormal it is often necessary to transform the variable to create a new variable with a normal distribution. This can lead to difficulties in interpretation, for example, what does the square root of problems with alcohol mean? Sometimes despite draconian transformation, variables cannot be forced into normal distributions. Sometimes a normal distribution is just not reasonable for a variables, for example drug use. In SEM if variables are nonnormally distributed it is entirely reasonable and perhaps even preferable to choose an estimation method that addresses the nonnormality instead of transforming the variable. Estimation methods for nonnormality are discussed in a later section.

Returning to Table 1, using a criteria of $p < .001$, it is clear that all of the variables are either significantly skewed and kurtotic or both skewed and kurtotic. Although these variables are significantly skewed, and in some instances also kurtotic, using a conservative *p* value, the sample size in this analysis is very large ($N = 736$). Significance is a function of sample size and with large samples, very small departures from normality may lead to significant skewness and kurtosis coefficients and rejection of the normality assumption. Therefore, with a sample size such as this, it is important to consider other criteria such as visual shape of the distribution and also measures of multivariate normality such as Mardia's coefficient. Although not included in this article, several distributions (e.g., physical health problems related to alcohol use, legal problems related to alcohol use) do appear to be nonnormal. In a multivariate analysis multivariate normality is also important. As seen in Figure 2, the normalized estimate of Mardia's coefficient = 188.6838. This is a *z* score and even with consideration of sample size this is very large and therefore indicates that the variables multivariate distribution is nonnormal, $p < .0001$. There were no multivariate or univariate outliers in this dataset.

Model Estimation Techniques and Test Statistics

After the model specification component is completed the population parameters are estimated and evaluated. In this section we briefly discuss a few of the popular estimation techniques and provide guidelines for selection of estimation technique and test statistic. The applied reader who would like to read more on selection of an estimation method may want to refer to Ullman (2006), readers with more technical interests may want to refer to Bollen (1989).

The goal of estimation is to minimize the difference between the structured and unstructured estimated population covariance matrices. To accomplish this goal a function, *F*, is minimized where,

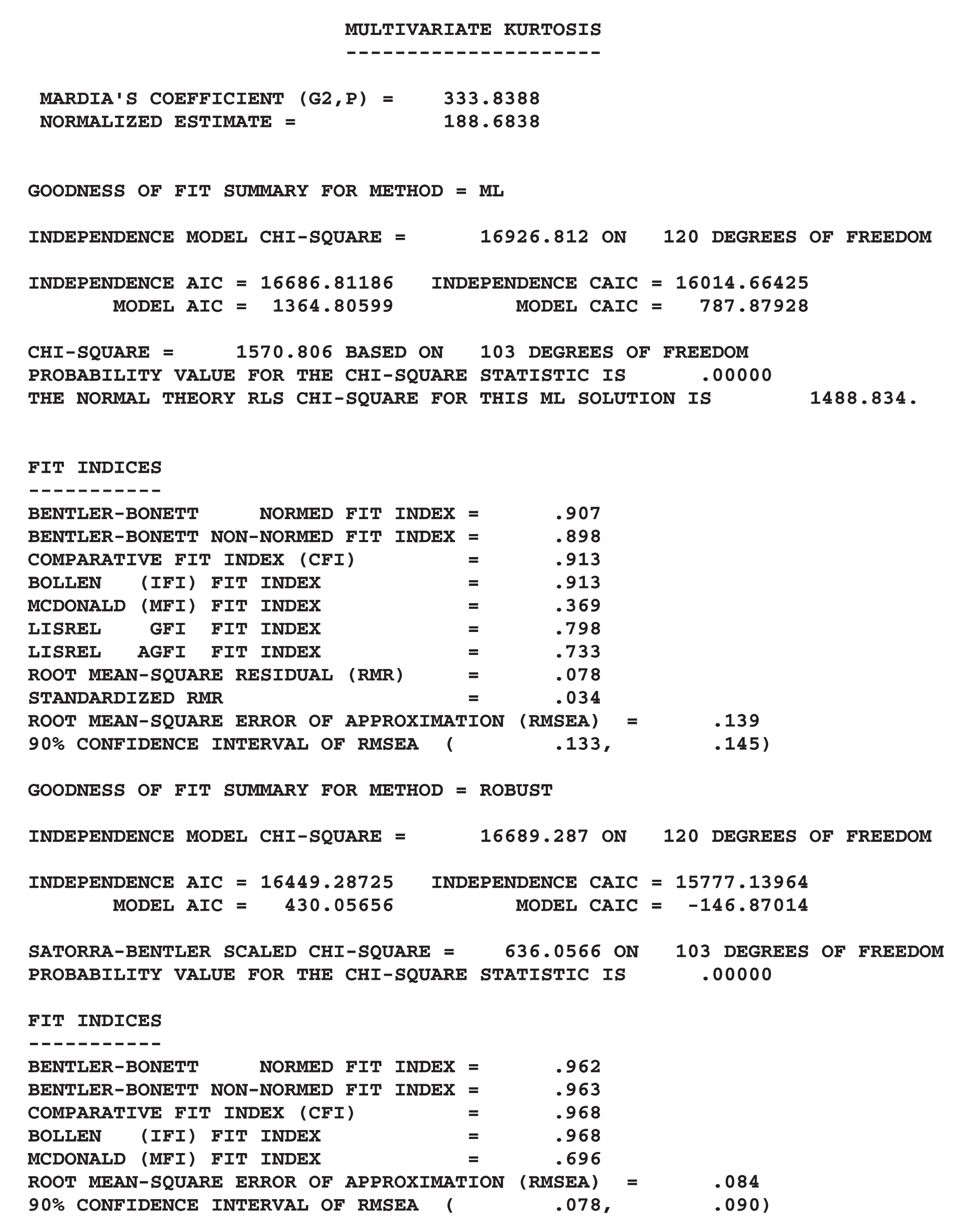
$$F = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\Theta}))\mathbf{W}(\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\Theta})), \qquad (4)$$

s is the vector of data (the observed sample covariance matrix stacked into a vector); $\boldsymbol{\sigma}$ is the vector of the estimated population covariance matrix (again, stacked into a vector) and $(\boldsymbol{\Theta})$ indicates that $\boldsymbol{\sigma}$ is derived from the parameters (the regression coefficients, variances and covariances) of the model. **W** is the matrix that weights the squared differences between the sample and estimated population covariance matrix.

In EFA the observed and reproduced correlation matrices are compared. This idea is extended in SEM to include a statistical test of the differences between the estimated structured and unstructured population covariance matrices. If the weight matrix, **W**, is chosen correctly, at the minimum with the optimal $\hat{\boldsymbol{\Theta}}$, F multiplied by $(N - 1)$ yields a chi-square test statistic.

There are many different estimation techniques in SEM, these techniques vary by the choice of **W**. Maximum likelihood (ML) is usually the default method in most programs because it yields the most precise (smallest variance) estimates when the data are normal. GLS (generalized least squares) has the same optimal properties as ML under normality. The ADF (asymptotically distribution free) method has no distributional assumptions and hence is most general (Browne, 1974; 1984), but it is impractical with many variables and inaccurate without very large sample sizes. Satorra and Bentler (1988, 1994, 2001) and Satorra (2000) also developed an adjustment for nonnormality that can be applied to the ML, GLS, or EDT chi-square test statistics. Briefly, the Satorra–Bentler scaled χ^2 is a correction to the χ^2 test statistic (Satorra & Bentler, 2001). EQS also corrects the standard errors for parameter estimates to adjust for the extent of nonnormality (Bentler & Dijkastra, 1985).

Some recommendations for selecting an estimation method. Based on Monte Carlo studies conducted by Hu, Bentler, and Kano (1992) and Bentler and Yuan (1999) some general guidelines can offered. Sample size and plausibility of the normality and independence assumptions need to be considered in selection of the appropriate estimation technique. ML, the Scaled ML, or GLS estimators may be good choices with medium (over 120) to large samples and evidence of the plausibility of the normality assumptions. ML estimation is currently the most frequently used estimation method in SEM.



```
                              MULTIVARIATE KURTOSIS
                              ---------------------

 MARDIA'S COEFFICIENT (G2,P) =    333.8388
 NORMALIZED ESTIMATE =            188.6838


GOODNESS OF FIT SUMMARY FOR METHOD = ML

INDEPENDENCE MODEL CHI-SQUARE =      16926.812 ON   120 DEGREES OF FREEDOM

INDEPENDENCE AIC = 16686.81186   INDEPENDENCE CAIC = 16014.66425
       MODEL AIC =  1364.80599          MODEL CAIC =   787.87928

CHI-SQUARE =     1570.806 BASED ON   103 DEGREES OF FREEDOM
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS      .00000
THE NORMAL THEORY RLS CHI-SQUARE FOR THIS ML SOLUTION IS        1488.834.


FIT INDICES
-----------
BENTLER-BONETT     NORMED FIT INDEX =      .907
BENTLER-BONETT NON-NORMED FIT INDEX =      .898
COMPARATIVE FIT INDEX (CFI)         =      .913
BOLLEN   (IFI) FIT INDEX            =      .913
MCDONALD (MFI) FIT INDEX            =      .369
LISREL    GFI  FIT INDEX            =      .798
LISREL   AGFI  FIT INDEX            =      .733
ROOT MEAN-SQUARE RESIDUAL (RMR)     =      .078
STANDARDIZED RMR                    =      .034
ROOT MEAN-SQUARE ERROR OF APPROXIMATION (RMSEA)  =      .139
90% CONFIDENCE INTERVAL OF RMSEA  (        .133,        .145)

GOODNESS OF FIT SUMMARY FOR METHOD = ROBUST

INDEPENDENCE MODEL CHI-SQUARE =      16689.287 ON   120 DEGREES OF FREEDOM

INDEPENDENCE AIC = 16449.28725   INDEPENDENCE CAIC = 15777.13964
       MODEL AIC =   430.05656          MODEL CAIC =  -146.87014

SATORRA-BENTLER SCALED CHI-SQUARE =    636.0566 ON   103 DEGREES OF FREEDOM
PROBABILITY VALUE FOR THE CHI-SQUARE STATISTIC IS      .00000

FIT INDICES
-----------
BENTLER-BONETT     NORMED FIT INDEX =      .962
BENTLER-BONETT NON-NORMED FIT INDEX =      .963
COMPARATIVE FIT INDEX (CFI)         =      .968
BOLLEN   (IFI) FIT INDEX            =      .968
MCDONALD (MFI) FIT INDEX            =      .696
ROOT MEAN-SQUARE ERROR OF APPROXIMATION (RMSEA)  =      .084
90% CONFIDENCE INTERVAL OF RMSEA  (        .078,        .090)
```

FIGURE 2 Heavily edited EQS 6.1 output for model estimation and test statistic information.

In medium (over 120) to large samples the scaled ML test statistic is a good choice with nonnormality or suspected dependence among factors and errors. In small samples (60 to 120) the Yuan–Bentler test statistic seems best. The test statistic based on the ADF estimator (without adjustment) seems like a poor choice under all conditions unless the sample size is very large (> 2,500). Similar conclusions were found in studies by Fouladi (2000), Hoogland (1999), and Satorra (2000). In this example the data are significantly nonnormal and our sample size is 736. Due to the nonnormality ML and GLS estimation are not appropriate. We have a reasonably large, but not huge (>2,500) sample, therefore we will use ML estimation with the Satorra–Bentler scaled chi-square.

Figure 2 shows heavily edited output for the model estimation and chi-square test. Several chi-square test statistics are given in the full output. In this severely edited output only the chi-squares associated with the Satorra–Bentler scaled chi-square and, for comparison, the ML chi-square are given. In the section labeled "Goodness of fit summary for method = robust, the independence model chi-square = 16689.287," with 120 degrees of freedom tests the hypothesis that the measured variables are orthogonal (independent). Therefore, the probability associated with this chi-square should be very small ($p <$ .05). The model chi-square test statistic is labeled "Satorra–Bentler scaled chi-square = 636.0566 based on 103 degrees of freedom" this tests the hypothesis that the difference between the estimated structured population covariance matrix and the estimated unstructured population covariance matrix (estimated using the sample covariance matrix) is not significant. Ideally, the probability associated with this chi-square should be large, greater than .05. In Figure 2 the probability associated with the model chi-square, p < .00001. (EQS reports probabilities only with 5 digits.) Strictly interpreted this indicates that the estimated population covariance matrix and the sample covariance matrix do differ significantly, that is, the model does not fit the data. However, the chi-square test statistic is strongly effected by sample size. The function minimum multiplied by $N-1$ equals the chi-square. Therefore we will examine additional measures of fit before we draw any conclusions about the adequacy of the model.

Model Evaluation

In this section three aspects of model evaluation are discussed. First, we discuss the problem of assessing fit in a SEM model. We then present several popular fit indexes. This section concludes with a discussion of evaluating direct effect estimates.

Evaluating the overall fit of the model. The model chi-square test statistic is highly dependent on sample size that is, the model chi-square test statistic is $(N - 1)F_{min}$. Where N is the sample size and F_{min} is the value of F_{min} Equation 4 at the function minimum. Therefore the fit of models estimated with large samples, as seen in the substance use problems model, is often difficult to assess. Fit indexes have been developed to address this problem. There are five general classes of fit indexes: comparative fit, absolute fit, proportion of variance accounted for, parsimony adjusted proportion of variance accounted for, and residual based fit indexes. A complete discussion of model fit is outside the scope of this article; therefore we will focus on two of the most popular fit indexes the comparative fit index (CFI; Bentler, 1990) and a residual based fit index, the root mean square error of approximation (RMSEA; Browne & Cudeck 1993; Steiger & Lind, 1980). For more detailed discussions of fit indexes see Ullman (2006), Bentler and Raykov (2000), and Hu and Bentler (1999).

One type of model fit index is based on a comparison of nested models. At one end of the continuum is the uncorrelated variables or independence model: the model that corresponds to completely unrelated variables. This model would have degrees of freedom equal to the number of data points minus the variances that are estimated. At the other end of the continuum is the saturated, (full or perfect), model with zero degrees of freedom. Fit indexes that employ a comparative fit approach place the estimated model somewhere along this continuum, with 0.00 indicating awful fit and 1.00 indicating perfect fit.

The CFI (Bentler, 1990) assesses fit relative to other models and uses an approach based on the noncentral χ^2 distribution with noncentrality parameter, τ_i. If the estimated model is perfect $\tau_i = 0$ therefore, the larger the value of τ_i, the greater the model misspecification.

$$\mathrm{CFI} = 1 - \frac{\tau_{\text{est.model}}}{\tau_{\text{indep.model}}}. \tag{5}$$

So, clearly, the smaller the noncentrality parameter, τ_i, for the estimated model relative to the τ_i, for the independence model, the larger the CFI and the better the fit. The τ value for a model can be estimated by,

$$\begin{aligned}\hat{\tau}_{\text{indep.model}} &= \chi^2_{\text{indep.model}} - df_{\text{indep.model}}\\ \hat{\tau}_{\text{est.model}} &= \chi^2_{\text{est.model}} - df_{\text{est.model}}\end{aligned} \tag{6}$$

where $\hat{\tau}_{\text{est.model}}$ is set to zero if negative.

For the example,

$$\begin{aligned}\tau_{\text{independence model}} &= 16689.287 - 120 = 16569.287\\ \tau_{\text{estimated model}} &= 636.0566 - 103 = 533.0566\\ \mathrm{CFI} &= 1 - \frac{533.0566}{16569.287} = .98.\end{aligned}$$

CFI values greater than .95 are often indicative of good fitting models (Hu & Bentler, 1999). The CFI is normed to the 0 to 1 range, and does a good job of estimating model fit even in small samples (Hu & Bentler, 1998, 1999). In this example the CFI is calculated from the Satorra–Bentler scaled chi-square due to data nonnormality. To clearly distinguish it from a CFI based on a normal theory chi-square this CFI is often reported as a "robust CFI".

The RMSEA (Steiger, 2000; Steiger & Lind, 1980) estimates the lack of fit in a model compared to a perfect or saturated model by,

$$\text{estimated RMSEA} = \sqrt{\frac{\hat{\tau}}{Ndf_{\text{model}}}}, \tag{7}$$

where $\hat{\tau} = \hat{\tau}_{\text{est.model}}$ as defined in Equation 6. As noted above, when the model is perfect, $\hat{\tau} = 0$, and the greater the model misspecification, the larger $\hat{\tau}$. Hence RMSEA is a measure of noncentraility relative to sample size and degrees of freedom. For a given noncentraility, large N and degrees of freedom imply a better fitting model, that is, a smaller RMSEA. Values of .06 or less indicate a close fitting model (Hu & Bentler, 1999). Values larger than .10 are indicative of poor fitting models (Browne & Cudeck, 1993). Hu and Bentler found that in small samples (< 150) the RMSEA over rejected the true model, that is, its value was too large. Because of this problem, this index may be less preferable with small samples. As with the CFI the choice of estimation method effects the size of the RMSEA.

For the example, $\hat{\tau} = 533.0566$, therefore,

$$\mathrm{RMSEA} = \sqrt{\frac{533.0566}{(736)(103)}} = .0838.$$

The Robust CFI values of .967 exceeds the recommended guideline for a good-fitting model however the RMSEA of .08 is a bit too high to confidently conclude that the model fits well. It exceeds .06 but is less than .10. Unfortunately, conflicting evidence such as found with these fit indexes is not uncommon. At this point it is often helpful to tentatively conclude that the model is adequate and perhaps look to model modification indexes to ascertain if a critical parameter has been overlooked. In this example the hypothesized model will be compared to a competing model that accounts for the variance due to common item wording. Therefore it is reasonable to continue interpreting the model. We can conclude that there is evidence that the constructs of Alcohol Use Problems and Drug Use Problems exist and at least some of the measured variables are significant indicators of the construct.

Another method of evaluating the fit of the model is to look at the residuals. To aid interpretability it is helpful to look at the standardized residuals. These are in a correlational metric and therefore can be interpreted as the residual correlation not explained by the model. Of particular interest is the average standardized variance residual and the average standardized covariance residual. In this example the average standardized variance residual = .0185, and the average standardized covariance = .021. These are correlations so that if squared they provide the percentage of variance, on average, not explained by the model. Therefore, in this example, the model does not explain .035% of the variance in the measured variable variances and .044% of the variance in the covariances. This is very small and is indicative of a good fitting model. There are no set guidelines for acceptable size of residuals, but clearly smaller is better. Given the information from all three fit indexes, we can tentatively conclude that our hypothesized simple structure factor model is reasonable. In reporting SEM analyses it is a good idea to report multiple-fit indexes, the three discussed here are good choices to report as they examine fit in different but related ways.

Interpreting parameter estimates. The model fits adequately, but what does it mean? To answer this question researchers usually examine the statistically significant relations within the model. Table 2 contains edited EQS output for evaluation of the regression coefficients for the example. If the unstandardized parameter estimates are divided by their respective standard errors, a z score is obtained for each estimated parameter that is evaluated in the usual manner,

$$z = \frac{\text{parameter estimate}}{SE \text{ for estimate}}. \tag{8}$$

EQS provides the unstandardized coefficient (this value is analogous to a factor loading from the pattern matrix in EFA but is an unstandardized item-factor covariance), and two sets of standard errors and z scores. The null hypothesis for

TABLE 2
EQS 6.1 Output of Standardized Coefficients for Hypothesized Model

```
AHLTHALC=V151=   .978*F1      + 1.000 E151
                 .040
                 24.534@
                 ( .047)
                 ( 20.718@
AHLTHDRG=V152=   1.324*F2     + 1.000 E152
                 .047
                 28.290@
                 ( .040)
                 ( 33.014@
AFAMALC =V153=   1.407*F1     + 1.000 E153
                 .043
                 32.874@
                 ( .034)
                 ( 41.333@
AFAMDRG =V154=   1.612*F2     + 1.000 E154
                 .047
                 34.284@
                 ( .024)
                 ( 66.637@
AATTALC =V155=   1.446*F1     + 1.000 E155
                 .042
                 34.802@
                 ( .030)
                 ( 48.416@
AATTDRG =V156=   1.600*F2     + 1.000 E156
                 .046
                 35.137@
                 ( .026)
                 ( 62.699@
AATTNALC=V157=   1.433*F1     + 1.000 E157
                 .042
                 33.929@
                 ( .033)
                 ( 42.817@
AATTNDRG=V158=   1.603*F2     + 1.000 E158
                 .046
                 34.488@
                 ( .026)
                 ( 61.806@
AWORKALC=V159=   1.361*F1     + 1.000 E159
                 .045
                 30.425@
                 ( .039)
                 ( 34.564@
AWORKDRG=V160=   1.575*F2     + 1.000 E160
                 .049
                 31.856@
                 ( .031)
                 ( 51.101@
AMONYALC=V161=   1.437*F1     + 1.000 E161
                 .046
                 31.055@
                 ( .036)
                 ( 40.009@
AMONYDRG=V162=   1.675*F2     + 1.000 E162
                 .051
                 33.159@
                 ( .022)
                 ( 74.658@
AARGUALC=V163=   1.394*F1     + 1.000 E163
                 .044
                 31.899@
                 ( .033)
                 ( 41.610@
```

(continued)

TABLE 2 Continued

AARGUDRG=V164=	1.546*F2 .047 32.888@ (.029) (52.726@	+ 1.000 E164		
ALEGLALC=V165=	1.088*F1 .043 25.317@ (.049) (22.003@	+ 1.000 E165		
ALEGLDRG=V166=	1.280*F2 .049 25.967@ (.045) (28.224@	+ 1.000 E166		
STANDARDIZED SOLUTION:				R^2
AHLTHALC=V151=	.767*F1	+ .642	E151	.588
AHLTHDRG=V152=	.842*F2	+ .539	E152	.710
AFAMALC =V153=	.923*F1	+ .385	E153	.852
AFAMDRG =V154=	.944*F2	+ .330	E154	.891
AATTALC =V155=	.952*F1	+ .305	E155	.907
AATTDRG =V156=	.957*F2	+ .291	E156	.915
AATTNALC=V157=	.939*F1	+ .343	E157	.882
AATTNDRG=V158=	.947*F2	+ .321	E158	.897
AWORKALC=V159=	.882*F1	+ .471	E159	.778
AWORKDRG=V160=	.906*F2	+ .424	E160	.820
AMONYALC=V161=	.893*F1	+ .450	E161	.797
AMONYDRG=V162=	.927*F2	+ .376	E162	.859
AARGUALC=V163=	.907*F1	+ .421	E163	.823
AARGUDRG=V164=	.922*F2	+ .386	E164	.851
ALEGLALC=V165=	.783*F1	+ .622	E165	.614
ALEGLDRG=V166=	.796*F2	+ .605	E166	.634

Note. Measurement equations with standard errors and test statistics. Statistics significant at the 5% level are marked with @. (Robust statistics in parentheses).

these tests is that the unstandardized regression coefficient = 0. Now, look at the equation for AHLTHALC V151 in Table 2, the unstandardized regression coefficient = .978. The standard error unadjusted for the nonnormality is on the line directly below = .04. The *z* score .978/.04 = 24.534 follows on the third line. These data however are nonnormal, so the correct standard error is the one that adjusts for the nonnormality, it appears on the fourth line, .047. The *z* score for the coefficient with the adjusted standard error = .978/.047 = 20.718. Typically this is evaluated against a *z* score associated with $p < .05$, $z = 1.96$. One can conclude that the Alcohol Use Problems construct significantly predicts problems with health (AHLTHALC). All of the measured variables that we hypothesized as predictors are significantly associated with their respective factors. When performing a CFA, or when testing the measurement model as a preliminary analysis stage, it is probably wise to drop any variables that do not significantly load on a factor and then reestimate a new, nonnested model.

Sometimes the unstandardized coefficients are difficult to interpret because variables often are measured on different scales; therefore, researchers often examine standardized coefficients. The standardized and unstandardized regression coefficients for the final model are in Table 2 and Figure 3. In Figure 3 the standardized coefficients are in parentheses. The standardized coefficients are given in the section that is labeled Standardized Solution of Table 2. Following each standardized equation is an R^2 value. This is the percentage of variance in the variable that is accounted for by the factor. This is analogous to a communality in EFA. In addition, although not shown in the table, the analysis revealed that the Alcohol Use Problems and Drug Use Problems significantly covary (covariance = .68, *z* score with adjusted standard error = 24.96, correlation = .68). Note the covariance and the correlation are the same value because the variance of both latent variables was fixed to 1.

Model Modification

There are at least two reasons for modifying a SEM model: to test hypotheses (in theoretical work) and to improve fit (especially in exploratory work). SEM is a confirmatory technique, therefore when model modification is done to improve fit the analysis changes from confirmatory to exploratory. Any conclusions drawn from a model that has undergone substantial modification should be viewed with extreme caution. Cross-validation should be performed on modified models whenever possible.

The three basic methods of model modification are the chi-square difference, Lagrange multiplier (LM), and Wald tests (Ullman, 2006). All are asymptotically equivalent under the null hypothesis but approach model modification differently. In this section each of these approaches will be discussed with reference to the problems with substance use example.

In CFA models in which the measurement structure is of particular interest, it may be the case that there are other characteristics of the items, the measured variables, that account for significant variance in the model. This will be demonstrated in this section by testing a competing theoretical measurement model. Return to Table 1 and notice wording and content of the items. Each domain area item is exactly the same except for substituting “drugs” or “alcohol” into the sentence. It is entirely reasonable to suggest there may be correlations among these like items even after removing the variance due to the construct, either Drug Use Problems or Alcohol Use Problems. Perhaps a better fitting model would be one that accounts for the common domains and wording across items. Said another way, it might be reasonable to estimate a model that allows covariances between the variance in each item that is not already accounted for by each construct. Before you read on, go back to the diagram and think about what exactly we need to covary to test this hypothesis. To test this hypothesis the residual variances between similarly worded items are allowed to covary. that is, the “E”s for each pair of variables, E151,E152. This represents covarying the variance in AHLTHALC and AHLTHDRG that is not accounted for by the two constructs. To demonstrate this a new model was estimated with these eight residual covariances added.

Chi-square difference test. Our initial model is a subset of this new larger model. Another way to refer to this is to say that our initial model is nested within our model that includes the residual covariances. If models are nested, that is, models are subsets of each other, the χ^2 value for the larger model is subtracted from the χ^2 value for the smaller, nested model and the difference, also χ^2, is evaluated with degrees of freedom equal to the difference between the degrees of freedom in the two models. When the data are normally distributed the chi-squares can simply be subtracted. However, when the data are nonnormal and the Satorra–Bentler scaled chi-square is employed an adjustment is required (Satorra, 2000; Satorra & Bentler, 2001) so that the S–B χ^2 difference test is distributed as a chi-square.

The model that included the correlated residuals was estimated and fit well, $\chi^2(N = 736, 95) = 178.25$, $p < .0001$, Robust CFI = .995, RMSEA = .035. In addition, the S–B χ^2 test statistic is smaller, the RMSEA is smaller, .035 versus .084, and the Robust CFI is larger, .995 versus .968, than the original model. Using a chi-square difference test we can ascertain if the model with the correlated residuals is significantly better than the model without. Had the data been normal we simply could have subtracted the chi-square test statistic values and evaluated the chi-square with the degrees of freedom associated with the difference between the models, in this case 103 – 95 = 8, the number of residual covariances we estimated. However, because the data were nonnormal and the S–B χ^2 was employed, an adjustment is necessary (Satorra, 2000; Satorra & Bentler, 2001). First a scaling correction is calculated,

$$\text{scaling correction} = \frac{(df\ \text{nested model})(\frac{\chi^2_{\text{MLnested model)}}}{\chi^2_{\text{S–Bnested model)}}})}{(df\ \text{nested model}}$$

$$-\frac{-\,(df\ \text{comparison model})(\frac{\chi^2_{\text{MLcomparison model)}}}{\chi^2_{\text{S–Bcomparison model)}}})}{-\,df\ \text{comparison model)}}$$

$$\text{scaling correction} = \frac{(103)(\frac{1570.806}{636.057}) - (95)(\frac{400.206}{178.249})}{(103 - 95)}$$

$$\text{scaling correction} = 5.13.$$

The scaling correction is then employed with the ML χ^2 values to calculate the S–B scaled χ^2 difference test statistic value,

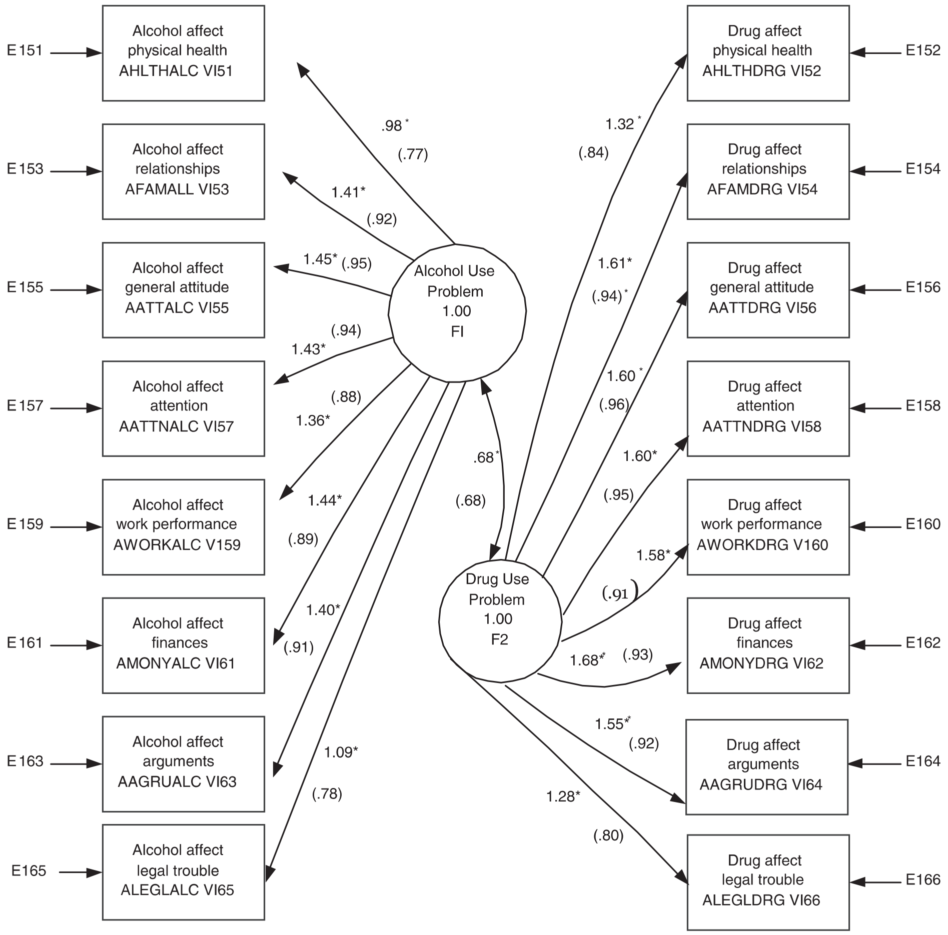

E151
Alcohol affect physical health AHLTHALC VI51
E153
Alcohol affect relationships AFAMALL VI53
E155
Alcohol affect general attitude AATTALC VI55
E157
Alcohol affect attention AATTNALC VI57
E159
Alcohol affect work performance AWORKALC V159
E161
Alcohol affect finances AMONYALC VI61
E163
Alcohol affect arguments AAGRUALC VI63
E165
Alcohol affect legal trouble ALEGLALC VI65
.98*
(.77)
1.41*
(.92)
1.45* (.95)
(.94)
1.43*
(.88)
1.36*
1.44*
(.89)
1.40*
(.91)
1.09*
(.78)
Alcohol Use Problem 1.00 FI
.68*
(.68)
Drug Use Problem 1.00 F2
1.32*
(.84)
1.61*
(.94)*
1.60*
(.96)
1.60*
(.95)
1.58*
(.91)
1.68* (.93)
1.55*
(.92)
1.28*
(.80)
Drug affect physical health AHLTHDRG VI52
E152
Drug affect relationships AFAMDRG VI54
E154
Drug affect general attitude AATTDRG VI56
E156
Drug affect attention AATTNDRG VI58
E158
Drug affect work performance AWORKDRG V160
E160
Drug affect finances AMONYDRG VI62
E162
Drug affect arguments AAGRUDRG VI64
E164
Drug affect legal trouble ALEGLDRG VI66
E166

FIGURE 3 Final substance use problem confirmatory factor analysis model is with unstandardized and standardized (in parentheses) parameter estimates.

$$\chi^2_{\text{S–B difference}} = \frac{\chi^2_{\text{ML nested model}} - \chi^2_{\text{ML comparison model}}}{\text{scaling correction}}$$
$$= \frac{1570.806 - 400.206}{5.13}$$
$$= 227.99.$$

The adjusted S–B $\chi^2_{\text{difference}}$ (N = 736, 8) = 227.99, $p < .01$. The chi-square difference test is significant. This means that the model was significantly improved by including the covariances between each commonly worded item.

Chi-square difference tests are popular methods for comparison of nested models especially when, as in this example, there are two nested a priori models of interest; however, a potential problem arises when sample sizes are small. Because of the relationship between sample size and χ^2, it is hard to detect a difference between models with small sample sizes.

LM test. The LM test also compares nested models but requires estimation of only one model. The LM test asks would the model be improved if one or more of the parameters in the model that are currently fixed were estimated. Or, equivalently, what parameters should be added to the model to improve the fit of the model?

There are many approaches to using the LM tests in model modifications. It is possible and often desirable to look only at certain parts of the model for possible change although it is also possible to examine the entire model for potential additions. In this example the only interest was whether a better-fitting model would include parameters to account for the common domain and wording variance, therefore, the only part of the model that was of interest for modification were the covariances among the "E"s, the residuals.

The LM test can be examined either univariately or multivariately. There is a danger in examining only the results of univariate LM tests because overlapping variance between parameter estimates may make several parameters appear as if their addition would significantly improve the model. All significant parameters are candidates for inclusion by the results of univariate LM tests, but the multivariate LM test identifies the single parameter that would lead to the largest drop in the model χ^2 and calculates the expected change in χ^2 if this parameter was added. After this variance is removed, the next parameter that accounts for the largest drop in model χ^2 is assessed, similarly. After a few candidates for parameter additions are identified, it is best to add these parameters to the model and repeat the process with a new LM test, if necessary.

Table 3 contains highly edited EQS output for the multivariate LM test. The eight parameters that would reduce the chi-square the greatest are included in the table. Notice that these eight LM parameter addition suggestions are exactly the eight parameters that we added. Within the table, the second column lists the parameter of interest. The next column lists the cumulative expected drop in the chi-square if the parameter and parameters with higher priority are added. Notice that, according the LM test, if all eight correlated residuals are added, the model chi-square would drop approximately by 1036.228. The actual difference between the two ML chi-squares was close to that approximated by the LM test (actual difference between ML χ^2 = 1170.60). Some caution should be employed when evaluating the LM test when the data are nonnormal. The model in this example was evaluated with a scaled ML chi-square and the path coefficient standard errors were adjusted for nonnormality. However the LM test is based on the assumption of normality and therefore the chi-squares given in the third and sixth columns refer to ML χ^2, not Satorra–Bentler scaled chi-squares. This means that conclusions drawn from the LM test may not apply to data after the model test statistics are adjusted for nonnormality. It is still worthwhile to use the LM test in the context of nonnormal data, but cautious use is warranted. It is a good idea to examine effect of adding parameters with a chi-square difference test in addition to the LM test.

TABLE 3
Edited EQS Output of Multivariate LM Test

	Cumulative Multivariate Statistics				*Univariate Increment*	
Step	*Parameter*	χ^2	*df*	*Probability*	χ^2	*Probability*
1	E160,E159	202.502	1	.000	202.502	.000
2	E166,E165	394.834	2	.000	192.332	.000
3	E152,E151	566.240	3	.000	171.406	.000
4	E164,E163	687.436	4	.000	121.196	.000
5	E162,E161	796.632	5	.000	109.196	.000
6	E154,E153	888.330	6	.000	91.698	.000
7	E156,E155	970.849	7	.000	82.519	.000
8	E158,E157	1036.228	8	.000	65.379	.000

Note. LM = Lagrange multiplier.

Wald test. While the LM test asks which parameters, if any, should be added to a model, the Wald test asks which, if any, could be deleted. Are there any parameters that are currently being estimated that could instead be fixed to zero? Or, equivalently, which parameters are not necessary in the model? The Wald test is analogous to backward deletion of variables in stepwise regression, in which one seeks a nonsignificant change in R^2 when variables are left out. The Wald test was not of particular interest in this example. However, had a goal been the development of a parsimonious model the Wald test could have been examined to evaluate deletion of unnecessary parameters.

Some caveats and hints on model modification. Because both the LM test and Wald test are stepwise procedures, Type I error rates are inflated but there are as yet no available adjustments as in analysis of variance. A simple approach is to use a conservative probability value (say, $p < .01$) for adding parameters with the LM test. There is a very real danger that model modifications will capitalize on chance variations in the data. Cross validation with another sample is also highly recommended if modifications are made.

Unfortunately, the order that parameters are freed or estimated can affect the significance of the remaining parameters. MacCallum (1986) suggested adding all necessary parameters before deleting unnecessary parameters. In other words, do the LM test before the Wald test.

A more subtle limitation is that tests leading to model modification examine overall changes in χ^2, not changes in individual parameter estimates. Large changes in $\chi^?$ are sometimes associated with very small changes in parameter estimates. A missing parameter may be statistically needed but the estimated coefficient may have an unintrepretable sign. If this happens it may be best not to add the parameter although the unexpected result may help to pinpoint problems with one's theory. Finally, if the hypothesized model is wrong, and in practice with real data we never know if our model is wrong, tests of model modification by themselves, may be insufficient to reveal the true model. In fact, the "trueness" of any model is never tested directly, although cross validation does add evidence that the model is correct. Like other statistics, these tests must be used thoughtfully. If model modifications are done in hopes of developing a good fitting model, the fewer modifications the better, especially if a cross-validation sample is not available. If the LM test and Wald tests are used to test specific hypotheses, the hypotheses will dictate the number of necessary tests.

CONCLUSIONS

The goal of this article was to provide a general overview of SEM and present an empirical example of the type of SEM analysis, CFA that might be useful in personality assessment. Therefore, the article began with a brief introduction to SEM analyses. Then, an empirical example with nonnormal data was employed to illustrate the fundamental issues and process of estimating one type of SEM, CFA. This analysis provided information about the two interrelated scales measuring problems with alcohol use and drug use across eight domains. The items in both the alcohol and the drug scales were strong indicators of the two underlying constructs of Alcohol Use Problems and Drug Use Problems. As expected the two constructs were strongly related. And, of measurement interest, the model was significantly improved after allowing the variance in the commonly worded items not accounted for by the constructs to vary. Finally, this example demonstrated one method to appropriately analyze nonnormally distributed data in the context of a CFA.

This analysis could be viewed as a first confirmatory analysis step for a newly developed scale. After demonstrating a strong measurement structure as in this example, further research could then use these constructs as predictors of future behavior. Or these constructs could be viewed as outcomes of personality characteristics or as intervening variables between personality dispositions and behavioral outcomes in a full SEM model.

SEM is a rapidly growing statistical technique with much research into basic statistical topics as model fit, model estimation with nonnormal data, and estimating missing data/sample size. Research also abounds in new application areas such latent growth curve modeling and multilevel model. This article presented an introduction and fundamental empirical example of SEM, but hopefully enticed readers to continue studying SEM, following the exciting growth in this area, and most important modeling their own data!

ACKNOWLEDGMENTS

This article was supported in part by NIDA Grant DA 01070–30. I thank Adeline Nyamathi for generously allowing me to use a subset of her data as an empirical example in this article. Adeline Nyamathi's grant is funded by NIH Grant MH 52029. I acknowledge Gisele Pham's editorial assistance. I also thank the two anonymous reviewers who provided helpful suggestions.

REFERENCES

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychology Bulletin*, *107*, 256–259.

Bentler, P. M. (2001). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis 6* (pp. 9–42). Amsterdam: North-Holland.

Bentler, P. M., & Raykov, T. (2000). On measures of explained variance in nonrecursive structural equation models. *Journal of Applied Psychology*, *85*, 125–131.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equation with latent variables. *Psychometrika, 45,* 289–308.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34,* 181–197.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8,* 1–24.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical Statistical Psychology,, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural models* (pp. 35–57). Newbury Park, CA: Sage.

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling, 7,* 356–410.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished Ph.D. dissertation, Rijksuniversiteit Groningen, The Netherlands.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structural equation modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics, 24,* 21–41.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods, 1,* 130–149.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* pp. 519–530.

Nyamathi, A. M., Stein, J. A., Dixon, E., Longshore, D., & Galaif, E. (2003). Predicting positive attitudes about quitting drug- and alcohol-use among homeless women. *Psychology of Addictive Behaviors, 7,* 32–41.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker* (pp. 233–247). Dordrecht, The Netherlands: Kluwer Academic.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the American Statistical Association,* 308–313.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571.

Steiger, J. H., & Lind, J. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Stein, J. A., & Nyamathi, A. M. (2000). Gender differences in behavioural and psychosocial predictors of HIV testing and return for test results in a high-risk population. *AIDS Care, 12,* 343–356.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

Ullman, J. B. (2006). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics*, (5th ed.; pp. 653–771). Boston: Allyn & Bacon.

APPENDIX

EQS 6.1 Syntax for Substance Use Problems Model

```
/TITLE
   Confirmatory Factor Analysis Model
/SPECIFICATIONS
   DATA='c:\data for JPA.ess';
   VARIABLES=216; CASES=736;
   METHODS=ML,ROBUST;
   MATRIX=RAW;
   ANALYSIS=COVARIANCE;
/LABELS
   V1=COUPLE; V2=AOWNHOME; V3=ASOBERLV; V4=AHOTEL;
V5=ABOARD;
   Lots of variable labels were deleted for this text
   V151=AHLTHALC; V152=AHLTHDRG; V153=AFAMALC;
V154=AFAMDRG; V155=AATTALC;
   V156=AATTDRG; V157=AATTNALC; V158=AATTNDRG;
V159=AWORKALC; V160=AWORKDRG;
   V161=AMONYALC; V162=AMONYDRG; V163=AARGUALC;
V164=AARGUDRG; V165=ALEGLALC;
   V166=ALEGLDRG; V167=CAGE1; V168=CAGE2; V169=CAGE3;
V170=CAGE4;
   Lots of variable labels were deleted for this text
   V216=SEX_AIDS; F1 = ALCOHOL_PROB; F2=DRUG_PROB;
/EQUATIONS
   ! Problems with Alcohol
   V151 = *F1 + E151;
   V153 = *F1 + E153;
   V155 = *F1 + E155;
   V157 = *F1 + E157;
   V159 = *F1 + E159;
   V161 = *F1 + E161;
   V163 = *F1 + E163;
   V165 = *F1 + E165;
   !Problems with Drugs
   V152 = *F2 + E152;
   V154 = *F2 + E154;
   V156 = *F2 + E156;
   V158 = *F2 + E158;
   V160 = *F2 + E160;
   V162 = *F2 + E162;
   V164 = *F2 + E164;
   V166 = *F2 + E166;
/VARIANCES
   F1,F2, = 1.00;
   E151 to E166 = *;
/COVARIANCES
   F1,F2=*;
   !E151,E152 =*;
   !E153,E154=*;
   !E155,E156=*;
   !E157,E158=*;
   !E159,E160=*;
   !E161,E162=*;
   !E163,E164=*;
   !E165,E166=*;
/LMTEST
   SET = PEE;
/PRINT
   FIT=ALL;
   TABLE=EQUATION;
 /END
```

Jodie B. Ullman
Department of Psychology
California State University, San Bernardino
5500 University Parkway
San Bernardino, CA 92407
Email: jullman@csusb.edu

Received August 8, 2005
Revised April 11, 2006