A Brief Guide to Structural Equation Modeling

Rebecca Weston
Southern Illinois University

Paul A. Gore Jr.
ACT, Inc.

To complement recent articles in this journal on structural equation modeling (SEM) practice and principles by Martens and by Quintana and Maxwell, respectively, the authors offer a consumer's guide to SEM. Using an example derived from theory and research on vocational psychology, the authors outline six steps in SEM: model specification, identification, data preparation and screening, estimation, evaluation of fit, and modification. In addition, the authors summarize the debates surrounding some aspects of SEM (e.g., acceptable sample size, fit indices), with recommendations for application. They also discuss the need for considering and testing alternative models and present an example, with details on determining whether alternative models result in a significant improvement in fit to the observed data.

Issues of interest to counseling psychologists are often complex and multidimensional in nature. For example, noting the mostly univariate nature of extant eating-disorder research, Tylka and Subich (2004) hypothesized that eating-disorder patterns in adult women were a function of personal, sociocultural, and relational factors. Furthermore, they offered hypotheses about how these factors interact in complex ways to explain symptom severity. We can draw another example from vocational psychology literature. Long, Kahn, and Schutz (1992) developed an integrative model of workplace stress and coping for employed women, which includes constructs such as human agency, status, coping, work-environment demand, and distress. Suspecting that the hypothesized relationships among these variables might differ across levels of employment prestige, Long (1998) compared the model's performance in two groups of employed women: a group of managers and a group of clerical workers (on testing multiple groups with discriminant analysis, see Sherry, 2006 [this issue]).

The authors wish to thank Kathie Chwalisz for her helpful comments on an earlier version of this article. Correspondence should be addressed to Rebecca Weston, Department of Psychology, Mailcode 6502, Southern Illinois University, Carbondale, IL 62901. E-mail: Weston@siu.edu.

THE COUNSELING PSYCHOLOGIST, Vol. 34 No. 5, September 2006 719-751
DOI: 10.1177/0011000006286345

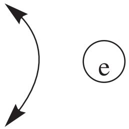
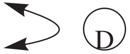
Just a few decades ago, researchers were dissuaded from asking (or answering) complex research questions such as these because statistical techniques did not easily allow for testing of multivariate models. Structural equation modeling (SEM) is a family of statistical techniques permitting researchers to test such models (Martens & Hasse, in press [*TCP* special issue, part 2]; Worthington & Whittaker, in press [*TCP* special issue, part 2]). We can think of SEM as a hybrid of factor analysis and path analysis. SEM's goal is similar to that of factor analysis: to provide a parsimonious summary of the interrelationships among variables (Kahn, 2006 [this issue]). SEM is also similar to path analysis in that researchers can test hypothesized relationships between constructs. Not surprisingly, SEM's popularity as an analytic tool in counseling psychology has grown since Fassinger's (1987) review (Tracey, 2002; Wei, Heppner, & Mallinckrodt, 2004). There has been a parallel increase in the development of new software for conducting analyses (e.g., Analysis of Moment Structures [AMOS], Equations [EQS], Mplus), and statisticians have explored new applications (e.g., growth-curve modeling, Meredith & Tisak, 1990; for details on SEM's use in scale-development research, see Worthington & Whittaker, in press).

Quintana and Maxwell (1999) discussed many of the recent technical and theoretical advances in SEM in a review in this journal. The reviews by Fassinger (1987) and by Quintana and Maxwell in some ways reflect the literature's bifurcation into practical application and statistical theory, respectively. Martens's (2005) examination of SEM practice in counseling psychology also suggests that there is a lack of consistency in the application of this technique. Given that Martens addressed many of the recent theoretical issues that are important for application, our goal in this review is to provide a guide to assist in SEM interpretation and to introduce some of the most important concepts and issues, including defining commonly used terms (Table 1) and highlighting current topics of debate (e.g., acceptable minimum sample size; for advanced applications of SEM, see Martens & Hasse, in press). In addition, as this guide is not intended as a complete SEM manual, we cite useful resources (e.g., Bollen, 1989; Byrne, 2001; Hoyle, 1995; Martens & Hasse, in press; Schumacker & Lomax, 2004) throughout for those interested in reading more in-depth discussions on topics presented here.

Relationship to Other Common Statistical Procedures

SEM is comparable to common quantitative methods, such as correlation, multiple regression, and analysis of variance (ANOVA). SEM is similar to these techniques in several ways. First, all four statistical procedures are general linear models. Second, all are valid only if specific assumptions

TABLE 1: Common Terms and Symbols in Structural Equation Modeling

Term Used Here	*Alternative Term(s)*	*Definition*	*Symbol*	*Examples in Figure(s)*
Latent variable	Factor, construct	Unobserved hypothetical variable (e.g., occupational interests).	⬭ ○	Interests
Indicator	Measured or manifest variable	Observed variable (e.g., Strong Interest Inventory).	▭ □	INT-1
Factor loading	Path loading	Correlation between latent variable and indicator.	→	Unidirectional arrow from Interests to INT-1
Direct effect	Path coefficient, path	Correlation between two latent variables.	→	Unidirectional arrow from Interests to Occupational Considerations in Figures 2 and 4
Nondirectional association	Covariance, correlation	Correlation between two latent variables.	↔	Bidirectional arrows between latent variables in Figure 1
Indicator error	Predictor error, measurement error	Error in indicator that is not accounted for by latent variable. Indicator error is also considered a latent variable.	e 	e associated with each indicator in Figure 2
Disturbance	Predictor error	Error in dependent latent variable not accounted for by predictors.	D 	D associated with each dependent latent variable in Figure 2
Explained variance		Percentage of variance in dependent latent variable accounted for by predictor(s).	R^2	1 – D^2 in Figure 4
Parameter	Path	Hypothesized association between two variables.	→, ↔	Arrows in all figures

(continued)

TABLE 1 (continued)

Term Used Here	*Alternative Term(s)*	*Definition*	*Symbol*	*Examples in Figure(s)*
Independent variable	Exogenous variable, predictor	Variable that is not dependent on or predicted by other latent variables or indicators.	—	Self-efficacy beliefs in Figures 2 and 4
Dependent variable	Endogenous variable, criterion	Variable that is predicted by other latent variables or indicators.	—	Predictor error in Figure 2; Outcome expectations in Figures 2 and 4; INT-1 in Figures 1 and 2
Set parameter	Constrained parameter; Fixed path	Parameter that is set at a constant and not estimated. Parameters fixed at 1.0 reflect an expected 1:1 association between variables. Parameters set at 0 reflect the assumption that no relationship exists.	Parameters set at nonzero values should be labeled: $\xrightarrow{1.0}$ Parameters set at 0 are omitted.	Parameter set at 1.0 from Interests to INT-1 in Figure 2; Parameter set at 0 from Self-Efficacy Beliefs to Occupational Considerations in Figure 2
Free parameter	Estimated parameter	Parameter that is not constrained and is to be estimated using observed data.	Represented with an asterisk or simply unlabeled.	Parameter from Interests to INT-2 in Figure 2
Covariance matrix	Sample matrix	Unstandardized associations between all pairs of variables.	Σ; S	Lower left diagonal of Table 2
Skewness	Asymmetry	Degree of asymmetry observed in the distribution for a variable.	—	—
Kurtosis	Flatness or peakedness	Degree of the peakedness of the distribution for a variable.	—	—

are met. Third, none of these techniques implies causality. Although causal relationships are hypothesized, causality cannot be determined by results of any of these techniques, only by the soundness of the underlying theory and research design. Finally, such as other statistical procedures, researchers can easily misuse SEM. Just as researchers are free (although not encouraged) to conduct several different multiple regression models until they find a model to their liking, they can also analyze models in SEM, identify and remove weaknesses in the model, and then present the revised model as if it were the originally hypothesized model. Most users would likely agree that SEM's true power lies in the fact that researchers must specify complex relationships a priori and then test whether those relationships are reflected in the sample data. Optimally, researchers will draw these hypothesized relationships from previous research or theory and will present the results with integrity. Should the researcher detect weaknesses in the proposed model, he or she should further explore them using a modified model in a new sample.

One difference between SEM and other methods, and an advantage of SEM, is its capacity to estimate and test the relationships among constructs. Compared with other general linear models, where constructs may be represented with only one measure and measurement error is not modeled, SEM allows for the use of multiple measures to represent constructs and addresses the issue of measure-specific error. This difference is important in that it allows researchers to establish the construct validity of factors (Hoyt, Warbasse, & Chu, in press [*TCP* special issue, part 2]). A second difference, which many new to SEM find frustrating, is that interpreting the significance of SEM results involves carefully evaluating many results. In SEM, researchers must evaluate multiple test statistics and a host of fit indices to determine whether the model accurately represents the relationships among constructs and observed variables (i.e., whether the model fits the data). To further complicate the issue, considerable controversy exists regarding what constitutes acceptable fit and recommendations found in introductory texts published less than a decade ago are now out of date. We address this, and other controversial issues, as we outline the steps involved in SEM.

Example Data

To provide simple examples of issues related to the development and testing of SEMs, we use models derived from a single data set throughout this article. The second author provided previously unpublished data from a small subsample of undergraduate college students who participated in a vocational psychology research project. Students in this sample completed instruments measuring self-efficacy beliefs, outcome expectations, career-related interests, and occupational considerations. Each instrument included heterogeneous items broadly representing the six person-environment categories Holland (1997) described. The data presented here represent students' responses to a subset of items that measure Holland's Social dimension. We refer readers wishing to more fully understand the theory and research guiding the models to Lent, Brown, and Hackett (1994) and Lent (2005).

Examples from others' application of SEM would be helpful in illustrating several of the issues discussed here. However, space limitations prohibit describing such studies in the amount of detail necessary. Readers interested in seeing additional examples of how SEM is being applied in the counseling psychology literature should refer to recent publications (e.g., Geurts, Kompier, Roxburgh, & Houtman, 2003; Heppner, Pretorious, Wei, Lee, & Wang, 2002; Long, 1998; Martens & Hasse, in press; Tylka & Subich, 2004).

Measurement and Structural Models

Thinking of SEM as a combination of factor analysis and path analysis sets the researcher up for thinking about SEM's two primary components: the measurement model and the structural model. The measurement model describes the relationships between observed variables (e.g., instruments) and the construct or constructs those variables are hypothesized to measure. In contrast, the structural model describes interrelationships among constructs. When the measurement model and the structural model are considered together, the model may be called the composite or full structural model.

Measurement model. The measurement model of SEM allows the researcher to evaluate how well his or her observed (measured) variables combine to identify underlying hypothesized constructs. Confirmatory factor analysis is used in testing the measurement model, and the hypothesized factors are referred to as latent variables. The measures chosen by the researcher define the latent variables in the measurement model. A latent variable is defined more accurately to the extent that the measures that define it are strongly related to one another. If, for example, one measure is only weakly correlated with two other measures of the same construct, then that construct will be poorly defined. This represents model misspecification or a misjudgment in the hypothesized relationships among variables. Development of the measurement model is one of several places where a researcher can misspecify his or her model.

The model in Figure 1 is an example of a measurement model. In this example, each of the four latent variables is represented by three measured

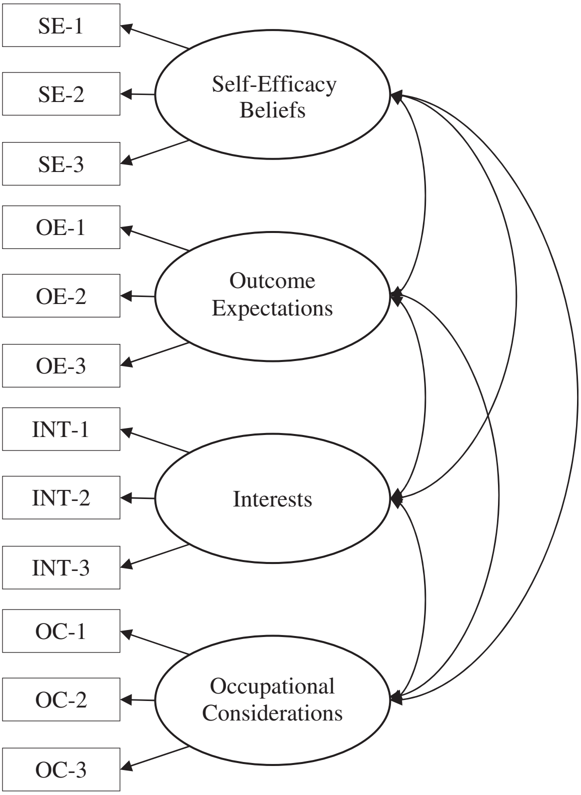


FIGURE 1. Confirmatory Factor Analysis Used to Test Measurement Model

variables, called indicators. Researchers are strongly discouraged from testing models that include constructs with single indicators (Bollen, 1989). Ideally, each of the three indicators would be a separate measure of the hypothesized latent variable, which, in combination, are a representation of the underlying construct. For example, the three measures of social career interests might include the Social General Occupational Theme scale score from the Strong Interest Inventory (Harmon, Hansen, Borgen, & Hammer, 1994), the Social score from Holland's Self-Directed Search (Holland, Fritzsche, & Powell, 1994), and the Helping Scale from the Campbell Interest and Skills Survey (Campbell, Hyne, & Nilsen, 1992). However, practical concerns often prevent researchers from using multiple measures. In such cases, investigators may elect to use item parcels (unique subsets of items from the same scale that are combined to form a type of item composite). The item-parceling strategy used here, although not optimal, allows researchers to include multiple indicators of a construct when limited measures of a construct exist or when practical issues preclude including multiple measures of a construct. Readers interested in learning more about parceling procedures should refer to several examples (Lent, Brown, & Gore, 1997; Lent, Lopez, Brown, & Gore, 1996; Mathieu & Farr, 1991; Quintana & Maxwell, 1999; Russell, Kahn, Spoth, & Altmaier, 1998).

Regardless of whether researchers use item parcels or separate scales, measures that are reliable and have little error will be better indicators of their respective latent variable, much as the items in a scale that most accurately represent the underlying construct have the highest factor loadings in a factor analysis. For example, if Self-Efficacy Beliefs–1 (SE-1) was a more reliable measure than SE-2, then SE-1 would be a better indicator of self-efficacy beliefs than would SE-2. The original measures used for this example had internal consistency reliability estimates that ranged from .85 to .95 (for more on reliability data, see Helms, Henze, Sass, & Mifsud, 2006 [this issue]).

Structural model. Equations in the structural portion of the model specify the hypothesized relationships among latent variables. We include one hypothesized structural model in the composite model in Figure 2. In this model, we hypothesize that occupational considerations are a function of an individual's interests for various occupations. Interests, in turn, are informed by an individual's self-efficacy beliefs for engaging in occupationally relevant activities and his or her beliefs in the outcomes of pursuing occupations in this career area. In other words, interests mediate the effects of self-efficacy beliefs and outcome expectations on occupational considerations.

We can describe relationships among latent variables as covariances, direct effects, or indirect (mediated) effects. Covariances are analogous to correlations in that they are defined as nondirectional relationships among independent latent variables. We indicate them pictorially using double-headed arrows. Because we did not anticipate any nondirectional relationships between the latent variables, we specified no covariances in the structural model in Figure 2.

Direct effects are relationships among measured and latent variables, similar to those found in ANOVA and multiple regressions. We indicate them pictorially using single-directional arrows (e.g., between self-efficacy beliefs and outcome expectations). It is important to note that although arrows imply directionality in SEM figures, researchers should not interpret relationships among latent variables as causal unless they analyze longitudinal or experimental data. The coefficients generated to describe the strength of these relationships are interpreted in much the same way as regression weights.

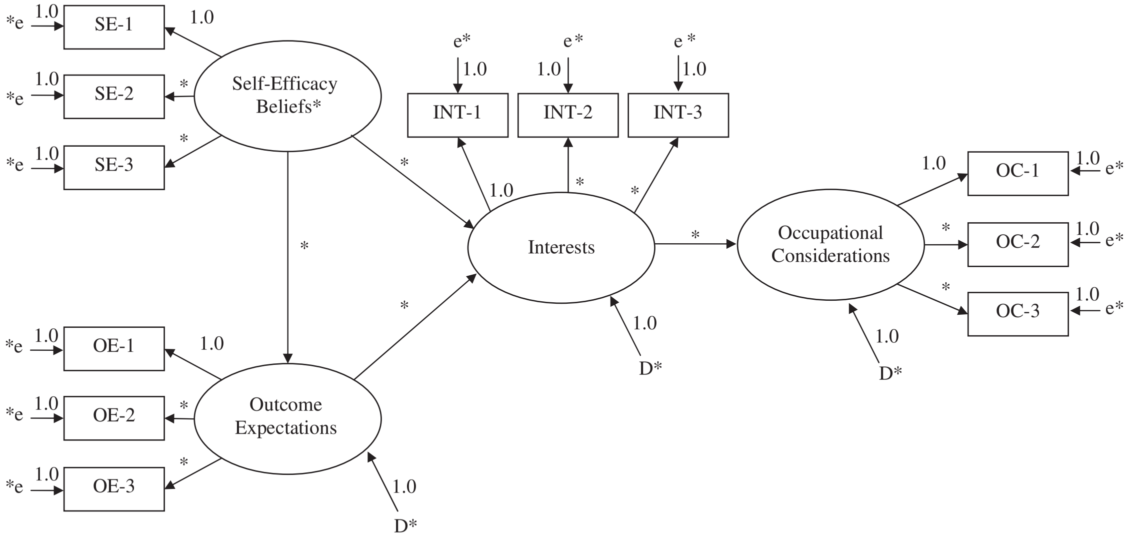


FIGURE 2. Fully Mediated Composite Model
NOTE: Asterisks represent parameters to be estimated.

An indirect effect is the relationship between an independent latent variable and a dependent latent variable that is mediated by one or more latent variables (Baron & Kenny, 1986). Mediation may be full or partial. In Figure 2, we hypothesize self-efficacy beliefs to have a direct effect on interests and an indirect effect on interests through outcome expectations. With both direct and indirect effects, outcome expectations partially mediate the impact of self-efficacy beliefs on interests. In contrast, the impact of self-efficacy beliefs on occupational considerations is fully mediated (by interests) because only an indirect effect is specified.

MODEL TERMINOLOGY AND REPRESENTATION

A model is a statistical statement, expressed with equations or a diagram, about the hypothesized relationships among variables based on theory and research (Hoyle, 1995). In Figure 2, self-efficacy beliefs, outcome expectations, interests, and occupational considerations are latent variables, diagrammatically indicated as such by ellipses. Rectangles represent measured variables, also called observed variables, manifest variables, or indicators. Because latent variables are thought to be the unobserved constructs that underlie indicators, unidirectional arrows indicate a direct effect of latent variables on measured variables. All relationships among variables in Figure 2 are directional.

Each of the measured and latent variables is exogenous (independent) or endogenous (dependent). In Figure 2, all of the indicators (e.g., SE-1, SE-2) are endogenous because they are dependent on (i.e., predicted by) their respective latent variables. Of the four latent variables, only self-efficacy beliefs is exogenous (i.e., not predicted by any variable); all other latent variables are dependent on another latent variable or variables.

From a classical test-theory perspective, variance of any observed measure consists of true scores and error. Reliable measures have less error and are considered a better measure of the underlying construct than are unreliable measures. This assumption is reflected in SEM when modeling error variance for dependent variables. The assumption is that dependent variables have some variance unexplained by the latent variable, thus error variance must also be modeled. In SEM, the latent variable would represent the underlying attribute associated with a true score, and error variance accounts for the variability not due to the true score. We specify error variance, often referred to as indicator error, as *e* for the 12 indicators in Figure 2. Error associated with dependent latent variables is referred to as disturbance and is represented with *D* for the three dependent latent variables. Fassinger (1987) offers a user-friendly description of the notation for structural models.

STEPS IN SEM

SEM experts agree on the six steps necessary in model testing. In addition to data collection, the steps are model specification, identification, estimation, evaluation, and modification (Hoyle, 1995; Kaplan, 2000; Kline, 2005; Schumacker & Lomax, 2004).

Model Specification

Model specification occurs when a researcher specifies which relationships are hypothesized to exist or not to exist among observed and latent variables. This distinction is important because any unspecified relationships among variables are assumed to be equal to zero. In Figure 2, the direct relationship between self-efficacy beliefs and occupational considerations is assumed to be equal to zero, although a mediated relationship (through interests) is hypothesized. Misspecification in this model will exist if the mediator (interests) does not fully account for the relationship between self-efficacy beliefs and occupational considerations. This is another point at which model misspecification can occur.

We have found it helpful in explaining model specification (and misspecification) to use a simple example. When considering this example, the reader should remember two points. First, all SEMs are built from raw data that are in the form of either a correlation matrix or a covariance matrix (an unstandardized correlation matrix). Second, researchers using SEM are required to specify hypothesized relationships among variables a priori. Let us suppose that the data we have to work with are represented in the correlation matrix in Figure 3. Furthermore, suppose we have hypothesized the measurement model describing the relationships among these four variables in Figure 3. As you can see, we have described a model that can account for the observed relationships between *a* and *b* and between *c* and *d* (e.g., both are indicators of their respective uncorrelated factors). You may also notice, however, that there is a moderate relationship between variables *b* and *d*. Our model does not account for this relationship. In essence, we have created a model that says the relationship between *b* and *d* is equal to 0 when it is equal to .57. Our model will suffer (e.g., be misspecified) to the extent that the relationships hypothesized in our models do not capture the observed relationships.

Relationships among variables (called parameters or paths) are either (a) set to a nonzero value and not estimated, (b) set to zero and not estimated, or (c) left free to be estimated. The first condition occurs most often when parameters are set to 1.0 to scale latent variables. Unlike regression, where the variables themselves define the scale of the predictors and criterion, latent variables have no inherent scale. To estimate the relationships among

	a	b	c	d
a	1.00	--	--	--
b	.81	1.00	--	--
c	.10	.09	1.00	--
d	.04	.57	.72	1.00

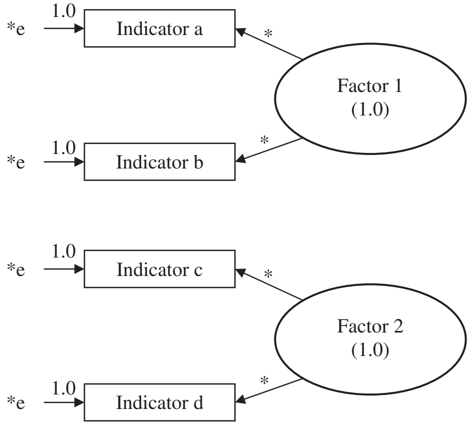


FIGURE 3. Example Correlation Matrix with Misspecified Model
NOTE: Asterisks represent parameters to be estimated.

latent variables, each latent variable must have an assigned scale. Researchers can address this problem either by setting the variance of the latent variable to 1.0 or by setting one factor loading (the parameter from a latent variable to an indicator) to 1.0. The model's overall fit is the same regardless of the option chosen. Researchers may also set parameters to other values based on past research.

Parameters set to zero are not typically included pictorially in models and are thus seldom considered. Researchers should carefully consider parameters set to zero as they reflect the hypothesized *lack* of a relationship between two variables. Our example earlier shows how setting a parameter to zero has implications that the researcher will realize later when inspecting his or her fit indices. We urge researchers to consider the implications of parameters that are not estimated in addition to considering implications of estimated parameters. For example, in Figure 2, if the direct paths from self-efficacy beliefs and outcome expectations to occupational considerations are indeed zero, the resulting fully mediated model has implications that differ from those of a partially mediated model where direct effects of self-efficacy beliefs and outcome expectations on occupational considerations are specified.

Parameter types. Three types of parameters can be specified: directional effects, variances, and covariances. Directional effects describe the relationships between latent variables and indicators (called factor loadings) and relationships between latent variables and other latent variables (called path coefficients). In Figure 2, we have set one factor loading for each latent variable at 1.0 to scale the latent variables. Free parameters, indicated with asterisks, are to be estimated for eight factor loadings between latent variables and indicators and four path coefficients between latent variables, for 12 total directional effects.

As mentioned earlier, one of SEM's advantages over other methods is its capacity to accommodate estimates of variance (recall that procedures such as regression assume variables are measured without error). Researchers can model the error associated with dependent observed and latent variables in two ways. One option is to estimate the variance for each of the error terms, while setting loadings of error terms on the dependent variables to 1.0 (the default on some software programs). This results in the estimation of parameters representing error variance. Alternatively, researchers may set the error variances to 1.0 and estimate the loadings. The second option results in standardizing the error term and estimating parameters that represent factor loadings. The option chosen does not affect model fit.

When a path loading for an independent latent variable is set to 1.0, as described earlier, the variance of the independent latent variable is estimated. This was the case for self-efficacy beliefs in the Figure 2 model. To summarize, variance was estimated for indicator error associated with the 12 observed variables, variance (referred to as disturbance) in the three endogenous latent variables, and variance in the single exogenous latent variable.

Finally, covariances are nondirectional associations among exogenous variables. If a researcher expected that two factors were associated but that a causal relationship did not exist, then he or she would specify a covariance between the factors. Given the theoretical background of the model in Figure 2, we included no covariances in the model. Thus, we specified 28 parameters for estimation.

Model Identification

For many new to SEM, model identification is a complex concept to understand. In fact, most researchers treat it not so much as a step in SEM but as a condition that they must consider prior to analyzing data. As our goal here is to provide a guide to interpreting others' analyses, we opted to present only the aspects most relevant to understanding an analysis. Therefore, we caution readers that they must consult additional resources (e.g., Bollen, 1989; Byrne, 2001; Kline, 2005; Schumacker & Lomax, 2004) prior to conducting any analysis (consulting meta-analytic reviews may also be helpful; see Quintana & Minami, in press [*TCP* special issue, in press]).

As in factor analysis, SEM's goal is to find the most parsimonious summary of the interrelationships among variables that accurately reflects the associations observed in the data. For example, the correlation matrix in Figure 3 summarizes the interrelationships between variables with 10 correlations. The Figure 3 model represents the hypothesis that these 10 correlations can be summarized in a reduced form. Specifically, the researcher is testing the hypotheses that variable *a* correlates so weakly with variables *c* and *d* that the associations are not significantly different from zero. A second set of hypotheses tests whether the correlations between variable *b* and variables *c* and *d* are significantly different from zero. By specifying fewer relationships between the variables than elements in the correlation matrix, researchers are able to test hypotheses about which relationships are significantly different from zero and which are not.

The model in Figure 3 is an example of an overidentified model because it represents a reduction of the correlation matrix. It has the potential to fit the data poorly, resulting in a failure to reject the null hypothesis. For example, because the model does not capture the correlation between variable *b* and variable *d*, the model is likely misspecified. If the model were to include all possible interrelationships between variables, it would be just-identified and would essentially reproduce the elements included in the correlation matrix. Because a just-identified model will always fit perfectly (being a summary of the observed data), it is of less interest to researchers. Finally, an underidentified model essentially requires more information than is available.

Determining whether the model in Figure 3 is over-, under-, or just-identified is a fairly straightforward process that involves determining the number of degrees of freedom. Researchers calculate the number of degrees of freedom in a model by subtracting the number of parameters to be estimated from the number of known elements (correlations) in the correlation matrix. Rather than counting the correlations in a correlation matrix, researchers can use the following formula:

$$(\text{no. observed variables } [\text{no. observed variables} + 1])/2.$$

Using the correlation matrix in Figure 3, which has four observed variables, the formula results in (4 [4+1])/2, or 10 known elements. The number of unknown, or estimated, parameters is captured in the model specification. In the Figure 3 model, parameters to be estimated include four error variances and four factor loadings, all marked with asterisks. Thus, eight parameters are unknown. Subtracting the 8 unknown parameters from the 10 known elements shows that the model has two degrees of freedom. When there are more than zero degrees of freedom, the model is overidentified.

If we were to specify two additional parameters to be estimated (e.g., include paths from Factor 1 to indicator *c* and from Factor 2 to indicator *b*), the model would include 10 unknown parameters to be estimated (the same 4 error variances and now 6 factor loadings). With 10 elements in the correlation matrix and 10 parameters to be estimated, there are zero degrees of freedom and the model is just-identified. A model with zero degrees of freedom is always just-identified and will fit the data perfectly.

If we were to continue adding parameters to the model (e.g., a path from Factor 1 to indicator *d* and from Factor 2 to indicator *a*), the model would be underidentified because there are more unknown parameters (4 error variances and 8 factor loadings = 12) than there are elements in the correlation matrix. Subtracting 12 (unknown parameters) from 10 (known elements) results in –2. When the number of degrees of freedom is negative, the model is underidentified and cannot be estimated. Although one would not likely encounter this in any publication, we note that underidentification is possible to emphasize the importance of specifying models before data collection. Determining whether the model is identified prior to data collection helps the researcher to avoid the costly mistake of collecting data only to find that he or she cannot test the specified model.

Applying this review of identification to the model in Figure 2, there are 12 observed variables (thus 78 elements in the variance and covariance matrix; 12(12+1)/2 = 78), and we have specified 28 parameters to be estimated, noted with asterisks. Subtracting the 28 parameters to be estimated from the 78 known parameters reveals that for this model, there are 50 degrees of freedom. In summary, the greater the degrees of freedom, the more parsimonious the model. Thus, when a parsimonious model fits the data well, researchers are able to demonstrate that associations between observed and latent variables are most important.

Issues Related to Data

Ideally, the model is specified and identified before data collection. Thus, researchers address issues related to sample size and data screening as the third step. However, it is not uncommon to use archival data, especially when testing complex models. Unfortunately, researchers using archival data are often limited by the available measures, which may not be optimal indicators for the latent variables of interest. Given the importance of including theoretically appropriate and reliable observed measures for the measurement model, a clear advantage of collecting data to test a specific model is the ability to select the best measures for the hypothesized model. Of course, the sample size necessary to test most SEMs is often seen as a cost.

Sample size. The issue of sample size is one of several where there is no consensus, except to suggest that missing or nonnormally distributed data require larger samples than do complete, normally distributed data. As a result, depending on the source, researchers may find conflicting information on what sample size is adequate for SEM. Previous guidelines (e.g., Kline, 1998) indicated that 10 to 20 participants per estimated parameter would result in a sufficient sample. Using this guideline, we would need at least 280 participants to test the model in Figure 2. However, others suggest that optimal sample size may depend on a number of additional issues.

Empirical research by MacCallum, Browne, and Sugawara (1996) suggests that sample-size requirements depend on the desired power, the null hypothesis being tested, and the overall model complexity. Not surprisingly, and everything else being equal, more participants yield more statistical power. MacCallum et al. also encourage researchers to use larger sample sizes when testing more complex models. To summarize MacCallum et al.'s extensive empirical research on factors affecting sample size, the minimum sample necessary tends to be smaller when (a) the researcher desires less rather than more power, (b) the researcher is testing whether the model approximates the data as opposed to testing whether the model exactly replicates the data (i.e., tests for close fit compared to exact fit), and (c) the model is less complex (i.e., has fewer parameters to be estimated) rather than more complex.

Not all researchers agree that sample size is model specific. For example, Jackson (2001, 2003) found only a small effect of sample size on model fit when he tested the hypothesis that an inadequate sample would result in poor-fitting models. Jackson's work suggests that the reliability of observed measures and the number of indicators per factor were important determinants of model fit.

Assuming the researcher anticipates no problems with data (e.g., missing data or nonnormal distributions), we recommend a minimum sample size of 200 for any SEM. When few acceptable measures of a construct exist, when multiple measures of a construct are not at least moderately related to each other, or when reliability of measures is low, careful researchers will use larger samples. Under these guidelines, the available sample of 403 participants is acceptable when testing the model in Figure 2 because multiple indicators per latent variable are specified, all of which have internal consistency reliabilities of .85 or greater.

Multicollinearity. Multicollinearity refers to situations where measured variables are so highly related that they are essentially redundant. This problem is a concern in SEM because researchers use related measures as indicators of a construct, and sometimes, measures are too highly related for certain statistical operations to function properly. A rough guideline to check for multicollinearity is screening bivariate correlations. Bivariate correlations higher than $r = .85$ can signal potential problems (Kline, 2005). When two observed variables are highly correlated, one solution is to remove one of the redundant variables. Table 2 shows two potential problems: Correlations both between Outcome Expectations–1 (OE-1) and OE-2 and between OE-2 and OE-3 are higher than .85. Although removing OE-2 would be reasonable, another option would be to note the potential for a problem and to reconsider removing the variable if problems were to arise in estimation.

Outliers. Researchers must also examine univariate and multivariate outliers. Participants' scores represent a univariate outlier if they are extreme on only one variable. When participants have two or more extreme scores or an unusual configuration of scores, they are multivariate outliers. Univariate outliers could be either transformed or changed to the next most extreme score, depending on the normality of the data. Researchers can recode (e.g., to a value equal to three or four standard deviations beyond the mean) or remove multivariate outliers.

Normality. Most statistics used in SEM assume that the multivariate distribution is normally distributed. Violating this assumption can be problematic because nonnormality will affect the accuracy of statistical tests. If a researcher tests a model with nonnormally distributed data, the results may incorrectly suggest that the model is a good fit to the data or that the model is a poor fit to the data, depending on the degree and type of the problem.

Testing whether the assumptions for multivariate normality are met is impractical as it involves examining an infinite number of linear combinations. One solution is to examine the distribution of each observed variable. This screening for univariate normality can inform researchers whether multivariate normality may be an issue.

To determine whether univariate normality exists, the researcher examines the distribution of each observed variable for skewness and kurtosis. Skewness is the degree to which a variable's distribution is asymmetrical, with positive skew describing a distribution where many scores are at the low end of a scale (e.g., the score distribution for a very difficult test). For the skewness index, absolute values greater than 3.0 are extreme (Chou & Bentler, 1995). Kurtosis is an index of the peak and tails of the distribution. Positive kurtosis reflects very peaked distributions (i.e., leptokurtic), with short, thick tails (also called heavy-tailed distributions) representing few outliers. Negative kurtosis exists when the distribution is quite flat (i.e., mesokurtic), with long, thin tails, indicating many outliers. Absolute values higher than 10.0 for the kurtosis index suggest a problem, and values higher than 20.0 are extreme (Kline, 2005). Univariate normality is especially

TABLE 2: Covariance and Correlation Matrices With Means and Standard Deviations for Observed Variables

	Occupational Considerations (OC)			*Outcome Expectations (OE)*			*Self-Efficacy Beliefs (SE)*			*Interests (INT)*		
	OC-1	*OC-2*	*OC-3*	*OE-1*	*OE-2*	*OE-3*	*SE-1*	*SE-2*	*SE-3*	*INT-1*	*INT-2*	*INT-3*
OC-1	**4.811**	0.791	0.774	0.633	0.637	0.612	0.542	0.521	0.449	0.437	0.606	0.457
OC-2	3.909	**5.074**	0.721	0.607	0.648	0.608	0.582	0.606	0.474	0.371	0.521	0.380
OC-3	3.179	3.041	**3.503**	0.599	0.658	0.626	0.444	0.476	0.517	0.470	0.568	0.379
OE-1	2.976	2.931	2.402	**4.590**	0.861	0.832	0.581	0.437	0.465	0.452	0.531	0.384
OE-2	2.743	2.867	2.418	3.626	**3.860**	0.889	0.543	0.538	0.515	0.427	0.559	0.436
OE-3	2.764	2.821	2.410	3.668	3.597	**4.237**	0.506	0.526	0.482	0.471	0.549	0.487
SE-1	3.260	3.592	2.279	3.413	2.926	2.852	**7.511**	0.748	0.799	0.330	0.431	0.231
SE-2	3.070	3.665	2.394	2.518	2.840	2.910	5.505	**7.220**	0.807	0.390	0.440	0.394
SE-3	2.596	2.817	2.553	2.630	2.671	2.618	5.773	5.716	**6.958**	0.362	0.405	0.284
INT-1	0.257	0.224	0.236	0.260	0.225	0.260	0.243	0.281	0.256	**0.072**	0.545	0.520
INT-2	0.336	0.297	0.269	0.288	0.278	0.286	0.299	0.299	0.270	0.037	**0.064**	0.538
INT-3	0.280	0.239	0.198	0.230	0.239	0.280	0.177	0.296	0.209	0.039	0.038	**0.078**
M	2.715	2.530	1.654	4.663	4.186	4.045	3.975	4.016	3.370	0.298	0.279	0.312
SD	2.193	2.252	1.871	2.142	1.964	2.058	2.741	2.687	2.637	0.268	0.252	0.279

NOTE: Covariances appear in the lower left of the matrix, with variances on the diagonal in bold. Correlations appear in the upper right of the matrix.

important to consider because distributions can vary from normality in at least four ways. For example, some variables may be positively skewed, whereas others are negatively skewed.

One method of increasing normality is transforming data. Common transformations include square root (when data are moderately positively skewed), logarithm (for more than moderate positive skew), and inverse (for severe positive skew, such as an *L*-shaped distribution). Deleting or transforming univariate or multivariate outliers enhances multivariate normality. Researchers should be aware, however, that transformation is not desirable if the variable is meaningful (e.g., height) or is widely used (e.g., IQ scores), because transformation would hinder interpretation.

Missing data. Rubin (1976) described three categories for missing data. Data may be missing completely at random, missing at random, or not missing at random (NMAR). Missing data in the first two conditions are less problematic than in the third because NMAR implies a systematic loss of data. For example, if participants were missing data on the interests construct because they have few interests and chose to skip those items, data would be NMAR. In longitudinal research, data missing due to attrition would be a concern when reasons for attrition are related to study variables (e.g., attrition due to death in a health study). Unfortunately, there is no procedure for determining whether data are missing randomly. Clearly, this is an issue, as methods for handling missing data vary according to the randomness of its missing. Allison's (2003) review of techniques such as listwise deletion, pairwise deletion, maximum likelihood (ML), and multiple imputation is quite informative for readers interested in details beyond the scope of this review. Minimally, we suggest that researchers address missing data by noting the extent of the problem and how they handled it.

Estimation

After specifying the model, determining that the model is identified, collecting data from a sufficiently large sample of participants, and addressing any problems with the data, researchers are finally at the point of estimating the model. Estimation involves determining the value of the unknown parameters and the error associated with the estimated value. As in regression, researchers include both unstandardized and standardized parameter values, or coefficients, as output. The unstandardized coefficient is analogous to a B weight in regression. Dividing the unstandardized coefficient by the standard error produces a *z* value that is analogous to the *t* value associated with each B weight in regression. The standardized coefficient is analogous to β in regression.

Researchers generate estimates of the free (unknown) parameters using an SEM software program. Many programs are available, including LISREL (Linear Structural Relationships; Jöreskog & Sörbom, 1996), AMOS (Analysis of Moment Structures; Arbuckle, 2003), PROC CALIS (Covariance Analysis and Linear Structural Equations; Hartmann, 1992), SAS (SAS Institute, 2000), EQS (Equations; Bentler, 1995), and Mplus (Muthén & Muthén, 1998-2004). Programs differ in their ability to compare multiple groups and estimate parameters for categorical indicators and in the specific fit indices provided as output. Completely discussing the advantages and disadvantages of each package is beyond the scope of this article, and we encourage readers to consult software manuals and SEM guides, some of which include student versions of software (e.g., Schumacker & Lomax, 2004), prior to selecting a software package. As we are most familiar with EQS, we used the most recent version (EQS 6.1) to test the covariance matrix in the lower diagonal of Table 2.

Types of estimators. There are several estimation procedures, including ML, least squares (LS), unweighted LS, generalized LS, and asymptotic distribution free (ADF). Researchers must select which estimation method to use prior to conducting their analysis. There are pros and cons with each estimation method, so we will briefly address the issues that researchers face in making this decision.

One deciding factor is whether the data are normally distributed. ML and generalized LS methods assume multivariate normality, whereas LS and ADF do not. LS estimation does not provide a valid inference to the population from the sample, but ADF does when the sample is sufficiently large. One of the most common techniques, ML, is robust to moderate violations of the normality assumption (Anderson & Gerbing, 1984), and many researchers opt to use ML when data are moderately nonnormal.

If the data are severely nonnormal, the researcher has three options (Kline, 2005). First, the researcher may analyze nonnormal data with corrected statistics, such as scaled goodness-of-fit tests and robust standard errors, to reduce bias (Kline, 2005). Researchers choose this option quite often with nonnormal data. Second, the researcher may transform the data (e.g., square root) and then analyze the data with ML or LS estimation. If the researcher transforms the original scale data, he or she typically retransforms the obtained parameter estimate to the original scale to allow for result interpretation. Kline (2005) notes that unweighted LS is sensitive to transformation and generally not effective with transformed data. Third, researchers may estimate nonnormal data with methods such as ADF, which do not assume multivariate normality. The disadvantage of ADF is that it needs very large samples (i.e., $n = 500$ or more) to generate accurate estimates for even the simplest models (Yuan & Bentler, 1998). In contrast, simple models estimated with ML require a sample size as small as 200 for accurate estimates.

Approaches to estimation. Anderson and Gerbing (1988) use confirmatory factor analysis to test the measurement model before estimating the full structural model. The confirmatory factor analysis tests whether indicators load on specific latent variables as proposed. After model estimation, researchers examine factor loadings to determine whether any indicators do not load as expected. Examples would be indicators that load on multiple factors when expected to load on only one factor or indicators that fail to load significantly on the expected factor. We recommend, as do others (e.g., Anderson & Gerbing, 1988; Kline, 2005), that researchers make reasonable necessary changes to the measurement model when encountering problems with the model. In the second step, researchers test the full structural model by estimating expected directional associations among latent variables, indicated with unidirectional arrows in Figure 2.

Mulaik and Millsap (2000) proposed a four-phase alternative to Anderson and Gerbing's (1988) approach. Mulaik and Millsap's first phase is the equivalent of estimating an exploratory factor analysis, with paths from each latent variable to all observed variables. For example, each of the four latent variables in Figure 1 would have a direct effect on SE-1 rather than on only self-efficacy beliefs. This step allows the researcher greater precision in determining potential problems with the measurement model. In the second phase, the researcher tests the confirmatory factor analysis as Anderson and Gerbing described. Third, the researcher tests the measurement and structural portions of the model simultaneously, equivalent to Anderson and Gerbing's second phase. The goal of the fourth phase is to test a priori hypotheses about parameters specified by the researcher. For example, this phase would involve testing whether the freely estimated path from self-efficacy beliefs to outcome expectations is significantly different from zero. By including their fourth phase, Mulaik and Millsap have explicitly addressed the need to consider potential alternatives to the proposed model.

We encourage researchers to follow one of the two multiphase approaches and to test the measurement portion of their models before testing the full model. In this example, we followed Anderson and Gerbing's (1988) approach, although we also considered an alternative model. However, to avoid duplication with another article in this issue addressing confirmatory factor analysis (Kahn, 2006), we discuss the estimation of the structural model only.

Estimation example. We estimated the model in Figure 2 with ML. Figure 4 shows the standardized results for the structural portion of the full model. (To save space, we have also included estimates for an alternative model with the two dashed paths. We discuss this alternative model later, and readers should ignore estimates in parentheses for now.) Most SEM software programs provide standardized and unstandardized output, which is analogous to standardized betas and unstandardized B weights (accompanied by standard

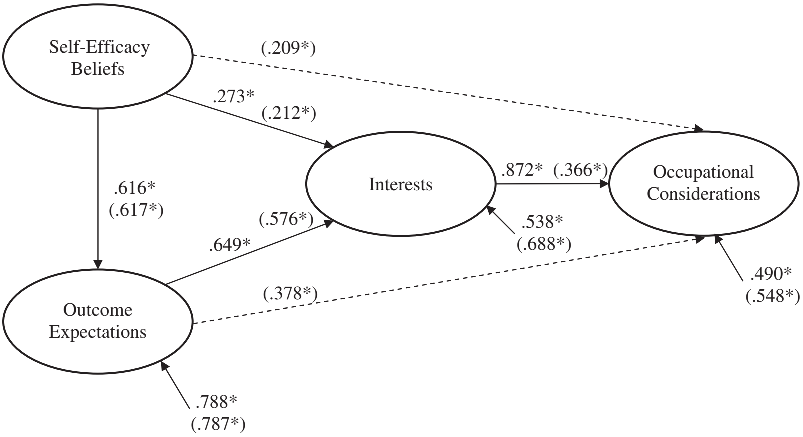


FIGURE 4. Standardized Parameter Estimates for Fully and Partially Mediated Models

NOTE: Standardized estimates for fully mediated model are listed above or to the left. Fit of fully mediated model: $\chi^2(50, N = 403) = 416.06$, $p < .05$; comparative fit index = .91; root mean square error of approximation (90% confidence interval) = .14 (.12 – .15); standardized root mean square residual = .05. *$p < .05$. Standardized estimates (in parentheses) for partially mediated model below or to the right of estimates for fully mediated model. Dashed lines represent paths only estimated in partially mediated model. Fit of partially mediated model: $\chi^2(48, N = 403) = 361.85$, $p < .05$; comparative fit index = .93; root mean square error of approximation (90% confidence interval) = .13 (.12 – .14); standardized root mean square residual = .04. *$p < .05$.

errors) in regression analysis. Researchers typically present standardized estimates but determine significance by examining the unstandardized portion of the output. For example, Figure 4 shows a significant relationship between self-efficacy beliefs and outcome expectations ($\beta = 0.616$, $p < .05$). We determined the significance of this path coefficient by examining the unstandardized output, which showed that the unstandardized coefficient (the B weight) was 0.498 and had a standard error of 0.039. Although the critical ratio (i.e., z score) is automatically calculated and provided with output in EQS and other programs, researchers can easily determine whether the coefficient is significant (i.e., $z \geq 1.96$ for $p \leq .05$) at a given alpha level by dividing the unstandardized coefficient by the standard error. Here, 0.498 divided by 0.039 is 12.77, which is greater than the critical z value (at $p = .05$) of 1.96, indicating that the parameter is significant.

Examining standardized estimates is also informative. Because different variables may have different scales, only by comparing standardized parameter estimates can a researcher determine which variable(s) has the greatest impact. In addition to providing standardized parameter estimates, most software programs include estimates of the proportion of variance explained (R^2) for all dependent variables. This information is helpful in determining which indicators contain the most (and least) measurement error.

Model Fit and Interpretation

Once estimated, the model's fit to the data must be evaluated. The objective is to determine whether the associations among measured and latent variables in the researcher's estimated model (e.g., the fully mediated model in Figure 4) adequately reflect the observed associations in the data (e.g., the data in Table 2). Statisticians agree that researchers should evaluate fit in terms of (a) significance and strength of estimated parameters, (b) variance accounted for in endogenous observed and latent variables, and (c) how well the overall model fits the observed data, as indicated by a variety of fit indices. Although addressing the first two criteria is a fairly straightforward task, there exists considerable disagreement over what constitutes acceptable values for global fit indices. Similar to sample size, this is a controversial issue in SEM.

Multiple indices are available to evaluate model fit. The most stringent concept of fit suggests that the model must exactly replicate the observed data. A second perspective is that models approximating the observed data are acceptable. For more information on the "exact versus close fit" debate, we refer readers to discussions by Quintana and Maxwell (1999) and by Marsh, Hau, and Wen (2004). Martens's (2005) study suggests that the perspective commonly taken by social scientists reflects the assumption that approximating observed data is acceptable and can result in important contributions to the literature. Hoyle and Panter (1995) have recommended that researchers report several indices of overall model fit, a practice that many have followed (Martens, 2005). We present the fit indices reported by most software programs, which have been shown to be the most accurate in a variety of conditions.

GFI and χ^2. Absolute fit indices directly assess how well a model fits the observed data (i.e., how well the model in Figure 4 describes the data in Table 2) and are useful in comparing models when testing competing hypotheses. Absolute fit indices include the goodness-of-fit index (GFI; Jöreskog & Sörbom, 1981), χ^2 (Bollen, 1989), and scaled χ^2 (Satorra & Bentler, 1994). GFI is analogous to R^2, used in regression to summarize the variance explained in a dependent variable, yet GFI refers to the variance accounted for in the entire model. However, researchers do not report GFI as consistently as χ^2. Both χ^2 values are actually tests of model misspecification.

Thus, a significant χ^2 suggests the model does not fit the sample data. In contrast, a nonsignificant χ^2 is indicative of a model that fits the data well. Although the most commonly reported absolute fit index is χ^2, two limitations exist with this statistic. First, this statistic tests whether the model is an exact fit to the data. Finding an exact fit is rare. Second, as with most statistics, large sample sizes increase power, resulting in significance with small effect sizes (Henson, 2006 [this issue]). Consequently, a nonsignificant χ^2 may be unlikely, although the model may be a close fit to the observed data. Despite these limitations, researchers report the χ^2 almost universally (Martens, 2005), and as we will show, it provides a means for testing whether two models differ in their fit to the data. Researchers typically consider additional fit indices to determine whether the model fit is acceptable.

Many additional indices are available, but not all software programs provide the same indices. This has resulted in the lack of a standard format for reporting fit and an inability to compare across studies. In addition, reviewers may prefer to see specific indices. Following the recommendations of Boomsma (2000), MacCallum and Austin (2000), and McDonald and Ho (2002), we present three indices in addition to the model χ^2: Bentler's (1990) Comparative Fit Index (CFI); Steiger's Root Mean Square Error of Approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), including the associated 90% confidence interval (90% CI); and the Standardized Root Mean Square Residual (SRMR).

CFI. Bentler's (1990) CFI is an example of an incremental fit index. This type of index compares the improvement of the fit of the researcher's model over a more restricted model, called an independence or null model, which specifies no relationships among variables. CFI ranges from 0 to 1.0, with values closer to 1.0 indicating better fit.

RMSEA. We also suggest the RMSEA (Steiger, 1990; Steiger & Lind, 1980) as an index of fit. This index corrects for a model's complexity. As a result, when two models explain the observed data equally well, the simpler model will have the more favorable RMSEA value. A RMSEA value of .00 indicates that the model exactly fits the data. A recent practice is to provide the 90% CI as well for the RMSEA, which incorporates the sampling error associated with the estimated RMSEA.

SRMR. The SRMR (Bentler, 1995) index is based on covariance residuals, with smaller values indicating better fit. The SRMR is a summary of how much difference exists between the observed data and the model. For example, we noted in Figure 3 of the model specification section, that indicator *b* has a stronger relationship with indicator *d* (r = .57 in the correlation matrix) than would be expected according to the figure. The actual correlation is greater than the model implies, resulting in error. The SRMR is the absolute mean of all differences between the observed and the model-implied correlations. A mean of zero indicates no difference between the observed data and the correlations implied in the model; thus, an SRMR of .00 indicates perfect fit. Regardless of whether researchers opt to report this index, we strongly recommend that researchers examine the standardized residuals, which are included as output, to identify any associations in the observed data that are not reflected in the estimated model.

Guidelines for fit. Previously, guidelines for acceptable fit included a nonsignificant χ^2, CFI greater than .90 (Hu & Bentler, 1995), RMSEA less than .10 with a maximum upper bound of the 90% CI of .10 (Browne & Cudek, 1993), and SRMR less than .10. (Bentler, 1995). Although many still follow these guidelines (as evidenced by the positive evaluation of models with fit indices at or near these values), readers should be aware that debate exists among statisticians (and likely among reviewers) regarding acceptable fit. Recent studies (e.g., Hu & Bentler, 1998, 1999) have suggested a minimum cutoff of .95 for CFI and a maximum cutoff of .06 for RMSEA. However, Hu and Bentler's work and that of others (e.g., Marsh et al., 2004) also indicate that sample size, model complexity, and degree of misspecification affect appropriate cutoff values. This body of research has shown that inappropriately applying the new cutoff criteria could result in the incorrect rejection of acceptable models when samples sizes are smaller than $n = 500$ and when models are not complex. On the other hand, using old criteria to evaluate complex models estimated with samples larger than $n = 500$ can result in the incorrect acceptance of misspecified models. Empirical research suggests that fit indices not meeting the previous guidelines (i.e., CFI ≥ .90, RMSEA ≤ .10, and SRMR ≤ .10) would likely not be acceptable and indicates that models with indices exceeding new criteria would be acceptable (i.e., CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08). This suggests that when CFI values between .90 and .95, RMSEA values between .05 and .10, and SRMR values between .08 and .15 are observed, readers should consider the sample size used to estimate the model (using more stringent criteria for samples larger than $n = 500$) and the model complexity (using more stringent criteria for less complex models).

We summarize fit indices for the fully mediated model in Figure 4 below the figure. As is occasionally the case, the fit indices contradict each other. The CFI and the SRMR indicate the model may be acceptable. However, the significant χ^2 and high RMSEA suggest the model may be a poor fit to the data. In addition, the 90% CI for the RMSEA indicates that there is little error in the point estimate of the fit index and that it is unlikely that the model in Figure 4 adequately describes the data in Table 2. Overall, more evidence suggests that some misspecification may exist than suggests that the model fits well.

Parameter estimates. In addition to considering overall model fit, it is important to consider the significance of estimated parameters, which are analogous to regression coefficients. As with regression, a model that fits the data quite well but has few significant parameters would be meaningless. Figure 4 includes standardized estimates for path coefficients, interpreted as regression coefficients. (We do not include path loadings for indicators here, because they would have been examined for significance when conducting the confirmatory factor analysis during the measurement-model testing.) Standardized estimates allow the relationships among latent variables to be compared. Figure 4 indicates a stronger relationship between outcome expectations and interests ($\beta = .649$) than between self-efficacy beliefs and interests ($\beta = .273$). However, the relationship between self-efficacy beliefs and interests is also partially mediated by outcome expectations, so two paths from self-efficacy beliefs to interests can be traced in the model (self-efficacy beliefs → interests and self-efficacy beliefs → outcome expectations → interests). Altogether, self-efficacy beliefs and outcome expectations explain 71.1% of the variance in interests. Researchers determine the amount of explained variance (R^2) by squaring the disturbance error associated with dependent latent variables (for interests, $D^2 = .538^2 = .289$) and subtracting the value from 1 ($R^2 = 1 - D^2$). Self-efficacy beliefs, as the sole predictor of outcome expectations, explain 37.9% of the variance in outcome expectations. The proportion of variance accounted for in occupational considerations was much higher (76%), with a strong, significant association between occupational considerations and interests.

All proposed parameters were significant and in the expected direction. One problem that can arise, but that was not observed here, is a negative value for an estimate of error variance associated with dependent latent variables or observed variables. Negative estimates of error variance are called Heywood cases and result from a standardized factor loading or path coefficient greater than 1.0. The presence of Heywood cases indicates a poorly specified model (Kline, 2005), and the researcher should consider problems in specifying the model that may have led to the Heywood case.

Model Modification

Rarely is a proposed model the best-fitting model. Consequently, modification (respecification) may be needed. This involves adjusting the estimated model by freeing (estimating) or setting (not estimating) parameters. Modification is a controversial topic, which has been likened to the debate about post hoc comparisons in ANOVA (Hoyle, 1995). Readers interested in specific aspects of the dispute should refer to Bollen and Long's (1993) edited volume, which is devoted entirely to the debate, as well as to discussions by Hoyle and Panter (1995), MacCallum and Austin (2000), and McDonald and Ho (2002). As with ANOVA and regression, problems with model modification include capitalization on chance and results that are specific to a sample because they are data driven. Although there is disagreement regarding the acceptability of post hoc model modification, statisticians and applied researchers alike emphasize the need to clearly state when there was post hoc modification rather than imply that analyses were a priori.

As Martens (2005) reports, researchers generally accomplish modification by using statistical search strategies (often called a specification search) to determine which adjustments result in a better-fitting model. The Lagrange Multiplier test identifies which of the parameters that the researcher assumed to be zero are significantly different from zero and should be estimated. The Wald test, in contrast, identifies which of the estimated parameters that were assumed to be significantly different from zero are not and should be removed from the model. Schumacker and Lomax (2004) and Kline (2005) provide detailed information on conducting specification searches using modification indices. Researchers who opt to engage in post hoc modification should be aware (and ready to address reviewers' concerns) that such capitalization on chance often results in estimating data-driven models that are potentially not generalizable across samples (e.g., Chou & Bentler, 1990; Green, Thompson, & Babyak, 1998). This problem is more likely when researchers (a) use small samples, (b) do not limit modifications to those that are theoretically acceptable, and (c) severely misspecify the initial model (Green et al., 1998). Careful researchers will modify their model within the limitations of their theory. For example, if a Wald test indicated the researcher should remove the freely estimated parameter from self-efficacy beliefs to outcome expectations, then the researcher would not include that modification, because the suggested relationship contradicts theory and research. Ideally, researchers should test model modifications suggested by Wald or Lagrange Multiplier tests on a separate (cross-validation) sample. Given the large samples necessary and the cost of collecting data for cross-validation, a more common practice is to carefully note how the model was modified and cautiously interpret modified models. As this practice may result in data-driven models, we emphasize the need for replicating models to advance the literature.

Testing Alternative Models

Statisticians proficient in SEM have commented that social scientists often fail to test alternatives to proposed models (e.g., Boomsma, 2000; Steiger, 2001). The best way for researchers to address this is to present the proposed model as well as one or more theoretically plausible models representing competing hypotheses. For example, the Figure 2 model specifies that interests fully mediate the effects of self-efficacy beliefs and outcome expectations on occupational considerations. Positing a fully mediated model has important implications that differ from those of a partially mediated model. Therefore, researchers should consider the possibility that self-efficacy beliefs and outcome expectations each have a direct effect on the outcome. The two dashed paths in Figure 4 show this alternative model.

The fit of the alternative model is compared with that of the proposed model in three ways: Researchers (a) evaluate paths by examining significance of parameter estimates, (b) consider the change in explained variance for occupational considerations, and (c) test significant improvement in model fit with a chi-square difference test and improvement in other fit indices. Examining the estimates in parentheses in Figure 4 shows that all original parameters and the additional two parameters are significant. This is the first indication that the alternative model may be acceptable. However, a decrease in explained variance is evident for occupational considerations (a 6% decrease to 69.9% explained). Ideally, model modifications would result in an increase in explained variance or, at least, no change. Third, the fit of the fully mediated model is compared with that of the alternative partially mediated model. The second set of fit indices below the figure shows that the χ^2 for the partially mediated model is smaller but still significant. There is little change in the RMSEA and associated 90% CI, but the CFI and SRMR show an improvement in fit.

Additional fit indices, used to determine which of two or more competing models is the best fitting, include the Akaike Information Criterion (AIC) and the Expected Cross-Validation Index (Browne & Cudek, 1993). Both are considered predictive fit indices and indicate how well models would be expected to fit sample data drawn from the same population. These indices are not informative in determining how well a single model fits the data but are generally used to choose between models. We report only the AIC here, as EQS does not include the Expected Cross-Validation Index as output. Smaller values indicate better-fitting models, with more parsimonious models favored. For the fully mediated model, AIC equals 321.31. In comparison, for the partially mediated model with the two dashed paths, AIC equals 265.85, indicating that the more complex, partially mediated model is a better fit.

The fully mediated model is more restricted than the alternative partially mediated model because the dashed paths from self-efficacy beliefs and outcome expectations to occupational considerations are set to zero and not estimated. Because the two paths remain free to be estimated (i.e., not set to any specific value) in the alternative partially mediated model, this model is nested within the more restrictive fully mediated model. Researchers can directly compare nested models to determine which provides a better fit. If the decrease in the χ^2 value is significant, then the less restrictive alternative model better fits the data. The difference is calculated by subtracting the χ^2 value of the less restrictive alternative model (361.85) from the χ^2 value of the more restrictive model (416.06). This calculation provides the χ^2 difference, which is 54.21. To determine the associated degrees of freedom, the value for the less restrictive alternative model ($df = 48$) is subtracted from the value for the more restrictive model ($df = 50$). Using the calculated degrees of freedom, the researcher would use a χ^2 table (found in the appendix of most univariate statistics texts) and find the critical value at the appropriate significance level. The critical value for a χ^2 with 2 degrees of freedom at $p < .05$ is 5.99. Our value of 54.21 exceeds the critical value. Thus, the χ^2 difference test suggests that the two additional parameters in the alternative partially mediated model have resulted in an improvement in fit. Taking this together with the improvement in fit for the CFI and SRMR, and the smaller AIC suggests that the added parameters are indeed significantly different from zero and should be included.

CONCLUSIONS

SEM clearly can allow counseling psychology researchers to ask more complex research questions and to test multivariate models in a single study. Applications beyond the models discussed here include the examination of growth trajectories over time through the analysis of repeated measures, comparison of multiple groups to test for factorial invariance, and analysis of hierarchical data with multilevel modeling. Although many new, easy-to-use software programs have increased the accessibility of this quantitative method, SEM is still a complex family of statistical procedures requiring a great deal of judgment on the part of the researcher to avoid misuse and misinterpretation. SEM users must make decisions regarding how many participants to use, how to normalize data, what estimation methods and fit indices to use, and how to evaluate the meaning of those fit indices. We offer several guidelines for answering these questions. SEM requires at least 200 participants, with larger samples needed for specifying more complex models. Guidelines for appropriate fit are a hot topic for debate among statisticians, resulting in some uncertainty about what constitutes a good fit in SEM applications. To address this issue, we recommend that researchers report the fit indices described here (i.e., χ^2, CFI, RMSEA with 90% CI, and SRMR) as well as the standardized parameter estimates with significance. In addition, we strongly recommend that researchers include (or reviewers request) a covariance matrix or correlation matrix with means and standard deviations, which allows others to duplicate the results and independently assess model fit.

Perhaps the most important suggestion we can offer those interested in understanding SEM is not to attempt to master any static set of decision rules. Instead, we emphasize the need to continue to seek out information on the appropriate application of this technique. As consensus emerges, best practices will likely change, affecting the way researchers make decisions.

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545-557.

Anderson, J. C., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155-173.

Anderson, J. C., & Gerbing, D. C. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423.

Arbuckle, J. L. (2003). *Amos user's guide*. Chicago: SmallWaters.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238-246.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17,* 303-316.

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7,* 461-483.

Browne, M. W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell Interest and Skills Inventory.* Minneapolis, MN: National Computer Systems.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among the likelihood ratio, Lagrange Multiplier, and Wald tests. *Multivariate Behavioral Research, 25,* 115-136.

Chou, C.-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 37-55). Thousand Oaks, CA: Sage.

Fassinger, R. E. (1987). Use of structural equation modeling in counseling psychology research. *Journal of Counseling Psychology, 34*, 425-436.

Geurts, S. A. E., Kompier, M. A. J., Roxburgh, S., & Houtman, I. L. D. (2003). Does work-home interference mediate the relationship between workload and well-being? *Journal of Vocational Behavior, 63*, 532-559.

Green, S. B., Thompson, M. S., & Babyak, M. A. (1998). A Monte Carlo investigation of methods for controlling Type I errors with specification searches in structural equation modeling. *Multivariate Behavioral Research, 33,* 365-383.

Harmon, L. W., Hansen, J. I. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong interest inventory: Applications and technical guide.* Palo Alto, CA: Consulting Psychologists Press.

Hartmann, W. M. (1992). *The CALIS procedure: Extended user's guide.* Cary, NC: SAS Institute.

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Treating Cronbach's alpha reliability as data in counseling research. *The Counseling Psychologist*, *34*, 630-660.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34*, 601-629.

Heppner, P. P., Pretorious, T. B., Wei, M., Lee, D., & Wang, Y. (2002). Examining the generalizability of problem-solving appraisal in Black South Africans. *Journal of Counseling Psychology, 49*, 484-498.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Holland, J. L., Fritzsche, B. A., & Powell, A. B. (1994). *Technical manual for the self-directed search.* Odessa, FL: Psychological Assessment Resources.

Hoyt, W. T., Warbasse, R. E., & Chu, E. Y. (in press [*TCP* special issue, part 2]). Construct validation in counseling psychology research. *The Counseling Psychologist*, *34*.

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues and applications.* Thousand Oaks, CA: Sage.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 158-176). Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

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

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

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8,* 205-223.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10,* 128-141.

Jöreskog, K. G., & Sörbom, D. (1981). *Analysis of linear structural relationships by maximum likelihood and least squares methods* (Research Report No. 81-8). Uppsala, Sweden: University of Sweden.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychologist*, *34*, 684-718.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Lent, R. W. (2005). A social cognitive view of career development and counseling. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 101-127). New York: John Wiley.

Lent, R. W., Brown, S. D., & Gore, P. A. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. *Journal of Counseling Psychology, 44*, 307-315.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79-122.

Lent, R. W., Lopez, F. G., Brown, S. D., & Gore, P. A. (1996). Latent structure of the sources of mathematics self-efficacy. *Journal of Vocational Behavior, 49*, 292-308.

Long, B. C. (1998). Coping with workplace stress: A multiple-group comparison of female managers and clerical workers. *Journal of Counseling Psychology, 45*, 65-78.

Long, B. C., Kahn, S. E., & Schutz, R. W. (1992). Causal model of stress and coping: Women in management. *Journal of Counseling Psychology, 39,* 227-239.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226.

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

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320-341.

Martens, M. P. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist, 33,* 269-298.

Martens, M. P., & Hasse, R. F. (in press [*TCP* special issue, part 2]). Advanced applications of structural equation modeling in counseling psychology research. *The Counseling Psychologist*, *34*.

Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology, 76*, 127-133.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64-82.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7,* 36-73.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Quintana, S. M., & Maxwell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. *The Counseling Psychologist, 27*, 485-527.

Quintana, S. M., & Minami, T. (in press [*TCP* special issue, part 2]). Guidelines for meta-analyses of counseling psychology research. *The Counseling Psychologist*, *34*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 61,* 581-592.

Russell, D. W., Kahn, J. H., Spoth, R., & Altmaier, E. M. (1998). Analyzing data from experimental studies: A latent variable structural modeling approach. *Journal of Counseling Psychology, 45,* 18-29.

SAS Institute. (2000). *SAS/ETS software: Changes and enhancements* (Release 8.1). Cary, NC: Author.

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

Schumaker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Sherry, A. (2006). Discriminant analysis in counseling psychology research. *The Counseling Psychologist*, *34*, 661-683.

Steiger, J. H. (1990). Structural model evaluation and modification: An internal estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Steiger, J. H, (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331-338.

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

Tracey, T. J. G. (2002). Development of interests and competency beliefs: A 1-year longitudinal study of fifth- to eighth-grade students using the ICA-R and structural equation modeling. *Journal of Counseling Psychology, 49,* 148-163.

Tylka, T. L., & Subich, L. M. (2004). Examining a multidimensional model of eating disorder symptomatology among college women. *Journal of Counseling Psychology, 51,* 314-328.

Wei, M., Heppner, P. P., & Mallinckrodt, B. (2004). Perceived coping as a mediator between attachment and psychological distress: A structural equation modeling approach. *Journal of Counseling Psychology, 50,* 438-447.

Worthington, R. L., & Whittaker, T. A. (in press [*TCP* special issue, part 2]). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51,* 289-309.