Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review

JAMES B. SCHREIBER
Duquesne University

FRANCES K. STAGE
New York University

JAMIE KING
Duquesne University

AMAURY NORA
University of Houston

ELIZABETH A. BARLOW
University of Houston

ABSTRACT The authors provide a basic set of guidelines and recommendations for information that should be included in any manuscript that has confirmatory factor analysis or structural equation modeling as the primary statistical analysis technique. The authors provide an introduction to both techniques, along with sample analyses, recommendations for reporting, evaluation of articles in *The Journal of Educational Research* using these techniques, and concluding remarks.

Key words: confirmatory factor analysis, reports statistical results, research methods, structural equation modeling

In many instances, researchers are interested in variables that cannot be directly observed, such as achievement, intelligence, or beliefs. In research methodology, authors use terms such as latent variables or factors to describe unobserved variables. We attempt to gain information about latent factors through observable variables. Factor analysis (exploratory and confirmatory) and structural equation modeling (SEM) are statistical techniques that one can use to reduce the number of observed variables into a smaller number of latent variables by examining the covariation among the observed variables.

In this article, we provide a general description of confirmatory factor analysis (CFA) and SEM, examples of both with a Results section, guidelines for evaluating articles with CFA and SEM as analysis techniques, and a brief review of CFA and SEM articles published in *The Journal of Educational Research* between 1989 and 2004.

Terminology for CFA and SEM

A discussion about CFA and SEM techniques must begin with the terminology and graphics typically used in these types of articles. With both techniques, we talk about observed and unobserved variables, but these distinct categories can incorporate a host of different names. *Observed variables* are also termed measured, indicator, and manifest, and researchers traditionally use a square or rectangle to designate them graphically (Figure 1). The response to a Likert-scaled item, ranging from 5 (*strongly agree*) to 1 (*strongly disagree*) is an example of an observed variable.

Unobserved variables are termed latent factors, factors, or constructs and are depicted graphically with circles or ovals (Figure 1). *Common factor* is another term used because the effects of unobserved variables are shared in common with one or more observed variables. In Figure 1, the circles at the top are the unobserved or latent variables; the circles at the bottom are the unique factors—measurement errors—in the variables. The unique factors differ from the latent factors because their effect is associated with only one observed variable. The straight line pointing from a latent variable to the observed variables indicates the causal effect of the latent variable on the observed variables. The curved arrow between latent variables indicates that they are correlated. If the curve were changed to a straight one-headed arrow, a hypothesized direct relationship between the two latent variables would be indicated. Also, the directional path would be considered a structural component of the model; this is discussed further in the SEM section.

CFA

CFA is a confirmatory technique—it is theory driven. Therefore, the planning of the analysis is driven by the theoretical relationships among the observed and unobserved variables. When a CFA is conducted, the researcher uses a hypothesized model to estimate a population covariance matrix that is compared with the observed covariance matrix. Technically, the researcher wants to minimize the difference between the estimated and observed matrices.

Figure 2 shows a CFA. The latent variables are deep processing (Deep) and knowledge is isolated facts (Isolated). In the example, each latent variable is measured with three observed variables. The six observed variables are responses to three statements from two Likert-based scales. The

Address correspondence to James B. Schreiber, Department of Foundations & Leadership, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282. (E-mail: schreiberj@duq.edu)

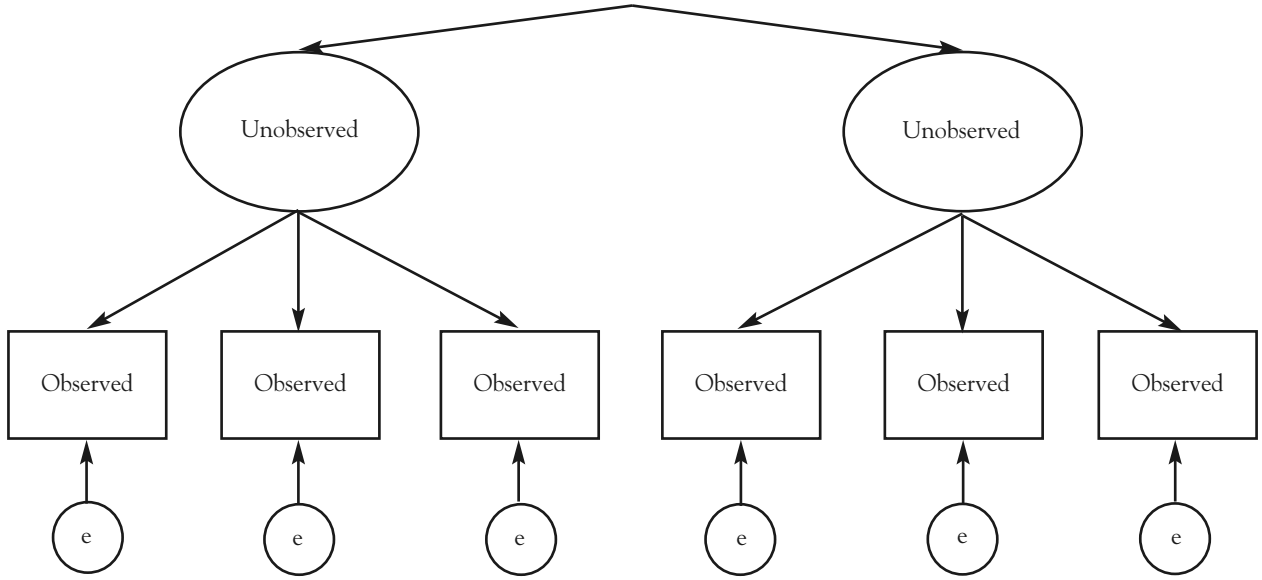

FIGURE 1. Generic example of a confirmatory factor analysis. e = error.

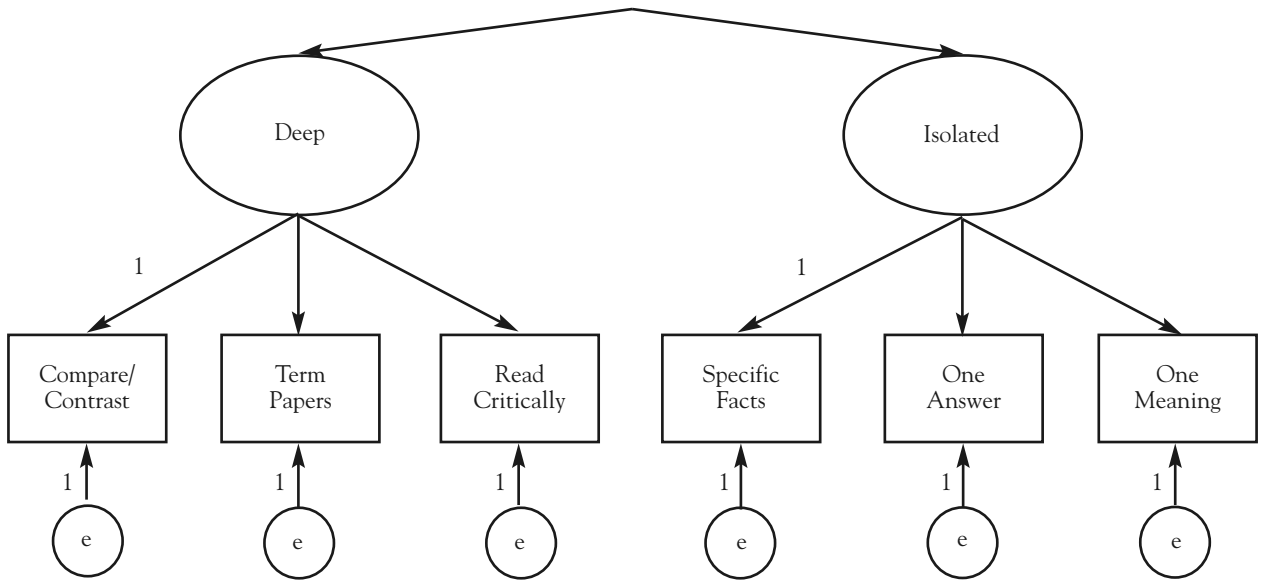

FIGURE 2. Example of a confirmatory factor analysis. e = error.

numbers "1" in the diagram indicate that the regression coefficient has been fixed to 1. Coefficients are fixed to a number to minimize the number of parameters estimated in the model. Values other than 1 can be chosen and will not change the overall fit of the model, but rather, affect the variance of the error. The graphic representation is the hypothesized model that is to be tested to see how well it fits the observed data. Mathematical equations exist that describe the pictured relationships, but presentation of these equations is beyond the scope of this article. Readers are referred to Long (1983a, 1983b) and Ullman (2001), which provide explanations of the mathematical models involved in CFA and SEM.

SEM

SEM has been described as a combination of exploratory factor analysis and multiple regression (Ullman, 2001). We

like to think of SEM as CFA and multiple regression because SEM is more of a confirmatory technique, but it also can be used for exploratory purposes. SEM, in comparison with CFA, extends the possibility of relationships among the latent variables and encompasses two components: (a) a measurement model (essentially the CFA) and (b) a structural model (Figure 3). In addition to the new terms, measurement and structural, two other terms are associated with SEM: *exogenous*, similar to independent variables and *endogenous*, similar to dependent or outcome variables. Exogenous and endogenous variables can be observed or unobserved, depending on the model being tested. Within the context of structural modeling, exogenous variables represent those constructs that exert an influence on other constructs under study and are not influenced by other factors in the quantitative model. Those constructs identified as endogenous are affected by exogenous and other endogenous variables in the model.

The measurement model of SEM is the CFA (see Figure 1) and depicts the pattern of observed variables for those latent constructs in the hypothesized model. A major component of a CFA is the test of the reliability of the observed variables. Moreover, researchers also use the measurement model to examine the extent of interrelationships and covariation (or lack thereof) among the latent constructs. As part of the process, factor loadings, unique variances, and modification indexes (should a variable be dropped or a path added) are estimated for one to derive the best indicators of latent variables prior to testing a structural model. The structural model (see Figure 4) comprises the other component in linear structural modeling. The structural model displays the interrelations among latent constructs and observable variables in the proposed model as a succession of structural equations—akin to running several regression equations.

Because of the confusion, misunderstanding, and disagreement regarding the use of the term "cause" or the phrase "causal modeling," we believe that one should simply discuss the direct, indirect, and total effects among latent constructs as dictated by theory or empirically based suppositions. A direct effect (Figure 4) represents the effect of an independent variable (exogenous) on a dependent variable (endogenous). For example, knowledge as isolated facts (Isolated) has a direct effect on course achievement, as does deep processing (Deep). An indirect effect (Figure 4) represents the effect of an independent variable on a dependent variable through a mediating variable (Baron & Kenny, 1986). Knowledge as isolated facts has a direct and an indirect effect (through deep processing) on achievement. The total effect for knowledge as isolated facts is the summation of the direct and indirect effects of this variable on course achievement. Also note in Figure 4 that Deep is exogenous and endogenous.

Although the focus of structural modeling is on estimating relationships among hypothesized latent constructs,

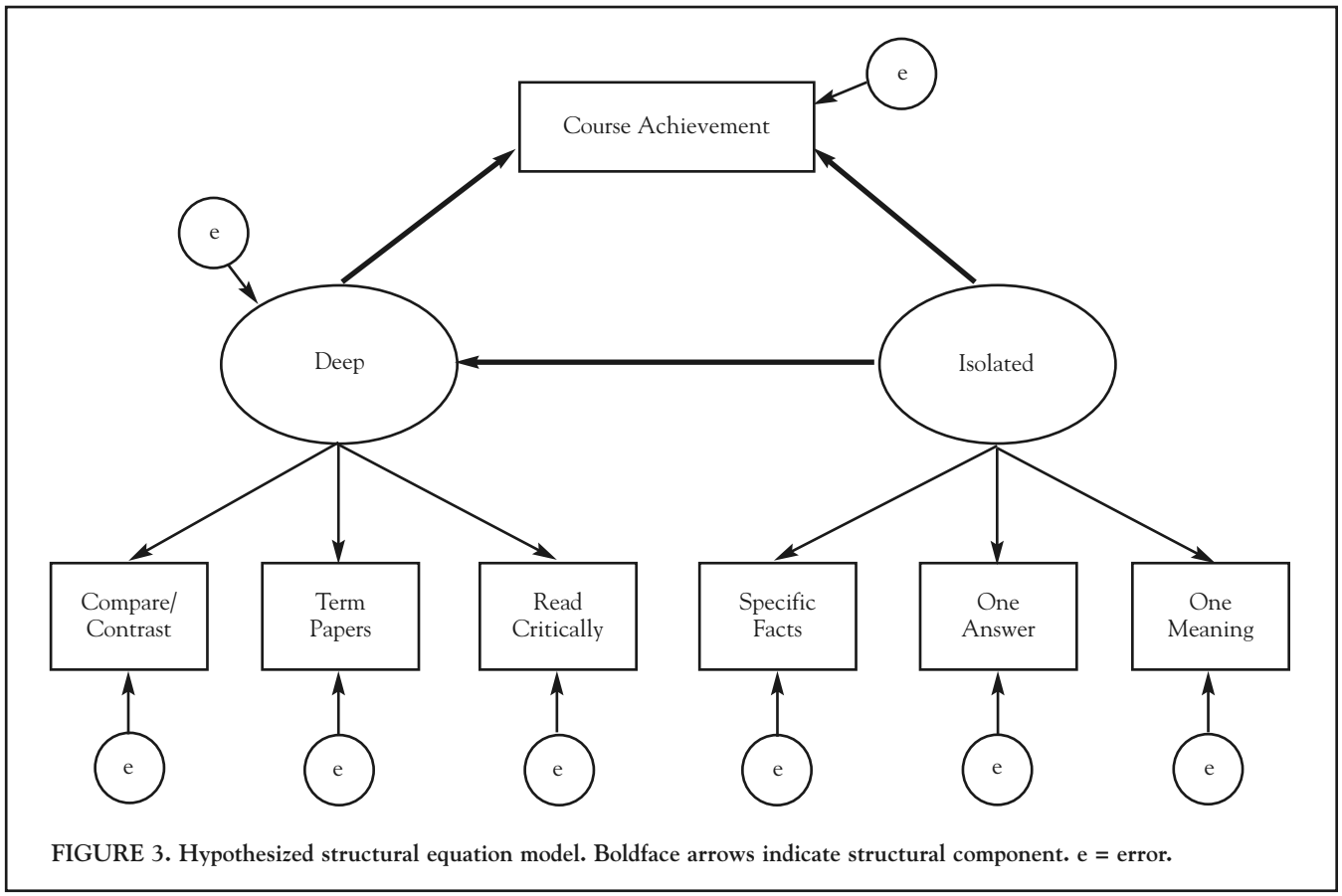

FIGURE 3. Hypothesized structural equation model. Boldface arrows indicate structural component. e = error.

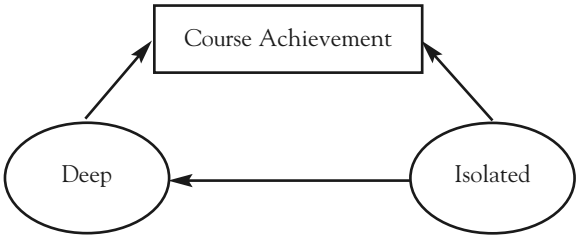

FIGURE 4. The structural model.

one can use structural modeling to test experimental data where one or more of the variables have been manipulated. In sum, SEM allows researchers to test theoretical propositions regarding how constructs are theoretically linked and the directionality of significant relationships.

Why Not Use Path Analysis?

Although the strength of path analysis lies in its ability to decompose the relationships among variables and to test the credibility of a theoretical perspective (or model), the use of such a statistical technique is predicated on a set of assumptions that are highly restrictive in nature (Pedhazur, 1982). Three of those postulations include the assumption that variables used in testing a causal model through path analysis should be measured without error, the assumption that error terms (or residuals) are not intercorrelated, and the supposition that the variables in the model flow are unidirectional (does not incorporate feedback loops among variables). Although those conditions are highly desirable, the reality is that the assumptions are rarely, if ever, found in educational settings in which nonexperimental research is more appropriate.

Almost all of the variables of interest in education research are not directly observable. Variables such as educational aspiration, test anxiety, student perceptions, and self-reported behaviors are latent constructs. The use of a single indicator to fully capture the complexities of such a construct as required in path analysis is impractical. Completely encapsulating the nature of those variables in path analysis requires that one use multiple indicators for each latent construct.

Another drawback of path analysis is that it does not permit the possibility of a degree of interrelationship among the residuals associated with variables used in the path model. Conceptually, this assumption is unsound in longitudinal studies in which individuals may be assessed at different points in time on identical variables. It is irrational to believe that error in the same variables for the same individuals at different times would not be interrelated.

Testing models that hypothesize a concurrent impact among variables is rare. The conceptualization of an investigation that centers on the feedback of one or more variables on each other is seldom, if ever, the intent of most education studies; the notion that there can only be an influence from one variable to another is unrealistic. Conceivably, academic experiences not only affect a student's academic performance but also the student's performance affects his or her academic experiences (e.g. studying, participating in study groups, accessing academic resources, engaging in classroom discussion). However, the use of path analysis for addressing such issues is not appropriate.

What Should I Look for in a CFA or SEM Article?

In this section, we provide a guide for evaluating the analysis section of a CFA or SEM article. We first describe nontechnical aspects of the article, many of which apply to other quantitative analyses. Next, we describe technical aspects of the article that we consider basic to the presentation of an analysis.

Nontechnical Evaluative Issues

We identify six nontechnical issues in evaluating a CFA or SEM article. They include (a) Research questions dictate the use of CFA or SEM; (b) a brief explanation or rationale for CFA or SEM is introduced in the method section; (c) sufficient information is provided on the measurement model's conceptual framework, structural framework, or both (i.e., the model is theoretically grounded); (d) tables and figures or text are appropriate and sufficient (i.e., descriptive statistics, such as correlation and mean tables); (e) a graphic display of the hypothesized or final models, or both, is provided; and (f) implications follow from the findings.

Technical Issues: Pre- and Postanalysis

In addition to nontechnical issues, several pre- and postanalyses technical issues must be provided within the text or tables of a CFA or SEM article. (See Table 1 detailing each article.) The first issue, sample size, is important because it relates to the stability of the parameter estimates. Are the results stable? Replication with multiple samples would demonstrate the stability of the results, but many times this is not feasible. Pohlmann (2004) argued that one could try to collect enough data to randomly split the data in half and estimate the model twice, then compare the results. For one sample analysis, there is no exact rule for the number of participants needed; but 10 per estimated parameter appears to be the general consensus. In our CFA example in a following paragraph, we specify 6 regressions, 1 covariance, and 6 variances, totaling 13 parameters that need to be estimated. Because we have an initial sample size of 203, we have an acceptable ratio of 15.6 participants to 1 parameter estimated.

Besides sample size, the report should include a thorough discussion of the handling of missing data (dropped pairwise, listwise, or estimated). One can then analyze missing

response patterns and can estimate missing data using full information maximum likelihood (FIML; Kline, 2005) or expectation–maximization (EM) algorithm (Muthén & Muthén, 1998). In general, pairwise deletion is not recommended, and listwise deletion is problematic unless the missing data have proved to be missing at random (MAR). Along with sample issues, a brief discussion concerning normality, outliers, linearity, and multicollinearity should be provided.

The final preanalysis components that one should include in the article are the software program and estimation method. Various programs provide slightly different pieces of information and can analyze only certain types of data (continuous vs. dichotomous); estimation methods are affected by sample size, normality, and the dependence of errors (Ullman, 2001).

The core of the postanalysis should be an examination of the coefficients of hypothesized relationships and should indicate whether the hypothesized model was a good fit to the observed data. An examination of the residuals should also be conducted as another indicator of model fit. Although examination and discussion of the coefficients are often secondary to the fit, they should not be. In other words, the researcher examines the significance of individual structural paths representing the impact of one latent construct on another or the latent construct on the observed variable, as is the case with CFA. The statistical significance of path coefficients is established through an examination of the *t* values or *z* values—depending on the software—associated with structural coefficients. The authors also could provide standard errors in combination with the unstandardized estimates.

In reference to model fit, researchers use numerous goodness-of-fit indicators to assess a model.[1] Some common fit indexes are the Normed Fit Index (NFI), Non-Normed Fit Index (NNFI, also known as TLI), Incremental Fit Index (IFI), Comparative Fit Index (CFI), and root mean square error of approximation (RMSEA; see Table 2). The popularity of fit-index research can be seen by the number of indexes that exist. We suggest that editors, reviewers, and consumers peruse research studies for an understanding of which indexes appear to work well with different samples sizes, types of data, and ranges of acceptable scores to decide whether a good fit exists (Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996; Yu, 2002). In general, the authors prefer the TLI, CFI, and RMSEA for one-time analyses. When modifications are made to the model after an initial analysis or multiple models are tested, one should use different indexes that are discussed in the following paragraphs.

We created a chart (Table 2) to help researchers with a basic understanding of fit indexes cutoff levels for determining model fit. In general, if the vast majority of the indexes indicate a good fit, then there is probably a good fit. Hu and Bentler (1999) suggested that for continuous data—RMSEA < .06, TLI > .95, CFI > .95, and standard root mean square residual (SRMR) < .08. For categorical outcomes, Yu (2002) reported that the above cutoff values are reasonable, except SRMR, and also suggested that weighted root mean square residual (WRMR) < .90 works well for continuous and categorical data.

MacCallum and colleagues (1996) provided a discussion on sample-size requirements for the RMSEA goodness of fit using model degrees of freedom and effect size as reference points. For example, a sample size of 231 with 45 degrees of freedom would have a power value of .80 (MacCallum et al. 1996, p. 144). Finally, for CFA, we want to know the reliability of the observed variables in relationship to the latent constructs, that is, the squared multiple correlations (SMC). For SEM, we want to know the proportion of variance accounted for in the endogenous variables.

In addition to the parameter estimates and goodness-of-fit examinations, authors also should discuss the standardized residuals. Software programs provide various types of outputs, such as a Q-Plot, but all provide some form of residual matrix that should be considered. A discussion of the standardized form of the residuals is preferred because it enables the researcher to determine the number of standard deviations of observed residuals from zero residuals that should exist if the causal model fits perfectly (Byrne, 1989). If a Q-plot is provided or discussed, standardized residuals that depart excessively from the Q-plot line indicate that the model is in some way misspecified (Byrne).

After examination of parameter estimates, fit indexes, and residuals, researchers can conduct model modifications to the original hypothesized model to have a better fitting or more parsimonious model. The software programs allow for the calculation of modification indexes because hypothesized models do not provide a perfect reproduction of the observed covariance matrix. Because those techniques are typically confirmatory in nature, any modification completed should make theoretical sense, not simply because of analyses indicated for addition or subtraction of a parameter. Therefore, the author should report (a) the modification test used (chi-square, Lagrange, or Wald), (b) why that test was used, and (c) whether the modification makes theoretical sense for the model. Otherwise, model modification simply becomes an exploratory journey and increases the likelihood of a Type 1 error.

If a model has been modified and reanalyzed, one should provide evidence that the modified model is statistically superior to the original model with a chi-square test. A model that has been modified, a trimmed model, is termed a nested or hierarchal model. In that case, one should have fit indexes and chi-square values from all models. It is imperative that the authors explain in detail from theoretical and statistical aspects why a modification was completed (Stage, 1990). Byrne (1989) advised that omitting paths not included in the original conceptualization of the model must be based on existing theoretical considerations or possibilities stating, "If the researcher is unhappy with the overall fit of the hypothesized model, he or she can re-specify a

TABLE 1. Evaluation Criteria From *The Journal of Educational Research* Articles

Measure	Herl, Baker, & Niemi (1996)	Hong (1998)	Kaplan, Liu, Xiaoru, & Kaplan (2000)	Kaplan, Peck, & Kaplan (1994)	Kaplan, Peck, & Kaplan (1997)	Keith & Benson (1992)	Quirk, Keith, & Quirk (2001)	Singh (1998)	Singh & Billingsley (1998)	Singh, Granville, & Dika (2002)	Wang & Staver (2001)	Watkins (1997)	Wong & Watkins (1998)	Cheung, Hattie, & Ng (2001)	Schommer-Aikins et al. (2000)	Loadman et al. (1999)
Pre-analysis technical																
Sample size	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Missing data	N	Y	Y	Y	Y	N	Y	Y	N	Y	N	N	N	Y	Y	Y
Normality	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Outliers	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Linearity/multi-collinearity	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Software & estimation method	Y	Y	Y	Y	Y	Y	Y	Y, no est.	Y	Y	Y	N	Y, no est.	N	N	N
Postanalysis technical																
Assessment of fit	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Model chi-square	Y	Y	Y	N	Y	Y	Y	N	Y	Y	N	Y	N	Y	Y	N
Multiple fit indices	N	Y	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N
Parameters estimated & significant tests	N	Y	Y	Y	Y	N	N	Y, no sig. test	Y, no sig. test	Y	Y	N	N	N	N	N
Squared multiple correlation/ variance accounted for	Y	N	N	N	N	N	N	N	N	Y	N	N	N	Y	N	N

Standardized & unstandardized estimates	N	N	Y	Y	N	N	N	N	N	N	N	N	N	N	N	N
Residual analysis	N	Y	N	N	N	Y	N	Y	Y	Y	Y	Y	N	N	N	N
Correlation means and tables	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	N
Modifications	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	Y	N
Rationale for modifications	N	Y	N	N	N	N	N	N	Y	N	N	Y	N	Y	Y	N
Lagrange (adding paths)	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Wald (dropping paths)	N	Y	N	N	N	N	N	N	N	N	N	Y	N	N	N	N
Correlation between estimated parameters	N	Y	N	N	N	N	Y	N	N	N	N	N	N	N	N	N/A
Diagram of final model	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N
Research questions	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
CFA/SEM introduced	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	N	N	Y	N
Theoretical justification	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	Y	Y	Y	Y
Adequate tables and figures	N	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	N	Y	Y	N
Graphic display	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
Implications in line	Y	Y	Y	N	Y	N	N	N	N	N	N	Y	N	Y	N	Y

Note. Y = yes; N = no; CFA = confirmatory factor analysis; SEM = structural equation modeling.

TABLE 2. Cutoff Criteria for Several Fit Indexes

Indexes	Shorthand	General rule for acceptable fit if data are continuous	Categorical data
Absolute/predictive fit			
Chi-square	χ^2	Ratio of χ^2 to $df \leq$ 2 or 3, useful for nested models/model trimming	
Akaike information criterion	AIC	Smaller the better; good for model comparison (nonnested), not a single model	
Browne–Cudeck criterion	BCC	Smaller the better; good for model comparison, not a single model	
Bayes information criterion	BIC	Smaller the better; good for model comparison (nonnested), not a single model	
Consistent AIC	CAIC	Smaller the better; good for model comparison (nonnested), not a single model	
Expected cross-validation index	ECVI	Smaller the better; good for model comparison (nonnested), not a single model	
Comparative fit		Comparison to a baseline (independence) or other model	
Normed fit index	NFI	$\geq$ .95 for acceptance	
Incremental fit index	IFI	$\geq$ .95 for acceptance	
Tucker–Lewis index	TLI	$\geq$ .95 can be 0 > TLI > 1 for acceptance	0.96
Comparative fit index	CFI	$\geq$ .95 for acceptance	0.95
Relative noncentrality fit index	RNI	$\geq$ .95, similar to CFI but can be negative, therefore CFI better choice	
Parsimonious fit			
Parsimony-adjusted NFI	PNFI	Very sensitive to model size	
Parsimony-adjusted CFI	PCFI	Sensitive to model size	
Parsimony-adjusted GFI	PGFI	Closer to 1 the better, though typically lower than other indexes and sensitive to model size	
Other			
Goodness-of-fit index	GFI	$\geq$ .95 Not generally recommended	
Adjusted GFI	AGFI	$\geq$ .95 Performance poor in simulation studies	
Hoelter .05 index		Critical *N* largest sample size for accepting that model is correct	
Hoelter .01 index		Hoelter suggestion, $N = 200$, better for satisfactory fit	
Root mean square residual	RMR	Smaller, the better; 0 indicates perfect fit	
Standardized RMR	SRMR	$\leq$ .08	
Weighted root mean residual	WRMR	< .90	< .90
Root mean square error of approximation	RMSEA	< .06 to .08 with confidence interval	< .06

model in which this parameter is set free; the model is then re-estimated" (p. 57). Once modifications have been completed, one must realize that the analysis has moved from confirmatory to exploratory. Obviously, researchers often respecify their model when parameter estimates are statistically nonsignificant. That procedure typically improves the fit of the model to the data. But, again, we caution that it must make sense theoretically. As MacCallum and colleagues (1992) warned, "when an initial model fits well, it is probably unwise to modify it to achieve even better fit because modifications may simply be fitting small idiosyncratic characteristics of the sample" (p. 501).

Our concern with modification indexes along with any modification is the abuse that occurs. Researchers sometimes become fascinated with the fit indexes. The best description we have seen of this is in Ullman's (2001) Footnote 14, where she states that adding post-hoc paths is like eating salted peanuts: "One is never enough" (Ulman, 2001, p. 750).

Sometimes, multiple models are analyzed because the researcher is testing competing theoretical models. From an evaluation perspective, we determine which model fits the data best, but sometimes the differences between the models appear small on the basis of the fit indexes. When comparing nonnested models, the AIC fit index is a good choice because the difference in the chi-square values among the models cannot be interpreted as a test statistic (Kline, 2005).

Results From CFA Example

Muthén and Muthén (1998A) used the SEM software MPlus 2.0 to perform a CFA, based on data from 206 undergraduate students enrolled in a teacher-education course at a public, midsized university. We chose maximum likelihood estimation because our data were normally distributed. The data came from six questions on two Likert-scale surveys measuring epistemological beliefs (Schommer, 1998) and learning processes (Schmeck, Ribich, & Ramanaiah,

1977). A correlation table with means and standard deviations is shown in Table 3; the theoretical model is presented in Figure 2. We hypothesized a two-factor model to be confirmed in the measurement portion of the model. We evaluated the assumptions of multivariate normality and linearity through SPSS 11.0. Using box plots and Mahalanobis distance, we observed no univariate or multivariate outliers. We removed the data from the data set entered by 3 participants because they did not follow directions. The final sample size was 203; there were no missing data. The comparative fit index (CFI) = .99, the Tucker-Lewis fit index (TLI) = .98, and the RMSEA = .05. Those values indicate a good fit between the model and the observed data. Standardized parameter estimates are provided in Figure 5; unstandardized estimates are shown in Table 4.

The squared multiple correlation (SMC) values also are provided in italics and indicate (lower bound) the reliability of the measure; read critically (.96) and one meaning (.03) have the highest and lowest, respectively. An interpretation of the example is that the construct deep processing accounts for 23% of the variance in term papers. No post-hoc modifications were indicated from the analysis because of the good-fit indexes, and the residual analysis did not indicate any problems.

Results From SEM Example

Our hypothesized SEM is described graphically in Figure 3. We show the measurement component by using thin lines and the structural component by using bolded lines,

TABLE 3. Correlations for CFA and SEM Analyses

Observed variable	1	2	3	4	5	6	7
1. Compare/contrast	1	—	—	—	—	—	—
2. Term papers	0.42	1	—	—	—	—	—
3. Read critically	0.81	0.47	1	—	—	—	—
4. Specific facts	–0.02	0.03	–0.09	1	—	—	—
5. One answer	–0.17	0.00	–0.24	0.5	1	—	—
6. One meaning	–0.07	–0.10	–0.16	0.12	0.14	1	—
7. Achievement	0.22	0.19	0.29	–0.10	–0.30	–0.20	1

Note. This table is essentially the same for the structural equation modeling (SEM) example. The difference is 9 individuals. Also, the variables were standardized to have a mean of 0 and a standard deviation of 1. CFA = confirmatory factor analysis. $N = 203$; $M = 0$; $SD = 1$.

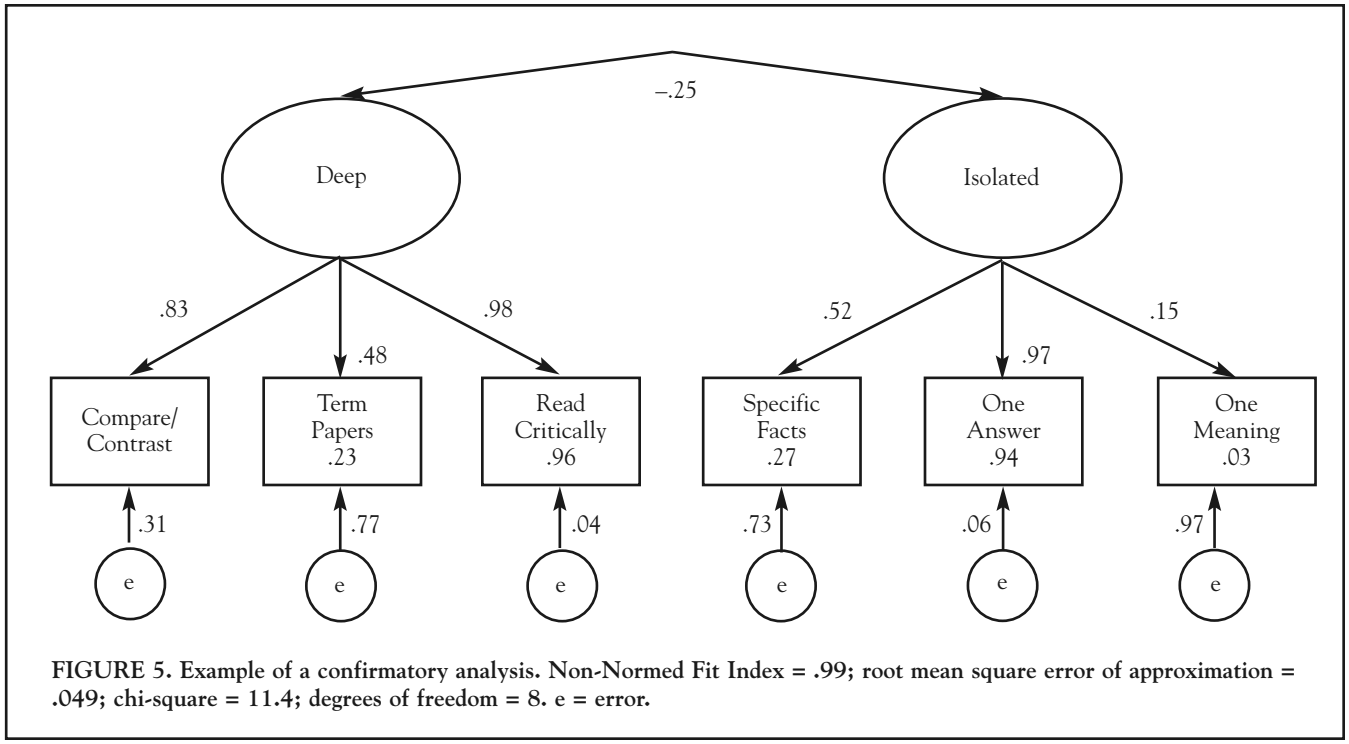

FIGURE 5. Example of a confirmatory analysis. Non-Normed Fit Index = .99; root mean square error of approximation = .049; chi-square = 11.4; degrees of freedom = 8. e = error.

for ease of distinguishing the components. We performed a SEM analysis based on data from 203 undergraduates at a midsized university with the AMOS 4.01 statistical package (Arbuckle, 1995–1999) on the six questions from two Likert-scale surveys measuring epistemological beliefs (Schommer, 1998) and learning processes (Schmeck et al. 1977). Circles represent latent variables and rectangles represent measure variables. A correlation table with means and standard deviations is shown in Table 3. We evaluated the assumptions of multivariate normality and linearity and observed nine multivariate outliers ($p < .001$). We removed the nine outliers from the subsequent analyses, leaving a final sample size of 194 (203 minus 9); there were no missing data. We chose maximum likelihood parameter estimation over other estimation methods (weighted least squares, two-stage least squares, asymptotically distribution-free [ADF]) because the data were distributed normally (Kline, 2005). (See Figure 6 and Table 5 for results.) The hypothesized model appears to be a good fit to the data. The CFI is .99; TLI is .98; and the RMSEA is .038. We did not conduct post-hoc modifications because of the good fit of the data to the model.

TABLE 4. Standardized and Unstandardized Coefficients for CFA Example

Observed variable	Latent construct	β	*B*	*SE*
Compare/contrast	Deep	0.83	1.00	
Term papers	Deep	0.48	0.58	0.08
Read critically	Deep	0.98	1.19	0.12
Specific facts	Isolated	0.52	1.00	
One answer	Isolated	0.97	1.76	0.68
One meaning	Isolated	0.15	0.29	0.14

Note. CFA = confirmatory factor analysis.

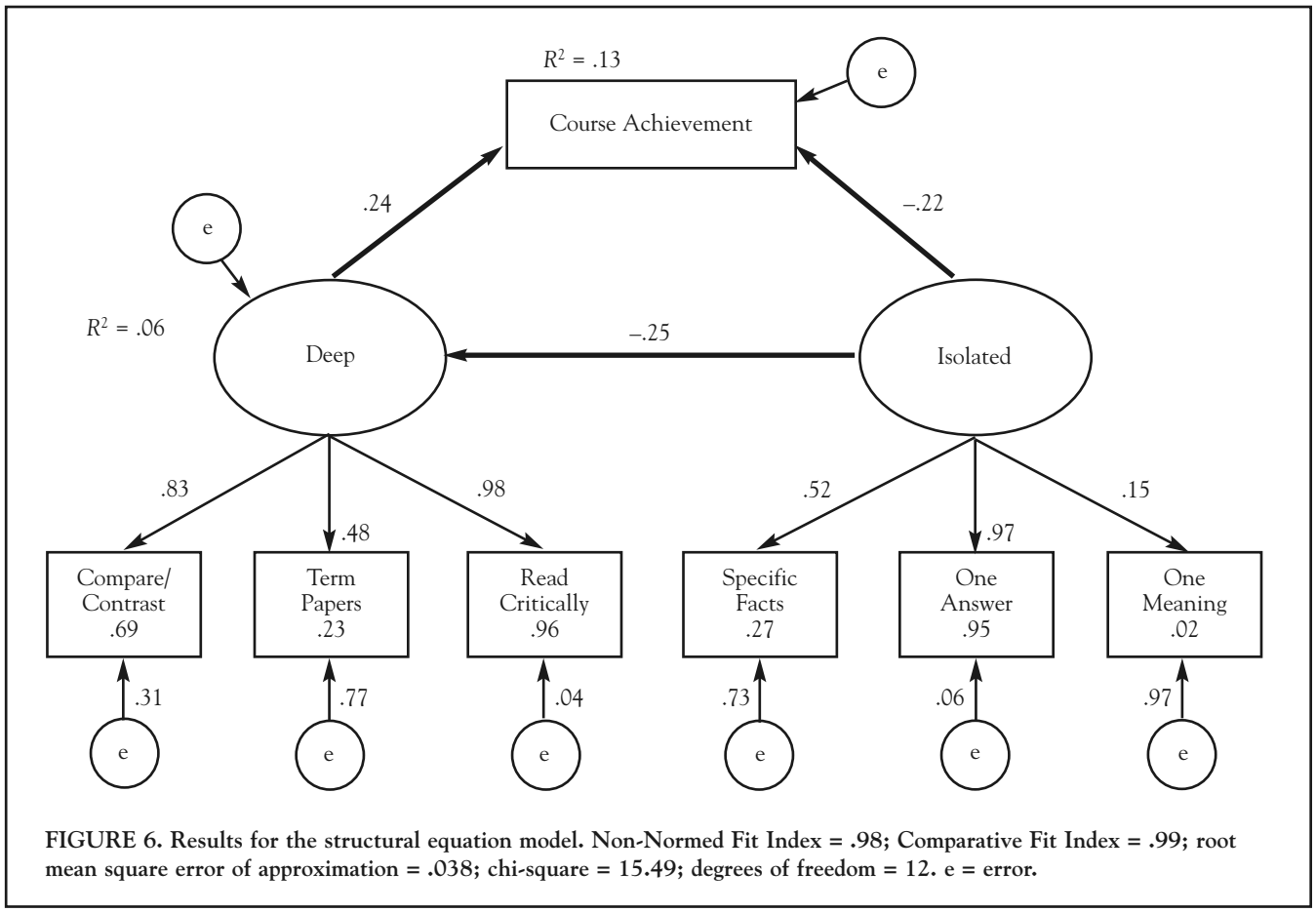

FIGURE 6. Results for the structural equation model. Non-Normed Fit Index = .98; Comparative Fit Index = .99; root mean square error of approximation = .038; chi-square = 15.49; degrees of freedom = 12. e = error.

TABLE 5. Results From Structural Equation Modeling Example

Model	β		*B*		*SE*	R^2
	Isolated	Deep	Isolated	Deep		
Direct						
Deep	–0.25		–0.47		.12	0.07
Achievement	–0.22	0.24	–0.42	0.24	.11	0.13
Indirect						
Deep	–0.06		–0.11			
Achievement						
Total						
Deep	–0.25		–0.47			
Achievement	–0.28	0.24	–0.54	0.24		

Direct Effects

Belief that knowledge is isolated facts (Isolated) was related negatively to deep processing (Deep) (standardized coefficient = –.25) and predictive of lower course achievement (standardized coefficient = –.22). Deep processing was predictive of greater course achievement (standardized coefficient = .24).

Indirect Effects

We hypothesized that the relationship between the belief that knowledge as isolated facts and course achievement was mediated has an indirect effect on course achievement, by deep processing. The result (standardized indirect coefficient = –.06, $p > .05$) was not statistically significant.

Obviously, in a journal article, research report, or dissertation more interpretation and discussion of the findings relative to other literature would be included in the Results sections. Here, we simply demonstrate what generally should be included for a reader to make a solid evaluative judgment on the merits of the analysis.

Structural Modeling: A Brief Review of Articles (1989–2004)

We used CFA and SEM in our review of 16 articles published in recent years in *The Journal of Educational Research* to assess the rigor of the application of the technique as used by researchers. The publication dates ranged from 1989–2002. The foci of the articles differed, but all addressed some form of attitude or behavior on the part of students at different levels of the K–16 continuum. For example, authors of one article examined students at the college level, whereas authors of all the other studies conducted their research on students at the secondary level or below. A variety of exogenous and endogenous variables was represented among the articles under investigation: motivation, general academic performance, performance in mathematics and science, academic engagement, cognitive structure professional support, employment, negative experiences in school, and self-regulation. The authors' objective in each study was to test the hypothesized quantitative model to capture the relationship among the variables specified in the model.

We used two sets of guidelines as rubrics in reviewing each article. The nontechnical rubric focused on reporting CFA and SEM to communicate effectively the responsible use of the technique and the full range of results necessary to verify the validity of models and individual parameters. We used the technical rubric to evaluate basic technical adequacy (see Table 1).

Research questions dictated the use of structural modeling. Assessing whether the research questions lent themselves to CFA and SEM was the first step in reviewing the articles because of the confirmatory nature of both methods. Most often, the criterion was readily satisfied when authors identified the relationships to be examined verbally or graphically, along with any mediating variables. Wang and Staver (2001) postulated direct, simultaneous influences of eight factors on a dependent variable, a relationship that could have been tested using a regression model. However, the authors used structural modeling appropriately in their study because of the multiple indicators for each of the latent constructs dictated by theoretical considerations. Herl, Baker, and Niemi (1996) examined how well different methods of measuring cognitive structure correspond. In their analysis, the researchers not only addressed the direct influence of different measures on a student's cognitive structure but also incorporated the possibility of correlated residuals for four of the variables in their model. Overall, specific research questions are preferred, but as models become more complicated, a diagram with expected relationships is an acceptable alternative.

CFA and SEM introduced before findings reported. Because of the complexity inherent with CFA and SEM techniques and the paucity of experience among many readers, authors should briefly introduce the combination of statistical procedures and should provide a justification for its use. Although Hong (1998) offered an informative and concise

explanation, most authors did not adequately introduce SEM to their readers. In one instance, SEM was explained as effective in examining nonexperimental data, but the authors gave no indication of what renders it effective. In another case, no introduction of the technique was provided; another article simply referred the reader to another publication for more in-depth details; and, finally, a third article offered a complete explanation of the approach but scattered the details in one-sentence fragments throughout the text. Given that pattern of weak general introductions of the technique, it was not surprising that we noted a corresponding absence of explanation that SEM generally entails two types of models. The pattern across the articles reviewed was clear. No explanation of the technique was provided, and a sound link of theory to the model proposed was lacking. With regard to CFA, it was not thoroughly introduced but simply mentioned that the technique would be used.

Sufficient theoretical justification provided. The studies revealed a trend that the theoretical discussion focused much more on the formation of constructs than on the configuration of the confirmatory or structural model. In some cases, the failure to detail the theoretical underpinning of the structural model was an indication that the examination of mediated relationships was not the primary interest of the researcher. Kaplan, Peck, and Kaplan (1997), for example, described their model as hypothesizing intervening variables between early negative academic experiences and later dropout behavior, but their "examination of the precise relationships among these variables was exploratory" (p. 338). SEM was deemed an appropriate technique for testing those relationships because the variables were measured at several points in time, but the model would be stronger if the underlying structural patterns among all latent variables were informed theoretically. When the theoretical framework is brief, it typically does not include an adequate discussion of the main theoretical constructs and their relationships. Often, the framework provides the appearance that the authors are "fishing" for statistically significant results. Hong's (1998) work did contain a complete conceptual explanation of the measurement model with a full presentation and discussion of the numerical results. Schommer-Aikins, Brookhart, and Hutter (2000) also provided a detailed historical and current theoretical model of the proposed CFA study.

Tables and figures—appropriate and sufficient. The inclusion of a graphic figure of at least one model in the articles presented was evident. For SEM, it is helpful to the reader to have a hypothesized model and a final model diagrammed. The hypothesized model in CFA usually can be provided in a table that displays the relationships between the observed and latent variables. The construction of the model varied somewhat in form from article to article. In one study, the hypothesized model was the only one found; all other studies advanced a final model marked with significant path coefficients.

All the articles included a correlation matrix; for the general reader, this may be the least useful information, but it is crucial for readers who wish to reanalyze the basic model presented. Because of the great deal of information generated through structural modeling, it is difficult to report everything. It is imperative that authors create concise tables or diagrams that contain key evaluative information, such as correlations and means and standard deviations, coefficients, fit indexes, and so forth. Although most researchers presented goodness-of-fit statistics in the quantitative model or in notes, Quirk, Keith, and Quirk (2001) presented a concise table with the goodness-of-fit results in an easy-to-evaluate form. The authors of only four of the articles that we examined presented direct, indirect, and total effects in an easy-to-read tabular form.

Implications in line with findings. Discussions centered on practice and policy were driven by the findings derived from the data analysis; however, at times we had difficulty assessing the appropriateness of those implications adequately without access to a full set of results. Reported path coefficients from one latent construct to another, the structural component, along with their corresponding implications, are difficult to evaluate without previous evidence of the validity of the latent measures used in the measurement model. Similarly, the relative importance of individual factors and their corresponding affects on one or more outcomes cannot be understood fully unless results have been reported in terms of direct, indirect, and total effects. In general, many of the topics provided in the discussions went beyond the conclusions reported in the analysis. Overgeneralizing has always been a problem, and we are as guilty as everyone else in this regard.

Sample size. Two issues that we found with sample size are (a) actual size of the sample and (b) missing data. Although the sample size needed is affected by the normality of the data and the estimation method that researchers use, the generally agreed-on value is 10 participants for every free parameter estimated. For example, Loadman, Freeman, Brookhart, Rahman, and McCague (1999), completed a CFA. On the basis of the text and the table, 51 free parameters would have been estimated; 45 for the factor loadings and 6 for the correlations among the latent factors. Using that rule, Loadman and colleagues would have needed 510 participants for this study—they had 1,687. That is a general rule, however, because as models become more complex or the data is more problematic, such as severe skewness, more data are needed. SEM is still a large sample analysis technique.

Although the problem of missing values is not unique to structural modeling, estimating a successful model necessitates the appropriate handling of missing data from a methodological, as well as conceptual, perspective. Reliance on pairwise deletion can result in a nonpositive covariance matrix, and other methods, including replacement with the mean, may result in heteroscedastic error (Schumaker & Lomax, 1996). It is important that the

researcher report the treatment of missing data so that results may be interpreted accordingly. Most of the studies reviewed did not address the issue of missing values or the way in which they were handled. Some studies did a pairwise or listwise deletion; in one study, authors described the percentage of missing data and used a mean imputation procedure. Missing data is a serious issue in SEM and must be discussed in any article. Also, given new technologies, more options can handle missing data, such as maximum likelihood estimation (Arbuckle, 1994–1999; Muthén & Muthén, 1998).

Basic assumptions. Essentially, authors provided no discussion concerning normality, outliers, linearity, or multicollinearity in the articles. About half of the articles reported the software used but not the version nor the estimation procedure.

Assessment of fit. Hong (1998) described the structure and goodness of fit of the initial measurement model, provided a description of, and theoretical justification for, changes in parameter constraints and presented the results of the final model. Results included chi-square and associated significance level, two goodness-of-fit indexes, plus factor correlations and standard residuals. Because the author provided comprehensive information, the reader can accept the judgment that the constructs used in the resulting structural model are sound and that the author's interpretation of results is appropriate. Authors of the remaining studies appear to have estimated a measurement model, but the results reported were insufficient to establish the validity of a set of multiple indicators.

All the articles reviewed provided goodness-of-fit indexes, although many simply gave numerical results, and few discussed with any clarity what standards were applied in determining a good fit. Wang and Staver (2001) discussed why several indexes were needed; Quirk and colleagues (2001) indicated the standards for a good fit in a Notes section; Singh, Granville, and Dika (2002) and Wong and Watkins (1998) listed the goodness-of-fit statistics in a table but gave no interpretation of their meaning; Singh and Billingsley (1998) and Kaplan, Peck, and Kaplan (1994) reported numbers and declared them indicative of a "reasonably good" (Singh & Billingsley, 1998) or "adequate" (Kaplan et al., 1994) fit. One disturbing aspect of a few studies includes the number of fit indexes below .90 that authors used to justify a good-fitting model. Even before Hu and Bentler's (1999) simulation work on continuous data and Yu's (2002) simulation work on categorical data, rules of thumb existed. The basic rule of thumb was that a fit index (e.g., IFI or NFI) had to be above .90. Some authors have questioned that level as the cutoff considered acceptable (Carlson & Mulaik, 1993). In this sample of articles, authors reported fit indexes as low as .85.

The validity of the final results of the structural model is dependent on capturing and establishing the reliability of the underlying constructs. The power of SEM is seen most fully when multiple indicators for each latent variable are first tested through CFA to establish the conceptual soundness of latent variables used in the final structural model. Without empirical evidence that such is the case, the relationships that the authors found significant in the structural model may be misleading. Singh and Billingsley (1998) were the only authors who mentioned unique variances or reliabilities of multiple indicators for latent constructs.[2]

Three of the reviewed studies represented the measurement and structural models in one quantitative model. Although authors in the three studies offered relatively complete tables and figures representing a combined measurement and structural model, they did not provide the information and discussion establishing the veracity of the latent variables. Factor loadings associated with observed variables are shown often in the models, but the unique coefficients (error) and the reliability of each observed variable are more often missing, as are the *t* values, or unstandardized coefficients with standard errors, for individual path coefficients for the latent constructs estimated in the model. In a longitudinal study examining the relationship between employment and academic performance (Quirk et al., 2001), family background is represented as a latent variable comprised of parents' highest level of education, parents' occupational status, and family income. Although the authors state "the latent variables are factors from a confirmatory factor analysis of the measured variables" (Quirk et al., p. 5), no information is proffered that establishes either the conceptual or statistical coherence of the three items forming a single construct. Wong and Watkins (1998) performed an exploratory and confirmatory factor analysis but did not report results to assure the reader of the validity of the constructs through a comparison of the two sets of values.

Although all studies gave some indication of the direct effects in the structural model, the format used to report results was inconsistent. Overall, several authors reported unstandardized and standardized coefficients along with standard errors or *t* values, others listed direct and indirect effects, and authors in four studies reported R^2 values for their endogenous variables. Most presentations and discussions focused on path coefficients, which provided accurate summary information. However, discussion of the results of a structural model is incomplete without consideration of indirect effects and the coefficients of determination (R^2) for each structural equation in the quantitative model. There was essentially no discussion of residual analysis.

Modifications

CFA and SEM can each be an iterative process by which modifications are indicated in the initial results, and parameter constraints altered to improve the fit of the model, if such changes are warranted theoretically. If a parameter is freed on the basis of a high modification index value, the researcher is called on to theoretically defend the change indicated so that the final model does not deviate from the initial theoretical model. Similarly, changes in parameter

constraints and the modification indexes indicating the changes should be reported.

Respecification of the structural model is driven most often by modification indexes, although authors in four of the studies reviewed did not discuss the reasons for changing the nature of the original hypotheses. Wang and Staver (2001) only briefly mentioned the total deletion of one latent construct simply because of "weak relationships" with other factors under examination. Although authors in four of the studies discussed altering their structural models, or parameters determined by comparison with alternate models, there was no mention of the conceptual or statistical standards by which changes were deemed appropriate. An assumption that all the authors shared was that the fit of the model was improved by adjusting the parameter constraints, but they did not express either the degree of improvement or any conceptual justification. Schommer and colleagues (2000) provided a theoretical rationale for model modification in the Discussion section of the article, but Hong (1998) appeared to provide the clearest delineation and justification of the modifications undertaken.

Other issues of concern included low-reliability values of latent variables according to the summation of several observed variable scores. Final sample sizes used in the analyses were not always clear. Specifically, we were concerned that, because of missing data, the sample size originally provided was not the exact sample size analyzed. None of the authors discussed very technical issues, such as the adequacy of the covariances or that the models were identified (Ullman, 2001). Finally, authors in four articles did mention performing cross-validation tests to examine the stability of the results of the model tested, but most authors did not discuss anything related to the stability of their results.

Summary

This article provides an introduction and description of CFA and SEM. Along with the Introduction, readers received a basic guideline for evaluating these types of articles. The guidelines included in this article could not cover every aspect of SEM because of its complexity, but they should provide a solid foundation for readers. We also hope that this review will help create a more consistent framework for authors who incorporate these techniques in the articles that are published in *The Journal of Educational Research*.

NOTES

In the discussion of CFA and SEM, several terms may be interpreted as synonymous when they are not: model, estimate(s), and predict or predictor. When we use the term model, we refer to the theoretical relationships among the observed and unobserved variables. At that point we may or may not know the precise relationship between the variables (i.e., path coefficient). The theoretical relationships are specified by mathematical models. The next step is to obtain estimates of the relationships among variables in the mathematics model; that is, we use a statistical principle, such as maximum likelihood to calculate the coefficients between the observed and unobserved variables. The values obtained from those calculations are known as parameter estimates.

1. For a full discussion of these and other indexes, see Hu and Bentler (1995) and Kaplan (2000). Hu and Bentler (1995) provide a detailed account of existing research on the behavior of all comparative fit indexes.

2. SEM estimates the degree to which a hypothesized model fits the data. In a CFA, goodness-of-fit indexes are estimated for each latent variable as a distinct structural model. Although it is wise and appropriate for one to measure items found in other studies to form a certain construct, it is not appropriate to assume that a certain group of items found to form a valid and reliable construct in another study will form an equally valid and reliable construct when measured in a different set of data. Similarly, constructs tested on a national data set are valid in a new study only in the rare instance when the new study uses the identical observations analysis in the same data with the same theoretical underpinning. Divergent choices addressing the problem of missing data will normally change construct validity results such that a new confirmatory analysis is appropriate.

REFERENCES

Arbuckle, J. L. (1994–1999). AMOS 4.01 [Software]. Chicago: SmallWaters.

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

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.

Carlson, M., & Mulaik, S. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research, 28*, 111–159.

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of U.S. history knowledge. *The Journal of Educational Research, 89*, 206–218.

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

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

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

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

Loadman, W. E., Freeman, D. J., Brookhart, S. M., Rahman, M. A., & McCague, G. J. (1999). Development of a national survey of teacher education program graduates. *The Journal of Educational Research, 93*, 76–82.

Long, J. S. (1983a). Confirmatory factor analysis: A preface to LISREL. Beverly Hills, CA: Sage.

Long, J. S. (1983b). Covariance structure models: An introduction to LISREL. Beverly Hills, CA: Sage.

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

Muthén, L, & Muthén, B. (1998). MPlus (Version 2.01) [Computer Software]. Los Angeles: Muthén & Muthén.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. New York: Holt, Rinehart and Winston, Inc.

Pohlmann, J. T. (2004). Use and interpretation of factor analysis in *The Journal of Educational Research*: 1992–2002. *The Journal of Educational Research, 98*, 14–23.

Quirk, K. J., Keith, T. Z., & Quirk, J. T. (2001). Employment during high school and student achievement: Longitudinal analysis of national data. *The Journal of Educational Research, 95*, 4–7.

Schmeck, R. R., Ribich, F. D., & Ramanaiah, N. (1977). Development of a self-report inventory for assessing individual differences in learning processes. *Applied Psychological Measurement, 1*, 413–431.

Schommer, M. (1998). The influence of age and education on epistemological beliefs. *The British Journal of Educational Psychology, 68*, 551–562.

Schommer-Aikins, M., Brookhart, S., & Hutter, R. (2000). Understanding middle students' beliefs about knowledge and learning using a multidimensional paradigm. *The Journal of Educational Research, 94*, 120–127.

Schumaker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.
Sivo, S. A., Xitao, F., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education, 74*, 267–288.
Stage, F. K. (1990). LISREL: An introduction and applications in higher education. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 427–466). New York: Agathon Press.
Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
Wang, J., & Staver, J. R. (2001). Examining relationships between factors of science education and student career aspiration. *The Journal of Educational Research, 94*, 312–319.
Wong, N., & Watkins, D. (1998). A longitudinal study of the psychosocial environmental and learning approaches in the Hong Kong classroom. *The Journal of Educational Research, 91*, 247–255.
Yu, C.-Y. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Unpublished dissertation. Retrieved January 5, 2005, from Mplus website http://www.statmodel.com/download/Yudissertation.pdf

LIST OF ARTICLES REVIEWED

Cheung, D., Hattie, J., & Ng, D. (2001). Reexamining the stages of concern questionnaire: A test of alternative models. *The Journal of Educational Research, 94*, 226–236.
Hong, E. (1998). Differential stability of state and trait self-regulation in academic performance. *The Journal of Educational Research, 91*, 148–158.
Kaplan, D. S., Liu, Xiaoru, & Kaplan, H. B. (2000). Family structure and parental involvement in the intergenerational parallelism of school adversity. *The Journal of Educational Research, 93*, 235–245.
Kaplan, D. S., Peck, B. M., & Kaplan, H. B. (1994). Structural relations model of self-rejection, disposition to deviance, and academic failure. *The Journal of Educational Research, 87*, 166–173.
Kaplan, D. S., Peck, B. M., & Kaplan, H. B. (1997). Decomposing the academic failure–dropout relationship: A longitudinal analysis. *The Journal of Educational Research, 90*, 331–343.
Keith, T., & Benson, M. J. (1992). Effects of manipulable influences on high school grades across five ethnic groups. *The Journal of Educational Research, 86*, 85–93.
Singh, K. (1998). Part-time employment in high school and its effect on academic achievement. *The Journal of Educational Research, 91*, 131–139.
Singh, K., & Billingsley, B. S. (1998). Professional support and its effects on teachers' commitment. *The Journal of Educational Research, 91*, 229–240.
Singh, K., Granville, M., & Dika, S. (2002). Mathematics and science achievement: Effects of motivation, interest, and academic engagement. *The Journal of Educational Research, 95*, 323–332.
Watkins, T. (1997). Teacher communications, child achievement, and parent traits in parent involvement models. *The Journal of Educational Research, 91*, 3–14.