32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 488

32

BEST PRACTICES IN STRUCTURAL EQUATION MODELING

RALPH O. MUELLER

GREGORY R. HANCOCK

Structural equation modeling (SEM) has evolved into a mature and popular methodology to investigate theory-derived structural/causal hypotheses. Indeed, with the continued development of SEM software packages such as AMOS (Arbuckle, 2007), EQS (Bentler, 2006), LISREL (Jöreskog & Sörbom, 2006), and Mplus (Muthén & Muthén, 2006), SEM "has become the preeminent multivariate method of data analysis" (Hershberger, 2003, pp. 43–44). Yet, we believe that many practitioners still have little, if any, formal SEM background, potentially leading to misapplications and publications of questionable utility. Drawing on our own experiences as authors and reviewers of SEM studies, as well as on existing guides for reporting SEM results (e.g., Boomsma, 2000; Hoyle & Panter, 1995; McDonald & Ho, 2002), we offer a collection of best practices guidelines to those analysts and authors who contemplate using SEM to help answer their substantive research questions. Throughout, we assume that readers have at least some familiarity with the goals and language of SEM as covered in any introductory textbook (e.g., Byrne, 1998, 2001, 2006; Kline, 2005; Loehlin, 2004; Mueller, 1996; Schumacker & Lomax, 2004). For those desiring even more in-depth or advanced knowledge, we recommend Bollen (1989), Kaplan (2000), or Hancock and Mueller (2006).

SETTING THE STAGE

The foundations of SEM are rooted in classical measured variable path analysis (e.g., Wright, 1918) and confirmatory factor analysis (e.g., Jöreskog, 1966, 1967). From a purely statistical perspective, traditional data analytical techniques such as the analysis of variance, the analysis of

Authors' Note: During the writing of this chapter, the first author was on sabbatical leave from The George Washington University and was partially supported by its Center for the Study of Language and Education and the Institute for Education Studies, both in the Graduate School of Education and Human Development. While on leave, he was visiting professor in the Department of Measurement, Statistics and Evaluation (EDMS) at the University of Maryland, College Park, and visiting scholar in its Center for Integrated Latent Variable Research (CILVR). He thanks the EDMS and CILVR faculty and staff for their hospitality, generosity, and collegiality. Portions of this chapter were adapted from a presentation by the authors at the 2004 meeting of the American Educational Research Association in San Diego.

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 489

covariance, multiple linear regression, canonical correlation, and exploratory factor analysis—as well as measured variable path and confirmatory factor analysis—can be regarded as special cases of SEM. However, classical path- and factor-analytic techniques have historically emphasized an explicit link to a theoretically conceptualized underlying causal model and hence are most strongly identified with the more general SEM framework. Simply put, SEM defines a set of data analysis tools that allows for the testing of theoretically derived and a priori specified causal hypotheses.

Many contemporary treatments introduce SEM not just as a statistical technique but as a *process* involving several stages: (a) initial model conceptualization, (b) parameter identification and estimation, (c) data-model fit assessment, and (d) potential model modification. As any study using SEM should address these four stages (e.g., Mueller, 1997), we provide brief descriptions here and subsequently use them as a framework for our best practices analysis illustrations and publication guidelines.

Initial Model Conceptualization

The first stage of any SEM analysis should consist of developing a thorough understanding of, and justification for, the underlying theory or theories that gave rise to the particular model(s) being investigated. In most of the traditional and typical SEM applications, the operationalized theories assume one of three forms:

- A *measured variable path analysis* (MVPA) model: hypothesized structural/causal relations among directly measured variables; the four-stage SEM process applied to MVPA models was illustrated in, for example, Hancock & Mueller, 2004.
- A *confirmatory factor analysis* (CFA) model: structural/causal relations between unobserved latent factors and their measured indicators; the four-stage SEM process applied to CFA models was illustrated in, for example, Mueller & Hancock, 2001.
- A *latent variable path analysis* (LVPA) model: structural/causal relations among latent factors. This type of SEM model is the focus in this chapter and constitutes a combination of the previous two. A distinction is made between the structural and the measurement portions of the model: While the former is concerned with causal relations among latent constructs and typically is the focus in LVPA studies, the latter specifies how these constructs are modeled using measured indicator variables (i.e., a CFA model).

More complex models (e.g., multisample, latent means, latent growth, multilevel, or mixture models) with their own specific recommendations certainly exist but are beyond the present scope. Regardless of model type, however, a lack of consonance between model and underlying theory will have negative repercussions for the entire SEM process. Hence, meticulous attention to theoretical detail cannot be overemphasized.

Parameter Identification and Estimation

A model's hypothesized structural and nonstructural relations can be expressed as population parameters that convey both magnitude and sign of those relations. Before sample estimates of these parameters can be obtained, each parameter—and hence the whole model—must be shown to be *identified;* that is, it must be possible to express each parameter as a function of the variances and covariances of the measured variables. Even though this is difficult and cumbersome to demonstrate, fortunately, the identification status of a model can often be assessed by comparing the total number of parameters to be estimated, t, with the number of unique (co)variances of measured variables,

$$u = p\left(p+1\right)/2,$$

where p is the total number of measured variables in the model. When $t > u$ (i.e., when attempting to estimate more parameters than there are unique variances and covariances), the model is *underidentified,* and estimation of some (if not all) parameters is impossible. On the other hand, $t \leq u$ is a necessary but not sufficient condition for identification, and usually parameter estimation can commence: $t = u$ implies that the model is *justidentified,* while $t < u$ implies that it is *overidentified* (provided that indeed all parameters are identified and any latent variables in the system have been assigned an appropriate metric; see Note 4).

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 490

SEM software packages offer a variety of parameter estimation techniques for models whose identification can be established. The most popular estimation method (and the default in most SEM software packages) is *maximum likelihood* (ML), an iterative large-sample technique that assumes underlying multivariate normality. Alternative techniques exist (e.g., generalized least squares [GLS], asymptotically distribution free [ADF; Browne, 1984], and robust estimators [Satorra & Bentler, 1994]), some of which do not depend on a particular underlying distribution of the data, but still, the vast majority of substantive studies use ML.

Data-Model Fit Assessment

A central issue addressed by SEM is how to assess the fit between observed data and the hypothesized model, ideally operationalized as an evaluation of the degree of discrepancy between the true population covariance matrix and that implied by the model's structural and nonstructural parameters. As the population parameter values are seldom known, the difference between an *observed,* sample-based covariance matrix and that *implied* by parameter estimates must serve to approximate the population discrepancy. For a justidentified model, the observed data will fit the model perfectly: The system of equations expressing each model parameter as a function of the observed (co)variances is uniquely solvable; thus, the sample estimate of the model-implied covariance matrix will, by default, equal the sample estimate of the population covariance matrix. However, if a model is overidentified, it is unlikely that these two matrices are equal as the system of equations (expressing model parameters as functions of observed variances and covariances) is solvable in more than a single way.

Abiding by a general desire for parsimony, overidentified models tend to be of more substantive interest than justidentified ones because they represent simpler potential explanations of the observed associations. While data-model fit for such models was initially conceived as a formal statistical test of the discrepancy between the true and model-implied covariance matrices (a chi-square test with $df = u–t$; Jöreskog, 1966, 1967), such a test now is often viewed as overly strict given its power to detect even trivial deviations of a proposed model from reality. Hence, many alternative assessment strategies have emerged (for a now classic review, see Tanaka, 1993) and continue to be developed. Data-model fit indices for such assessment can be categorized roughly into three broad classes (with recommended indices in italics):

- *Absolute indices* evaluate the overall discrepancy between observed and implied covariance matrices; fit improves as more parameters are added to the model and degrees of freedom decrease: for example, the *standardized root mean square residual* (*SRMR*), the *chi-square test* (recommended to be reported mostly for its historical significance), and the goodness-of-fit index (GFI).

- *Parsimonious indices* evaluate the overall discrepancy between observed and implied covariance matrices while taking into account a model's complexity; fit improves as more parameters are added to the model, as long as those parameters are making a useful contribution: for example, the *root mean square error of approximation* (*RMSEA*) with its associated confidence interval, the *Akaike information criterion* (*AIC*) for fit comparisons across nonnested models, and the adjusted goodness-of-fit index (AGFI).

- *Incremental indices* assess absolute or parsimonious fit relative to a baseline model, usually the null model (a model that specifies no relations among measured variables): for example, the *comparative fit index* (*CFI*), the normed fit index (NFI), and the nonnormed fit index (NNFI).

If, after considering several indices, data-model fit is deemed acceptable (and judged best compared to competing models, if applicable), the model is retained as tenable, and individual parameters may be interpreted. If, however, evidence suggests unacceptable data-model fit, the next and often final stage in the SEM process is considered: modifying the model to improve fit in hopes of also improving the model's correspondence to reality.

Potential Model Modification

In a strict sense, *any* hypothesized model is, at best, only an approximation to reality; the remaining question is one of degree of that misspecification. With regard to external specification errors—when irrelevant variables were included in the model or substantively important ones were left out—remediation can only occur by respecifying

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 491

the model based on more relevant theory. On the other hand, internal specification errors—when unimportant paths among variables were included or when important paths were omitted—can potentially be diagnosed and remedied using *Wald statistics* (predicted increase in chi-square if a previously estimated parameter were fixed to some known value, e.g., zero) and *Lagrange multiplier statistics* (also referred to as *modification indices;* estimated decrease in chi-square if a previously fixed parameter were to be estimated). As these tests' recommendations are directly motivated by the data and not by theoretical considerations, any resulting respecifications must be viewed as exploratory in nature and might not lead to a model that resembles reality any more closely than the one(s) initially conceptualized.

Best Practices in SEM Data Analysis: A Set of Illustrations

Using the four-stage SEM process as a framework, we turn to an illustration of best practices in the most common type of SEM analyses. We chose to focus on a set of hypothesized models involving structural/causal relations among latent factors (i.e., LVPA models) to demonstrate our preference for using a two-phase approach (i.e., a measurement phase followed by a structural phase) over a single-phase, all-in-one analysis. We conclude this section by illustrating the statistical comparison of hierarchically related or *nested* models (occurring, for example, when one model's parameters are a proper subset of another model's parameters) and addressing the disattenuation (i.e., purification and strengthening) of structural parameter estimates obtained from an LVPA when compared with those obtained from an analysis of the same overall structure but one that uses measured variables only.

Suppose an educational researcher is interested in investigating the structural effects of girls' reading and mathematics self-concept (Read-SC and Math-SC, respectively) on mathematics proficiency (Math-Prof), as potentially mediated by task-goal orientation (Task-Goal). More specifically, the investigator might have strong theoretical reasons to believe that at least one of three scenarios is tenable: In Model 1 (Figure 32.1a), it is hypothesized that the effects of Read-SC and Math-SC on Math-Prof are both completely mediated by Task-Goal. In Model 2 (Figure 32.1b), only the effect of Read-SC on Math-Prof is completely mediated by Task-Goal, while Math-SC affects Math-Prof not only indirectly via Task-Goal but also directly without other intervening variables. Finally, in Model 3 (Figure 32.1c), Read-SC and Math-SC are thought to affect Math-Prof directly as well as indirectly via Task-Goal. To illustrate the testing of the tenability of these three competing models, multivariate normal data on three indicator variables for each of the four constructs were simulated for a sample of $n = 1{,}000$ ninth-grade girls. Table 32.1 describes the 12 indicator variables in more detail, while Table 32.2 contains relevant summary statistics.[1]

At this point, it is possible and might seem entirely appropriate to address the research questions implied by the hypothesized models through a series of multiple linear regression (MLR) analyses. For example, for Model 2 in Figure 32.1b, two separate regressions could be conducted: (1) An appropriate surrogate measure of Math-Prof could be regressed on proxy variables for Math-SC and Task-Goal, and (2) a suitable indicator of Task-Goal could be regressed on proxies for Read-SC and Math-SC. If the researcher would choose items ReadSC3, MathSC3, TG1, and Proc from Table 32.1 as surrogates for their respective constructs, MLR results would indicate that even though all hypothesized effects are statistically significantly different from zero, only small amounts of variance in the dependent variables TG1 and Proc are explained by their respective predictor variables ($R^2_{TG1} = 0.034$, $R^2_{Proc} = 0.26$; see Table 32.6 for the unstandardized and standardized regression coefficients obtained from the two MLR analyses[2]). As we will show through the course of the illustrations below, an appropriately conducted LVPA of the models in Figure 32.1 and the data in Table 32.2 will greatly enhance the utility of the data to extract more meaningful results that address the researcher's key questions.

SEM Notation

As the three alternative structural models depicted in Figure 32.1 are at the theoretical/latent construct level, we followed common practice and enclosed the four factors of Read-SC, Math-SC, Task-Goal, and Math-Prof in ellipses/circles. On the other hand, a glance ahead at the operationalized model in Figure 32.2 reveals that the now included measured variables

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 492

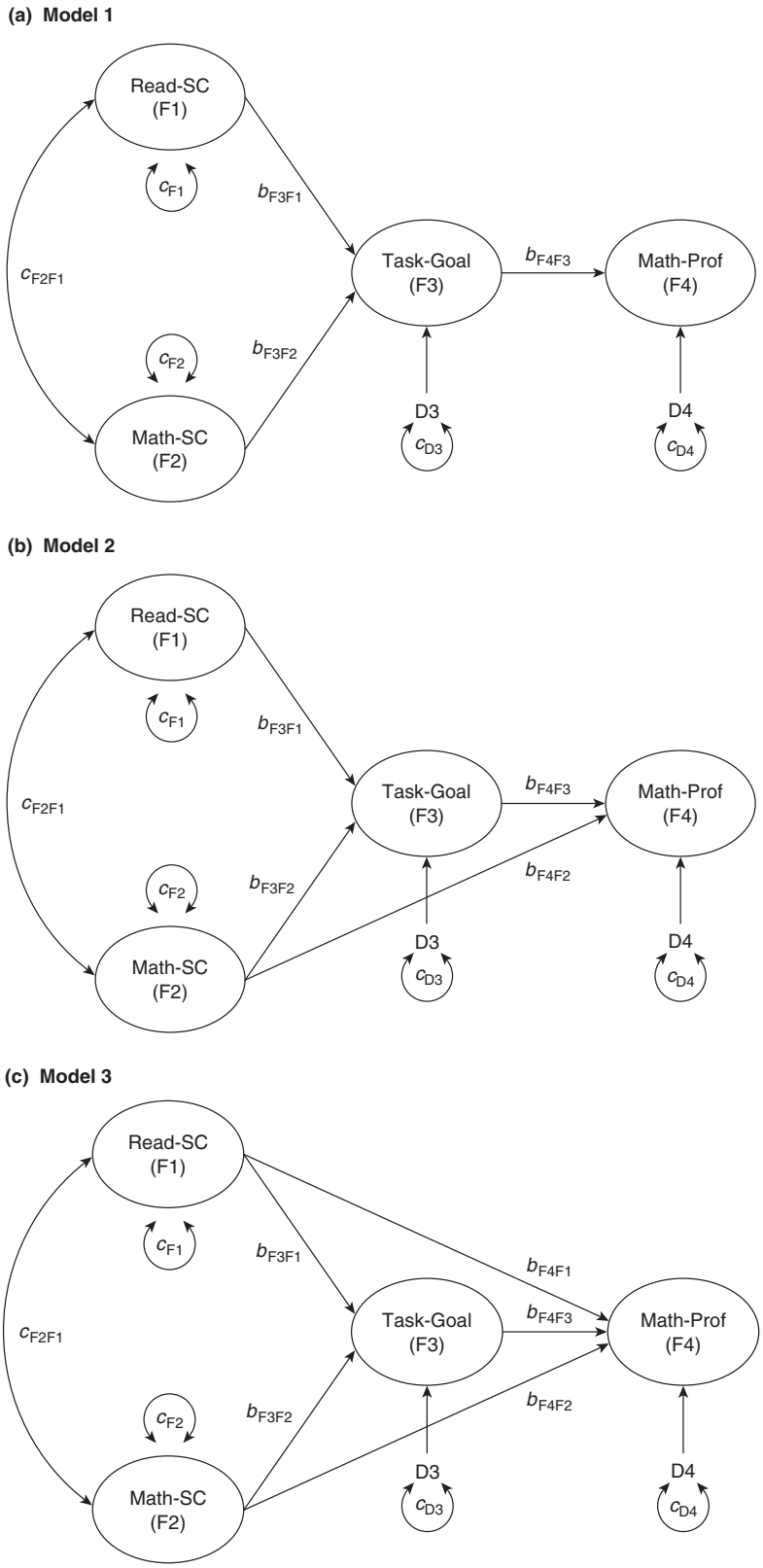

Figure 32.1 The theoretical models.

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 493

Table 32.1 Indicator Variable/Item Description

Construct	*Variable Label*	*Item*	*Scores*
Read-SC (F1)			
	RSC1 (V1)	"Compared to others my age, I am good at reading."	1 (*false*) to 6 (*true*)
	RSC2 (V2)	"I get good grades in reading."	
	RSC3 (V3)	"Work in reading class is easy for me."	
Math-SC (F2)			
	MSC1 (V4)	"Compared to others my age, I am good at math."	1 (*false*) to 6 (*true*)
	MSC2 (V5)	"I get good grades in math."	
	MSC3 (V6)	"Work in math class is easy for me."	
Task-Goal (F3)			
	TG1 (V7)	"I like school work that I'll learn from, even if I make a lot of mistakes."	1 (*false*) to 6 (*true*)
	TG2 (V8)	"An important reason why I do my school work is because I like to learn new things."	
	TG3 (V9)	"I like school work best when it really makes me think."	
Math-Prof (F4)			
	Math (V10)		Mathematics subtest scores of the Stanford Achievement Test 9
	Prob (V11)		Problem Solving subtest scores of the Stanford Achievement Test 9
	Proc (V12)		Procedure subtest scores of the Stanford Achievement Test 9

(items RSC1 to RSC3, MSC1 to MSC3, TG1 to TG3, Math, Prob, and Proc) are enclosed in rectangles/squares. Using the Bentler-Weeks "VFED" labeling convention (V for measured *V*ariable/item, F for latent *F*actor/construct, E for *E*rror/measured variable residual, D for *D*isturbance/latent factor residual), the latent and measured variables in the current models are labeled F1 through F4 and V1 through V12, respectively. The hypothesized presence or absence of relations between variables in the model is indicated by the presence or absence of arrows in the corresponding path diagram: One-headed arrows signify direct structural or causal effects hypothesized from one variable to another, while two-headed arrows denote hypothesized covariation and variation without structural specificity. For example, for Model 1 in Figure 32.1a, note (a) the hypothesized covariance between Read-SC and Math-SC and the constructs' depicted variances (two-headed arrows connect the factors to each other and to themselves, given that a variable's variance can be thought of as a covariance of the variable with itself), (b) the hypothesized structural effects of these two factors on Task-Goal (one-headed arrows lead from both to Task-Goal), but (c) the absence of such hypothesized direct effects on Math-Prof (there are no one-headed arrows directly leading from Read-SC and Math-SC to Math-Prof; the former two constructs are hypothesized to affect the latter only

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 494

Table 32.2 Correlations and Standard Deviations of Simulated Data

	READSC1 (V1)	*READSC2 (V2)*	*READSC3 (V3)*	*MATHSC1 (V4)*	*MATHSC2 (V5)*	*MATHSC3 (V6)*
READSC1	1.000					
READSC2	0.499	1.000				
READSC3	0.398	0.483	1.000			
MATHSC1	0.206	–0.148	–0.123	1.000		
MATHSC2	–0.150	0.244	–0.095	0.668	1.000	
MATHSC3	–0.121	–0.091	0.308	0.633	0.641	1.000
GOALS1	0.141	0.150	0.123	0.140	0.143	0.167
GOALS2	0.123	0.151	0.134	0.163	0.180	0.145
GOALS3	0.161	0.199	0.160	0.147	0.151	0.158
SATMATH	–0.049	–0.007	0.003	0.556	0.539	0.521
SATPROB	–0.031	–0.009	0.023	0.544	0.505	0.472
SATPROC	–0.025	–0.029	0.006	0.513	0.483	0.480
SD	1.273	1.353	1.285	1.396	1.308	1.300
	GOALS1 (V7)	*GOALS2 (V8)*	*GOALS3 (V9)*	*SATMATH (V10)*	*SATPROB (V11)*	*SATPROC (V12)*
READSC1						
READSC2						
READSC3						
MATHSC1						
MATHSC2						
MATHSC3						
GOALS1	1.000					
GOALS2	0.499	1.000				
GOALS3	0.433	0.514	1.000			
SATMATH	0.345	0.385	0.337	1.000		
SATPROB	0.304	0.359	0.281	0.738	1.000	
SATPROC	0.259	0.330	0.279	0.714	0.645	1.000
SD	1.334	1.277	1.265	37.087	37.325	45.098

indirectly, mediated by Task-Goal). Finally, because variation in dependent variables usually is not fully explainable by the amount of variation or covariation in their specified causes, each dependent variable has an associated residual term. For example, in the operationalized model in Figure 32.2, D3 and D4 denote the prediction errors associated with the latent factors F3 (Task-Goal) and F4 (Math-Prof), while E1 through E3 indicate the residuals associated with the measured indicator variables (V1 to V3) of the latent construct Read-SC.

For purposes of labeling structural and non-structural parameters associated with the connections between measured and/or latent variables in a path diagram, we used the *abc* system[3] (Hancock & Mueller, 2006, pp. 4–6). Structural effects from one variable (measured or latent) to another are labeled $b_{\text{to, from}}$, with the subscripts indicating the *to* and *from* variables (e.g., in Figure 32.2, b_{F3F1} indicates the path

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 495

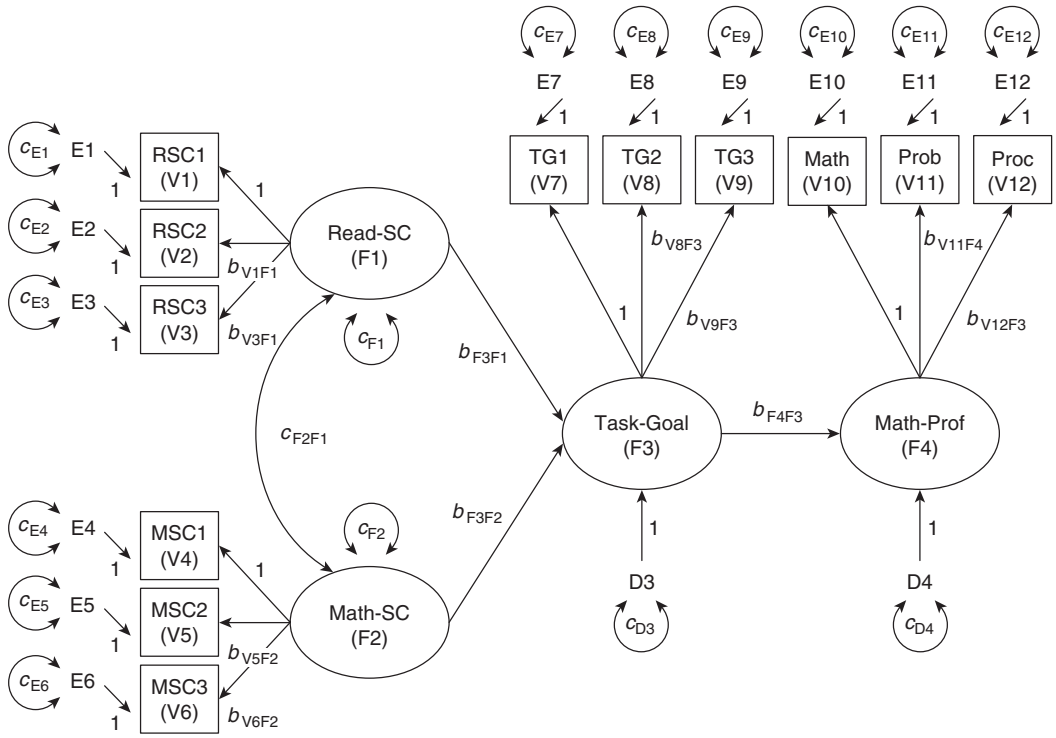

Figure 32.2 Initially operationalized Model 1.

to F3 from F1, and b_{V2F1} denotes the path/factor loading to item V2 from factor F1). On the other hand, variances and covariances are labeled by the letter *c* (e.g., in Figure 32.2, c_{F1} denotes the variance of the latent construct F1, while c_{F2F1} represents the covariance between the factors F2 and F1).

All-in-One SEM Analysis—Generally Not Recommended

Although we generally do not recommend the analytic strategy outlined in this section, it nevertheless will prove pedagogically instructive and will motivate arguments in later sections. With a hypothesized structure among latent constructs in place and associated measured indicators selected, Model 1 in Figure 32.1a can be operationalized as illustrated in Figure 32.2. This path diagram implies a set of 14 structural equations, one for each dependent variable: two equations from the *structural portion* of the model (i.e., the part that specifies the causal structure among latent constructs) and 12 equations from the *measurement portion* of the model (i.e., the part that links each of the indicator variables with the designated latent constructs). Table 32.3 lists all 14 structural equations and their associated *endogenous* (dependent) and *exogenous* (independent) variables that together specify the model in Figure 32.2 (variables are assumed to be mean-centered, thus eliminating the need for intercept terms; items V1, V4, V7, and V10 are used as reference variables for their respective factors, and thus their factor loadings are not *free* to be estimated but *fixed* to 1.0; also see Note 4).

Though it might seem that the statistical estimation of the unknown coefficients in the structural equations (the *b* and *c* parameters) should be the focus at this stage of the analysis, a prior assessment of the data-model fit is more essential as it allows for an overall judgment about whether the data fit the structure as hypothesized (indeed, should evidence materialize that the data do not fit the model, interpretations of individual parameter estimates might be useless). As can be verified from the path diagram in Figure 32.2 by counting one- and

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 496

Table 32.3 Structural Equations Implied by the Path Diagram in Figure 32.2

Structural Portion		
Endogenous Variable	*Structural Equations*	*Exogenous Variables*[a]
Task-Goal (F3)	F3 = b_{F3F1} F1 + b_{F3F2} F2 + D3	Read-SC (F1) Math-SC (F2)
Math-Prof (F4)	F4 = b_{F4F3} F3 + D4	Task-Goal (F3)
Measurement Portion		
Endogenous Variable	*Structural Equations*	*Exogenous Variables*[b]
RSC1 (V1)	V1 = (1)F1 + E1	Read-SC (F1)
RSC2 (V2)	V2 = b_{V2F1} F1 + E2	
RSC3 (V3)	V3 = b_{V3F1} F1 + E3	
MSC1 (V4)	V4 = (1)F2 + E4	Math-SC (F2)
MSC2 (V5)	V5 = b_{V5F2} F2 + E5	
MSC3 (V6)	V6 = b_{V6F2} F2 + E6	
TG1 (V7)	V7 = (1)F3 + E7	Task-Goal (F3)
TG2 (V8)	V8 = b_{V8F3} F3 + E8	
TG3 (V9)	V9 = b_{V9F3} F3 + E9	
Math (V10)	V10 = (1)F4 + E10	Math-Prof (F4)
Prob (V11)	V11 = b_{V11F4} F4 + E11	
Proc (V12)	V12 = b_{V12F4} F4 + E12	

a. Residuals D, though technically independent, are not listed.

b. Residuals E, though technically independent, are not listed.

two-headed arrows labeled with *b* or *c* symbols, the model contains $t = 28$ parameters to be estimated:[4] two variances of the independent latent constructs and one covariance between them, two variances of residuals associated with the two dependent latent constructs, three path coefficients relating the latent constructs, eight factor loadings, and 12 variances of residuals associated with the measured variables. Furthermore, the 12 measured variables in the model produce $u = 12\ (12 + 1)/2 = 78$ unique variances and covariances; the model is overidentified ($t = 28 < u = 78$), and it is likely that some degree of data-model misfit exists (i.e., the observed covariance matrix will likely differ, to some degree, from that implied by the model). To assess the degree of data-model misfit, various fit indices can be obtained and then should be compared against established cutoff criteria (e.g., those empirically derived by Hu & Bentler, 1999, and listed here in Table 32.4). Though here LISREL 8.8 (Jöreskog & Sörbom, 2006) was employed, running any of the available SEM software packages will verify the following data-model fit results for the data in Table 32.2 and the model in Figure 32.2 (because the data are assumed multivariate normal, the maximum likelihood estimation method was used): $\chi^2 = 3624.59$ ($df = u - t = 50$, $p < .001$), SRMR = 0.13, RMSEA = 0.20 with CI_{90}: (0.19, 0.20), and CFI = 0.55.

As is evident from comparing these results with the desired values in Table 32.4, the current data do not fit the proposed model; thus, it is not appropriate to interpret any individual parameter estimates as, on the whole, the model in Figure 32.2 should be rejected based on the current data. Now the researcher is faced with the question of what went wrong: (a) Is the source of the data-model misfit indeed primarily a flaw in the underlying structural theory (Figure 32.1a), (b) can the misfit be attributed to misspecifications in the measurement portion of the model with the hypothesized structure among latent

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 497

Table 32.4 Target Values for Selected Fit Indices to Retain a Model by Class

Index Class		
Incremental	*Absolute*	*Parsimonious*
NFI ≥ 0.90		
NNFI ≥ 0.95	GFI ≥ 0.90	AGFI ≥ 0.90
CFI ≥ 0.95	SRMR ≤ 0.08	RMSEA ≤ 0.06
Joint Criteria		
NNFI, CFI ≥ 0.96 and SRMR ≤ 0.09		
	SRMR ≤ 0.09 and RMSEA ≤ 0.06	

SOURCE: Partially taken from Hu and Bentler (1999).

NOTE: CFI = comparative fit index; NFI = normed fit index; NNFI = nonnormed fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

constructs actually having been specified correctly, or (c) do misspecifications exist in both the measurement and structural portions of the model? To help address these questions and prevent potential confusion about the source of observed data-model misfit, we do *not* recommend that researchers conduct SEM analyses by initially analyzing the structural and measurement portions of their model simultaneously, as was done here. Instead, analysts are urged to follow a two-phase analysis process, as described next.

Two-Phase SEM Analysis—Recommended

Usually, the primary reason for conceptualizing LVPA models is to investigate the tenability of theoretical causal structures among latent variables. The main motivation for recommending a two-phase process over an all-in-one approach is to initially separate a model into its measurement and structural portions so that misspecifications in the former, if present, can be realized and addressed first, before the structure among latent constructs is assessed.[5] This approach will simplify the identification of sources of data-model misfit and might also aid in the prevention of nonconvergence problems with SEM software (i.e., when the iterative estimation algorithm cannot converge upon a viable solution for parameter estimates).

Consider the path diagram in Figure 32.3a. It is similar to the one depicted in Figure 32.2 as it involves the same measured and latent variables but differs in two important ways: Not only are Read-SC and Math-SC now explicitly connected to Math-Prof, but all structural links among latent variables have been changed to nonstructural relations (note in Figure 32.3a the two-headed arrows between all latent constructs that are now labeled with *c* symbols). That is, latent constructs are allowed to freely covary without an explicit causal structure among them. In short, Figure 32.3a represents a CFA model of the latent factors Read-SC, Math-SC, Task-Goal, and Math-Prof, using the measured variables in Table 32.1 as their respective effect indicators.[6]

Measurement Phase. An analysis of the CFA model in Figure 32.3a constitutes the beginning of the *measurement phase* of the proposed two-phase analysis process and produced the following data-model fit results: $\chi^2 = 3137.16$ ($df = 48$, $p < .001$), SRMR = 0.062, RMSEA = 0.18 with CI_{90}: (0.17, 0.18), and CFI = 0.61. These values signify a slight improvement over fit results for the model in Figure 32.2. To the experienced modeler, this improvement was predictable given that the model in Figure 32.2 is more restrictive than, and a special case of, the CFA model in Figure 32.3a (with the paths from Read-SC and Math-SC to Math-Prof fixed to zero); that is, the former model is *nested* within the latter, a topic more fully discussed in the next section. Irrespective of this minor improvement, however, why did the data-model fit remain unsatisfactory (as judged by the criteria listed in Table 32.4)? Beginning to analyze and address this misfit constitutes a move toward the fourth and final phase in the general SEM process, potential post hoc model modification.

First, reconsider the list of items in Table 32.1. While all variables certainly seem to "belong" to the latent factors they were selected to indicate, note that for the reading and mathematics self-concept factors, corresponding items are identical except for one word: The word *reading* in items RSC1 through RSC3 was replaced by the word *math* to obtain items MSC1 through MSC3. Thus, it seems plausible that individuals' responses to corresponding reading and mathematics self-concept items are influenced by some of the same or related causes. In fact, the model in Figure 32.3a explicitly posits that two such related causes are

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 498

the latent constructs Read-SC and Math-SC. However, as the specification of residual terms (E) indicates, responses to items are influenced by causes other than the hypothesized latent constructs. Those other, unspecified causes could also be associated. Thus, for example, the residual terms E1 and E4 might covary to some degree, particularly since both are associated with items that differ by just one word. Based on similar theoretical reasoning, a nonzero covariance might exist between E2 and E5 and also between E3 and E6. In sum, it seems theoretically justifiable to modify the CFA model in Figure 32.3a to allow residual terms of corresponding reading and mathematics self-concept items to freely covary, as shown in Figure 32.3b. In fact, with enough foresight, these covariances probably should have been included in the initially hypothesized model.

Second, as part of the analysis of the initial CFA model in Figure 32.3a, Lagrange multiplier (LM) statistics may be consulted for empirically based model modification suggestions. These statistics estimate the potential improvement in data-model fit (as measured by the estimated decrease in chi-square) if a previously fixed parameter were to be estimated. Here, the three largest LM statistics were 652.0, 567.8, and 541.7, associated with the fixed parameters c_{E5E2}, c_{E4E1}, and c_{E6E3}, respectively. Compared with the overall chi-square value of 3137.16, these estimated chi-square decreases seem substantial,[7] foreshadowing a statistically significant improvement in data-model fit. Indeed, after respecifying the model accordingly (i.e., freeing c_{E5E2}, c_{E4E1}, and c_{E6E3}; see Figure 32.3b) and reanalyzing the data, fit results for the modified CFA model improved dramatically: $\chi^2 = 108.04$ ($df = 45$, $p < .001$), SRMR = 0.018, RMSEA = 0.037 with CI_{90}: (0.028, 0.046), and CFI = 0.99.

Though the degree of improvement might have been a pleasant surprise, *some* improvement was again to be expected: When compared with the initial CFA model, the modified model places fewer restrictions on parameter values (hence, better data-model fit) by allowing three error covariances to be freely estimated. Once again, the two CFA models are nested, and their fit could be statistically compared with a chi-square difference test, as discussed in the next section. For now, suffice it to say that an informal, descriptive comparison of the fit results for the initial and modified CFA models (e.g., a decrease from $\chi^2_{initial} = 3137.16$ to $\chi^2_{mod} = 108.04$, a drop from $SRMR_{initial} = 0.062$ to $SRMR_{mod} = 0.018$, and a reduction from $RMSEA_{initial} = 0.18$ to $RMSEA_{mod} = 0.037$) seems to indicate that the data fit the modified model much better. In more absolute terms, comparing the fit results from the modified model to the target values in Table 32.4 suggests that the data fit the model very well (due to the large sample size of $n = 1{,}000$, relatively little weight should be placed on the still significant $\chi^2_{(45)} = 108.04$). An examination of the remaining modification indices indicated that, even though further modifications would continue to slightly improve fit, none of the suggested modifications were theoretically justifiable (e.g., the largest estimated drop in chi-square, 23.0, could be obtained by freeing c_{E8E2}, the error covariance associated with the items TG2 and RSC2). Thus, no further modifications to the measurement model seem warranted. Figure 32.4 lists partial results for this final CFA model: Standardized factor loadings were statistically significant and sizable, estimated correlations among error terms and among latent factors were significant and of moderate size (except for the nonsignificant correlation between the latent factors Read-SC and Math-Prof), item reliabilities (the squared standardized factor loadings, ℓ^2) ranged from $\ell^2 = .36$ for RSC3 to $\ell^2 = .81$ for Math, and construct reliabilities[8] for the latent factors ranged from $H = .74$ for Read-SC to $H = .89$ for Math-Prof. Thus, with solid evidence of a quality measurement model, we now are ready to proceed to the second phase of the analysis.

Structural Phase. With a final measurement model in place, the *structural phase* consists of replacing the nonstructural covariances among latent factors with the hypothesized structure that is of main interest (currently, Model 1 in Figure 32.1a) and reanalyzing the data. When comparing these new data-model fit results ($\chi^2 = 601.85$ with $df = 47$, $p < .001$, SRMR = 0.12, RMSEA = 0.10 with CI_{90}: [0.096, 0.11], and CFI = 0.93) to those from the final CFA model, we learn that the introduction of two key restrictions—namely, the a priori specified zero paths from Read-SC and Math-SC to Math-Prof—significantly eroded data-model fit.[9] Consulting cutoff criteria in Table 32.4, we may conclude that the data do not fit the conceptual model in Figure 32.1a. Having conducted a two-phase analysis, however, we now know something we could not glean from the all-in-one analysis: The observed data-model misfit must largely be due to misspecifications in the structural portion

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 499

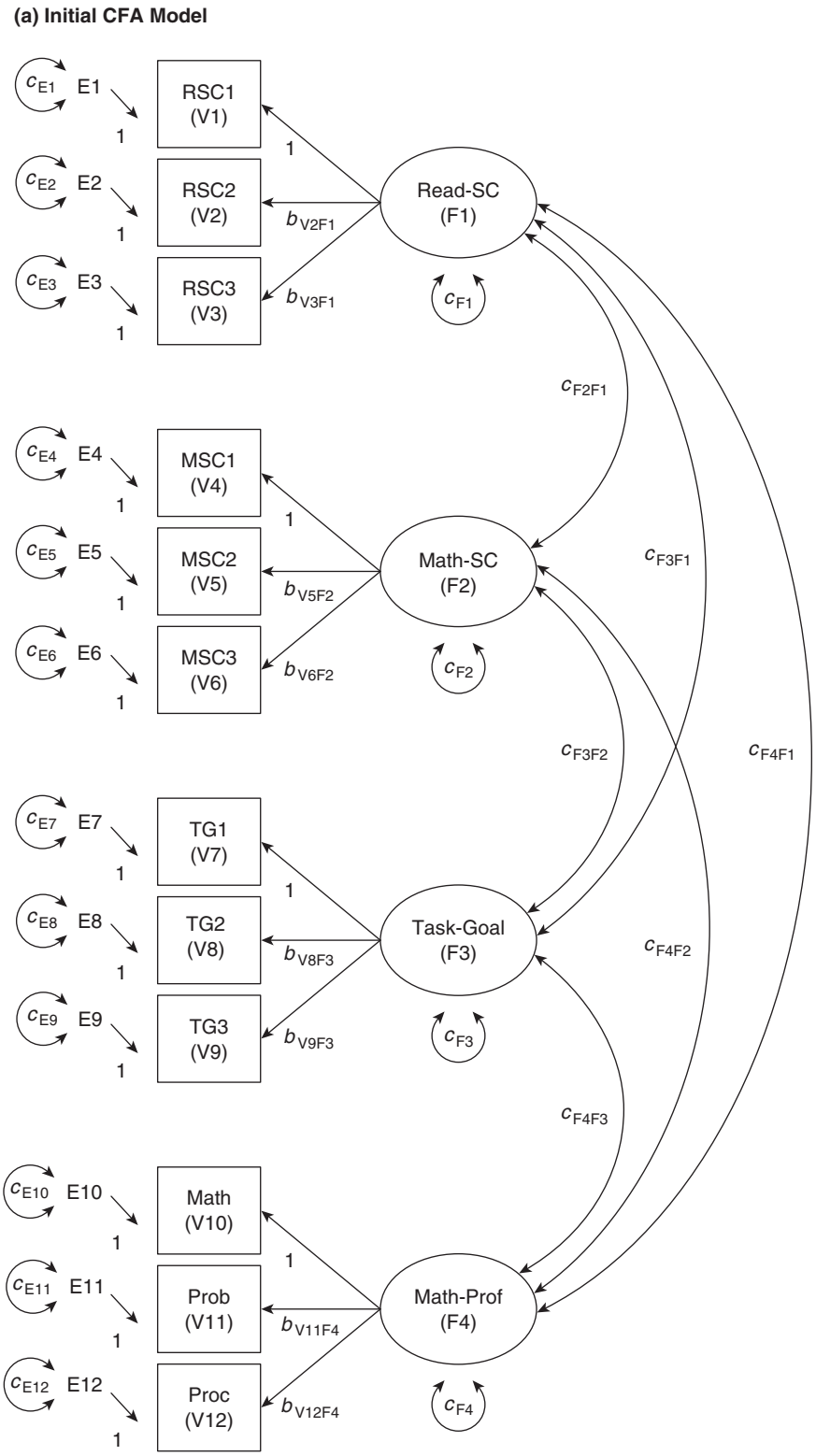

Figure 32.3 *(Continued)*

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 500

(b) Modified CFA Model

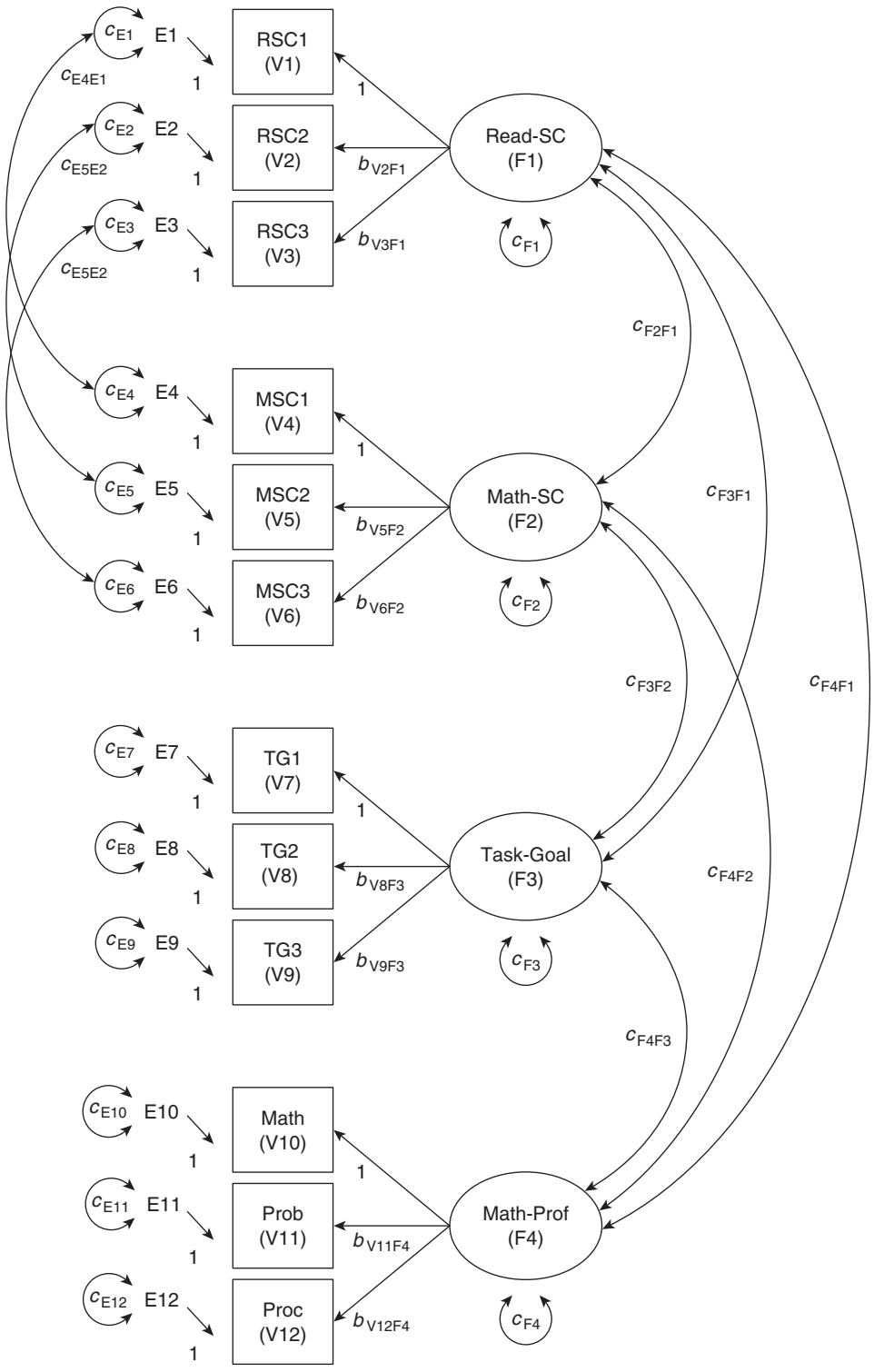

Figure 32.3 The measurement (CFA) models.

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 501

of the model since modifications to the measurement portion of the model (freeing error covariances for corresponding reading and mathematics self-concept items) led to a CFA model with no evidence of substantial data-model misfit. Having not yet reached a state where the data fit the hypothesized structure to an acceptable degree, we forego an interpretation of individual parameter estimates for a while longer in favor of illustrating how to compare and choose among the current model (Model 1) and the two remaining a priori hypothesized structures (Models 2 and 3) in Figure 32.1.

Choosing From Among Nested Models

Thus far in the illustrations, the comparison of models with respect to data-model fit could be accomplished only by descriptively weighing various fit index values across models. However, in the special case when two models, say Model 1 and Model 2, are nested (such as when the estimated parameters in the former are a proper subset of those associated with the latter), fit comparisons can be accomplished with a formal chi-square difference test. That is, if Model 1 (with df_1) is nested within Model 2 (with df_2), their chi-square fit statistics may be statistically compared by $\Delta\chi^2_{(df_1 - df_2)} = \chi^2_{(df_1)} - \chi^2_{(df_2)}$, which is distributed as a chi-square distribution with $df = df_1 - df_2$ (under conditions of multivariate normality).

Now reconsider the three theoretical models in Figure 32.1, all now incorporating the final measurement model in Figure 32.3b. As the fit information in Table 32.5 shows, Models 2 and 3 seem to fit well,[10] while Model 1 does not, as previously discussed. Furthermore, note that Model 1, the most parsimonious and restrictive model, is nested within both Models 2 and 3 (letting $b_{F4F2} \neq 0$ in Model 1 leads to Model 2; allowing both $b_{F4F2} \neq 0$ and $b_{F4F1} \neq 0$ in Model 1 leads to Model 3) and that Model 2 is nested in Model 3 (permitting $b_{F4F1} \neq 0$ in Model 2 leads to Model 3). Thus, the chi-square fit statistics for the three competing models can easily be compared by chi-square difference tests. Based on the three possible chi-square comparisons shown in Table 32.5, we glean that out of the three alternatives, Model 2 is the preferred structure (when weighing chi-square fit and parsimony):

1. Both Models 2 and 3 are chosen over Model 1 (they both exhibit significantly better fit, $\Delta\chi^2_{(1)} = 492.56$, $p < .001$; $\Delta\chi^2_{(2)} = 493.81$, $p < .001$; respectively), and
2. Model 2 is favored over Model 3 (even though it is more restrictive—but hence more parsimonious—the erosion in fit is nonsignificant, $\Delta\chi^2_{(1)} = 1.25$, $p = .264$).

Having chosen Model 2 from among the three alternative models and judging its data-model fit as acceptable (Table 32.5), what remains is an examination and interpretation of the structural parameter estimates that link the latent constructs (see Table 32.6; interpretations of results from the measurement phase are listed in Figure 32.4 and were examined earlier). Note that the latent factors Read-SC and Math-SC explained 20% of the variance in Task-Goal ($R^2 = 0.20$) and that those three factors explained more than 70% of the variance in latent mathematics proficiency ($R^2 = 0.71$). All structural estimates were statistically significant and can be interpreted in a manner similar to regression coefficients, but now with a focus on structural direction, given the specific causal nature of the underlying hypothesized theory. For example, considering the standardized path coefficients,[11] one might expect from within the context of Model 2 that a one standard deviation increase in ninth-grade girls' latent reading self-concept causes, on average, a bit more than a third (0.36) of a standard deviation increase in their latent task-goal orientation; similarly, a one standard deviation increase in girls' task-goal orientation leads, on average, to a 0.38 standard deviation increase in latent mathematics proficiency. Given the hypothesized structure, the effect of reading self-concept on mathematics proficiency is completely mediated by Task-Goal, with an estimated standardized indirect effect of $0.36 \times 0.38 = 0.14$ ($p < .05$, as indicated by SEM software).

Finally, recall from the beginning of this section the two separate multiple linear regression analyses for a somewhat crude initial attempt at addressing coefficient estimation for the structure in Model 2. In addition to LVPA results, Table 32.6 also lists R^2 values and the unstandardized and standardized regression coefficients associated with the two implied structural equations (using the proxy variables ReadSC3 for latent reading self-concept, MathSC3 for latent mathematics self-concept, TG1 for latent task-goal orientation, and Proc for latent mathematics proficiency). First, compare the two regression

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 502

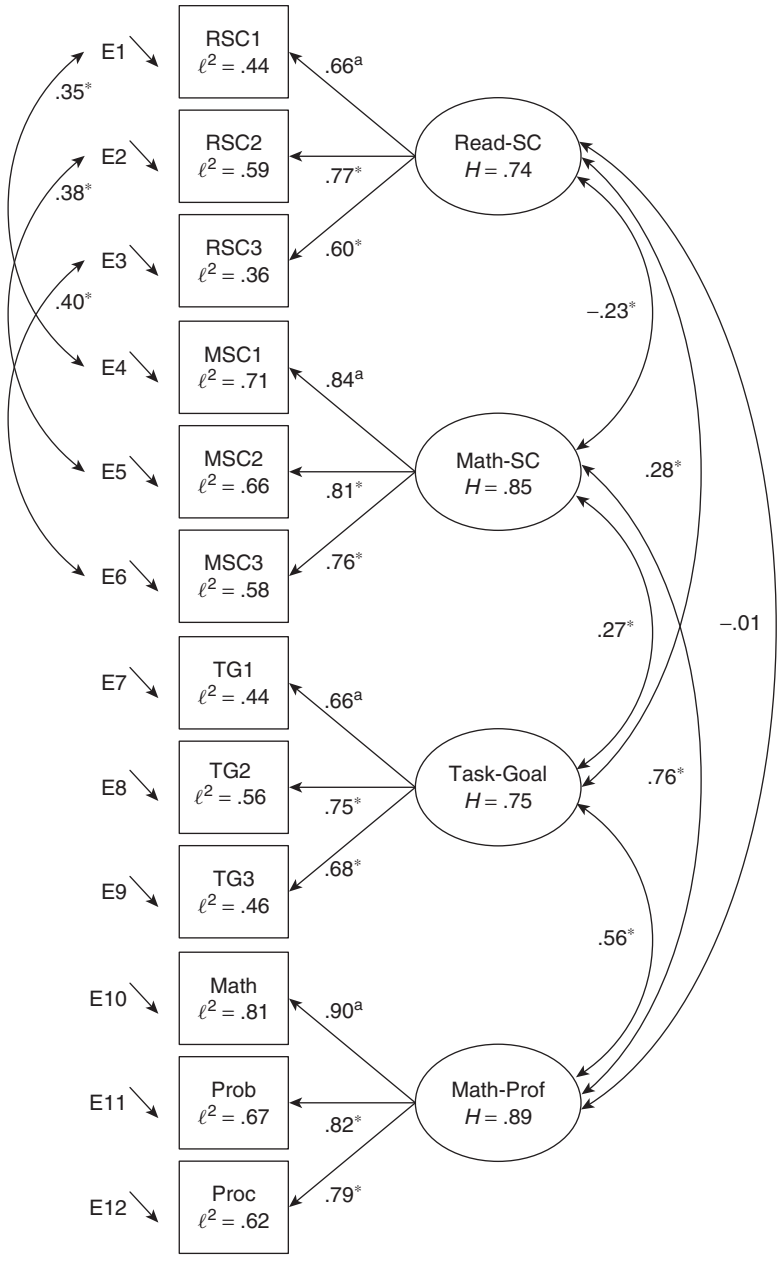

Figure 32.4 Standardized results for final CFA model.

NOTE: ℓ^2 is indicator reliability computed as the squared standardized loading. *H* is Hancock and Mueller's (2001) measure of construct reliability.

a. Unstandardized loading fixed to 1.0; thus, no standard error computed.

*$p < .05$.

R^2 values with those obtained from the SEM analysis and appreciate the huge increases observed with the LVPA approach: Explained variability in task-goal orientation jumped from 3.4% to 20%; in mathematics proficiency, it improved from 26% to 71%. Second, compare the MLR to the LVPA coefficients and note how in each case, the standardized LVPA estimates are higher/stronger

Table 32.5 Data-Model Fit and Chi-Square Difference Tests for Nested Models

				Model 2				Model 3			
				$\chi^2(df, p)$	*SRMR*	*RMSEA* CI_{90}	*CFI*	$\chi^2(df, p)$	*SRMR*	*RMSEA* CI_{90}	*CFI*
				109.29 (46, < 0.001)	0.017	0.036 (0.027, 0.045)	0.99	108.04 (45, < 0.001)	0.018	0.037 (0.028, 0.046)	0.99
Model 1				$\Delta\chi^2_{(1)} = \chi^2_{M1} - \chi^2_{M2} = 601.85 - 109.29 = 492.56$				$\Delta\chi^2_{(2)} = \chi^2_{M1} - \chi^2_{M3} = 601.85 - 108.04 = 493.81$			
$\chi^2(df, p)$	*SRMR*	*RMSEA* CI_{90}	*CFI*	($df = 1$, $p < .001$)				($df = 2$, $p < .001$)			
601.85 (47, < 0.001)	0.12	0.10 (0.096, 0.11)	0.93								
Model 2								$\Delta\chi^2_{(1)} = \chi^2_{M2} - \chi^2_{M3} = 109.29 - 108.04 = 1.25$			
								($df = 1$, $p = .264$)			

NOTE: CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 503

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 504

than their MLR counterparts. This disattenuation of relations among variables is due to LVPA's ability to "cleanse" estimates of the "noise" in the system that was introduced by the indicator variables' inevitable measurement error. Herein, then, lies one of the strengths of latent variable SEM approaches: Structural parameter estimates become purer, untangled from errors that originated in the measurement portion of the system, thus generating a higher portion of explained variability in dependent constructs.

Best Practices in SEM: Sharing the Study With Others

The four-stage SEM process and pointers gleaned from the above analysis illustrations map nicely onto the broadly accepted manuscript sections of introduction, methods, results, and discussion. What follows are some brief and general best practices guidelines on communicating research that uses SEM.

Introduction Section

Early in a manuscript—in the introduction and background to the problem under study—authors should convey a firm, overall sense of what led to the initial model conceptualization and of the specific model(s) investigated. The existence of latent constructs, as well as the hypothesized causal relations among them, must be justified clearly and convincingly based on theoretical grounds. Often, the articulation of competing, alternative models strengthens a study as it provides for a more complete picture of the current thinking in a particular field. The justification of measured and/or latent variables and models is accomplished by analyzing and synthesizing relevant literature by authors who proposed particular theories or empirically researched the same or similar models as the one(s) under current investigation. Path diagrams often are helpful in expressing the hypothesized structural links relating the measured and/or latent variables. Especially in complex models involving many variables, the hypothesized causal structure among latent variables can be easily illustrated in a diagram, while the psychometric details of how each latent variable was modeled may be left for the method section.

Table 32.6 Regression/Structural Coefficients for Model 2

		Analysis Method	
Effect From/To		*MLRs*	*LVPA*
Read-SC		0.08*	0.37*
	Task-Goal	0.08	0.36
Math-SC		0.15*	0.25*
		0.14	0.35
Math-SC		15.59*	18.41*
	Math-Prof	0.45	0.65
Task-Goal		6.22*	14.64*
		0.18	0.38
R^2		$R^2_{TG1} = 0.034$	$R^2_{Task\text{-}Goal} = 0.20$
		$R^2_{proc} = 0.26$	$R^2_{Math\text{-}Prof} = 0.71$

NOTE: Top coefficient is unstandardized; bottom coefficient is standardized.

*$p < .05$.

Method Section

In addition to specific information on participants, instruments, and procedures, this section should include a reference to the specific version of the SEM software package used since results can vary not only across programs but also across versions of a single package (mainly due to differences and continual refinements in estimation algorithms). Given the complexities of demonstrating each parameter's identification—and its necessity for parameter estimation—it is generally accepted to omit detailed discussions of identification issues from a manuscript, unless some unique circumstances warrant their inclusion. However, since the accuracy of parameter estimates, associated standard errors, and the overall chi-square value all depend on various characteristics of the indicators chosen and the data collected, authors should address issues such as multivariate distributions, sample size, missing data, outliers, and potential multilevel data structures, if applicable.

In most applied studies, ML estimates are presented that assume underlying multivariate normality and continuity of the data. Studies have shown that if sample size is sufficiently large, ML parameter estimates are quite robust against violations of these assumptions, though their associated standard errors and the overall chi-square might

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 505

not be (e.g., West, Finch, & Curran, 1995). Some have suggested, as a rough guideline, a 5:1 ratio of sample size to number of parameters estimated in order to trust ML parameter estimates (but associated standard errors and the model chi-square statistic might still be compromised; e.g., Bentler & Chou, 1987). We hesitate to endorse such a "one-size-fits-all" suggestion for three reasons. First, if data are not approximately normal, then alternate strategies such as the Satorra-Bentler rescaled statistics should be employed that have larger sample size requirements. Second, even under normality, methodological studies have illustrated that for models with highly reliable factors, quite satisfactory solutions can be obtained with relatively small samples, while models with less reliable factors might require larger samples (e.g., Gagné & Hancock, 2006; Marsh, Hau, Balla, & Grayson, 1998). Third, such general sample size recommendations ignore issues of statistical power to evaluate models as a whole or to test parameters within those models (see, e.g., Hancock, 2006).

If the model posits latent constructs, the choice of indicators usually is justified in an instrumentation subsection. First, for each construct modeled, the reader should be able to determine if effect or cause indicators were chosen: Only the former operationalize the commonly modeled latent factors; the latter determine latent composites (see Note 6). In some SEM analyses, emergent constructs are erroneously treated as latent, implying a mismatch between the modeled and the actual nature of the construct, hence leading to the potential for incorrect inferences regarding the relations the construct might have with other portions of the model. Second, each latent construct should be defined by a sufficient number of psychometrically sound indicators: "Two *might* be fine, three is better, four is best, and anything more is gravy" (Kenny, 1979, p. 143). Doing so can prevent various identification and estimation problems as well as ensure satisfactory construct reliability (since latent constructs are theoretically perfectly reliable but are measured by imperfect indicators, numerical estimates of construct reliability are likely to be less than 1.0 but can be brought to satisfactory levels with the inclusion of quality indicator variables; see, e.g., Hancock & Mueller, 2001). Finally, the scale of the indicator variables should be accommodated by the estimation method, where variables clearly yielding ordinal data might warrant the use of estimation strategies other than ML (see Finney & DiStefano, 2006).

Results Section

How authors structure the results section obviously is dictated by the particular model(s) and research questions under study. Notwithstanding, it is the researcher's responsibility to provide access to data in order to facilitate verification of the obtained results: If moment-level data were analyzed, a covariance matrix (or correlation matrix with standard deviations) should be presented in a table or appendix; if raw data were used, information on how to obtain access should be provided.

When analyzing LVPA models, results from both the measurement and structural phases should be presented. For overidentified models, judging the overall quality of a hypothesized model usually is presented early in the results section. Given that available data-model fit indices can lead to inconsistent conclusions, researchers should consider fit results from different classes so readers can arrive at a more complete picture regarding a model's acceptability (Table 32.4). Also, a comparison of fit across multiple, a priori specified alternative models can assist in weighing the relative merits of favoring one model over others. As illustrated, when competing models are nested, a formal chi-square difference test is available to judge if a more restrictive—but also more parsimonious—model can explain the observed data equally well, without a significant loss in data-model fit (alternative models that are not nested have traditionally been compared only descriptively—relative evaluations of AIC values are recommended, with smaller values indicating better fit—but recent methodological developments suggest statistical approaches as well; see Levy & Hancock, 2007).

If post hoc model modifications are performed following unacceptable data-model fit from either the measurement or structural phase of the analysis, authors owe their audience a detailed account of the nature and reasons (both statistical and theoretical) for the respecification(s), including summary results from Lagrange multiplier tests and revised final fit results. If data-model fit has been assessed and deemed satisfactory, with or without respecification, more detailed results are presented, usually in the form of individual unstandardized and standardized parameter estimates for each structural equation of interest, together with associated standard errors and/or test statistics and coefficients of determination (R^2). When latent variables are part of a model, estimates of their construct reliability should be presented, with values ideally falling above .70 or .80 (see Hancock & Mueller, 2001).

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 506

Discussion Section

In the final section of a manuscript, authors should provide a sense of what implications the results from the SEM analysis have on the theory or theories that gave rise to the initial model(s). Claims that a well-fitting model was "confirmed" or that a particular theory was proven to be "true," especially after post hoc respecifications, should be avoided. Such statements are grossly misleading given that alternative, structurally different, yet mathematically equivalent models always exist that would produce identical data-model fit results and thus would explain the data equally well (see Hershberger, 2006). At most, a model with acceptable fit may be interpreted as *one* tenable explanation for the associations observed in the data. From this perspective, a SEM analysis should be evaluated from a *dis*confirmatory, rather than a confirmatory, perspective: Based on unacceptable data-model fit results, theories can be shown to be false but not proven to be true by acceptable data-model fit (see also Mueller, 1997).

If evidence of data-model misfit was presented and a model was modified based on statistical results from Lagrange multiplier tests, readers must be made aware of potential model overfitting and the capitalization on chance. Statistically rather than theoretically based respecifications are purely exploratory and might say little about the true model underlying the data. While some model modifications seem appropriate and theoretically justifiable (usually, minor respecifications of the measurement portion are more easily defensible than those in the structural portion of a model), they only address internal specification errors and should be cross-validated with data from new and independent samples.[12]

Finally, the interpretation of individual parameter estimates can involve explicit causal language, *as long as this is done from within the context of the particular causal theory proposed* and the possibility/probability of alternative explanations is raised unequivocally. Though some might disagree, we think that explicit causal statements are more honest than implicit ones and are more useful in articulating a study's practical implications; after all, is not causality the ultimate aim of science (see Shaffer, 1992, p. x)? In the end, SEM is a powerful disconfirmatory tool at the researcher's disposal for testing and interpreting theoretically derived causal hypotheses from within an a priori specified causal system of observed and/or latent variables. However, we urge authors to resist the apparently still popular belief that the main goal of SEM is to achieve satisfactory data-model fit results; rather, it is to get one step closer to the "truth." If it is true that a proposed model does not reflect reality, then reaching a conclusion of *mis*fit between data and model should be a desirable goal, not one to be avoided by careless respecifications until satisfactory levels of fit are achieved.

Conclusion

Throughout the sections of this chapter, we have attempted to provide an overview of what we believe should be considered best practices in typical SEM applications. As is probably true for the other quantitative methods covered in this volume, a little SEM knowledge is sometimes a dangerous thing, especially with user-friendly software making the mechanics of SEM increasingly opaque to the applied user. Before embracing SEM as a potential analysis tool and reporting SEM-based studies, investigators should gain fundamental knowledge from any of the introductory textbooks referenced at the beginning of this chapter. In an effort to aid in the conduct and publication of appropriate, if not exemplary, SEM utilizations, we offered some best practices guidelines, except one, saving it for last. While SEM offers a general and flexible methodological framework, investigators should not hesitate to consider other analytical techniques—many covered in the present volume—that potentially address research questions much more clearly and directly. As it was explained to the second author several years ago, "Just because all your friends are doing this 'structural equation modeling' thing doesn't mean *you* have to. If all your friends jumped off a cliff. . . ." (Marta Foldi,[13] personal communication, 1992).

Notes

1. Several indicator variables are rating scales that could be argued to provide ordinal-level rather than interval-level data. Given the relatively high number of scale points, however, we analyzed these data as if they were interval (see Finney & DiStefano, 2006).

2. Using different proxy variables, or even composites of indicators, would still yield attenuated

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 507

results as none of the options filter the inherent measurement error.

3. Only *b* and *c* coefficients are used here; the letter *a* denotes intercept and mean terms in the analysis of mean structures.

4. To help ensure identification and to provide a metric for each latent factor, *reference variables* were specified (i.e., one factor loading for each latent construct was fixed to 1.0 as indicated in Figure 32.2).

5. Here it is assumed that measured variables of "high quality" were chosen to serve as indicator variables of the latent constructs (i.e., measured variables with relatively high factor loadings). Somewhat paradoxically, the use of low-quality indicators in the measurement portion of the model can erroneously lead to an inference of acceptable data-model fit regarding the structural portion (see Hancock & Mueller, 2007).

6. *Effect* indicators are measured variables that are specified to be the structural effects of the latent constructs that are hypothesized to underlie them (e.g., Bollen & Lennox, 1991). The analysis of models involving *cause* indicators—items that contribute to composite scores to form *emergent* factors, or latent composites—is theoretically different but also possible, albeit more difficult (see Kline, 2006).

7. Typically, LM statistics are not additive; that is, the chi-square statistic is *not* expected to drop by 1761.5 (= 652.0 + 567.8 + 541.7). When a fixed parameter is estimated in a subsequent reanalysis, LM statistics for the remaining fixed parameters usually change. Hence, theoretically justifiable model modifications motivated by LM results should usually occur one parameter at a time unless there is a clear theoretical reason for freeing multiple parameters at once, as is the case here.

8. One way to assess construct reliability is through Hancock and Mueller's (2001, p. 202) coefficient *H*. *H* is a function of item reliabilities, ℓ_i^2, and is computed by the equation

$$H = 1 \Big/ \left[1 + \left(1 \Big/ \sum_{i=1}^{k} \ell_i^2 / \left(1 - \ell_i^2\right) \right) \right],$$

where *k* is the number of measured variables associated with a given latent factor.

9. Such a statistical comparison is possible with a chi-square difference test since the current model is nested within the final measurement model, as explained next.

10. Perceptive readers will have noticed that fit results for Model 3 equal those previously discussed for the final measurement model. Indeed, this is no coincidence but an illustration of two *equivalent* models, that is, models that differ in structure but exhibit identical data-model fit for any data set (see Hershberger, 2006).

11. For a given sample, only the interpretation of standardized coefficients is meaningful as the latent factor metrics are arbitrary (different choices for reference variables could lead to different latent metrics); unstandardized coefficients can be useful in effect comparisons across multiple samples or studies.

12. If this is impractical or impossible, a cross-validation index could be computed (see Browne & Cudeck, 1993).

13. Mrs. Foldi is the second author's mother; she has no formal training in SEM or in any other statistical technique.

References

Arbuckle, J. L. (2007). AMOS (Version 7) [Computer software]. Chicago: SPSS.

Bentler, P. M. (2006). EQS (Version 6.1) [Computer software]. Encino, CA: Multivariate Software.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural equation modeling. *Sociological Methods & Research, 16,* 78–117.

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

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal, 7,* 461–483.

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

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

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS.* Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS.* Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.

Gagné, P. E., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65–83.

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O.

32-Osborne (Best)-45409.qxd 10/9/2007 1:00 PM Page 508

Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International, Inc.

Hancock, G. R., & Mueller, R. O. (2004). Path analysis. In M. Lewis-Beck, A. Brymann, & T. F. Liao (Eds.), *Sage encyclopedia of social science research methods* (pp. 802–806). Thousand Oaks, CA: Sage.

Hancock, G. R., & Mueller, R. O. (Eds.). (2006). *Structural equation modeling: A second course.* Greenwich, CT: Information Age Publishing.

Hancock, G. R., & Mueller, R. O. (2007, April). *The reliability paradox in structural equation modeling fit indices.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling: A Multidisciplinary Journal, 10,* 35–46.

Hershberger, S. L. (2006). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 13–41). Greenwich, CT: Information Age Publishing.

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

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

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31,* 165–178.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443–482.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL (Version 8.80) [Computer software]. Lincolnwood, IL: Scientific Software International.

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

Kenny, D. A. (1979). *Correlation and causation.* New York: John Wiley.

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

Kline, R. B. (2006). Formative measurement and feedback loops. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 43–68). Greenwich, CT: Information Age Publishing.

Levy, R., & Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research, 42,* 33–66.

Loehlin, J. C. (2004). *Latent variable models* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

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

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS.* New York: Springer-Verlag.

Mueller, R. O. (1997). Structural equation modeling: Back to basics. *Structural Equation Modeling: A Multidisciplinary Journal, 4,* 353–369.

Mueller, R. O., & Hancock, G. R. (2001). Factor analysis and latent structure, confirmatory. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 5239–5244). Oxford, UK: Elsevier.

Muthén, B. O., & Muthén, L. K. (2006). Mplus (Version 4.1) [Computer software]. Los Angeles: Author.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 285–305). Thousand Oaks, CA: Sage.

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

Shaffer, J. P. (Ed.). (1992). *The role of models in nonexperimental social science: Two debates.* Washington, DC: American Educational Research Association.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Wright, S. (1918). On the nature of size factors. *Genetics, 3,* 367–374.

33-Osborne (Best)-45409.qxd 10/9/2007 5:49 PM Page 509

33

Introduction to Bayesian Modeling for the Social Sciences

Gianluca Baio

Marta Blangiardo

In the context of statistical problems, the *frequentist* (or empirical) interpretation of probability has historically played a predominant role in modern statistics. In this approach, probability is defined as the limiting frequency of occurrence in an infinitely repeated experiment. The underlying assumption is that of a "fixed" concept of probability, which is unknown but can be theoretically disclosed by means of repeated trials, under the same experimental conditions.

However, although the frequentist approach still plays the role of the standard in various applied areas, many other possible conceptualizations of probability characterize different philosophies behind the problem of statistical inference. Among these, an increasingly popular one is the *Bayesian* (also referred to as *subjectivist*), originated by the posthumous work of Reverend Thomas Bayes (1763)—see Howie (2002), Senn (2003), or Fienberg (2006) for a historical account of Bayesian theory.

The main feature of this approach is that probability is interpreted as a subjective degree of belief in the occurrence of an event, representing the individual level of uncertainty in its actual realization (cf. de Finetti, 1974, probably the most comprehensive account of subjective probability). One of the main implications of subjectivism is that there is no requirement that one should be able to specify, or even conceive of, some relevant sequence of repetitions of the event in question, as happens in the frequentist framework, with the advantage that events of the "one-off" type can be assessed consistently.

In the Bayesian philosophy, the probability assigned to any event depends also on the individual whose uncertainty is being taken into account and on the state of background information underlying this assessment. Varying any of these factors might change the probability. Consequently, under the subjectivist view, there is no assumption of a unique, correct (or "true") value for the probability of any uncertain event (Dawid, 2005). Rather, each individual is entitled to his or her own subjective probability, and according to the evidence that becomes sequentially available, individuals tend to update their beliefs.[1]

33-Osborne (Best)-45409.qxd 10/9/2007 5:49 PM Page 510

Bayesian methods are not new to the social sciences—from Phillips (1973) to Iversen (1984), Efron (1986), Raftery (1995), Berger (2000), and Gill (2002)—but they are also not systematically integrated into most research in the social sciences. This may be due to the common perception among practitioners that Bayesian methods are "more complex."

In fact, in our opinion, the apparent higher degree of complexity is more than compensated by at least the two following consequences. First, Bayesian methods allow taking into account, through a formal model, all the available information, such as the results of previous studies. Moreover, the inferential process is straightforward, as it is possible to make probabilistic statements directly on the quantities of interest (i.e., some unobservable feature of the process under study, typically represented by a set of parameters).

Despite their subjectivist nature, Bayesian methods allow the practitioner to make the most of the evidence: In just the situation of "repeated trials," after observing the outcomes (successes and failures) of many past trials (assuming no other source of information), the individuals will be drawn to an assessment of the probability of success on the next event that is extremely close to the observed proportion of successes so far. However, if past data are not sufficiently extensive, it may be reasonably argued that there should indeed be scope for interpersonal disagreement as to the implications of the evidence. Therefore, the Bayesian approach provides a more general framework for the problem of statistical inference.[2]

In order to facilitate comprehension, we shall present two worked examples and switch between theory and practice in every section. In the first part of the chapter, we consider data about nonattendance at school for a set of Australian children, with additional information about their race (Aboriginal, White) and age band also included. We use this data set to present the main feature of Bayesian reasoning and to follow the development of the simplest form of models (conjugated analysis). In the last section of the chapter, we describe a more realistic representation for the analysis of SAT score data. The main objective of this analysis is to develop a more complex model combining information for a number of related variables, using the simulation techniques of Markov chain Monte Carlo methods.

Conditional Probabilities and Bayes Theorem

A fundamental concept in statistics, particularly within the Bayesian approach, is that of *conditional probability* (for a technical review, see Dawid, 1979). Given two events A and B, we can define the event $A \mid B$ (read "A given B") as the occurrence of the event A under the circumstance that the event B has already occurred.

In other words, by considering the conditional probability, we are in fact changing the reference population; the probability of the event A, $\Pr(A)$, is generally defined over the space Ω, which contains all the possible events under study, including A. Conversely, when considering the conditional probability $\Pr(A \mid B)$, we are restricting our attention to the subspace of Ω where both the events A and B can occur. Such a subspace is indicated as $(A \,\&\, B)$. Moreover, the basis of our comparison will not be Ω but just its subspace where B is possible. Consequently, the probability of the occurrence of the event A conditional on the event B is formally defined as

$$\Pr(A \mid B) = \frac{\Pr(A\&B)}{\Pr(B)}. \qquad (1)$$

Example: Nonattendance at School in Australia

Paul and Banerjee (1998) studied Australian educational data on the days of nonattendance at school for 146 children by race (Aboriginal, White) and age band (primary, first form, second form, third form). The observed average value of nonattendance days is 16, which will be used as a cutoff threshold for our analysis.

Suppose we are interested in the probability that a student accumulates more than the average number of nonattendance days, conditional on his or her race. We can define the events of interest as follows:

$$H = \{>16 \text{ nonattendance days}\}$$

$$W = \{\text{White race}\}$$

(where H stands for *high* nonattendance) and their complement as