In: Kline, R. B. 2011: Principles and Practice of Structural Equation Modeling. 3. Auflage. New York: Guilford.

9

Measurement Models and Confirmatory Factor Analysis

This is the first of two chapters about the analysis of core latent variable models in SEM, in this case measurement models as analyzed in CFA. The multiple-indicator approach to measurement of CFA represents literally half the basic rationale of analyzing covariance structures in SEM—the analysis of structural models is the other half—so CFA is a crucial technique. It is also a primary technique for many researchers, especially those who conduct assessment-related studies. Also introduced in this chapter is multiple-sample CFA, in which a measurement model is fitted simultaneously to data from more than one group. The results provide a test of measurement invariance, or of whether a set of indicators has the same measurement properties across the groups. If you know something about CFA, then it is easier to learn about structural regression (SR) models, which have features of both path models and CFA models. The next chapter covers SR models.

NAMING AND REIFICATION FALLACIES

The specification and identification of CFA models were introduced in, respectively, Chapters 5 and 6 using model diagrams where factors were designated with letters, such as *A* and *B* (e.g., Figure 5.6). In real analyses, researchers usually assign meaningful labels to factors such as sequential processing (Figure 5.7). However, it is important to avoid two logical errors concerning factor names. The first is the **naming fallacy**: Just because a factor is named does not mean that the hypothetical construct is understood or even correctly labeled. Factors require some type of designation, though, if for no other reason than communication of the results. Although verbal labels are more "reader friendly" than more abstract symbols, such as *A* or ε (xi, a symbol from LISREL's

matrix notation for exogenous factors), they should be viewed as conveniences and not as substitutes for critical thinking.

The second logical error to avoid is **reification**: the belief that a hypothetical construct *must* correspond to a real thing. For example, a general ability factor, often called *g*, is a hypothetical construct. To automatically consider g as real instead of a concept, however, is a potential error of reification. Along these lines, Gardner (1993) reminded educators not to assume that "intelligence" corresponds to a single domain that is adequately measured by IQ scores. Instead, he argued that intelligence is multifaceted and includes not only academic skills but also social, artistic, and athletic domains.

ESTIMATION OF CFA MODELS

This discussion assumes that all indicators are continuous variables. This is most likely to happen when each indicator is a *scale* that generates a total score over a set of items. A later section deals with the analysis of models where *items* are specified as indicators.

Interpretation of Estimates

Parameter estimates in CFA are interpreted as follows:

1. Factor loadings estimate the direct effects of factors on indicators and are interpreted as regression coefficients. For example, if the unstandardized factor loading is 4.0 for the direct effect $A \rightarrow X_1$, then we expect a four-point difference in indicator X_1 given a difference of 1 point on factor *A*. Loadings fixed to 1.0 to scale the corresponding factor remain so in the unstandardized solution and are not tested for statistical significance because they have no standard errors.
2. For indicators specified to load on a single factor, standardized factor loadings are estimated correlations between the indicator and its factor. Thus, squared standardized loadings are proportions of explained variance, or R^2_{smc}. If a standardized loading is .80, for example, the factor explains $.80^2 = .64$, or 64.0% of the variance of the indicator. Ideally, a CFA model should explain the majority of the variance ($R^2_{\text{smc}} > .50$) of each indicator.
3. For indicators specified to load on multiple factors, standardized loadings are interpreted as beta weights that control for correlated factors. Because beta weights are not correlations, one cannot generally square their values to derive proportions of explained variance.
4. The ratio of an unstandardized measurement error variance over the observed variance of the corresponding indicator equals the proportion of unexplained variance, and one minus this ratio is the proportion of explained variance. Suppose that the variance of X_1 is 25.00 and that the variance of its error term is 9.00. The proportion of unexplained variance is 9.00/25.00 = .36, and the proportion of explained variance is $R^2_{\text{smc}} = 1 - .34 = .64$.

5. Estimates of unanalyzed associations between either a pair of factors or measurement errors are covariances in the unstandardized solution. These estimates are correlations in the standardized solution.

The estimated correlation between an indicator and a factor is a **structure coefficient**. If an indicator loads on a single factor, its standardized loading is a structure coefficient; otherwise, it is not. Graham, Guthrie, and Thompson (2003) remind us that the specification that a direct effect of a factor on an indicator is zero does *not* mean that the correlation between the two must be zero. That is, a zero pattern coefficient (factor loading) does *not* imply a zero structure coefficient. This is because the factors in CFA models are assumed to covary, which implies nonzero correlations between each indicator and all factors. However, indicators should have higher estimated correlations with the factors they are specified to measure.

Problems

Failure of iterative estimation in CFA can be caused by poor start values; suggestions for calculating start values for measurement models are presented in Appendix 9.A. Inadmissible solutions include Heywood cases such as negative variance estimates or estimated absolute correlations greater than 1.0. Results of some computer studies indicate that nonconvergence or improper solutions are more likely when there are only two indicators per factor or the sample size is less than 100–150 cases (Marsh & Hau, 1999). The authors just cited give the following suggestions for analyzing CFA models when the sample size is not large:

1. Use indicators with good psychometric characteristics that will each also have relatively high standardized factor loadings (e.g., > .70). Models with indicators that have relatively low standardized loadings are more susceptible to Heywood cases (Wothke, 1993).
2. Estimation of the model with equality constraints imposed on the unstandardized loadings of indicators of the same factor may help to generate more trustworthy solutions. This assumes that all indicators have the same metric.
3. When the indicators are items instead of continuous total scores, it may be better to analyze them in groups (parcels) rather than individually. Recall that the analysis of parcels is controversial because it requires a very strong assumption, that the items of a parcel are unidimensional (Chapter 7).

Solution inadmissibility can also occur at the parameter matrix level. Specifically, the computer estimates in CFA a factor covariance matrix and an error covariance matrix. If any element of either parameter matrix is out of bounds, then that matrix is nonpositive definite. Causes of nonpositive definite parameter matrices include the following (Wothke, 1993):

1. The data provide too little information (e.g., small sample, two indicators per factor).
2. The model is overparameterized (too many parameters).
3. The sample has outliers or severely non-normal distributions (poor data screening).
4. There is empirical underidentification concerning factor covariances (e.g., Figure 6.4).
5. The measurement model is misspecified.

Empirical Checks for Identification

It is theoretically possible for the computer to generate a converged, admissible solution for a model that is not really identified, yet print no warning message. However, that solution would not be unique. This is most likely to happen in CFA when analyzing a model with both correlated errors and complex indicators for which the application of heuristics cannot prove identification (Chapter 6). Described next are empirical checks for solution uniqueness that can be applied when analyzing any type of structural equation model, not just CFA models. These checks concern necessary but insufficient requirements. That is, if any of these checks is failed, then the solution is not unique, but passing them does not prove identification:

1. Conduct a second analysis of the same model but use different start values than in the first analysis. If estimation converges to a different solution working from different initial estimates, the original solution is not unique and the model is not identified.
2. This check applies to overidentified models only: Use the model-implied covariance matrix from the first analysis as the input data for a second analysis of the same model. If the second analysis does not generate the same solution as the first, the model is not identified.
3. Some SEM computer programs optionally print the matrix of estimated correlations among the parameter estimates. Although parameters are fixed values that do not vary randomly in the population, their estimates are considered random variables with their own distributions and covariances. Estimates of these covariances are based on the **information matrix**, which is associated with full-information methods such as ML. If the model is identified, this matrix has an inverse, which is the matrix of covariances among the parameter estimates. Correlations among the parameters are derived from this matrix. An identification problem is indicated if any of these absolute correlations is close to 1.0, which indicates linear dependency. See Bollen (1989, pp. 246–251) for additional necessary-but-insufficient empirical checks based on linear algebra methods.

DETAILED EXAMPLE

This example concerns the analysis of the measurement model for the first edition Kaufman Assessment Battery for Children (KABC-I) introduced in Chapter 5. The two-factor,

eight-indicator theoretical model for this test is presented in Figure 9.1. Briefly, the first three subtests are specified to load on one factor (sequential processing) and the other five subtests on a second factor (simultaneous processing).[1] The data for this analysis are summarized in Table 9.1, which are from the test's standardization sample for 10-year-old children (N = 200).

Test for a Single Factor

When theory is not specific about the number of factors, this is often the first step in a series of analyses: if a single-factor model cannot be rejected, there is little point in evaluating more complex ones. Even when theory is more precise about the number of factors (e.g., two for the KABC-I), it should be determined whether the fit of a simpler, one-factor model is comparable. I submitted to Mplus 5.2 the covariance matrix based on the data in Table 9.1 for ML estimation of a single-factor CFA model. The unstandardized loading of the Hand Movements indicator was fixed to 1.0 to scale the single factor. With v = 8 indicators, there are 36 observations available to estimate a total of 16 free parameters, including nine variances of exogenous variables (of the single factor and eight measurement errors) and seven factor loadings, so df_M = 20. Estimation in Mplus converged to an admissible solution. Values of selected fit indexes for the one-factor model are listed next (Mplus does not print the GFI). The 90% confidence interval associated with the RMSEA is reported in parentheses:

$$\chi^2_M(20) = 105.427, \quad p < .001$$
$$\text{RMSEA} = .146\ (.119\text{–}.174), \quad p_{\text{close-fit } H_0} < .001$$
$$\text{CFI} = .818; \quad \text{SRMR} = .084$$

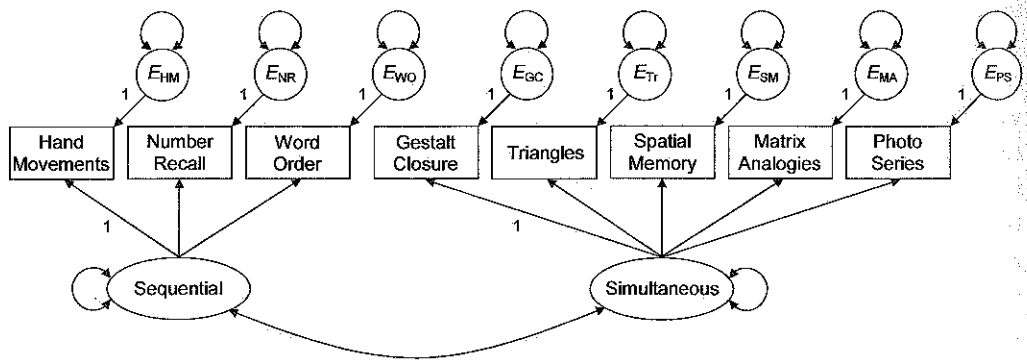

FIGURE 9.1. A confirmatory factor analysis model of the Kaufman Assessment Battery for Children, first edition.

[1]Keith (1985) suggested alternative names for the factors in the KABC-I's theoretical model, including "short-term memory" instead of "sequential processing" and "visual-spatial reasoning" instead of "simultaneous processing." One reason is that all three sequential tasks are immediate recall tasks, and all five simultaneous tasks involve responding to visual stimuli.

TABLE 9.1. Input Data (Correlations, Standard Deviations) for Analysis of a Two-Factor Model of the Kaufman Assessment Battery for Children

Variable	1	2	3	4	5	6	7	8
Sequential scale								
1. Hand Movements	1.00							
2. Number Recall	.39	1.00						
3. Word Order	.35	.67	1.00					
Simultaneous scale								
4. Gestalt Closure	.21	.11	.16	1.00				
5. Triangles	.32	.27	.29	.38	1.00			
6. Spatial Memory	.40	.29	.28	.30	.47	1.00		
7. Matrix Analogies	.39	.32	.30	.31	.42	.41	1.00	
8. Photo Series	.39	.29	.37	.42	.58	.51	.42	1.00
SD	3.40	2.40	2.90	2.70	2.70	4.20	2.80	3.00

Note. Input data are from Kaufman and Kaufman (1983); N = 200.

The overall fit of a one-factor model to the data in Table 9.1 is obviously poor, so it is rejected.

The test for a single factor is relevant not just for CFA models. For example, Kenny (1979) noted that such models could also be tested as part of a path analysis. The inability to reject a single-factor model in this context would mean the same thing as in CFA: the observed variables do not show discriminant validity; that is, they seem to measure only one domain. I conducted a test for single-factoredness for the five variables of the Roth et al. (1989) path model of illness factors analyzed in Chapter 8 (see Figure 8.1). A single-factor model for these variables was fitted to a covariance matrix based on the data summarized in Table 3.4 with the ML method of Mplus 5.2. Values of selected fit statistics clearly show that a single-factor model poorly explains the Roth et al. (1989) data and provide a "green light" to proceed with evaluation of a path model:

$$\chi^2_M(5) = 60.549, \quad p < .001$$
$$\text{RMSEA} = .173\ (.135\text{–}.213), \quad p_{\text{close-fit } H_0} < .001$$
$$\text{CFI} = 644; \quad \text{SRMR} = .096$$

Two-Factor Model

There are also 36 observations available for the analysis of the two-factor model of the KABC-I in Figure 9.1. To scale the two factors, the unstandardized loadings of the Hand Movements task and the Gestalt Closure task on their respective factors were each fixed to 1.0. A total of 17 free parameters remain to be estimated, including 10 variances (of two factors and eight error terms), one factor covariance, and six factor loadings (two on the first factor, four on the second), so df_M = 19.

I used Mplus 5.2 to fit the two-factor model of Figure 9.1 to the data in Table 9.1

with ML estimation. You can download the Mplus syntax, data, and output files for this analysis from this book's website (see p. 3) and all EQS and LISREL computer files for the same analysis, too. The analysis in Mplus converged to an admissible solution. Presented in Table 9.2 are the parameter estimates for the two-factor model. Note in the table that unstandardized loadings of the reference variables (Hand Movements, Gestalt Closure) equal 1.0 and have no standard errors. The other six unstandardized loadings were freely estimated, and their values are all statistically significant at the .01 level. Although statistical significance of the unstandardized estimates of factor variances is indicated in the table, it is expected that these variances are not zero (i.e., there are individual differences). It would be senseless in most analyses to get worked up about the statistical significance of these terms, but results of significance tests for

TABLE 9.2. Maximum Likelihood Estimates for a Two-Factor Model of the KABC-I

Parameter	Unstandardized	*SE*	Standardized
	Factor loadings		
Sequential factor			
Hand Movements	1.000[a]	—	.497
Number Recall	1.147	.181	.807
Word Order	1.388	.219	.808
Simultaneous factor			
Gestalt Closure	1.000[a]	—	.503
Triangles	1.445	.227	.726
Spatial Memory	2.029	.335	.656
Matrix Analogies	1.212	.212	.588
Photo Series	1.727	.265	.782
	Measurement error variances		
Hand Movements	8.664	.938	.753
Number Recall	1.998	.414	.349
Word Order	2.902	.604	.347
Gestalt Closure	5.419	.585	.747
Triangles	3.425	.458	.472
Spatial Memory	9.998	1.202	.570
Matrix Analogies	5.104	.578	.654
Photo Series	3.483	.537	.389
	Factor variances and covariance		
Sequential	2.839	.838	1.000
Simultaneous	1.835	.530	1.000
Sequential ↔ Simultaneous	1.271	.324	.557

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. Standardized estimates for measurement errors are proportions of unexplained variance.

[a]Not tested for statistical significance. All other unstandardized estimates are statistically significant at $p < .01$.

factor covariances may be of greater interest. Note that Mplus can automatically print correct standard errors for standardized estimates, but these values are not reported in Table 9.2.

Of greater interpretive import are the standardized factor loadings in Table 9.2. Because each indicator loads on a single factor, the square of each standardized loading equals R^2_{smc} for the corresponding indicator. Some standardized loadings are so low that convergent validity seems doubtful. For example, the loadings of Hand Movements and Gestalt Closure on their respective factors are both only about .50, and $R^2_{smc} < .50$, for a total of four out of eight indicators. That is, the model in Figure 9.1 explains the minority of the observed variance for exactly half of the indicators. On the other hand, the estimated factor correlation (.557) is only moderate in size, which suggests discriminant validity.

Reported in Table 9.3 are values of structure coefficients (estimated factor-indicator correlations) for all eight indicators in the two-factor model of Figure 9.1. Coefficients presented in boldface in the table are also standardized factor loadings for indicators specified to measure either factor. For example, the Hand Movements is not specified to measure simultaneous processing (Figure 9.1). Therefore, the pattern coefficient for the Hand Movements-simultaneous processing correspondence is zero. However, the structure coefficients for the Hand Movements task are .497 and .277 (Table 9.3). The former, .497, equals the standardized loading of this indicator on the sequential processing factor (Table 9.2). The latter, .277, is the model-implied correlation between the Hand Movements task and the simultaneous processing factor. It is calculated using the tracing rule as the product of the standardized loading for the Hand Movements task and the estimated correlation between the factors, or .497 (.557) = .277. Exercise 1 asks you to derive the other structure coefficients in Table 9.3 using the tracing rule. The results in the table clearly show that the structure coefficients are not typically zero for corresponding zero pattern coefficients when the factors are substantially correlated.

TABLE 9.3. Structure Coefficients for a Two-Factor Model of the KABC-I

	Factor	
Indicator	Sequential	Simultaneous
Hand Movements	**.497**	.277
Number Recall	**.807**	.449
Word Order	**.808**	.450
Gestalt Closure	.280	**.503**
Triangles	.404	**.726**
Spatial Memory	.365	**.656**
Matrix Analogies	.328	**.588**
Photo Series	.436	**.782**

Note. KABC-I, Kaufman Assessment Battery for Children, first edition.

Tests for Multiple Factors

The chi-square reported by Mplus for the two-factor model in Figure 9.1 is $\chi^2_M(19) = 38.325$, $p = .005$. Thus, this model fails the chi-square test, and so it is necessary to investigate the magnitude and patterns of discrepancies between model and data. We will do so momentarily, but at this point we can infer that the fit of the two-factor model is better than that of the one-factor model fitted to the same data based on their respective χ^2_M values. Now, can we compare the relative fits of these two models with the chi-square difference test? Recall that this test is only for hierarchical models (Chapter 8). Is this true of the one- and two-factor models of the KABC-I?

Yes, and here is why: the one-factor model is actually a restricted version of the two-factor model. Look again at Figure 9.1. If the correlation between the two factors is fixed to equal 1.0, then the two factors will be identical, which is the same thing as replacing both factors with just one. The results of the chi-square difference test are

$$df_{M_{1\ factor}} - df_{M_{2\ factors}} = 20 - 19 = 1$$

$$\chi^2_D(1) = \chi^2_{M_{1\ factor}} - \chi^2_{M_{2\ factors}} = 105.427 - 38.325$$

$$= 67.102, \quad p < .001$$

which says that the fit of the two-factor model is statistically better than that of the single-factor model. The meaning of this particular result is not clear at this point because the fit of the more complex two-factor model is problematic. However, the comparison just described can be generalized to models with more factors. With a four-factor model, for instance, fixing all factor correlations to 1.0 generates a single-factor model that is nested under the unrestricted model. Merging any two factors (and their indicators) in a four-factor model into a single factor results in a three-factor model that is nested under the original model, and so on.

Assessment of Model Fit

Reported in Table 9.4 are values of fit statistics and results of model-level power analyses for the two-factor model of the KABC-I (Figure 9.1). As mentioned, the model chi-square is statistically significant, so the exact-fit hypothesis is rejected. Results for the RMSEA are mixed. The lower bound of the 90% confidence interval for this statistic is .037, so the close-fit hypothesis is not rejected ($p = .171$). However, the upper bound exceeds .10, so the poor-fit hypothesis cannot be rejected. Values of the CFI and SRMR are, respectively, .959 and .072. Levels of statistical power estimated in the Power Analysis module of STATISTICA 8 Advanced for tests of the close-fit hypothesis and the not-close-fit hypothesis are both low (respectively, .440 and .302). Minimum sample sizes of over twice that of the actual size for this analysis ($N = 200$) would be needed in order for power to be at least .80 (see Table 9.4).

TABLE 9.4. Values of Fit Statistics and Power Estimates for a Two-Factor Model of the KABC-I

Fit statistics		Power estimates	
Statistic	Result	Statistic or test	Result
χ^2_M	38.325	N	200
df_M	19	df_M	19
p	.005	Power	
RMSEA (90% CI)	.071 (.038–.104)	Close-fit test[a]	.440
$p_{\text{close-fit } H_0}$	.132	Not-close-fit test[b]	.302
CFI	.959	Minimum N[c]	
SRMR	.072	Close-fit test	455
		Not-close-fit test	490

Note. KABC-I, Kaufman Assessment Battery for Children, first edition; CI, confidence interval.

[a]H_0: $\varepsilon \leq .05$, $\varepsilon_1 = .08$, $\alpha = .05$.
[b]H_0: $\varepsilon \geq .05$, $\varepsilon_1 = .01$, $\alpha = .05$.
[c]Sample size rounded up to closest multiple of 5 required for power ≥ .80.

Reported in Table 9.5 are the correlation residuals (calculated in EQS) for the two-factor model. Many of these residuals (shown in boldface in the table) exceed .10 in absolute value. Most of the larger residuals concern one of the indicators of sequential processing, Hand Movements, and most of the indicators of simultaneous processing. All of these residuals are positive, which means that the two-factor model generally underestimates correlations between Hand Movements and those specified to measure the other factor. Based on all the results described so far, the fit of the two-factor model in Figure 9.1 is unacceptable. Exercise 2 will ask you to use an SEM computer tool to derive the standardized residuals (z statistics) for this analysis.

TABLE 9.5. Correlation Residuals for a Two-Factor Model of the KABC-I

Variable	1	2	3	4	5	6	7	8
Sequential scale								
1. Hand Movements	0							
2. Number Recall	−.011	0						
3. Word Order	−.052	.018	0					
Simultaneous scale								
4. Gestalt Closure	.071	**−.116**	−.066	0				
5. Triangles	**.119**	−.057	−.037	.015	0			
6. Spatial Memory	**.218**	−.005	−.015	−.030	−.007	0		
7. Matrix Analogies	**.227**	.056	.035	.014	−.007	.024	0	
8. Photo Series	**.174**	−.061	.018	.027	.012	−.003	−.040	0

Note. KABC-I, Kaufman Assessment Battery for Children, first edition.

RESPECIFICATION OF MEASUREMENT MODELS

In the face of adversity, the protagonist of Kurt Vonnegut's novel *Slaughterhouse-Five* often remarks, "So it goes." And so it often goes in CFA that an initial model does not fit the data very well. The respecification of a CFA model is even more challenging than that of a path model because there are more possibilities for change. For example, the number of factors, their relations to the indicators, and patterns of measurement error correlations are all candidates for modification. Given so many potential variations, respecification of CFA models should be guided as much as possible by substantive considerations. Otherwise, the specification process could put the researcher in the same situation as the sailor in this adage attributed to Leonardo da Vinci: One who loves practice without theory is like a sailor who boards a ship without a rudder and compass and never knows where he or she may be cast.

Two general classes of problems can be considered in respecification. The first concerns the indicators. Sometimes the indicators fail to have substantial standardized loadings (e.g., < .20) on the factors to which they were originally assigned. One option is to specify that the indicator measures a different factor. Inspection of the correlation residuals can help to identify the other factor to which the indicator's loading may be switched. Suppose that an indicator is originally specified to measure factor *A*, but the correlation residuals between it and the indicators of factor *B* are large and positive. This would suggest that the indicator may measure factor *B* more than it does factor *A*. Note that an indicator can have relatively high loadings on its own factor but also have high residual correlations between it and the indicators of another factor. The pattern just described suggests that the indicator in question measures more than one construct (i.e., allow it to load on > 1 factor). Another possibility consistent with this same pattern is that these indicators share something that is unique to them, such as a particular method of measurement. This possibility would be represented by allowing that pair of measurement errors to covary.

The second class of problems concerns the factors. For example, the researcher may have specified the wrong number of factors. On the one hand, poor discriminant validity as evidenced by very high factor correlations may indicate that the model has too many factors. On the other hand, poor convergent validity within sets of indicators of the same factor suggests that the model may have too few factors.

A starting point for respecification often includes inspection of the correlation residuals and modification indexes. Earlier we examined the correlation residuals in Table 9.5 for the two-factor model of the KABC-I. Most of the large and positive residuals are between the Hand Movements task and tasks specified to measure the other factor. Because the standardized loading of the Hand Movements task on its original factor is at least moderate (.497; Table 9.2), it is possible that this task may measure both factors. Reported in Table 9.6 are the 10 largest modification indexes computed by Mplus for factor loadings and error covariances that are fixed to zero in the original model (Figure 9.1). Note in the table that the χ^2 (1) statistics for the paths

TABLE 9.6. Ten Largest Modification Indexes for a Two-Factor Model of the KABC-I

Path	MI
Simultaneous → Hand Movements	20.091**
$E_{WO} \curvearrowright E_{NR}$	20.042**
Simultaneous → Number Recall	7.010**
$E_{HM} \curvearrowright E_{WO}$	7.015**
$E_{HM} \curvearrowright E_{SM}$	4.847*
$E_{HM} \curvearrowright E_{MA}$	3.799
Sequential → Matrix Analogies	3.247
$E_{NR} \curvearrowright E_{PS}$	3.147
Sequential → Gestalt Closure	2.902
$E_{MA} \curvearrowright E_{PS}$	2.727

Note. KABC-I, Kaufman Assessment Battery for Children, first edition; MI, modification index; HM, Hand Movements; WO, Word Order; SM, Spatial Memory; MA, Matrix Analogies; PS, Photo Series.

*$p < .05$; **$p < .01$.

Simultaneous → Hand Movements and $E_{WO} \curvearrowright E_{NR}$

are nearly identical (respectively, 20.091 and 20.042). Thus, either allowing Hand Movements to also load on the simultaneous processing factor or adding an error covariance between the Word Order and Number Recall tasks would reduce the value of χ^2_M by about 20 points. Among other changes suggested by the modification indexes, two have nearly the same χ^2 (1) value: allow Number Recall to also load on the sequential processing factor (7.010), or allow the errors of the Hand Movements and Word Order tasks to covary (7.015). The researcher needs a rationale for choosing among these potential respecifications. Based on my knowledge of the KABC-I (e.g., Kline, Snyder, & Castellanos, 1996) and results of other factor-analytic studies (e.g., Keith, 1985), allowing the Hand Movements task to load on both factors is plausible.

SPECIAL TOPICS AND TESTS

Different types of score reliability coefficients (test–retest, internal consistency, etc.) for individual indicators were described in Chapter 3. There are also a few different coefficients for estimating the reliability of construct (factor) measurement through all its indicators in CFA. Two of these coefficients are described in Topic Box 9.1. Values of one of these coefficients for the two-factor model of the KABC-I in Figure 9.1 are reported in the box.

For CFA models fitted to data from a single sample, the choice between analyzing factors in unstandardized versus standardized form (e.g., Figure 6.1) usually has no impact on model fit. Steiger (2002) describes an exception called **constraint interac-**

TOPIC BOX 9.1

Reliability of Construct Measurement

Raykov (1997, 2004) describes coefficients that estimate the reliability of factor (construct) measurement. This coefficient is the **factor rho coefficient**, which is a ratio of explained variance over total variance that can be expressed in terms of CFA parameters. It can also be computed for factors in SR models. For factors with no error covariances that involve their indicators (i.e., uncorrelated measurement errors), the rho coefficient is estimated in the unstandardized solution as follows:

$$\hat{\rho}_{X_i X_i} = \frac{\left(\Sigma\hat{\lambda}_i\right)^2 \hat{\phi}}{\left(\Sigma\hat{\lambda}_i\right)^2 \hat{\phi} + \Sigma\hat{\theta}_{ii}} \qquad (9.1)$$

where $\Sigma\hat{\lambda}_i$ is the sum of the estimated unstandardized factor loadings among indicators of the same factor, $\hat{\phi}$ is the estimated factor variance, and $\Sigma\hat{\theta}_{ii}$ is the sum of the unstandardized error variances of those indicators. A different formula is needed for factors with indicators that share at least one error covariance:

$$\hat{\rho}_{X_i X_i} = \frac{\left(\Sigma\hat{\lambda}_i\right)^2 \hat{\phi}}{\left(\Sigma\hat{\lambda}_i\right)^2 \hat{\phi} + \Sigma\hat{\theta}_{ii} + 2\Sigma\hat{\theta}_{ij}} \qquad (9.2)$$

where $\Sigma\hat{\theta}_{ij}$ is the sum of the nonzero unstandardized error covariances. Raykov (2004) describes variations of these equations for the standardized solution.

Calculation of $\hat{\rho}_{X_i X_i}$ for the sequential processing construct of the two-factor model in Figure 9.1 is demonstrated next. The errors of the three indicators of this factor are independent, so we need Equation 9.1. From Table 9.2 we obtain these numerical results: The unstandardized factor loadings are 1.000, 1.147, and 1.388. The unstandardized error variances are 8.644, 1.998, and 2.902; the estimated variance of the sequential factor is 2.839; the sum of the factor loadings is 3.535; and the sum of the error variances is 13.564. Given these totals, the estimated reliability for measurement of the sequential processing factor is

$$\hat{\rho}_{X_i X_i} = [3.535^2\ (2.839)]/[3.535^2\ (2.839) + 13.564] = .723$$

which is not a terrible result, but still the evidence for convergent validity among the indicators of this factor is questionable (see Table 9.2). The estimated reliability for measurement of the simultaneous processing factor by its five indicators (Figure 9.1) is somewhat higher, $\hat{\rho}_{X_i X_i} = .786$. See Hancock and Mueller (2001) for information about other factor reliability coefficients; Byrne (2006) describes factor reliability coefficients printed by EQS.

tion that can occur for CFA models where some factors have only two indicators *and* a **cross-factor equality constraint** is imposed on the loadings of indicators on different factors. In some cases the value of $\chi^2_D(1)$ for the test of the equality constraint depends on how the factors are scaled. Constraint interaction probably does not occur in most applications of CFA, but you should know something about this phenomenon in case it ever crops up in your own work. See Appendix 9.B for more information.

Some other kinds of tests with CFA models are briefly described. Whether a set of indicators is congeneric, tau-equivalent, or parallel can be tested in CFA by comparing hierarchical models with the chi-square difference test (Chapter 8). **Congeneric indicators** measure the same construct but not necessarily to the same degree. The CFA model for congenerity does not impose any constraints except that a set of indicators is specified to load on the same factor. If this model fits reasonably well, one can proceed to test the more demanding assumptions of tau equivalence and parallelism. **Tau-equivalent indicators** are congeneric and have equal true score variances. This hypothesis is tested by imposing equality constraints on the unstandardized factor loadings (i.e., they are all fixed to 1.0). If the fit of the tau equivalence model is not appreciably worse than that of the congenerity model, then additional constraints can be imposed that test for parallelism. Specifically, **parallel indicators** have equal error variances. If the fit of this model with equality-constrained residuals is not appreciably worse than that of the model for tau equivalence, the indicators may be parallel. All these models assume independent errors and must be fitted to a covariance matrix, not a correlation matrix; see Brown (2006, pp. 238–252) for examples.

It was noted earlier that fixing all factor correlations to 1.0 in a multifactor model generates a single-factor model that is nested under the original. In the factor analysis literature, the comparison with the chi-square difference test just described is referred to as the **test for redundancy**. A variation is to fix the covariances between multiple factors to zero, which provides a **test for orthogonality**. If the model has only two factors, this procedure is not necessary because the statistical test of the factor covariance in the unconstrained model provides the same information. For models with three or more factors, the test for orthogonality is akin to a multivariate test for whether all the factor covariances together differ statistically from zero. Note that each factor should have at least three indicators for the redundancy test; otherwise, the constrained model may not be identified; see Nunnally and Bernstein (1994, pp. 576–578) for examples.

Remember that estimates of equality-constrained factor loadings are equal in the unstandardized solution, but the corresponding standardized coefficients are typically unequal. This will happen when the two indicators have different variances. *Thus, it usually makes no sense to compare standardized coefficients from equality-constrained factor loadings.* If it is really necessary to constrain a pair of standardized loadings to be equal, then one option is to fit the model to a correlation matrix using the method of constrained estimation (Chapter 7).

ITEMS AS INDICATORS AND OTHER METHODS FOR ANALYZING ITEMS

There are examples of "successful" CFA analyses where the indicators are Likert scale items instead of scales with continuous total scores (e.g., Harris, 1995), but there are potential problems. One is that default ML estimation is not generally appropriate for Likert-type items, which are ordinal variables. Some special methods for ordinal indicators were described in Chapter 7, including robust WLS estimation. These special methods can be more difficult to apply than ML estimation.

Another problem is that item-level data tend to be "noisy." Specifically, people's responses to individual items may be unstable, so item reliabilities can be low. Items in exploratory factor analysis (EFA) often have relatively high secondary loadings (e.g., about .30) on factors other than the one on which they have primary loadings (e.g., > .50). Secondary loadings in EFA often account for relatively high proportions of the variance, so constraining them to zero in CFA may be too conservative. Consequently, the more restrictive CFA model may not fit the data. This is one reason the specification of a CFA model based on EFA outcomes and analyzed with the same data may lead to the rejection of the CFA model (van Prooijen & van der Kloot, 2001). That is, CFA does not generally "confirm" the results of EFA.

An alternative to analyzing items as indicators with special estimators is to analyze parcels with a normal theory method, such as ML. Recall that (1) a parcel is a total score across a set of homogeneous items and (2) parceling is controversial because it requires items that are unidimensional for each parcel. If this assumption is not tenable, then the results may be misleading (Chapter 7).

In some situations, other statistical methods for item-level analyses are better alternatives than CFA. When constructing a scale, the derivation of classical items statistics, such as item-total correlations and item difficulties (the proportion of respondents who responded correctly), with procedures in general statistical programs for analyzing scales, such as the Reliability procedure in SPSS, offers more flexibility. This is also true for EFA, which analyzes unrestricted models where each item is allowed to load on every factor. A more sophisticated alternative is the generation of **item characteristic curves** (ICC) according to **item response theory** (IRT). Briefly, the analysis of ICC yields detailed estimates about characteristics of individual items, including their difficulty, discrimination (i.e., the degree to which an item discriminates among persons in different regions on a latent variable), and susceptibility to guessing. It is also assumed in IRT that relations between items and factors as represented by the ICC are nonlinear. For example, the probability of correctly answering a particular item may be slight for low-ability examinees but increases geometrically at increasingly higher levels of ability before leveling off. In contrast, CFA assumes linear associations between indicators (items in this case) and underlying factors. The IRT method is also oriented toward the development of **tailored tests**, subsets of items that may optimally assess a particular person based on the correctness of their previous responses. If the examinee fails initial items, for instance, then the computer presents easier ones. Testing stops when more difficult items are consistently failed. See Reise, Widaman, and Pugh (1993) for a comparison of CFA and IRT for item-level analyses. Noar (2007) considers the role of SEM in test development, and Kamata and Bauer (2008) compare the specification of two-parameter IRT models and factor analysis models for dichotomous items. The use of IRT/ICC analysis as an alternative to CFA for estimating measurement invariance at the item level is considered later.

ESTIMATED FACTOR SCORES

When raw data are analyzed, it is possible to calculate factor scores for each case. Because factors are measured not directly but instead through their indicators, such scores are only estimates of the cases' relative standings on the factor. There is more than one way to calculate factor scores, however, and although scores derived using different methods tend to be highly correlated, they generally do not all yield identical rank orderings of the cases. For example, given structure coefficients, multiple regression (MR) can be used to derive estimated factor scores that are weighted combinations of the indicators and the factor. The weights derived in MR are those that lead to the closest correspondence between the underlying factor(s) and the estimated factor scores. An alternative to empirically derived weights is simply to add the scores for each case across the indicators, which weights each variable equally. The application of equal weights is called **unit weighting**. This method has the advantage of simplicity and less susceptibility to sample-specific variation, but unit weights may not be optimal ones within a particular sample. Given that there is more than one way to derive estimated factor scores, Bollen's (1989) perspective on this matter is relevant: researchers should probably refrain from making too fine a comparison on estimated factor scores.

EQUIVALENT CFA MODELS

There are two sets of principles for generating equivalent CFA models—one for models with multiple factors and another for single-factor models. As an example of the former, consider the two-factor model of self-perception of ability and achievement by Kenny (1979) presented in Figure 9.2(a) without measurement errors to save space. I used the method of constrained ML estimation in the SEPATH module of STATISTICA 9 Advanced to fit this model to the correlation matrix reported in a sample of 556 Grade 8 students that is presented in Table 9.7. Values of selected fit statistics indicate acceptable overall model fit:

$$\chi^2_M(8) = 9.256, \quad p = .321$$
$$\text{RMSEA} = .012\ (.017–.054)$$
$$\text{GFI} = .994; \quad \text{CFI} = .999; \quad \text{SRMR} = .012$$

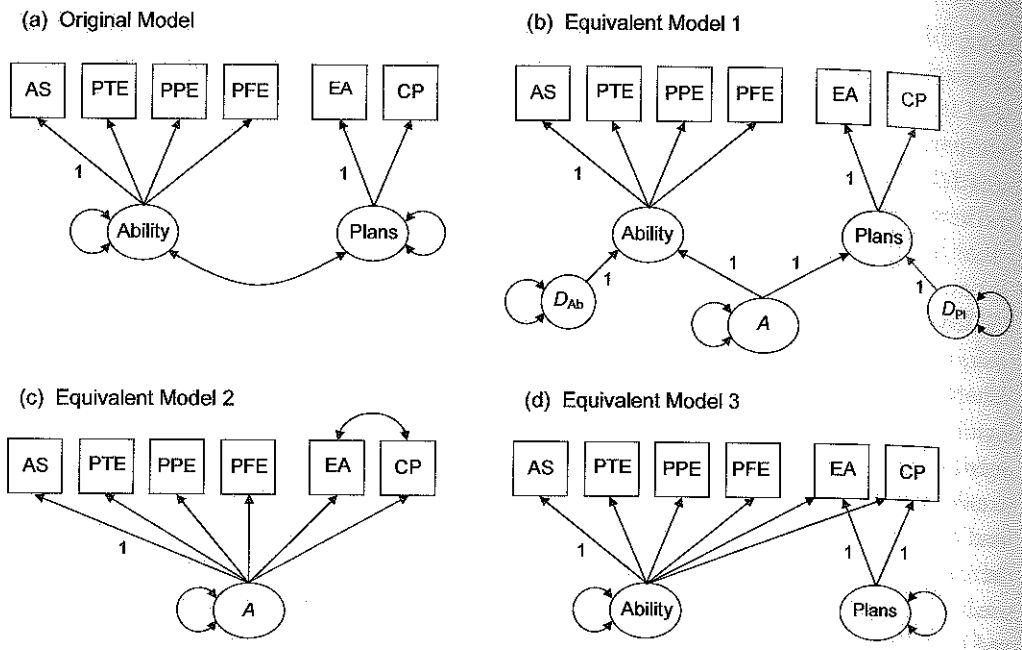

FIGURE 9.2. Four equivalent measurement models of self-perceived ability and educational plans. Measurement errors are omitted. The symbol for an unanalyzed association in (c) represents an error correlation between the corresponding pair of indicators. AS, Ability Self-Concept; PTE, Perceived Teacher Evaluation; PPE, Perceived Parental Evaluation; PFE, Perceived Friends' Evaluation; EA, Educational Aspiration; CP, College Plans.

The other three CFA models presented in Figure 9.2 are equivalent versions of the original model that yield the same values of fit statistics and predicted correlations. The equivalent model of Figure 9.2(b) is a hierarchical CFA model in which the unanalyzed association between the factors of the original model is replaced by a second-order factor (*A*), which has no indicators and is presumed to have direct effects on the first-order factors (ability, plans). This specification provides a specific account of *why* the two lower-order factors (which are endogenous in this model) covary. Because the second-

TABLE 9.7. Input Data (Correlations) for Analysis of Two-Factor Model of Perceived Ability and Educational Plans

Variable	1	2	3	4	5	6
1. Ability Self-Concept	1.00					
2. Perceived Parental Evaluation	.73	1.00				
3. Perceived Teacher Evaluation	.70	.68	1.00			
4. Perceived Friends' Evaluation	.58	.61	.57	1.00		
5. Education Aspiration	.46	.43	.40	.37	1.00	
6. College Plans	.56	.52	.48	.41	.71	1.00

Note. Input data are from Kenny (1979); $N = 556$.

order factor has only two indicators, it is necessary to constrain its direct effects on the first-order factors to be equal; that is:

$$A \rightarrow \text{Ability} = A \rightarrow \text{Plans} = 1.0$$

The other two equivalent versions are unique to models wherein some factors have only two indicators. The equivalent model in Figure 9.2(c) features the substitution of the plans factor with a correlation between the measurement error of its indicators. The equivalent model in Figure 9.2(d) features replacement of the correlation between the ability and plans factor with the specification that some indicators are multidimensional. Although the factors are assumed to be orthogonal in this model, all six indicators have loadings on a common factor, which explains the sample correlations just as well as the original model. Note that because the factors are specified as independent in the model of Figure 9.2(d), it is necessary to constrain the factor loadings of the educational aspiration and college plans indicators to be equal in order to identify this model.

For two reasons, the situation regarding equivalent versions of CFA models with multiple factors is even more complex than suggested by the last example. First, it is possible to apply the Lee–Hershberger replacing rules (Chapter 8) to substitute factor covariances (unanalyzed associations) with direct effects, which makes some factors endogenous. The resulting model is not a CFA model. It is an SR model, but it will fit the data equally well. For example, substitution of the factor covariance Ability ↖↗ Plans in the original model of Figure 9.2(a) with the direct effect Ability → Plans generates an equivalent SR model. Second, Raykov and Marcoulides (2001) show that there is actually a set of infinitely many equivalent models for standard multifactor CFA models. For each equivalent model in this set, the factor covariances are eliminated (orthogonality is specified) and replaced by one or more factors not represented in the original model with fixed unit loadings (1.0) on all indicators. These models with additional factors explain the data just as well as the original.

Equivalent versions of single-factor CFA models can be derived using Hershberger's (1994) **reversed indicator rule**, which involves the specification of one of the observed variables as a cause (formative) indicator while the rest remain as effect (reflective) indicators. Consider the hypothetical single-factor model of reading presented in Figure 9.3(a). The effect indicators represent different tasks, including word recognition, word attack, and phonics skills. An equivalent version is presented in Figure 9.3(b), and it features phonics skill as a cause of reading. Note that the factor in this equivalent model is no longer exogenous: because a casually prior variable (phonics skill) has a direct effect on it, the factor here is endogenous and thus has a disturbance. Also, the phonics skill indicator is exogenous in Figure 9.3(b). Thus, this equivalent model is actually an SR model. A total of three other equivalent models could potentially be generated, one with each of the remaining indicators specified as causes, Not all of these equivalent versions may be theoretically plausible, but at least the one with phonics skill as a cause indicator is logical (e.g., Wagner, Torgeson, & Rashotte, 1994).

The factor in Figure 9.3(b) is an example of a **multiple indicators and multiple**

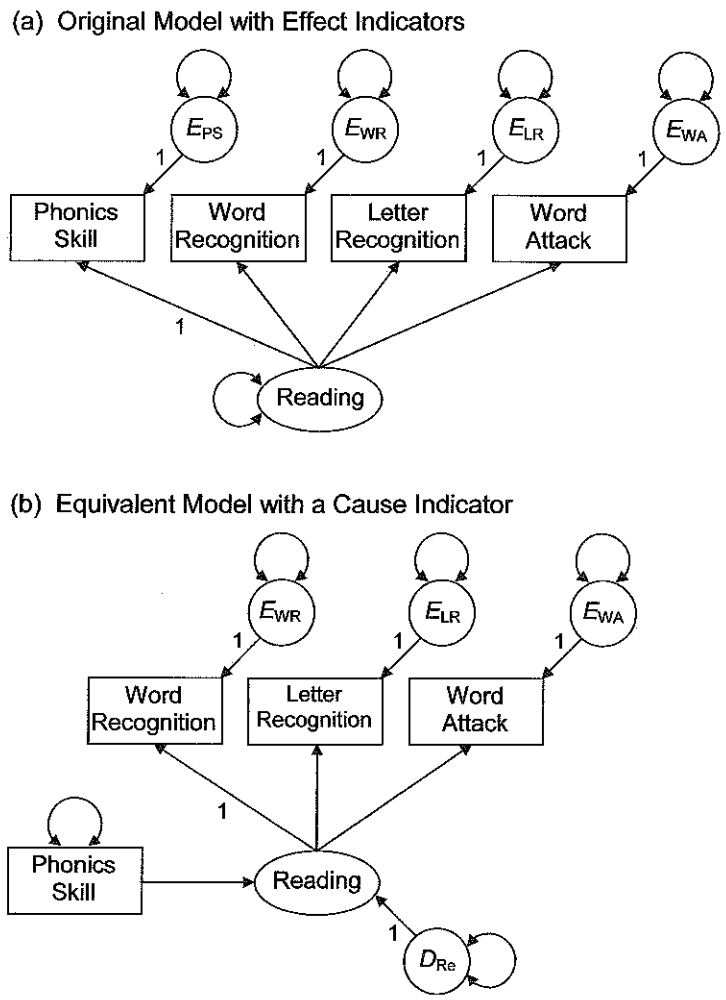

FIGURE 9.3. Application of the reversed indicator rule to generate an equivalent one-factor model of reading.

causes (MIMIC) factor. A MIMIC factor has both cause indicators and effect indicators, and they can be continuous as in the previous example or categorical. A categorical cause indicator represents group membership. We will see in Chapter 11 that a MIMIC model with a cause indicator is a special case of the SEM approach to estimating group differences on latent variables.

HIERARCHICAL CFA MODELS

It is possible to represent hypotheses about hierarchical relations among constructs through the specification of higher-order factors with presumed direct causal effects on

lower-order factors. For example, the hierarchical CFA model in Figure 9.4 represents the hypotheses that (1) indicators X_1–X_3 measure verbal ability, X_4–X_6 reflect visual-spatial ability, and X_7–X_9 depend on memory ability; and (2) each of these **first-order factors** has two direct causes. One is a **second-order factor**, which represents a general ability construct (g) with no indicators. This is because second-order factors are measured indirectly through the indicators of the first-order factors. The specification of g as a common cause of the first-order factors implies that associations between the latter are spurious. The other presumed direct cause of each first-order factor is a disturbance, which represents factor variance not explained by g. Thus, the disturbances and g are exogenous, but the first-order factors are endogenous in Figure 9.4.

To identify a hierarchical CFA model, there must be at least three first-order factors. Otherwise, the direct effects of the second-order factor on the first-order factors or the disturbance variances may be underidentified. Each first-order factor should have at least two indicators. The model in Figure 9.4 satisfies both of these requirements. There are two ways to scale the second-order factor g in the figure. One way is to fix any one of g's unstandardized direct effects on a first-order factor to 1.0. This tactic corresponds to the specification

$$g \rightarrow \text{Verbal Ability} = 1.0$$

in Figure 9.4. A second option is to fix the variance of g to 1.0 (standardize it). This approach leaves all three direct effects of g on the first-order factors as free parameters.

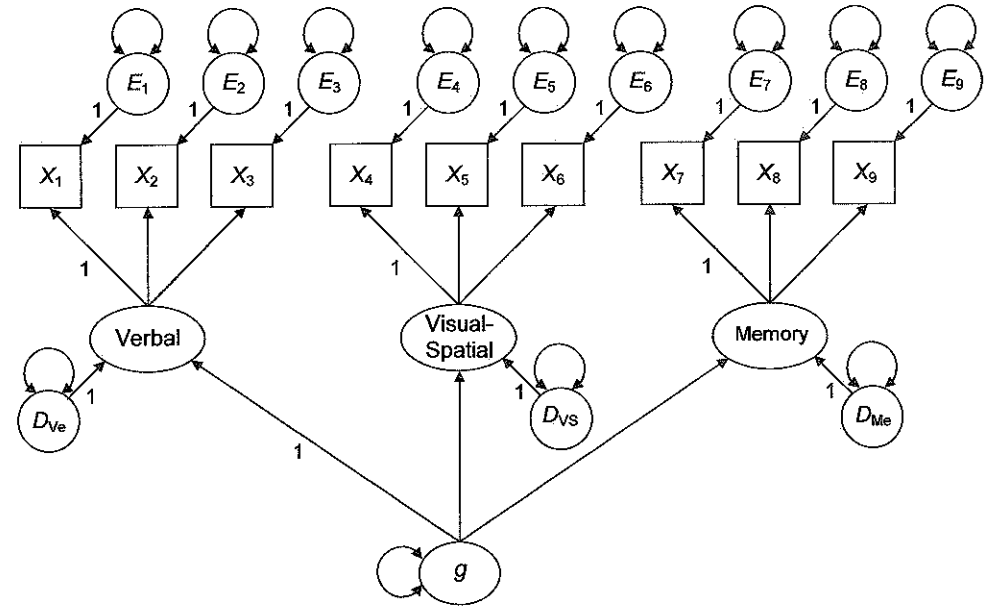

FIGURE 9.4. A hierarchical confirmatory factor analysis model of the structure of cognitive ability.

constraints on an initially restricted model, such as one represented by $H_{\Lambda, \Theta}$ (equal loadings and error variances–covariances), are gradually released (e.g., next test H_{Λ} by allowing error variances–covariances to be freely estimated in each group). The goal of both approaches is the same: find the most restricted model that still fits the data and respects theory. That theory may dictate which hypothesis testing approach, model trimming or building, is best.

Cheung and Rensvold (2002) remind us that the chi-square difference test is affected by overall sample size. In invariance testing with very large samples, this means that χ^2_D could be statistically significant, even though the absolute differences in parameter estimates are of trivial magnitude. That is, the outcome of the chi-square difference test could indicate the lack of measurement invariance when the imposition of cross-group equality constraints makes relatively little difference in model fit. One way to detect this outcome is to compare the unstandardized parameter estimates across the two solutions. Another is to inspect changes in values of approximate fit indexes, but there are few guidelines for doing so in invariance testing. In two-group computer simulation analyses, Cheung and Rensvold (2002) studied the characteristics of changes in the values of 20 different approximate fit indexes when invariance constraints were added. Changes in most indexes were affected by model characteristics, including the number of factors or the number of indicators per factor. That is, model size and complexity were generally confounded with changes in approximate fit indexes. An exception is the Bentler CFI, for which Cheung and Rensvold (2002) suggested that change in CFI values less than or equal to .01 (i.e., $\Delta CFI \leq .01$) indicate that the null hypothesis of invariance should *not* be rejected. Of course, this suggested threshold is not a golden rule, nor should it be treated as such. Specifically, it is unknown whether this rule of thumb would generalize to other models or data sets not directly studied by Cheung and Rensvold (2002). A second approximate fit index that performed relatively well in Cheung and Rensvold's (2002) simulations is McDonald's (1989) **noncentrality index** (NCI).[2]

Meade, Johnson, and Braddy (2008) extended the work of Cheung and Rensvold (2002) by studying the performance of several approximate fit indexes in generated data with different levels of lack of measurement invariance, from trivial to severe. Types of lack of measurement invariance studied by Meade et al. (2008) included different factor structures (forms), factor loadings, and indicator intercepts across two groups. In very large samples studied by Meade et al. (2008), such as $n = 6{,}400$ per group, the χ^2_D statistic indicated lack of measurement invariance most of the time when there were just slight differences in measurement model parameters across the groups. In contrast, values of approximate fit indexes were generally less affected by group size and also by the number of factors and indicator than the chi-square difference test in large samples. The Bentler CFI was among the best performing approximate fit indexes along with the McDonald NCI. Based on their results, Meade et al. (2008) suggested that change in CFI values less than or equal to .002 (i.e., $\Delta CFI \leq .002$) may indicate that deviations from perfect measurement invariance are functionally trivial. These authors also provide a table of values for changes in the NCI that vary depending on the number of factors and indicators (Meade et al., 2008, p. 586). Again, these suggested thresholds are not golden rules, but results by Cheung and Rensvold (2002) and Meade et al. (2008) indicate that researchers working with very large samples should look more to approximate fit indexes than statistical tests to establish measurement invariance.

[2]$NCI = \exp[-\frac{1}{2}(\chi^2_M - df_M) / N]$ where "exp" is the exponential function e^x and e is the natural base, approximately 2.71828. The range of the NCI is 0–1.0 where 1.0 indicates the best fit. Mulaik (2009) notes that values of the NCI tend to drop off quickly from 1.0 with small increases in lack of fit.

Empirical Example

Sabatelli and Bartle–Haring (2003) administered to each spouse in a total of 103 married heterosexual couples three indicators of family-of-origin experiences (FOE) and two indicators of marital adjustment. The indicators of FOE are retrospective measures of the perceived quality of each spouse's relationship with his or her own father or mother and of the relationship between the parents while growing up. The marital adjustment indicators are ratings of problems and intimacy in the marital relationship. Higher scores on all variables indicate more positive reports of FOE or marital adjustment. Presented in Table 9.8 are descriptive statistics for these variables in the samples of husbands and wives. Note that means are reported in the table, but they are not analyzed here.[3]

TABLE 9.8. Input Data (Correlations, Standard Deviations) for a Two-Factor Model of Family-of-Origin Experiences and Marital Adjustment Analyzed across Samples of Husbands and Wives

							Husbands	
Variable		1	2	3	4	5	*M*	*SD*
Marital adjustment indicators								
1. Problems		—	.658	.288	.171	.264	155.547	31.168
2. Intimacy		.740	—	.398	.295	.305	137.971	20.094
Family-of-origin experiences indicators								
3. Father		.265	.422	—	.480	.554	82.764	11.229
4. Mother		.305	.401	.791	—	.422	85.494	11.743
5. Father–Mother		.315	.351	.662	.587	—	81.003	13.220
Wives	*M*	161.779	138.382	86.229	86.392	85.046		
	SD	32.936	22.749	13.390	13.679	14.382		

Note. These data are from S. Bartle-Haring (personal communication, June 3, 2003); $n_1 = 103$ husbands (above diagonal), $n_2 = 103$ wives (below diagonal). Means are reported but not analyzed for the model in Figure 9.7, but means are analyzed for the model in Figure 11.5.

[3]It could be argued that the samples in this analysis—husbands and wives—are not really independent groups because each spousal pair is "linked" across the two samples. An alternative way to view this data set is that individuals are nested under pairs (couples); that is, the data are hierarchical and thus amenable to a multilevel analysis. This possibility is not pursued in this pedagogical example.

Scaling Factors in Multiple-Sample Analyses

The two factor, five-indicator model for this example is presented in Figure 9.7. The best way to scale the factors in a multiple-sample analysis is to select the same reference variable for each factor in each group. Here, the unstandardized loadings of the father indicator and the problems indicator were fixed to 1.0 in order to scale their respective factors in both samples. However, there are two potential complications: First, loadings fixed to 1.0 in both groups cannot be tested for statistical significance. The second complication follows from the first: because fixed loadings are excluded from tests of measurement invariance, it must be assumed a priori that the reference variables measure their factors equally well over groups. This assumption means that if the researcher decides to fix the loading of an indicator that is not metric invariant across the groups, then the subsequent results may be inaccurate. One way to address this dilemma is to reanalyze the model after fixing the loadings of other indicators to 1.0. If the unstandardized factor loadings that were originally fixed are comparable in the new analysis in which they are free parameters, then that indicator may be metric invariant. See Reise et al. (1993) for more information about factor scaling when testing for measurement invariance. Little, Slegers, and Card (2006) describe a method to scale factors in a multiple-group analysis that involves neither the arbitrary selection of a reference variable nor the standardization of factors. This method may be specially well suited to applications of CFA where group differences on factor means (i.e., the model has both a covariance structure and a mean structure) are also estimated (Chapter 11).

Invariance Testing

With five indicators in each of two samples, there are a total of 5(6)/2 × 2, or 30 observation for the analysis. Because the samples consist of married couples who share many experiences, the initial model assumed a strict form of invariance—one that corresponds

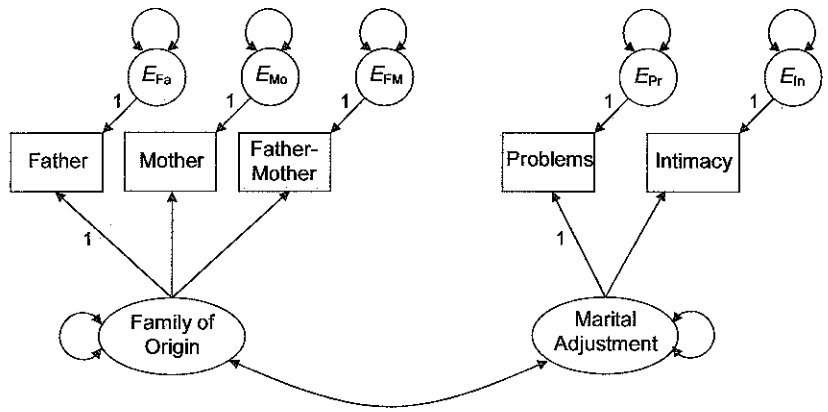

FIGURE 9.7. A measurement model of family-of-origin experiences and marital adjustment evaluated across samples of husbands and wives.

to $H_{\Lambda, \Phi, \Theta}$, or equivalence of factors loadings, factor variances–covariance, and error variances–covariances for husbands and wives. This means that cross-group equality constraints were imposed on the estimates of three factor loadings (those not already fixed to 1.0), seven variances (of two factors and five measurement errors), and one factor covariance (see Figure 9.7). There are no error covariances in the initial model, so it is assumed that all of these values are zero in both samples. Because only one estimate of each free parameter was required when equality was assumed across the samples, a total of 11 parameters require estimates across both samples, so $df_M = 30 - 11 = 19$.

I used the ML method of EQS 6.1 to simultaneously fit the model of Figure 9.7 with cross-group equality constraints to the covariance matrices for husbands and wives based on the data in Table 9.8. The program printed this warning:

```
Do not trust this output
Iterative process has not converged
Maximum number of iterations was reached
30 iterations have been completed and the program stopped
```

That is, a converged solution was not reached after 30 iterations, the default limit in EQS. In a second run with EQS, I increased its iteration limit to 100. In this second analysis, EQS generated a converged and admissible solution. Reported in Table 9.9 are values of selected fit statistics for the test of $H_{\Lambda, \Phi, \Theta}$. Because the group sizes in this analysis are not large (n = 103), we focus on the chi-square difference test when comparing nested models. To summarize, the initial model passes the chi-square test (χ^2_M (19) = 23.190, p = .229), so the hypothesis of exact fit is not rejected. Values of some approximate fit indexes seem favorable (GFI = .959, CFI = .990), but the upper bound of the 90% confidence interval based on the RMSEA (.103) just exceeds .10. The result for the SRMR, .127, is not favorable (Table 9.9). Across both samples, there were a total of 16 absolute correlation residuals > .10 (9 for husbands, 7 for wives). This is a terrible result; therefore, the initial model is rejected.

TABLE 9.9. Values of Selected Fit Statistics for Hypotheses about Measurement Invariance for a Two-Factor Model of Family-of-Origin Experiences and Marital Adjustment Analyzed across Samples of Husbands and Wives

Hypothesis	χ^2_M	df_M	χ^2_D	df_D	RMSEA (90% CI)	GFI	CFI	SRMR
$H_{\Lambda, \Phi, \Theta}$	23.190[a]	19	—	—	.047 (0–.103)	.959	.990	.127
$H_{\Lambda, \Theta}$	16.127[b]	16	7.063[c]	3	0 (0–.092)	.970	.999	.037
$H_{\Lambda, \Theta}$ except $E_{Fa} \curvearrowright E_{Mo}$ in both groups	7.097[d]	14	9.030[e]	2	0 (0–.028)	.987	1.000	.026

Note. CI, confidence interval; $H_{\Lambda, \Phi, \Theta}$, equal loadings, factor variances–covariances, and measurement error variances–covariances.

[a]p = .229; [b]p = .444; [c]p = .070; [d]p = .931; [e]p = .011.

In the next analysis, the factor variances–covariance for the model in Figure 9.7 were freely estimated in each sample (i.e., the corresponding cross-group equality constraints were dropped). This respecified model corresponds to the invariance hypothesis $H_{\Lambda, \Theta}$, which assumes equal factor loadings and measurement error variances only. This second analysis converged to an admissible solution, and values of selected fit statistics are reported in Table 9.9. The second model passes the chi-square test (χ^2_M (16) = 16.127, p = .444), and the improvement in overall fit due to dropping the equality constraint on the factor variances-covariance is almost statistically significant (χ^2_D (3) = 7.063, p = .070). The value of the SRMR is better for the second model (.037) compared with that of the original model (.127). The largest absolute correlations are −.094 for husbands and .066 for wives, both for the association between the father and mother indicators of the FOE factor. The only statistically significant modification indexes in both samples were for the error covariances between the indicators just mentioned: husbands: χ^2 (1) = 7.785, p < .01; wives: χ^2 (1) = 7.959, p < .01.

Because it is plausible that reports about quality of relationships with one's parents may have common omitted causes, the third CFA model was respecified so that the error covariances between the father and mother indicators of the FOE factor ($E_{Fa} \leftrightarrow E_{Mo}$; Figure 9.7) were freely estimated in each sample. Values of selected fit statistics for this third model are reported in Table 9.9, and their values are generally favorable. For example, the improvement in overall fit compared with the second model without error covariances is statistically significant (χ^2_D (2) = 9.030, p = .011), and values of approximate fit indexes are generally good (e.g., RMSEA = 0). Furthermore, all absolute correlation residuals in both samples are < .10.

Based on these results, the third CFA model was retained as the final measurement model. To summarize, this model assumes that all factor loadings and measurement error variances are equal for husbands and wives. In contrast, the factor variances and covariance and the error covariance between the father and mother indicators were freely estimated in each sample. Overall, it seems that the five indicators represented in Figure 9.7 measure the same two factors in similar ways for both husbands and wives. You can download from the website for this book (see p. 3) the EQS syntax and output files for this analysis. Computer files for the same analysis but in LISREL and Mplus are also available for download from the site, too.

Parameter Estimates

Reported in the top part of Table 9.10 are ML parameter estimates for the final measurement model that were freely estimated in each sample. Wives may be somewhat more variable than husbands on both factors. For example, the estimated variance of the marital adjustment factor is 583.685 among wives but 452.140 among husbands. Although the estimated factor covariance is also somewhat greater for wives than for husbands (139.534 vs. 93.067, respectively), the estimated factor correlation in both samples is about .50. These correlations are consistent with discriminant validity in factor measurement because their values are not too high. Although neither error covariance between the father and mother indicators of the FOE factor is statistically significant for husbands or wives, their values have opposite signs, negative for husbands (−12.617) but positive for wives (16.351).

Reported in the bottom part of Table 9.10 are estimates for parameters of the measurement model constrained to have equal unstandardized values across the samples. Because the sizes of the groups are the same (n = 103), the standard errors of these estimates are also equal for husbands and wives. The pattern of standardized factor loadings is generally similar within each sample and consistent with convergent validity in factor measurement. Note in the table that, although the unstandardized factor loadings are equal for every indicator across the two samples, such as .885 for the mother indicator of the FOE factor, the corresponding standardized factor loadings are not equal. For example, the standardized loading of the mother indicator is .698 for husbands and .779 for wives (Table 9.10). This pattern is expected because EQS derives standardized estimates based on the separate variances and covariances within each group. If the groups

TABLE 9.10. Maximum Likelihood Parameter Estimates for a Two-Factor Model of Family-of-Origin Experiences and Marital Adjustment Analyzed across Samples of Husbands and Wives

	Husbands			Wives		
Parameter	Unst.	*SE*	St.	Unst.	*SE*	St.
			Unconstrained estimates			
Factor variances and covariance						
FOE	87.896	21.438	1.000	143.102	30.412	1.000
Mar Adj	452.140	105.126	1.000	583.685	146.837	1.000
FOE ↔ Mar Adj	93.067	27.853	.467	139.534	40.774	.483
Measurement error covariance						
$E_{Fa} \leftrightarrow E_{Mo}$	−12.617[a]	15.364	−.246	16.351[a]	15.634	.319
			Equality-constrained estimates			
Factor loadings						
Mar Adj → Probs	1.000[b]	—	.685	1.000[b]	—	.730
Mar Adj → Intim	.933	.146	.988	.933	.146	.991
FOE → Father	1.000[b]	—	.841	1.000[b]	—	.893
FOE → Mother	.885	.079	.698	.885	.079	.779
FOE → Fa-Mo	.897	.143	.648	.897	.143	.735
Measurement error variances						
E_{Pr}	510.199	88.407	.530	510.199	88.407	.466
E_{In}	9.687[a]	63.179	.024	9.687[a]	63.179	.019
E_{Fa}	36.249[c]	16.928	.291	36.249[c]	16.928	.202
E_{Mo}	72.411	16.533	.513	72.411	16.533	.392
$E_{Fa\text{-}Mo}$	97.868	16.264	.580	97.868	16.264	.459

Note. Unst., unstandardized; St., standardized; FOE, family-of-origin experiences. Standardized estimates for measurement errors are proportions of unexplained variance.

[a]$p \geq .05$; [b]Not tested for statistical significance; [c]$p < .05$; for all other unstandardized estimates, $p < .01$.

do not have the same variances and covariances (likely), then one cannot directly compare standardized estimates across the groups (Chapter 2).

Note that LISREL can optionally print up to *four* different standardized solutions in a multiple-sample analysis, including the *within-group standardized solution* and the *within-group completely standardized solution*. Both are derived from standardizing the within-group variances–covariance matrices except that only the factors are standardized in the former solution versus all variables in the latter solution. The third is LISREL's *common metric standardized solution* where the factors only are automatically scaled so that the weighted average of their covariance matrices across the samples is a correlation matrix. In contrast, all variables are so scaled in the fourth solution, the *common metric completely standardized solution*. The common metric standardized estimates may be more directly comparable across the groups than are the within-group standardized estimates, but the unstandardized estimates are still preferred for this purpose. Check the documentation of your SEM computer tool to find out how it calculates a standardized solution in a multiple-sample analysis. Raykov and Marcoulides (2000) describe a method for comparing completely standardized estimates across equal-size groups based on analyzing a correlation structure using the method of constrained estimation (Chapter 7).

Any type of structural equation model—path models, SR models, and so on—can be tested across multiple samples. The imposition of cross-group equality constraints on certain parameters allows for tests of group differences on these parameters, just as in testing for measurement invariance in CFA. In Chapter 11, I will show you how to compare two groups on latent variable means in SEM.

Alternative Methods for Item-Level Analysis of Measurement Invariance

The indicators in the empirical example just described are scales, not items. When the indicators are items instead of scales, however, IRT/ICC analysis may be a better alternative in some cases than CFA. Results of a recent computer simulation study by Meade and Lautenschlager (2004) are relevant in this regard. These authors studied the relative capabilities of CFA and IRT/ICC analysis to detect differential item functioning across groups in generated data sets for samples of three different sizes (N = 150, 500, and 1,000) and for a six-item scale that measured a single factor (i.e., unidimensional items). Neither CFA nor IRT/ICC analysis performed well in the smallest sample size, but these results were expected. In larger samples, the CFA technique was generally inadequate at detecting items with differences in discrimination parameters. The CFA methods were also generally unable to detect items with differences in difficulty parameters. In contrast, the IRT/ICC methods were generally better at detecting items with either type of differential functioning just mentioned. As noted by Meade and Lautenschlager (2004), however, the application of IRT/ICC methods to multiscale tests where different sets of items are assigned to different scales (i.e., multiple factors are measured) is problematic compared with CFA. This is because IRT/ICC methods provide no information about

covariances between factors, which may be of interest in testing for measurement invariance at the scale level. Accordingly, Meade and Lautenschlager (2004) suggested that both techniques could be applied in the same analysis: IRT/ICC methods for item-level analyses within each scale, and CFA methods for scale-level analyses, both of measurement invariance. Along similar lines, Stark, Chernyshenko, and Drasgow (2006) describe and test in computer simulations a common strategy for identifying differential item functioning using either IRT/ICC of CFA.

Power in Multiple-Sample CFA

The relatively small group sizes (n = 103) in this example analysis limits the statistical power to detect lack of measurement invariance. In a recent computer simulation study, Meade and Bauer (2007) found that the power of tests to detect group differences in factor loadings was uniformly low (e.g., < .40) when the group size was 100. In contrast, power was generally high when the group size was 400, but power estimates for an intermediate group size of 200 were highly variable. This is because the power of tests for measurement invariance is affected not just by sample size but also by model and data characteristics, including the number of indicators per factor and the magnitudes of factor intercorrelations. Accordingly, Meade and Bauer's (2007) results did not indicate a single rule of thumb regarding a ratio of group size to the number of indicators that could ensure adequate power to detect the absence of measurement invariance when the group size is not large. In any event, large group sizes are typically needed in order to have reasonable statistical power when testing for measurement invariance.

SUMMARY

Many types of hypotheses about measurement can be tested with standard CFA models. For example, the evaluation of a model with multiple factors that specifies unidimensional measurement provides specific tests of both convergent and discriminant validity. Respecification of a measurement model can be challenging because many possible changes could be made to a given model. Another problem is that of equivalent measurement models. The only way to deal with both of these challenges is to rely more on substantive knowledge than on statistical considerations in model evaluation. When analyzing structural equation models across multiple samples, it is common to impose cross-group equality constraints on certain unstandardized parameter estimates. In multiple-sample analyses, cross-group equality constraints are typically imposed to test hypotheses of measurement invariance. There are degrees of measurement invariance, but a common tactic is to constrain just the unstandardized factor loadings to be equal across the groups. If the fit of the measurement model with constrained factor loadings is much worse than that of the unconstrained model, then one may conclude that the indicators measure the factors in different ways across the groups.