Either means of scaling *g* in a single-sample analysis is probably fine. In a multiple-sample analysis, however, it is typically inappropriate to standardize factors. See Neuman, Bolin, and Briggs (2000), who analyzed a hierarchical model of cognitive ability similar to that represented in Figure 9.4 for a group-administered test.

MODELS FOR MULTITRAIT–MULTIMETHOD DATA

The method of CFA can also be used to analyze data from a **multitrait–multimethod** (MTMM) study, the logic of which was first articulated by Campbell and Fiske (1959). In an MTMM study, two or more traits are measured with two or more methods. Traits are hypothetical constructs that concern cognitive abilities, personality attributes, or other stable characteristics. Methods refer to multiple test forms, occasions, methods (e.g., self-report), or informants (e.g., parents) (Marsh & Grayson, 1995). The main goals are to (1) evaluate the convergent and discriminant validity of tests that vary in their measurement method and (2) derive separate estimates of the effects of traits versus methods on the observed scores.

The earliest procedure for analyzing data from an MTMM study involved inspection of the correlation matrix for all variables. For example, convergent validity would be indicated by the observation of high correlations among variables that supposedly measure the same trait but with different methods. If correlations among variables that should measure different traits but use the same methods are relatively high, then **common method effects** are indicated. This would imply that correlations among different variables based on the same method may be relatively high even if they measure unrelated traits.

The CFA method offers a more systematic way to analyze data from an MTMM study. When first applied to the problem in the 1970s, researchers typically specified CFA models like the one presented in Figure 9.5, a **correlated trait-correlated method** (CTCM) model. Such models have separate trait and method factors that are assumed to covary, but method factors are assumed to be independent of trait factors. In the figure, indicators X_1–X_3 are based on one method, X_4–X_6 are based on another method, and X_7–X_9 are based on a third method. This model also specifies that the set of indicators (X_1, X_4, X_7) measures one trait but that each of the other two sets, (X_2, X_5, X_8) and (X_3, X_6, X_9), measures different traits. Given these specifications, relatively high loadings on trait factors would suggest convergent validity, high loadings on method factors would indicate common method effects, and moderate correlations (not too high) between the factors would indicate discriminant validity.

There are reports of "successful" analyses of CTCM models (e.g., Villar, Luengo, Gómez-Fraguela, & Romero, 2006), but others have found that such analyses tend to yield inadmissible or unstable solutions. For example, Marsh and Bailey (1991) found in computer simulation studies that illogical estimates were derived about three-quarters of the time for CTCM models. Kenny and Kashy (1992) noted part of the problem: CTCM models are not identified if the loadings on the trait or method factors are equal. If the loadings are different but similar in value, then CTCM models may be empirically underidentified.

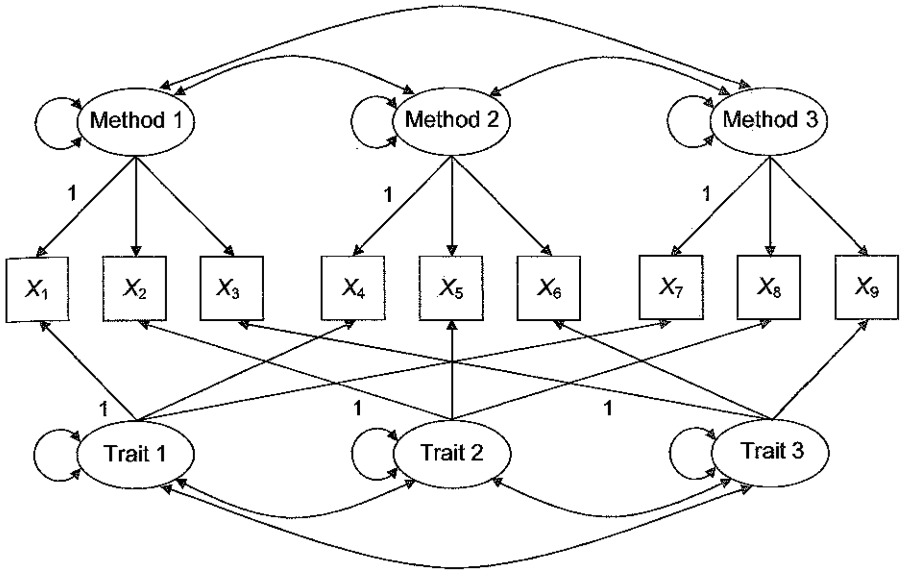

FIGURE 9.5. A correlated-trait correlated method (CTCM) model for multitrait–multimethod data. Measurement errors are omitted and assumed to be independent.

Some simpler alternatives to CTCM models have been proposed, including those with multiple but uncorrelated method factors, a single-method factor specified to affect all the indicators, and a model like the one in Figure 9.6, which is a **correlated uniqueness** (CU) model (Marsh & Grayson, 1995). This model has measurement error correlations among indicators based on the same method instead of separate method factors. That is, method effects are assumed to be a property of each indicator, and relatively high correlations among their residuals are taken as evidence for common method variance. Note that the similarity of methods for different traits is only one possible explanation for high measurement error correlations in CU models. Saris and Alberts (2003) evaluated alternative CFA models that could account for correlated residuals in CU models, including models that represented response biases, effects due to relative answers (when respondents compare their answers), and method effects. See Brown (2006, chap. 6) and Eid et al. (2008) for more information about MTMM analyses with CFA.

MEASUREMENT INVARIANCE AND MULTIPLE-SAMPLE CFA

Broadly defined, **measurement invariance (equivalence)** concerns whether scores from the operationalization of a construct have the same meaning under different conditions (Meade & Lautenschlager, 2004). These different conditions could involve consistency of measurement over populations, time of measurement, or methods of test administration

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

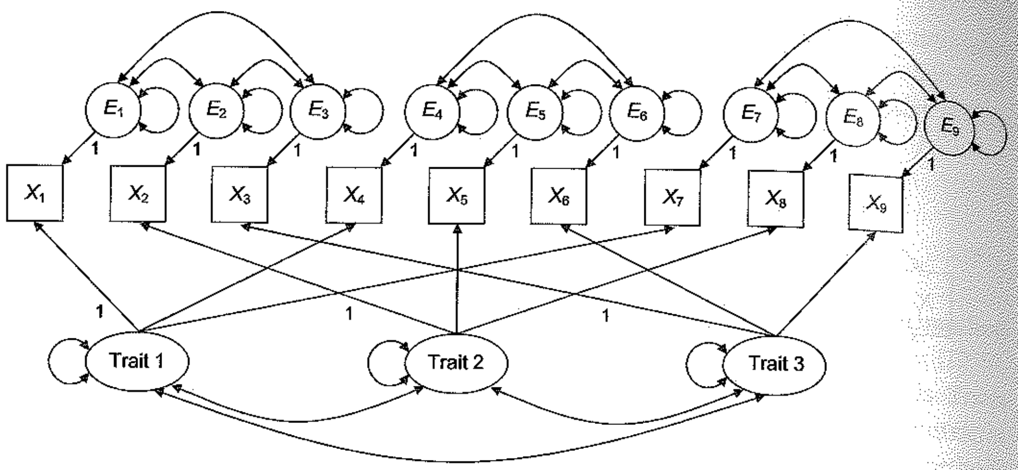

FIGURE 9.6. A correlated uniqueness (CU) model for multitrait–multimethod data.

(e.g., computer administered vs. paper-and-pencil format). Stability over time is referred to as **longitudinal measurement invariance**, and it concerns whether a set of indicators has the same factor structure across different occasions in a longitudinal design. If so, then measurement is invariant over time. Invariance over populations is related to the concept of **construct bias**, which implies that a test measures something different in one group (e.g., men) than in another (women). If not (i.e., there is no evidence for construct bias), then measurement is invariant over groups. The CFA technique is widely used to test hypotheses about measurement invariance over groups. Because the basic logic of invariance testing over groups is the same as for invariance testing over time or modes of test administration, only the former is described next. See Brown (2006, pp. 252–266) for an example of testing for longitudinal measurement invariance. See also Whitaker and McKinney (2007), who studied the invariance of job satisfaction ratings as a function of administration method (Internet vs. paper-and-pencil format) and respondent age and gender.

Testing Strategy

Hypotheses about measurement invariance over groups are tested in **multiple-sample CFA** where a measurement model is simultaneously fitted to the covariance matrices from at least two independent samples. The most basic form of measurement invariance is **configural invariance** or **equal form invariance**. It is tested by specifying the same measurement model across the groups. In this model, both the number of factors and the factor-indicator correspondence are the same, but all parameters are freely estimated within each sample. If this model does not fit the data, then measurement invariance does not hold at any level. Otherwise, the configural invariance hypothesis, H_{form}, is retained. If so, then the researcher could conclude that the same constructs are manifested in somewhat different ways in each group. These "different ways" refer to the

unstandardized factor loadings, which were freely estimated in each group. This means that if factor scores were calculated, a different weighing scheme would be applied to the indicators in each group.

A stronger form of measurement invariance is **construct-level metric invariance** or **equal factor loadings**, which means that the unstandardized factor loadings of each indicator are equal across the groups. If the equal factor loadings hypothesis, or H_Λ, is retained, then the researcher could conclude that the constructs are manifested the same way in each group. This implies that if factor scores were calculated, the same weighing scheme could be applied across all groups. The hypothesis H_Λ is tested by (1) imposing cross-group equality constraints on the factor loadings and (2) comparing with the chi-square difference test two hierarchical models, one that corresponds to H_Λ and the other corresponds to H_{form}, which was estimated with no equality constraints. This assumes that H_{form} was not rejected.

If χ^2_D for the comparison just described is not statistically significant, then the fit of the model with equality-constrained factor loadings is not appreciably worse than that of the model without these constraints. That is, H_Λ is retained. If so, the researcher can go on to test even stronger forms of measurement invariance, described momentarily. If H_Λ is rejected, though, the less strict hypothesis of **partial measurement invariance**, or H_λ, can be tested by releasing some, but not all, of the cross-group equality constraints on the unstandardized factor loadings. The goal is to locate the indicator(s) responsible for metric noninvariance at the construct level (Cheung & Rensvold, 2002). In subsequent analyses, the unstandardized loadings of these indicators are freely estimated in each sample, but the loadings of the remaining indicators are constrained to be equal across the groups. Indicators with appreciably different loadings across groups are **differential functioning indicators (items)**, and the pattern where some, but not all indicators have equal loadings in every group is **indicator-level metric invariance** (i.e., H_λ). The hypothesis of partial measurement invariance is tested by χ^2_D for the comparison of the less restricted model represented by H_{form} with the more restricted model represented by H_λ.

One can also test additional hypotheses about even stricter forms of invariance. The hypotheses described next all generally assume that H_Λ (equal factor loadings hypothesis) was not rejected. For example, the **equivalence of construct variances and covariances** hypothesis, or $H_{\Lambda,\,\Phi}$, assumes that the factor variances and covariances are equal across the groups. The **equivalence of residual variances and covariances** hypothesis, or $H_{\Lambda,\,\Theta}$, assumes that the measurement error variance for each indicator and all corresponding error covariances (if any) are equal across the groups. Each of these hypotheses is tested by comparing with χ^2_D the less restricted model implied by H_Λ with the more restricted model represented by $H_{\Lambda,\,\Phi}$ or $H_{\Lambda,\,\Theta}$. See Cheung and Rensvold (2002) for more information about measurement invariance hypotheses.

The testing strategy just outlined corresponds to model trimming where an initial unconstrained model (represented by H_{form}) is gradually restricted by adding constraints (e.g., next test H_Λ by constraining factor loadings to be equal across groups). It is also possible to test for measurement invariance through model building where 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., ΔCFI ≤ .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., ΔCFI ≤ .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.

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.

[2] NCI = exp[–½ ($\chi^2_M - df_M$) / 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.

[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 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 $df_M = 30 - 11 = 19$.

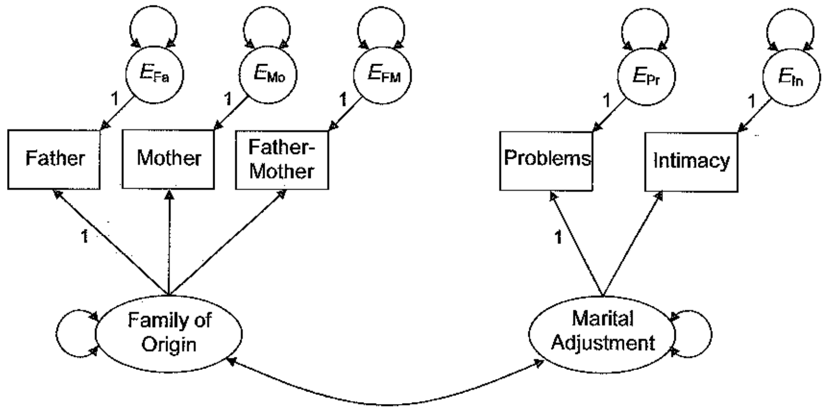

FIGURE 9.7. A measurement model of family-of-origin experiences and marital adjustment evaluated across samples of husbands and wives.

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 ($\chi^2_M(19) = 23.190$, $p = .229$), 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 $E_{Fa} \leftrightarrow E_{Mo}$ 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 ($\chi^2_M(16) = 16.127$, $p = .444$), and the improvement in overall fit due to dropping the equality constraint on the factor variances-covariance is almost statistically significant ($\chi^2_D(3) = 7.063$, $p = .070$). 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: $\chi^2(1) = 7.785$, $p < .01$; wives: $\chi^2(1) = 7.959$, $p < .01$.

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 ($E_{\text{Fa}} \leftrightarrow E_{\text{Mo}}$; 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 ($\chi^2_D(2) = 9.030$, $p = .011$), 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 $\leftrightarrow$ Mar Adj	93.067	27.853	.467	139.534	40.774	.483
Measurement error covariance						
$E_{\text{Fa}} \leftrightarrow E_{\text{Mo}}$	–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
$E_{\text{Fa-Mo}}$	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.

RECOMMENDED READINGS

The book by Brown (2006) is an excellent resource for CFA. It also includes many examples of Amos, CALIS, EQS, LISREL, and Mplus syntax for analyzing measurement models. The shorter work by Harrington (2009) is less technical and intended for social work researchers, but readers from other disciplines would be familiar with the substantive examples. The accessible presentation by Thompson (2004) deals with both EFA and CFA. Schmitt and Kuljanin (2008) describe issues in the evaluation of measurement invariance in the human resource management area.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford Press.

Harrington, D. (2009). *Confirmatory factor analysis.* New York: Oxford University Press.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and limitations. *Human Resource Management Review, 18,* 210–222.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

EXERCISES

1. Reproduce the values of the structure coefficients in Table 9.3 using the tracing rule for the model in Figure 9.1 and the parameter estimates in Table 9.2.
2. Use an SEM computer tool to derive the standardized residuals for the corresponding correlation residuals in Table 9.5 for the model in Figure 9.1 and the data in Table 9.1.
3. Show the calculation of $\hat{\rho}_{X_i X_i} = .786$ for the simultaneous processing factor in Figure 9.1 with the parameter estimates in Table 9.2. (See Topic Box 9.1.)
4. Evaluate the fit of a respecified version of the model in Figure 9.1 but with a direct effect from the simultaneous processing factor to the Hand Movements task against the data in Table 9.1.
5. Derive df_M for the hierarchical CFA model in Figure 9.4.
6. Use an SEM computer tool to test the hypothesis H_{form} for the model in Figure 9.7 with the data in Table 9.8; do not include any error covariances in this analysis. Look *carefully* through the output. What did you find?
7. Why would it be incorrect to scale the factors in a multiple-sample CFA by fixing their variances to 1.0 in all samples?

APPENDIX 9.A

Start Value Suggestions for Measurement Models

These recommendations concern measurement models, whether those models are CFA models or part of an SR model. Unstandardized variables, including the factors, are assumed. Initial estimates of factor variances should probably not exceed 90% of that of the observed (sample) variance for the corresponding reference variable. Start values for factor covariances follow the initial estimates of their variances. That is, they are the product of each factor's standard deviation (the square root of the initial estimates of their variances) and the expected correlation between them. If the indicators of the same factor have similar variances to that of the reference variable, then initial estimates of their factor loadings can also be 1.0. If the reference variable is, say, one-tenth as variable as another indicator of the same factor, the initial estimate of the other indicator's factor loading could be 10.0. Conservative start values for measurement error variances could be 90% of the observed variance of the associated indicator, which assumes that only 10% of the variance will be explained. Bentler (1995) suggests that it is probably better to overestimate the variances of exogenous variables than to underestimate them. This advice is also appropriate for Heywood cases of the type where a variance estimate is negative: in the reanalysis of the model, try a start value that is higher than that in the previous run.