In Chapters 4 and 8, we discussed the internal structure, or dimensionality, of a psychological test. Recall that a test’s internal structure and dimensionality have to do with the number and nature of psychological constructs that are assessed by a test’s items. For example, in Chapter 4, we presented a hypothetical 6-item personality test, and we discussed the possibility that those 6 items (talkative, assertive, outgoing, creative, imaginative, and intellectual) might reflect two psychological traits—extraversion and openness to experience. Similarly, in Chapter 8, we described the 10-item Rosenberg Self-Esteem Inventory (RSEI), and we noted that its items are usually seen as reflecting only one psychological trait—global self-esteem. Thus, dimensionality and internal structure refer to the way a test’s items cohere together and thereby represent one or more psychological constructs.

As a quick refresher, take a moment to recall the importance of a test’s dimensionality. As we have discussed in earlier chapters, a test’s dimensionality has fundamental implications for the test’s development, reliability, validity, and use. In terms of test development, internal structure should be a major consideration when constructing a new test or scale. For example, a test might be designed to match a specific dimensionality (e.g., to reflect five uncorrelated personality traits or a single dimension of self-esteem), and examinations of internal structure can reveal the degree to which the test being developed actually corresponds to this dimensionality. In terms of reliability, a test’s dimensionality or internal structure reflects the test’s internal consistency. That is, evaluation of a test’s dimensionality reveals which items are consistent with which other items (e.g., are all items roughly equally consistent with each other, or are there sets of items that are particularly consistent with each other?). In terms of validity, a test’s internal structure is important because the appropriate interpretation of a test’s scores depends on the match between its actual internal structure and the internal structure of its intended construct(s). For example, if we discovered that the RSEI included two uncorrelated dimensions, then it would be invalid to interpret the RSEI solely in terms of a unitary global self-esteem
trait. Following from its implications for test development, reliability, and validity, a test’s internal structure has robust implications for its use. That is, internal structure should guide the way a test is scored, producing one or more scores that are meaningful, both psychologically and quantitatively (see Figure 4.1).

In our earlier discussions of dimensionality, we highlighted exploratory factor analysis (EFA) as a statistical tool that is often used to evaluate a test’s dimensionality. As we noted, EFA is most appropriate when one has few, if any, hypotheses about a scale’s internal structure. Again, for example, we began our discussion of the six personality items in Chapter 4 began by assuming that we had no idea about the number of dimensions reflected in those items.

In contrast, confirmatory factor analysis (CFA) is useful when there are clear hypotheses about a test’s dimensionality. That is, CFA is designed to examine a test’s dimensionality when a test developer or test evaluator has clear expectations about the number of factors or dimensions underlying a test’s items, the links between the items and factors, and the association between the factors.

In this chapter, we introduce several important issues in CFA. Specifically, we discuss basic issues in the logic of CFA, the process of conducting a CFA, and the key results that are obtained from a CFA. In addition, we discuss the way CFA can be used to examine fundamental psychometric issues—dimensionality (of course), reliability, and validity. Although the information obtained from CFA is most directly relevant to a test’s internal structure, it also can be used to examine a test’s internal consistency, and it can be used to evaluate convergent and discriminant evidence.

Our presentation, like the two chapters that follow, is intended to be relatively nontechnical and intuitive—we emphasize the basic logic of CFA and the psychological interpretation of its key results. However, CFA is a complex multivariate statistical procedure, with many technical considerations and problems. We cannot address many such issues in this chapter, and we recommend that interested readers consult other specialized sources for a more focused coverage (e.g., Brown, 2006; Hoyle, 2011; Thompson, 2004).

On the Use of EFA and CFA

The Frequency and Roles of EFA and CFA

Our emphasis on EFA in the earlier chapters reflects two facts about the typical use of factor analysis in psychometric evaluation. First, EFA has been used much more frequently than CFA. This discrepancy is at least partly due to the fact that EFA has been integrated much more seamlessly into statistical packages that are user-friendly and widely used. For example, EFA has long been integrated into the popular SPSS statistical package, but the ability to conduct CFA in SPSS is only recently becoming available and even now still requires additional software components.

Despite this long-standing difference in the frequency with which EFA and CFA have been used, CFA seems to be enjoying emerging interest and application.
This emergence is likely due to the increasing availability and user-friendliness of statistical software that is capable of conducting CFA. For example, the popular statistical packages SAS and SPSS now include modules that are capable of conducting CFA (though again, these might not be included in all versions of either package). Moreover, when using some software (e.g., the AMOS package in SPSS), one can conduct CFA without doing much more than creating figures to represent hypotheses about a test’s dimensionality (e.g., Figures 4.2, 4.3, or 4.4).

A second issue in the use of factor analysis is that EFA and CFA are most appropriate for different phases of the test development and evaluation process. EFA is perhaps most appropriate for the early phases of test use. That is, EFA is most useful when test developers might still be clarifying their understanding of the constructs and of the test itself. In contrast, CFA is more appropriate in later phases of test development—after the initial evaluations of item properties and dimensionality and after any significant revisions of test content. Some psychologists have suggested a more integrated approach, pointing out that EFA can, in fact, be used in a somewhat hypothesis-driven way and that CFA is often, in fact, used in a way that is somewhat exploratory. Interested readers are directed to other sources for in-depth discussion of these points (e.g., Hopwood & Donnellan, 2010).

**Using CFA to Evaluate Measurement Models**

As noted earlier, when using CFA, we evaluate hypotheses about specific “measurement models” regarding the dimensionality or internal structure of a test. That is, CFA allows test developers and test evaluators to understand the degree to which their hypothesized measurement models are consistent with actual data produced by respondents. For example, based on previous research and theory, we might hypothesize that the RSEI has a unidimensional structure (i.e., we might hypothesize a one-factor measurement model). Using CFA, we can test this “model” formally—we can collect responses to the RSEI items and examine the degree to which those actual responses produce data in which the RSEI items cohere into a single factor.

Moreover, we can—if necessary and appropriate—alter our hypothesized model in a way that makes it more consistent with the actual structure of responses to the test. For example, some information produced by our CFA of the RSEI might indicate that the actual responses are, in fact, not consistent with a unidimensional model. Thus, we might examine other information from our CFA and discover that a two-dimensional structure is much more consistent with the data. In this way, we might modify our hypothesis regarding the true structure of the RSEI.

Furthermore, we can examine several measurement models in a series of CFAs, using the results to discover the one that best matches participants’ actual responses. For example, we could formally test the RSEI’s “fit” with both a unidimensional model and a two-dimensional model, and we could formally test which model is in fact more consistent with the actual test responses that we have obtained.
The Process of CFA for Analysis of a Scale’s Internal Structure

In this section, we describe the process of conducting a CFA and interpreting its main results. Our goal is to provide a conceptual perspective on the logic, process, and meaning of CFA. With this goal in mind, we do not provide details about the way to conduct CFA in any particular statistical package. Rather, we lay a conceptual foundation that helps readers become informed consumers of CFA-based psychometric analysis. This conceptual foundation should also be valuable for readers who also wish to be producers of CFA-based psychometric information; indeed we hope that our conceptual coverage provides a solid foundation for additional training in the “how to” of CFA with respect to a variety of specific statistical packages (e.g., Byrne, 2001, 2006; Diamantopoulos & Siguaw, 2000; Hatcher, 1994).

Overview of CFA and Example

The typical process of conducting a CFA is summarized in Figure 12.1. In the following discussion, we present the steps in this process, and we discuss the logic of each step and the psychometric information that is obtained and usually reported. We describe the steps in which test developers and test evaluators have active roles (the unshaded boxes in Figure 12.1) and the steps that are carried out by statistical software (in the shaded boxes). As illustrated in Figure 12.1, CFA can be an iterative, back-and-forth process. The process begins when we articulate and evaluate a specific measurement model, but the process often does not end there. Often, after initially evaluating a specific measurement model, we revise the model and then evaluate the revised model. In fact, this revision and reevaluation can occur multiple times as we learn more and more from the CFA process.

After carrying out a CFA of a test, test developers and test evaluators usually report key information about the model-testing process. This information includes any revisions that have been made to the model(s). Usually, the information highlights the measurement model that is most consistent with the test’s actual internal structure, as discovered via the CFA.

To illustrate and explain this process, we discuss a CFA of the Authenticity Scale (Wood, Linley, Maltby, Baliousis, & Joseph, 2008). The Authenticity Scale was intended to measure the degree to which a person “knows himself or herself” and “acts accordingly,” and it was based on a conceptual model that identifies three dimensions of authenticity (with example items):

1. **Self-alienation** was defined (roughly) as the degree to which a person really understands himself or herself (e.g., “I don’t know how I really feel inside”).

2. **Authentic living** was defined as the degree to which a person behaves and expresses emotion in a way that is an honest reflection of his or
her self-perception (e.g., “I think it is better to be yourself than to be popular”).

3. Accepting external influence was defined as the degree to which a person understands that other people can influence one’s life and conforms to these influences (e.g., “I am strongly influenced by the opinions of others”).

Considering this conceptual basis, the Authenticity Scale was intended to have a three-dimensional structure, with a subscale for each of the three dimensions. Each item is phrased as a self-relevant statement, and respondents are asked to rate their level of agreement with each item, using response options ranging from 1 (“Does not describe me at all”) to 7 (“Describes me very well”).
Before conducting a CFA, there are at least three key preliminary steps. First, of course, is clarification of the psychological construct to be assessed and initial development of the test items. Indeed, Wood et al. (2008) describe a conceptually based process of writing a large number of items, an initial analysis of some responses to those items (e.g., an EFA), and a selection of 4 items for each of the three intended dimensions. This produced a total of 12 items that passed some initial tests of psychometric quality.

A second preliminary step is the collection of a large number of responses to the test. The appropriate sample size for a CFA is an important but complex issue, and experts have offered a variety of recommendations. In terms of absolute numbers of respondents, recommendations range from a minimum of 50 people (for simple measurement models and “clean” conditions) to 400 people or more. Other recommendations are made in terms of the ratio of respondents to items, with suggestions ranging from 5 respondents per item to 20 (or more) respondents per item. The bottom line is that, as a complex multivariate procedure, CFA requires a large number of responses—it is not uncommon to have much more than 200 respondents in a CFA. For example, Wood and his colleagues recruited three samples totaling more than 550 people to respond to the 12-item test.

A third preliminary step is to reverse score any negatively keyed items. This ensures that all items are keyed in the same direction, and it avoids any confusion arising from items that might otherwise seem highly inconsistent with each other.

**Step 1: Specification of Measurement Model**

After these preliminary steps, we translate our hypothesized measurement model into a statistical software package designed to conduct CFA. Using contemporary software packages such as AMOS/SPSS, SAS, EQS, and LISREL, the process can be quite straightforward. Such packages allow us simply to draw a figure to represent or “specify” the measurement model. The packages translate these drawings into statistical equations, which it uses to conduct the CFA. Of course, these packages also allow us to create those equations ourselves, rather than by drawing a figure. However, it is likely that most people opt to begin with the drawing capabilities.

For example, Figure 12.2 illustrates two measurement models evaluated in the CFA of the Authenticity Scale (Wood et al., 2008), and it shows at least three elements of a measurement model that need to be specified—as summarized in Table 12.1. First, we must specify the number of dimensions, factors, or latent variables (represented by ovals) that are hypothesized to underlie the test’s items (represented by rectangles). For example, Figure 12.2a presents a unidimensional measurement model in which the scale’s 12 items load on a single Authenticity factor. Note that this unidimensional model does not correspond to the three-factor structure for which the Authenticity Scale was designed. However, Wood and his colleagues tested this model as a comparison with their main hypothesized model, which is presented
in Figure 12.2b. This figure presents a “hierarchical” multidimensional measurement model in which the scale’s items load on three “lower-level” factors (i.e., Self-Alienation, Authentic Living, Accepting External Influence), all of which load on a single, more fundamental “higher-order” general Authenticity factor.

A second element of the measurement model to be specified is the links between items and factors. That is, we must specify which items are linked to (i.e., load on) each factor. In the typical factor-analytic figure (e.g., Figures 12.2a and 12.2b), a pathway (i.e., arrow) between an item and a factor indicates that the item is hypothesized to load on a factor. There are two general guidelines that are generally followed when specifying these links. First, in most measurement models, at least one item is linked to each factor. For example, in Figure 12.2a, all items are hypothesized to load on the sole factor. Similarly, there are four items loading on each of the three lower-level factors in Figure 12.2b. Note that there are no items loading on the general Authenticity factor in Figure 12.2b. This is typical for such higher-level factors (i.e., factors on which other factors load)—we will return to this later.

Figure 12.2 Example: Two Measurement Models Examined by Wood et al. (2008)
Table 12.1 Facets of the Measurement Model to be Specified

<table>
<thead>
<tr>
<th>Facets</th>
<th>Required specifications:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Number of factors</td>
</tr>
<tr>
<td></td>
<td>2. The associations between items and factors</td>
</tr>
<tr>
<td></td>
<td>3. The potential associations between factors (if more than one factor is hypothesized)</td>
</tr>
<tr>
<td>Examples of some additional specification options:</td>
<td>4. Exact values of one or more parameters (e.g., specific factor-loading values)</td>
</tr>
<tr>
<td></td>
<td>5. Equality of parameters (e.g., two factor loadings constrained to be equal)</td>
</tr>
</tbody>
</table>

A second general guideline is that each item is typically linked to only one latent variable. Typically, we create each test item to reflect one and only one psychological characteristic, though different items might reflect different characteristics. With this in mind, we generally hypothesize that each item loads on one and only one factor. For example, the second item in the Authenticity Scale was written to reflect Self-Alienation, and the first item was written to reflect Authentic Living. Thus, Figure 12.2b hypothesizes that Item 2 should load only on the Self-Alienation factor, and Item 1 should load only on the Authentic Living factor.

If a measurement model is multidimensional (e.g., Figure 12.2b), then we must specify a third element of the model—the possible associations between factors. We have two possible ways of indicating that factors are associated with each other. First, we can specify that the factors load on a higher-order factor. Indeed, Figure 12.2b is a hierarchical measurement model in which three lower-order factors load on a higher-order factor. This suggests that a person’s levels of Self-Alienation, Authentic Living, and Acceptance of External Influence are hypothesized to be affected by his or her level of general Authenticity. This hierarchical relationship suggests that the three lower-order factors are associated with each other. That is, if two or more lower-order factors load on a higher-order factor, then by extension, those lower-order factors should be associated with each other. For example, Figure 12.2b indicates that people who have a high level of Authentic Living will also tend to have a high level of Acceptance of External Influence. A second way in which we can indicate that factors are associated with each other is by having them simply be correlated with each other, as represented by a two-way arrow between them. In contrast to one-way arrows (which indicate that one thing affects another thing), two-way arrows simply indicate that two things are associated with each other, without any implication of causality. With this in mind, we might choose to omit the general Authenticity factor in Figure 12.2b and instead draw three two-way arrows—one arrow for each pair of the remaining factors (e.g., between Self-Alienation and Authentic Living).

When thinking about the associations between factors, we might also hypothesize that factors are not associated with each other. Consider the multidimensional test with uncorrelated factors that we discussed in Chapter 4 (e.g., Figure 4.4).
To indicate the hypothesis that two factors are not correlated with each other, we simply omit any connection between those factors. That is, we would not include a higher-order factor that connects them, and we would not include a two-way arrow between them.

Typically, when specifying the hypothesized item–factor links and any factor–factor associations, our hypotheses are fairly simple—we simply hypothesize that a link/association exists or not. That is, we hypothesize either that an item is associated with a particular factor or not and we hypothesize that factors are associated with each other or not. The statistical software will then estimate the precise associations. In other words, we simply hypothesize that peoples’ responses to particular items are affected by their levels of particular psychological characteristics. We then rely on the statistical software to estimate the actual magnitude of those effects. Essentially, the presence of a pathway between an item and a factor (or between two factors) indicates a hypothesized nonzero association; in contrast, the absence of a pathway indicates a hypothesized zero association. In sum, we generally specify the model in a way that the associations are “freely estimated”; that is, we essentially state that “these parameters are probably not zero” and allow the software to estimate the parameters’ precise values.

It is worth noting that the models in Figure 12.2 are somewhat simplified and that you might encounter more complex representations. Specifically, sometimes figures are drawn so that each item also has a unique random error term associated with it. However, such complexity is not always included in published psychometric work. Thus, we will keep the description as simple as possible for our immediate purposes, and we will return to this additional complexity in the “Reliability” section of this chapter.

**Step 2: Computations**

After we have specified the hypothesized measurement model, we then ask our statistical software to conduct the CFA based on those specifications and on the data that we have collected (i.e., actual responses to the test). Although the statistical computations are conducted “behind the scenes” by the software, a quick overview of the statistical process can enhance insight into CFA and its main results. At a rather simplified level, the basic CFA computations have four phases.

*Phase 1: Actual Variances and Covariances.* In the first phase of the CFA computations, the collected data are used to compute the items’ variances and the covariances among the items. That is, the software computes the actual degree of variance for each item and the actual level of covariance between each pair of items. For example, in the analysis of the Authenticity Scale, the software computes 12 variances and 66 covariances (i.e., $66 = 12(12 –1)/2$). These values are used in subsequent phases.

*Phase 2: Parameter Estimates (and Inferential Tests).* Indeed, in the second phase of computations, the items’ actual variances and covariances are used to estimate values for the “parameters” as specified by the researcher. Parameters are quantitative
values that are related to specific elements of the measurement model, and they are an important part of the results produced by CFA. There are several key types of parameters to be estimated. One type of parameter is the factor loading(s) for each item. As discussed earlier, when we hypothesize that an item loads on a factor, we usually do not hypothesize a specific value for that loading (i.e., we do not usually hypothesize that the item loads on the factor to any specific degree). Instead, we usually rely on the statistical software to compute a value for that parameter. For example, in the analysis of the model presented in Figure 12.2b, the software will compute 12 parameter estimates for the factor loadings—one for each item. A second type of parameter that often occurs in multidimensional tests are those that connect factors to each other—either correlations between factors or the loading of lower-order factors on higher-order factors. Again, when we hypothesize that one factor loads on another or that two factors are correlated, we usually do not hypothesize specific values. Thus, for example, in a CFA of the model in Figure 12.2b, the software will compute three parameter estimates for the loadings of the three lower-order factors on the higher-order factor. There are other types of parameters that can be estimated (e.g., error variances), but the item–factor parameters and the factor–factor parameters are typically of most interest.

To compute the parameter estimates, the software begins with the actual variances and covariances from the first phase of computations. For example, in the analysis of the model in Figure 12.2b, it would use, in part, the actual covariance (i.e., association) between Item 2 and Item 7 to estimate the loadings that those items might have on the Self-Alienation factor. Indeed, if the two items are strongly associated with each other (i.e., if they have a robust covariance—if people who endorse Item 2 are very likely to endorse Item 7), then they might have something strongly in common with each other. Thus, the statistical software might compute strong loadings for both items on the Self-Alienation factor, which they would have in common (i.e., the Self-Alienation factor is the “thing” that the items have in common). The actual variances and covariances are thus used to compute values for all factor loadings, interfactor correlations, error variances, and so on.

Importantly, CFA software also computes an inferential statistic (i.e., significance test) for each parameter estimate. We will return to this later—for now we will note that the typical null hypothesis for a given parameter is that the parameter’s estimated value is 0 in the population from which the sample was drawn.

Phase 3: Implied Variances and Covariances. In the third phase of computations, the software uses its estimated parameter values (from Phase 2) to compute “implied” item variances and covariances. That is, the software computes item variances and covariances, as implied by the estimated parameters. For example, if both Item 2 and Item 7 have very strong loadings on the Self-Alienation factor, then this implies that they will be strongly correlated with each other (i.e., they will have a relatively large covariance). However, if both items have very weak loadings on the Self-Alienation factor, then this implies that they are weakly correlated with each other.

This third phase might seem circular—indeed, you might suspect that the implied variances and covariance should exactly match the actual variances and
covariances; after all, the implied variances and covariances are based on parameter estimates, which are themselves computed from the actual variances and covariances! In fact, there will not be exact matches (except under very specific conditions), and moreover, there can be substantial mismatches between actual and implied variances/covariances. The mismatches can happen because the parameter estimates are based on the software’s attempt to account for a great amount of information when computing each parameter estimate. For example, we noted earlier that software would use, in part, the actual covariance (i.e., association) between Item 2 and Item 7 to estimate the loadings that those items might have on the Self-Alienation factor. This is true, but it oversimplifies the process. In fact, the software’s estimate of Item 2’s factor loading is based on the variances and covariances of all the items. After all, the Self-Alienation factor directly involves four items (see Figure 12.2b), which means that there are at least 6 covariances to be considered and balanced in computations related to Self-Alienation. Actually, it is even more complicated, because Self-Alienation is hypothesized to load on Authenticity, which in turn is hypothesized to affect the other two factors. Thus, the model implies that the Self-Alienation factor will be correlated with the other two factors. This in turn implies that the four Self-Alienation items will be correlated with the other eight items on the scale (i.e., 32 additional covariances). So when computing each and every parameter estimate, the software must balance and weigh a huge amount of information. Sometimes, when attempting to find solutions to the problem of balancing all of this information, the CFA produces parameter estimates that are not good representations of each single piece of information.

Thus, it is important to consider the degree of match or mismatch between the actual variances/covariances and the implied variances/covariances. If the hypothesized model is good (i.e., if it is a good approximation of the true model underlying the scale’s items), then the implied variances and covariances will match closely the actual variances and covariances computed in the first phase of analysis. However, if the model is poor, then the implied values will differ greatly from the actual values. This important issue is the focus of the next computational phase.

**Phase 4: Indices of Model Fit.** In the fourth phase, the software produces information regarding the general adequacy or “fit” of the hypothesized model. To do this, it compares the implied variances/covariances to the actual variances/covariances, and it computes indices of “model fit” and modification. If the comparison between implied and actual values reveals only minor discrepancies or mismatch, then the software produces indices of “good fit.” This would indicate that the hypothesized measurement model adequately reflects the actual pattern of responses to the test. In contrast, if the comparison between implied and actual values reveals large discrepancies or mismatches, then the software produces indices of “poor fit.” This, of course, would indicate that the hypothesized measurement model does not adequately reflect the actual pattern of responses to the test. In the next section, we will discuss the interpretation of these values in more depth.

Going further, CFA software can compute “modification indices” that indicate specific ways in which the measurement model could be improved. These values reveal potential modifications that would bring the model closer to the factor
structure that truly may be underlying the test’s items (as they were responded to in the sample). For example, a modification index derived from analysis of Figure 12.2b might indicate that Item 2 loads on the Authentic Living factor in addition to the Self-Alienation factor. Because the original hypothesized model in Figure 12.2b does not hypothesize that Item 2 loads on Authentic Living, we might consider modifying our hypothesis to fit this suggestion.

**Step 3: Interpreting and Reporting Output**

After collecting responses to the test, specifying a measurement model that we believe underlies those responses, and computing parameter estimates and fit indices, we interpret the results. CFA produces many types of output addressing a variety of psychometric and statistical issues, and this section describes some of the most important and commonly reported results.

As shown in Figure 12.1, the particulars of this step and the next depend on the results that are obtained. Depending on some of the results, we might examine other results. Moreover, depending on what we find, we might conclude our analysis and report our findings, or we might modify our hypothesized model and rerun the analysis. Ideally, we will find that our hypothesized measurement model is a good match to the actual responses to the test. In that case, we might examine only two sets of results.

*Fit Indices.* Typically, we first examine the fit indices that address the overall adequacy of our hypothesized measurement model. As described earlier, “good fit” indicates that the hypothesized measurement model is consistent with the actual responses to the test, and this supports the validity of the model. That is, if the fit indices are good, then we can have initial confidence in interpreting the test’s dimensionality as we had hypothesized. However, “poor fit” indicates that the hypothesized dimensionality is not consistent with the actual responses to the test. This is usually seen as evidence against the validity of the hypothesized measurement model, and we should not interpret the test’s dimensionality as we had initially hypothesized.

Most CFA programs will compute and present many fit indices. For example, the statistical package SAS computes approximately 20 fit indices as part of its “calis” procedure, which is used to conduct CFA. Although many fit indices are available, most published reports of a CFA present only a few. Unfortunately, there is no clear consensus regarding best fit indices to interpret and report, thus different reports will present different sets of fit indices.

That said, there are a few that you might be most likely to encounter in reports of CFA. The computational details for these indices are beyond the scope of our current discussion; the important goal at this point is to be familiar with some of these indices and with their general interpretations. The chi-square statistic is probably the most commonly reported fit index, and it actually indicates the degree of “poorness of fit” or “misfit” of the model. That is, large, significant chi-square values are evidence of poor fit, whereas small, nonsignificant chi-square values indicate
good fit, providing support for a hypothesized measurement model. Note that this “significant is bad” interpretation of chi-square is quite the opposite of the typical perspective on statistical significance, in which we generally hope to find that a test is statistically significant. Although chi-square values are usually included in reports of CFA, it is important to note that sample size affects the chi-square values: All else being equal, large samples will produce large chi-square values, which produce statistical significance.

Interestingly, this is a bit of a paradox for CFA. On one hand, we want to have large samples so that we can obtain robust, reliable parameter estimates. On the other hand, large samples increase the chance that we will obtain significant chi-square values, which would seem to indicate that our hypothesized model is invalid.

Partly because of this paradox, reports of CFA usually include additional fit indices. These alternative indices include the goodness-of-fit index (GFI), the incremental fit index (IFI), the normed fit index (NFI), the comparative fit index (CFI), the nonnormed fit index (NNFI, also known as the Tucker-Lewis Index, or TLI), the root mean square of approximation (RMSEA), the root mean square residual (RMR), the standardized root mean square residual (SRMR), and the Akaike information criterion (AIC), to name but a few. Note that these indices, unlike the chi-square index, do not include a formal test of statistical significance.

Importantly, these fit indices have different scales and norms for indicating model fit. For example, the CFI ranges from 0 to 1.0, with larger values indicating a good fit. In contrast, the RMR has a lower bound of 0 but an upper bound that depends on the test’s scale of measurement, with smaller values indicating good fit. As another example, the SRMR ranges from 0 to 1.0 (like the CFI), but lower values indicate good fit (like the RMR). Many sources are available for guidance in interpreting indices (e.g., Hu & Bentler, 1999; Kline, 2010).

For example, in an article describing the CFA of the Authenticity Scale (Wood et al., 2008), fit indices are the first results that are presented. In that report, the researchers present four fit indices—chi-square, SRMR, CFI, and RMSEA—related to their analysis of the model in Figure 12.2b (their primary model of interest). Considering the possibility that some readers might not be familiar with these particular indices, the researchers note specific values that have been recommended by experts as indicative of “good fit” as represented by these indices (i.e., SRMR values ≤ .08, CFI values ≥ .95, and RMSEA values ≤ .06), although they also noted that these recommendations are a bit more conservative than what is often taken as indicating an adequate fit. Based on a CFA of responses from 213 students, the fit indices are presented in the “12.2b” column of Table 12.2. On finding CFI, SRMR, and RMSEA values very close to the conservative recommendations, the researchers concluded that the model (as shown in Figure 12.2b) “provided a good fit” (p. 393) to the Authenticity Scale’s responses. They also reported a significant chi-square but reminded readers of the link between large sample sizes and statistical significance, and they ignored the chi-square’s apparent indication of misfit. Such dismissal of a significant chi-square value is common in CFA reports.

To more fully evaluate the adequacy of their main model of interest, Wood et al. (2008) compared this model with the alternative, unidimensional measurement
Table 12.2  Fit Indices for Two Measurement Models Examined by Wood et al. (2008): Sample 3 (N = 213)

<table>
<thead>
<tr>
<th>Fit Index</th>
<th>Measurement Model</th>
<th>Cited Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Figure 12.2b</td>
<td>Figure 12.2a</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>90.06*</td>
<td>353.45*</td>
</tr>
<tr>
<td>SRMR</td>
<td>.08</td>
<td>.16</td>
</tr>
<tr>
<td>CFI</td>
<td>.94</td>
<td>.53</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.06</td>
<td>.16</td>
</tr>
</tbody>
</table>

NOTE. Benchmarks are those cited by Wood et al. (2008). SRMR = standardized root mean square residual; CFI = comparative fit index; RMSEA = root mean square of approximation.

* df = 51.

b df = 54.

*p < .05.

model in Figure 12.2a. The results indicated a much poorer fit for this alternative model—see the “Figure 12.2a” column in Table 12.2. Specifically, poor fit is indicated by a CFI value well below the .95 benchmark cited by Wood and colleagues, by SRMR and RMSEA values well above their benchmarks, and by a large, significant chi-square value. The comparison of this model with the three-dimensional model (in Figure 12.2b) strengthened the researchers’ confidence that the Authenticity Scale’s internal structure is indeed well represented by the model in Figure 12.2b. As shown in Figure 12.1, the examination of fit indices can lead in two possible directions. First, if the fit indices suggest that the model is adequate, then we will examine parameter estimates to evaluate the more specific psychometric qualities of the test. Second and alternatively, if the fit indices instead suggest that the model is inadequate, then we will likely examine the modification indices and consider ways in which the model could be revised.

Parameter Estimates and Significance Tests. After deciding that a measurement model has an adequate overall fit, we then examine a variety of parameter estimates. We obtain an estimated value for each parameter, including values for the items’ factor loadings and the interfactor associations. Parameter estimates are an important facet of a test’s overall dimensionality and, as we shall see, other psychometric properties.

As described earlier, an item’s loading on a factor represents the degree to which differences among peoples’ responses to an item are determined by differences among their levels of the underlying psychological construct that is assessed by that item. If an item is hypothesized to load on a particular factor, then we expect to discover a large, positive, and statistically significantly factor loading. If we do indeed find such results, then we are likely to conclude that the item is a good reflection of the underlying psychological dimension. Thus, we are likely to keep that item on the test. However, if we find that the item’s factor loading is small and/or nonsignificant, then we are likely to conclude that the item is unrelated to the
psychological dimension. In this case, we are likely to remove the item from the test. We then might respecify the model to accommodate this change (i.e., eliminating the item from the model) and rerun the computations to evaluate the revised test. We realize that this scenario might seem paradoxical—how could we have a generally well-fitting model (as indicated by the fit indices) but a weak factor loading? The answer is that fit indices represent the overall adequacy or fit of the entire measurement model and that a model can have generally good support despite having some weak specific aspects.

Once again turning to the CFA of the Authenticity Scale (Wood et al., 2008), Figure 12.3 presents some key parameter estimates obtained in the analysis of the three-dimensional measurement model. Recall that the fit indices indicated good support for this model (Table 12.2), so the researchers interpreted and reported the parameter estimates for this model. Figure 12.3 presents standardized factor loadings and interfactor associations. Note that as standardized factor loadings, these values are interpretable like factor loadings from an EFA. That is, they are interpreted in terms of correlations or standardized regression weights—generally ranging between −1 and +1. As these values indicate, all 12 items loaded robustly on their hypothesized factors—the weakest factor loading was Item 1’s loading of .60 on the Authentic Living factor. Figure 12.3 also indicates that each lower-order factor loads strongly on the higher-order factor—the weakest loading being the −.58 loading of Accepting External Influence on the higher-order Authenticity factor. Finally, although the researchers do not explicitly state this, the magnitudes of these 15 parameter estimates lead us to assume that all of them are statistically significant.

![Figure 12.3](image-url)
As done in the report of the Authenticity Scale’s CFA, researchers usually highlight these particular parameters—the items’ factor loadings and any interfactor associations (Wood et al., 2008). They sometimes also present parameters reflecting items’ error variances or items’ variance-explained values. However, because these values are implied by the factor loadings themselves, researchers usually choose not to include this extra information. Indeed, the report of the Authenticity Scale’s CFA did not report present these values.

In sum, as Figure 12.1 suggests, CFA interpretation and the subsequent steps depend on several important issues. The most important issue is perhaps the overall adequacy of an initial hypothesized measurement model, as indicated by the fit indices. If the fit indices suggest that the model fits well, then we generally move to an examination of the parameter estimates. If we find that those parameter estimates provide further support for the model (e.g., there are no weak, nonsignificant factor loadings), then we will likely finish the CFA by concluding that the model is a good representation of the test’s internal structure. However, things do not always proceed so smoothly. Indeed, if the fit indices indicate that the model fits poorly, then we will likely revise our hypothesis about the model’s internal structure. This brings us to the next potential step in a CFA—model modification and reanalysis, with the goal of improving our understanding of the test’s true dimensionality.

Step 4: Model Modification and Reanalysis (If Necessary)

As Figure 12.1 illustrates, the results of a CFA sometimes, perhaps often, force us to consider modifying a hypothesized measurement model. If we obtain fit indices suggesting that a model is inconsistent with the actual responses to the test items, then we will likely examine modification indices to find hints about the revisions that we can make to the model.

When conducting a CFA, we obtain a large number of modification indices; indeed, each one represents a parameter that was left out of (i.e., set to zero in) the initial measurement model. For example, the model in Figure 12.2b implies that Item 2 loads on the Self-Alienation factor but not on the Authentic Living factor. Thus, Wood et al. (2008) initially allowed the software to estimate Item 2’s loading on Self-Alienation (see Figure 12.3), but they set or “fixed” the loading of Item 2 on the Authentic Living factor to 0, indicating that Item 2 has no direct association with Authentic Living. When we examine modification indices for this model, we thus would find a value referring to this “fixed” parameter—the “fixed-to-zero” factor loading of Item 2 on Authentic Living. Indeed, we would find modification values for every parameter that was initially set to zero—for example, the association between Item 2 and Authentic Living, the association between Item 2 and Acceptance of External Influence, the association between Item 1 and Self-Alienation, the association between Item 1 and Acceptance of External Influence, and so on.

The magnitude of a modification index reflects the potential impact of revising the relevant parameter. For example, a CFA of the model in Figure 12.2b might produce a relatively large modification index related to Item 2’s potential loading on the Authentic Living factor. To anthropomorphize a bit, this would indicate that Item 2
“wants” to load on the Authentic Living factor in addition to the Self-Alienation factor. That is, it would indicate that peoples’ responses to Item 2 are affected by both their degree of Self-Alienation and their degree of Authentic Living. More statistically, it would indicate that if we allowed Item 2 to load on both factors, then the fit indices would be improved. Thus, by examining the modification indices and making changes based on the largest indices, we learn about the real dimensionality of the test, and our measurement model becomes a better reflection of this reality.

With this in mind, as Figure 12.1 indicates, after examining modification indices, we might change one or more parameters and then rerun the analysis. Analysis of the revised model will produce new output—new fit indices, new parameter estimates, and so on. We then evaluate the adequacy of the revised model, and we either examine parameter estimates (if the revised model fits well) or examine the new modification indices (if the model still fits poorly).

We should note a few cautions regarding modification in a CFA. First, modification begins to obscure the difference between confirmatory analysis and exploratory analysis. This reiterates our earlier comment that CFA can be used in a semi-exploratory manner. Second, a test developer or test evaluator should be hesitant to perform many modifications in a CFA, with particular hesitancy about modifications that lack a clear conceptual basis. Such modifications might arise from response patterns that are idiosyncratic or unique to that sample of people, and not be representative of other test takers more generally. Thus, if more than one or two modifications are made to a model, then test developers and test evaluators should strongly consider evaluating the revised model in a different sample of test takers (i.e., a “cross-validation sample”) before drawing strong inferences about the “true” internal structure of the test in a way that generalizes to a broad range of people.

Comparing Models

Earlier, we noted that the analysis of the Authenticity Scale included a comparison of the two models in Figure 12.2. Although the three-dimensional model was the main one of interest, the researchers contrasted it against the unidimensional model (Wood et al., 2008).

Indeed, when conducting a CFA of a test, test developers and test evaluators often evaluate competing measurement models. The point of such comparisons is to identify which model is the best representation of the test’s true internal structure. Rather than evaluating a single possible model’s fit, test developers and test evaluators can learn even more by evaluating and comparing several reasonable potential measurement models. All else being equal (e.g., in terms of theoretical basis), we would prefer measurement models with a relatively good fit. That is, we would identify the model having the most supportive fit indices and conclude that it reflects the test’s true dimensionality. This can provide strong insight into the test’s actual properties.

A full discussion of strategies for comparing models is beyond the scope of this chapter. However, interested readers can find such a discussion in other sources (e.g., Brown, 2006; Hoyle, 2011).
Summary

As a psychometric tool, CFA can provide great insight into the internal structure of psychological tests. By allowing us to test hypotheses about specific measurement models, CFA is an important complement to EFA. However, the value of CFA goes beyond the evaluation of dimensionality, and the following sections describe the way in which it can be extended to the evaluation of test reliability and validity.

CFA and Reliability

Coefficient alpha is the most widely used method of estimating reliability (see Chapter 4), but its accuracy depends on psychometric assumptions that may not be valid in some applications of behavioral research (Miller, 1995; Zinbarg, Revelle, Yovel, & Li, 2005). Indeed, alpha’s accuracy as an estimate of reliability is determined by the pattern and nature of items’ psychometric properties (e.g., do the items have correlated errors, do they have equal factor loadings?).

These issues have led psychometricians to use factor analysis (or principal components analysis) for estimating reliability (e.g., Armor, 1974). Although CFA is not currently used widely in this way, experts have recently developed CFA-based procedures for this purpose. In this section, we describe a CFA-based procedure for estimating the reliability of unidimensional scales; additional details, examples, and alternatives can be found in several sources (e.g., Brown, 2006, pp. 337–351; Raykov, 2004; Zinbarg et al., 2005).

When using CFA, we can estimate reliability through a two- or three-step process, depending on the need to modify the initial measurement model. First, we use CFA to evaluate the test’s basic measurement model. Consider, for example, the Interaction Anxiousness scale (IAS; Leary, 1983). The IAS is a 15-item personality scale designed to reflect the tendency to experience social anxiety in interactions in which an “individual’s responses are contingent on the responses of other interactants” (e.g., a one-on-one conversation, as opposed to a speech delivered to an audience; Leary, 1983, p. 68). Each item presents a self-relevant statement (e.g., “I often feel nervous even in casual get-togethers”), and respondents are asked to rate their level of agreement/endorsement of each.

Figure 12.4a presents a simple unidimensional model hypothesized to reflect the dimensionality of the IAS—all 15 items load on a single Interaction Anxiousness factor, with no additional links among items. Note that this figure is a bit more complex than Figure 12.2a, discussed earlier. Specifically, in this figure, each item is affected by a unique error term that represents the effect of random measurement error on responses to each item. These error terms are part of all CFA models, but they are often omitted from graphical presentations of CFA models (e.g., see Figures 12.2 and 12.3). We collected responses to the IAS from a relatively small sample of respondents ($n = 107$, smaller than is ideal), and a CFA indicated that the
unidimensional model did not fit the IAS well in these respondents—\(X^2(90) = 224.30, p < .05\); NNFI = .74, CFI = .78; RMSEA = .12, and SRMR = .09.

In the second step of the process, we modify and reanalyze the measurement model if necessary. As in the case of our CFA of the IAS, if the initial hypothesized model fits poorly, then we identify useful revisions to the model (via the modification indices). Specifically, we focus mainly on potential associations among the items' error terms. For example, we examined modification indices from the poorly fitting model in Figure 12.4a, and we identified six useful modifications. Not only

\[\text{Figure 12.4 Initial Model, Modified Model, and Parameter Estimates for the Interaction Anxiousness Scale}\]
did the modification indices provide statistical support for these changes but the changes largely were psychologically reasonable as well. For example, the modification indices suggested that Items 4 and 14 “wanted to be correlated with each other.” Indeed, both items refer to interactions with people in positions of authority (i.e., “I get nervous when I must talk to a teacher or boss” and “I get nervous when I speak to someone in a position of authority”). Thus, these two items shared something that was a bit different from the remaining items, none of which refer explicitly to authority figures. Similarly, the results suggested that Items 6 and 11 were statistically linked, and indeed, these are the only two items that explicitly include the word shy. Thus, Items 4 and 14 share an “authority” commonality, whereas Items 6 and 11 share a “shyness” commonality.

Considering the relative magnitudes of several modification indices and the conceptual connections between pairs of items, we modified the model by adding six parameters—the pairwise associations between items’ error terms. Our reanalysis revealed a much improved fit for the modified model ($X^2_{(84)} = 122.76, p < .05$; NNFI = .92, CFI = .94; RMSEA = .07, and SRMR = .07). Figure 12.4b presents the modified model and unstandardized parameter estimates, including each item’s factor loading, the error variance of each item, and the six covariances between error terms.

In the final step of the process, we use these unstandardized parameter estimates to estimate the test’s reliability. Recall that reliability is defined as

$$\text{Reliability} = \frac{\text{True variance}}{\text{True variance + Error variance}}.$$ 

The parameter estimates obtained via CFA can be used to estimate the true variance and error variance, and thus they can be used to estimate reliability (Brown, 2006):

$$\text{Estimated reliability} = \frac{\left(\sum \lambda_i\right)^2}{\left(\sum \lambda_i\right)^2 + \sum \theta_i + 2\sum \theta_{ij}}.$$ (12.1)

In this equation, $\lambda_i$ refers to an item’s factor loading, $\theta_i$ refers to an item’s error variance, and $\theta_{ij}$ refers to the covariance between the error terms of two items (this is zero for models without correlated error terms). In terms of our earlier discussion of reliability, $\left(\sum \lambda_i\right)^2$ reflects the variance of true scores (i.e., signal), because the factor loadings reflect the links between the items and the “true” psychological attribute. Similarly, the sum $\sum \theta_i$ reflects random error variance (i.e., noise), because these terms reflect the unique aspects that affect each item but that are not related to the underlying psychological attribute of interest. Thus, Equation 12.1 represents the theoretical definition of reliability as the ratio of true score variance to total observed score variance (with observed variance being the sum of true score variance and error variance; see Chapter 5). For the results in Figure 12.4, true score variance is estimated as 101.81:
Similarly, error variance is estimated as 14.52:

\[
\left( \sum \lambda_i \right)^2 = (0.66+0.70+0.51+0.60+0.64+0.68+0.58+0.37 \\
+1.10+0.55+0.83+0.85+0.75+0.60+0.67)^2 = 101.81.
\]

\[
\sum \theta_\varepsilon = (0.63+0.70+0.72+0.97+0.73+1.15+0.75+0.99+0.87 \\
+1.00+0.78+0.66+0.72+0.85+0.84) = 12.36,
\]

\[
\sum \theta_\varepsilon = (0.29 + 0.48 + (-0.25) + (-0.20) + 0.50 + 0.26) = 1.08,
\]

\[
\sum \theta_\varepsilon + 2 \sum \theta_\varepsilon = 12.36 + 2(1.08) = 14.52.
\]

Thus, the estimated reliability of the IAS is

\[
\frac{101.81}{101.81 + 14.52} = .87.
\]

For these responses to the IAS, the CFA-based reliability estimate is only somewhat smaller than the reliability estimate obtained via coefficient alpha, which is \(\alpha = .89\). Although the difference between the two estimates is not large in this case, it can be much more dramatic, and it reflects the alpha’s tendency to misestimate reliability (Miller, 1995).

CFA is a very flexible tool for examining reliability, going well beyond the relatively simple analysis of a unidimensional test such as the IAS. For example, it can be used to estimate reliability for multidimensional scales, to estimate group differences in reliability, and to obtain confidence intervals around estimates of reliability (e.g., Raykov, 2004).

**CFA and Validity**

Going even further, CFA can be a useful tool for evaluating validity in several ways. First, as implied by our relatively in-depth discussion of CFA’s ability to test a specific hypothesis about a test’s internal structure, CFA offers insight into the “internal-structure” aspect of validity (i.e., does the actual structure of responses to the test items fit the structure that is implied by the theoretical basis of the intended construct?).

Second, if test responses are collected along with measures of related constructs or criteria, then we can evaluate the test’s association with those variables. Whether we view this associative evidence in terms of convergent/discriminant validity, concurrent validity, criterion validity, predictive validity, or external validity, it provides important information about the psychological meaning of test scores. There are at least two ways in which we can use CFA to examine these facets of validity.

One way in which we use CFA for evaluating convergent and discriminant validity is by applying it to multitrait–multimethod (MTMM) matrices. As discussed in Chapter 5, an MTMM study includes multiple traits/constructs (e.g., social skill, impulsivity, conscientiousness, and emotional stability), each of which is assessed...
through multiple methods (e.g., self-report, acquaintance ratings, and interviewer ratings). When we examine the associations among the entire set of scores, we can evaluate convergent validity, discriminant validity, method effects, and other important validity information. There are several CFA-based methods for evaluating an MTMM matrix, though a full discussion is beyond the scope of this chapter. Interested readers can consult other good sources for this information (e.g., Brown, 2006, chap. 6; Marsh & Grayson, 1995).

A second way to use CFA to examine convergent validity (and potentially discriminant validity) is through the focused examination of a test along with one or more criterion variables. For example, we could examine the construct validity of the IAS by collecting individuals’ responses to the IAS along with other measures of their tendency to experience social anxiety. Indeed, the people who completed the IAS, as described in the previous section, also completed a measure of “situational social anxiety.” That is, they completed a survey that described 11 different social situations and asked the participants to rate the level of anxiety that they would likely experience in each of those situations. Presumably, people who rate themselves as likely to experience a high level of social anxiety, on average, across the 11 situations will also have high scores on the IAS.

With such information, we can address several important psychometric questions: Does the IAS have a unidimensional structure as hypothesized? To what degree do participants’ anxiety ratings from the 11 situations reflect a single “situational social anxiety” factor? To what degree is the IAS factor associated with a potential “situational anxiety” factor? Using CFA, we can evaluate all of these issues.

For example, we used CFA (actually, structural equation modeling, or SEM) to examine the model illustrated in Figure 12.5. First, as described earlier, we evaluated the internal structure of the IAS, to properly represent its internal structure. Second, we similarly evaluated the internal structure of the situational anxiety questionnaire, discovering that two of the situations did not load on a core “situational anxiety” factor. Thus, we dropped those two items, and we discovered some correlated error terms among the remaining nine items. Third, after clarifying

![Figure 12.5](attachment:image.png)
the measurement models for the two questionnaires, we examined the model in Figure 12.5. As shown in this figure, we were primarily interested in the correlation between the two factors. And as shown in the figure, the correlation was extremely strong, and it was statistically significant. (Note that we include only this parameter estimate in the figure, for the sake of simplicity.)

By conducting such an analysis, we can evaluate validity evidence in terms of the association between a test and a relevant criterion, while also accounting for measurement error in both the test and the criterion. This is an important advantage of CFA over many alternative analytic strategies (e.g., the zero-order correlation between scale scores and criterion scores). Such models and similar ones (e.g., Figure 2 in McArdle, 1996) extend CFA to SEM, but the basic principles described in this chapter apply to SEM as well as to CFA. Again, readers interested in additional details on SEM are directed to a variety of useful sources (e.g., Hoyle, 2011).

CFA is a useful and increasingly accessible tool that we can use in the development and/or evaluation of psychological tests. It provides power and flexibility when evaluating a test’s dimensionality, reliability, and validity, and it has important advantages over other statistical techniques (e.g., EFA and regression) and indices (e.g., coefficient alpha). To be sure, it requires more thought and careful attention than do some other psychometric tools; however, its important advantages and its increasing accessibility make it a very useful tool for the examination of psychometric quality.

Summary