15

Confirmatory Factor Analysis I

FACTOR ANALYSIS: THE MEASUREMENT MODEL

This chapter will focus in more detail on the *measurement model* of latent variable structural equation modeling, more generally known as *confirmatory factor analysis.* At its most basic level, factor analysis is a reduction technique, a method of reducing many measures into fewer measures. The methodology works by placing scales or items that correlate highly with each other on one factor, while placing items that correlate at a low level with each other

on different factors. Because one primary reason items correlate highly with one another is that they measure the same construct, factor analysis provides insights as to the common constructs measured by a set of scales or items. Because it helps answer questions about the constructs measured by a set of items, factor analysis is a major method of establishing the internal validity of tests, questionnaires, and other measurements. You can also think of factor analysis as a method of establishing convergent and divergent validity: items that measure the same thing form a factor (converge), whereas items that measure different constructs form a separate factor (diverge).

With *exploratory* factor analysis (not covered in this text), one analyzes a set of items or scales that presumably measures a smaller set of abilities, traits, or constructs. Decisions are made concerning the method of factor extraction to use, the method for deciding the number of factors to retain, and the method of factor rotation to use. Given these choices and the data, the results of the analysis will suggest that the items measure a smaller number of factors. For example, factor analysis of 13 scales may suggest that these scales measure four constructs. The output from the analysis will include factor loadings of each scale on the four factors and, if oblique rotation is used, the correlations of the factors with each other. The researcher then decides on names for the factors based on the constructs they presumably reflect, a decision based on the loadings of the variables on the factors, relevant theory, and previous research.

With *confirmatory* factor analysis one uses previous research and relevant theory to decide in advance what the factors or constructs are that underlie the measures. Just as in path analysis, we propose a model that underlies the variables of interest. The fit statistics then provide feedback concerning the adequacy of the model in explaining the data. I hope it is obvious why the methods are termed exploratory versus confirmatory factor analysis. With the first, we examine the results and decide what the various scales are measuring, whereas with the second we decide what the various scales are measuring and then examine the results to find out how accurate our predictions were. This dichotomy is an obvious simplification—we can use exploratory factor analysis in a confirmatory fashion and can use confirmatory factor analysis in an exploratory fashion—but it is still a useful distinction.

The development of factor analysis is inexorably linked with development of theories of intelligence and intelligence tests. Early intelligence researchers developed the methods of factor analysis to understand the nature and measurement of intelligence, and factor analysis continues to be a major method of supporting and challenging the validity of intelligence tests. For this reason, I will illustrate the method of confirmatory factor analysis using intelligence test data. Note that this is one of two chapters on the topic of CFA; we will return to more advanced CFA topics after learning more about latent variable SEM.

AN EXAMPLE WITH THE DAS-II

The Differential Ability Scales, Second Edition (DAS-II; Elliott, 2007) is among the most commonly administered individual intelligence tests for children. The DAS-II includes a series of short verbal and nonverbal subtests and is appropriate for children and youth ages 2½ to 18. The DAS-II is a common portion of a broader psychological evaluation for children and adolescents who are having learning, behavioral, or adjustment problems. It may be used to help evaluate children for special programs (e.g., those for children with learning disabilities and gifted programs); diagnose learning, behavioral, and neurological problems; or provide information relevant to an intervention to ameliorate such problems.

Structure of the DAS-II

Although the DAS-II includes different tests for children at different ages, all 21 tests from the battery were standardized for children ages 5–8. We will analyze data for 12 of these tests designed to measure four underlying constructs. The test names and a portion of the theoretical structure of the DAS-II are shown in Figure 15.1. Although I will not describe the subtests in detail, they measure a variety of verbal and nonverbal skills. For example, the Word Similarities subtest requires children to explain the construct shared by three words. In contrast, Pattern Construction requires the child to construct, from pictures, geometric designs using two-colored foam squares and blocks. According to the author, the DAS-II measures verbal reasoning (Verbal Ability), nonverbal, inductive reasoning (Nonverbal Reasoning), visual–spatial reasoning (Spatial), and short-term memory (Working Memory) (and some other skills not discussed here; you may also see these abilities referred to as Crystallized Intelligence, or Gc; Fluid Intelligence, Gf; Visual Processing, Gv; and Short-Term Memory, Gsm). The figure shows which subtests are designed to measure which skills. This structure is reflected in the actual scoring of the test. For children 7 and older, for example, scores for two tests per construct are added together to form Verbal, Nonverbal Reasoning, Spatial, and Working Memory composite scores.

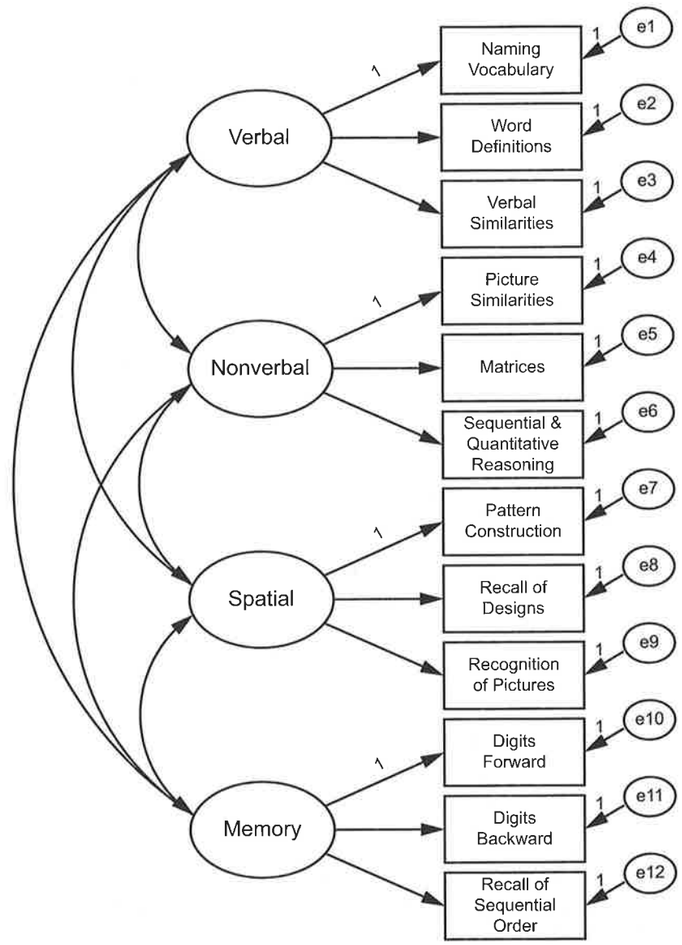


Figure 15.1 Initial DAS-II model. Does the DAS-II measure verbal, nonverbal, and spatial reasoning skills, along with short-term memory?

The Initial Model

Figure 15.1 is also the setup for a confirmatory factor model (indeed, the figure is the input for analysis in Amos), with the constructs underlying the DAS-II shown in ovals as latent variables and the eight subtests (the actual measurements we obtain) shown in rectangles as measured variables. The arrows in the figure make explicit the causal assumptions underlying such testing and models. The paths point from the constructs to the subtests in recognition of the implicit assumption that each person's level of verbal reasoning ability is the primary influence on his or her score on the Word Definitions subtest, for example, whereas each person's level of visual spatial ability is the primary influence on his or her score on the Pattern Construction subtest. Although the constructs the test is designed to measure are the primary influence on individuals' scores on the subtests, you know from the last chapter that individuals' scores on each subtest are also influenced by unreliability and by the unique characteristics of each test. This latter statement makes sense intuitively as well. Although Pattern Construction and Recall of Designs (in which children draw complex designs from memory) obviously both require visual and spatial skills, they also both obviously require different specific skills, such as the mental translation of a two-dimensional picture into three-dimensional form versus visual and spatial memory skills. These unique skills and unreliability are represented by the small latent variables pointing to each subtest labeled e1 through e12. e7, for example, represents all influences on children's scores on the Pattern Construction subtest other than Spatial Ability.

You will recall that latent variables have no set scale, and we must set the scale of each latent variable to estimate the model. Recall also that one way to set the scale of a latent variable is to set a path from each latent variable to one measured variable at 1.0. This is done in Figure 15.1. The Verbal factor's scale is set to be the same as that for the Naming Vocabulary subtest. The choice of which measured variable to use is arbitrary; I have simply set the scale of each factor to be the same as the first variable that measures this factor. Without these constraints to set the scales of the latent variables the model would be underidentified. Kline (2011, p. 127) calls this method of setting the scales of latent variables "unit loading identification," or ULI. The scales for the unique–error variances are also set to the same scale as their corresponding subtests: e1 is set to have the same scale as Naming Vocabulary, e2 as Word Definitions, and so on. Alternatively, we could also set the scale of the factors by setting the variance of each factor to 1.0 (we will come back to this point).

The model shown in Figure 15.1 also includes correlations among each construct thought to be measured by the DAS-II. It is commonly recognized that cognitive tests and cognitive factors are positively correlated (Carroll, 1993). The model shown in the figure is on the Web site (www.tzkeith.com) in the folder for this chapter under the name "das 2 first order 1.amw"; Mplus script is also available.

The DAS-II manual includes tables of correlations among the subtests for each age level 2½ through 17 (along with means and standard deviations). The averaged covariance matrix for these subtests for children 5–8 is shown in Table 15.1; this matrix was produced as a by-product of CFA analyses designed to determine whether the DAS-II measures the same

constructs across its age levels (Keith, Low, Reynolds, Patel, & Ridley, 2010). The matrix of covariances among the twelve subtests was used to estimate the model shown in Figure 15.1. The covariance matrix is also contained in the Excel file "DAS 2 cov.xls" and the SPSS file "DAS 2 cov.sav." The sample size for the analyses was 800.

Standardized and Unstandardized Results: The Initial Model

Figure 15.2 shows standardized results of the initial analysis of the DAS-II model. First, focus on the fit indexes. The Root Mean Square Error of Approximation (RMSEA) was .046, lower (better) than our rule of thumb for good models of .05. The Standardized Root Mean Square Residual (SRMR) was .027, meaning that the average difference between the

Table 15.1 Average Covariance Matrix for the DAS-II for Ages 5 through 8

*rowtype_*	*varname_*	*wdss*	*vsss*	*sqss*	*soss*	*rpss*	*rdss*	*psss*	*pcss*	*nvss*	*mass*	*dfss*	*dbss*
cov	wdss	91.52											
cov	vsss	58.43	104.34										
cov	sqss	42.21	53.06	94.14									
cov	soss	50.34	54.85	54.15	113.43								
cov	rpss	27.88	36.19	44.16	40.00	102.09							
cov	rdss	31.19	44.29	49.98	48.46	48.19	99.74						
cov	psss	36.86	41.62	39.46	37.48	33.56	41.31	106.38					
cov	pcss	37.21	48.52	54.01	48.53	40.82	55.81	38.72	84.74				
cov	nvss	53.94	59.64	44.16	52.13	33.62	44.43	39.34	46.83	102.13			
cov	mass	41.67	47.50	60.40	54.75	40.01	41.38	39.48	47.34	40.05	104.59		
cov	dfss	44.45	51.76	46.54	61.56	32.91	46.32	37.07	44.28	49.54	39.66	121.52	
cov	dbss	41.78	50.76	52.77	62.90	37.51	47.28	36.98	47.58	43.47	51.09	56.20	103.25
n		800	800	800	800	800	800	800	800	800	800	800	800
mean		50.03	50.21	49.99	49.94	50.01	49.75	49.92	50.07	50.22	50.26	50.02	49.65

Note: Variable names: wdss = Word Definitions; vsss = Verbal Similarities; sqss = Sequential &Quantitative Reasoning; soss = Recall of Sequential Order; rpss = Recognition of Pictures; rdss = Recall of Designs; psss = Picture Similarities; pcss = Pattern Construction; nvss = Naming Vocabulary; mass = Matrices; dfss = Digits Forward; dbss = Digits Backward.

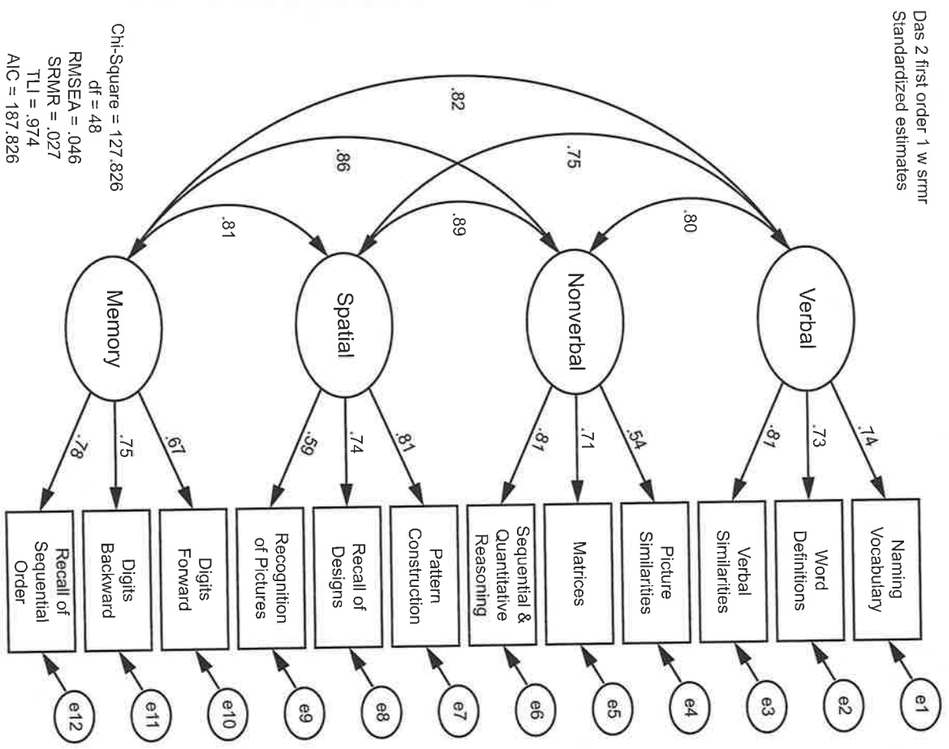


Figure 15.2 Standardized estimates for the initial DAS-II four-factor model

actual and the implied correlation matrices was only .027. The TLI (and the CFI, not shown) were above our target for a good model (.95). By these criteria, it appears that the DAS-II model fits the data well. In other words, the model that underlies the DAS-II indeed could indeed have produced the correlations and covariances we observed among the DAS-II subtests, and the theoretical structure of the DAS-II is supported. Note, however, that the χ^2 is statistically significant (127.355 [47], $p < .001$), which, in contrast to the other indices, suggests a lack of fit of the model to the data. We will examine possible sources of misfit later in this chapter. Focusing on the model itself, it appears that most subtests provided relatively strong measures of the appropriate ability or construct; the factor loadings for most subtests on the Verbal, Nonverbal Reasoning, Spatial, and Memory factors were .6 or higher. The exceptions to these larger loadings were the Picture Similarities and Recognition of Pictures subtests on the Nonverbal and Spatial factors (loadings of .54 and .59). Although the detailed printout shows that these loadings were statistically significant, they are lower than for the other factors. Within factors, most subtests had fairly equivalent loadings on the factor they supposedly measure, although there are clearly subtests that have stronger loadings (e.g., Verbal Similarities, Sequential and Quantitative Reasoning, Pattern Construction). This difference in loadings suggests that the common construct measured by these tests is better measured by, for example, Sequential & Quantitative Reasoning than by Picture Similarities. The results also show that the latent factors correlate substantially with each other, with factor correlations ranging from .75 to .89.

Figure 15.3 shows the unstandardized estimates ("Regression Weights") of the factor loadings, standard errors, z values (critical ratio, or CR), and p values (all less than .001). Note that the loadings used to set the scales of the latent variables, the ones that were set to 1, were not tested for statistical significance. Estimated values are tested for statistical significance; constrained parameters are not. In the second section of the figure are the standardized loadings ("Standardized Regression Weights") followed by the covariances and correlations among latent factors. Note that all estimated paths (factor loadings) and covariances were statistically significant ($z > 2$), and that the standardized loadings match those in the figural display of the model.

Testing a Standardized Model

It is also possible to set the scale of the latent factors in the model by setting the factor variances to 1.0 (instead of setting one factor loading per factor to 1.0). The setup for such a standardized model—also known as unit variance identification, or UVI (Kline, 2011, p. 128)—for the DAS-II is shown in Figure 15.4. Although less common, and less consistent with SEM, than the method of setting factor loadings, the factor variance method has the advantage of producing *standardized* covariances (i.e., correlations) among the factors. Recall that a correlation matrix is a standardized covariance matrix, the result of standardizing the variables in the matrix (i.e., setting their variances to 1.0). Alternatively, you can think of a correlation matrix as just another variance–covariance matrix, but with all variances set to 1.0. Thus, when we set the variances of the factors in a CFA to 1.0, we have standardized the covariance matrix of factors. Figure 15.5 shows the *unstandardized* output for the UVI analysis just described. Note that the covariances (correlations) in this figure are the same as the correlations from the standardized output shown in Figure 15.2. The factor loadings, however, are still in an unstandardized metric (although a different unstandardized metric than previously).

The advantage of having the factor covariances standardized comes into play when we wish to compare competing models. Note the high correlation between the Nonverbal Reasoning and Spatial factors (.89). We may wonder if this correlation is statistically significantly

Regression Weights

			Estimate	S.E.	C.R.	P
nvss	<---	Verbal	1.0000			
wdss	<---	Verbal	.9418	.0489	19.2542	***
vsss	<---	Verbal	1.0996	.0526	20.8866	***
psss	<---	Nonverbal	1.0000			
mass	<---	Nonverbal	1.3056	.0926	14.0950	***
sqss	<---	Nonverbal	1.4059	.0936	15.0205	***
pcss	<---	Spatial	1.0000			
rdss	<---	Spatial	.9822	.0468	20.9828	***
rpss	<---	Spatial	.7949	.0485	16.3777	***
dfss	<---	Memory	1.0000			
dbss	<---	Memory	1.0346	.0576	17.9493	***
soss	<---	Memory	1.1187	.0609	18.3809	***

Standardized Regression Weights

			Estimate
nvss	<---	Verbal	.7395
wdss	<---	Verbal	.7333
vsss	<---	Verbal	.8052
psss	<---	Nonverbal	.5414
mass	<---	Nonverbal	.7102
sqss	<---	Nonverbal	.8082
pcss	<---	Spatial	.8139
rdss	<---	Spatial	.7370
rpss	<---	Spatial	.5906
dfss	<---	Memory	.6692
dbss	<---	Memory	.7535
soss	<---	Memory	.7780

Covariances

			Estimate	S.E.	C.R.	P
Verbal	<-->	Nonverbal	33.4641	3.0790	10.8684	***
Verbal	<-->	Spatial	42.0275	3.2851	12.7934	***
Verbal	<-->	Memory	45.4283	3.6929	12.3016	***
Nonverbal	<-->	Spatial	37.2672	3.2043	11.6305	***
Nonverbal	<-->	Memory	35.2301	3.2871	10.7175	***
Spatial	<-->	Memory	44.8726	3.5411	12.6721	***

Correlations

			Estimate
Verbal	<-->	Nonverbal	.8049
Verbal	<-->	Spatial	.7509
Verbal	<-->	Memory	.8239
Nonverbal	<-->	Spatial	.8922
Nonverbal	<-->	Memory	.8561
Spatial	<-->	Memory	.8100

Figure 15.3 Unstandardized and standardized text output for the initial DAS-II four-factor model.

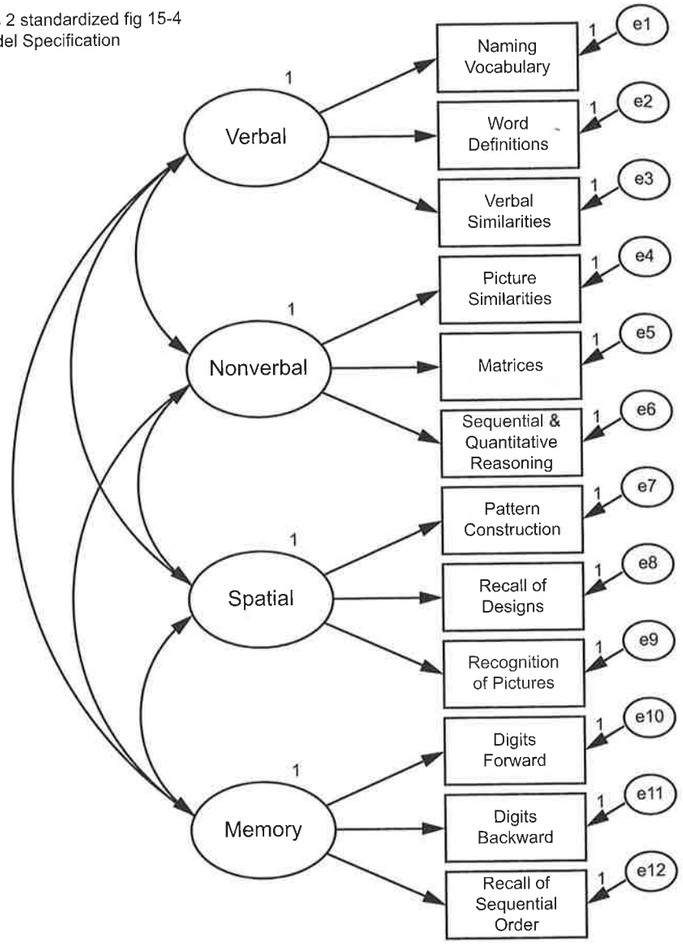


Figure 15.4 An alternative standardized method of specifying the initial DAS-II model. With this method, we set the scale of the latent variables by setting their variances to 1 instead of constraining factor loadings.

different from 1.0, meaning that the factors may be statistically indistinguishable. We could test this supposition by setting the factor correlation to 1.0 and comparing the fit of this model with the original model. However, model constraints apply to the *unstandardized* model only. Thus, if we wish to set a factor correlation to 1.0 (or some other value), we need to make the factor correlations equivalent to the factor covariances, using this standardized model. (As will be shown, a few other constraints are also needed to test the distinguishability of factors).

Although the primary results of a CFA—notably the fit indexes and the standardized output—will generally be the same whichever method is used, it is possible for some results

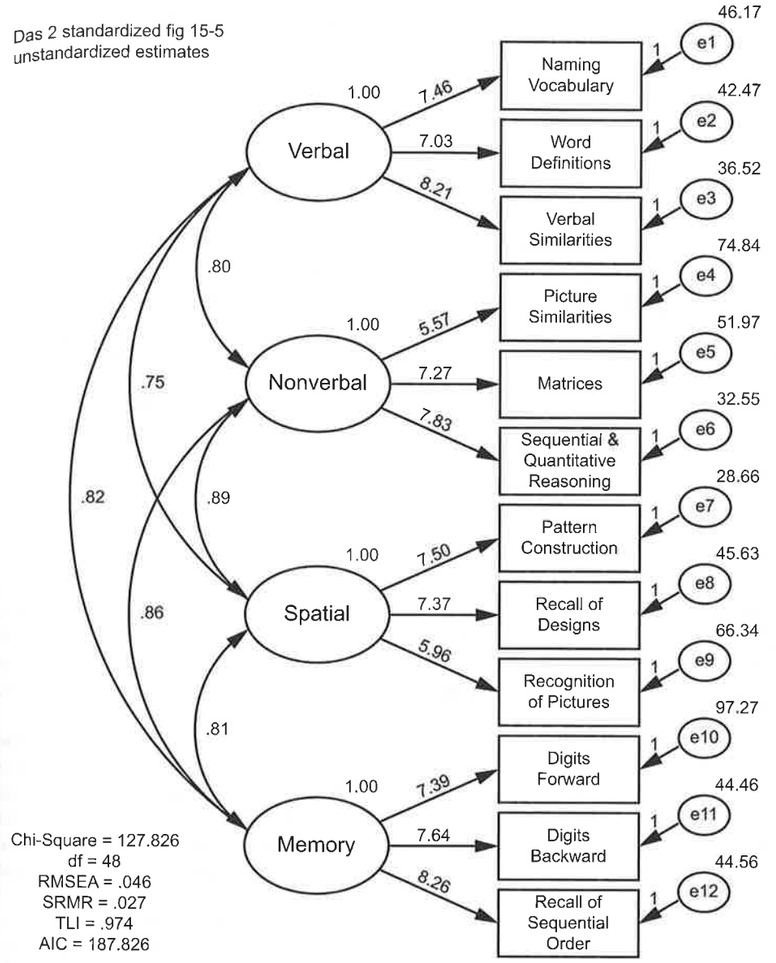


Figure 15.5 Unstandardized solution using the standardized model. Note that the factor covariances are now equivalent to the factor correlations from Figure 15.2.

to change slightly depending on whether the ULI (factor loading set to 1) or the UVI (factor variance set to 1) method is used. Likewise, results generally do not change—but sometimes do—depending on which factor loading is set to 1 using the ULI method. In particular, the unstandardized parameter estimates and the standard errors may change across the two methods, and the resulting *z* values (critical ranges) may change as well. What this means is that it is possible for a factor loading or factor covariance to be statistically significant using one method but not statistically significant using the other. (For more information, see Millsap, 2001). This article also shows that with complex models, where tests load on multiple factors, the fit of models can change depending on which factor-to-test path is set to 1.0.)

Before moving to the next topic, notice the numbers beside the unique and error variances: 46.17 for e1, 42.47 for e2, and so on. These numbers are the estimates of the combined unique and error variances of the various subtests. You can compare them to the variances of the variables shown in the diagonals of the variance–covariance matrix (Table 15.1). It appears that close to one half of the variation in the Word Definitions subtest is error and unique variance.

TESTING COMPETING MODELS

This initial example has tested the adequacy of a single confirmatory model. As in SEM, however, a more powerful use of the methodology is to compare alternative and competing models. I will briefly illustrate this method using the DAS-II example.

Note that in the models shown thus far each subtest has been assumed to measure one and only one underlying common ability or factor. But the constructs measured by tests may be and often are much more complex than this; indeed, it seems likely that some of the DAS-II subtests may measure more than one underlying ability. For example, the Recall of Designs subtest requires children to draw from memory designs they have seen a few seconds earlier. Doesn't it make sense to assume that this test requires short-term memory skills in addition to (or instead of) visual-spatial reasoning?

Testing Plausible Cross-Loadings

Figure 15.6 shows a model that tests this possible cross-loading by allowing Recall of Designs to load on both the Spatial and Memory factors. Note that this model and the initial model (e.g., Figure 15.1) are nested, because the initial model can be derived from this model by constraining the path (loading) from Memory to Recall of Designs to zero. Thus the model in Figure 15.1 is nested within the model shown in Figure 15.6. Figure 15.7 shows the standardized loadings for this model, along with some of the fit indices.

The DAS-II alternative cross-loading model fits the data well. Our primary stand-alone fit index, the RMSEA, suggests that the two-factor model explains well the test standardization data. The other stand-alone fit indexes (SRMR, TLI) also suggest a good fit of the model to the data. If we focus only on the fit of each model in isolation, we conclude that this model fits well, as does the earlier four-factor model. Our primary interest, however, is *relative* fit of the two models. In particular, we are interested in how this three-factor model compares to the initial model that did not include any cross-loadings. The cross-loaded model is less parsimonious than the initial model, with 47 degrees of freedom shown in Figure 15.7 versus 48 for the initial model in Figure 15.2. Degrees of freedom represent parameters that are constrained to some value, rather than freely estimated, and thus each additional degree of freedom means an increase in parsimony. Thus, if the two models fit equally well, we will prefer the initial (more parsimonious) model. Do the models fit equally well? To answer this question, we need to focus on the fit indexes appropriate for comparing competing models.

In Chapter 13 I argued that $\Delta\chi^2$ was a good method for comparing competing models that were nested, that is, when one model can be derived from the other by fixing one or more parameters. The two models are indeed nested; to derive the model shown in Figure 15.1 from that in Figure 15.6 we would only need to constrain the loading of Recall of Designs on the Memory factor to 0.

Table 15.2 shows the $\Delta\chi^2$ comparing these two models. According to the χ^2, the initial model fit slightly worse than did model 2 (the model with Recall of Designs loaded on two factors). But if two models are nested, the more constrained model (the model with the larger *df*) will always fit worse than the less constrained model according to χ^2. The question,

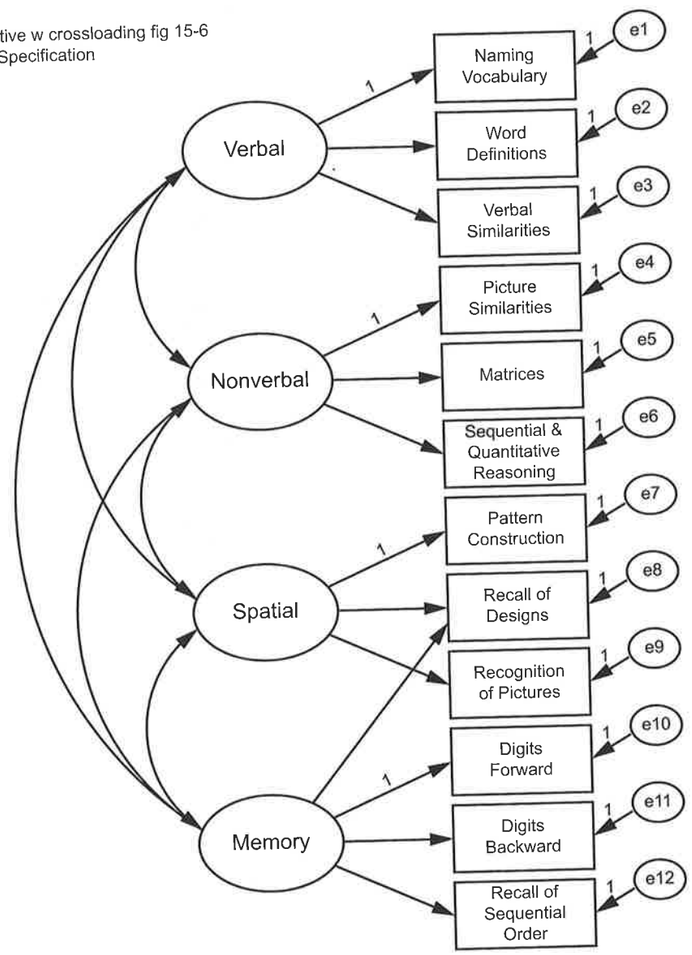


Figure 15.6 An alternative model testing whether Recall of Designs measures both visual-spatial and short-term memory skills. This model and the initial model are nested.

then, is how much worse is the fit? Is it trivial or is it large enough so that we say that it is not worth the extra degrees of freedom we gain? The common way to judge whether the fit-worsening constraint is "worth it" is to test the $\Delta\chi^2$ for statistical significance. This has also been done in the table. As shown, when the extra path/loading was added to the second model, χ^2 decreased by only .491, and this difference is not statistically significant ($p = .483$) (And recall that we need a $\Delta\chi^2$ of approximately 3.9 for statistical significance with 1 *df* and $p < .05$.) What does this mean? Recall also our rule that if $\Delta\chi^2$ is not statistically significant that we prefer the more constrained model, the one with more *df*. This means that we would tentatively accept the Initial four-factor model over the cross-loaded model, and that we

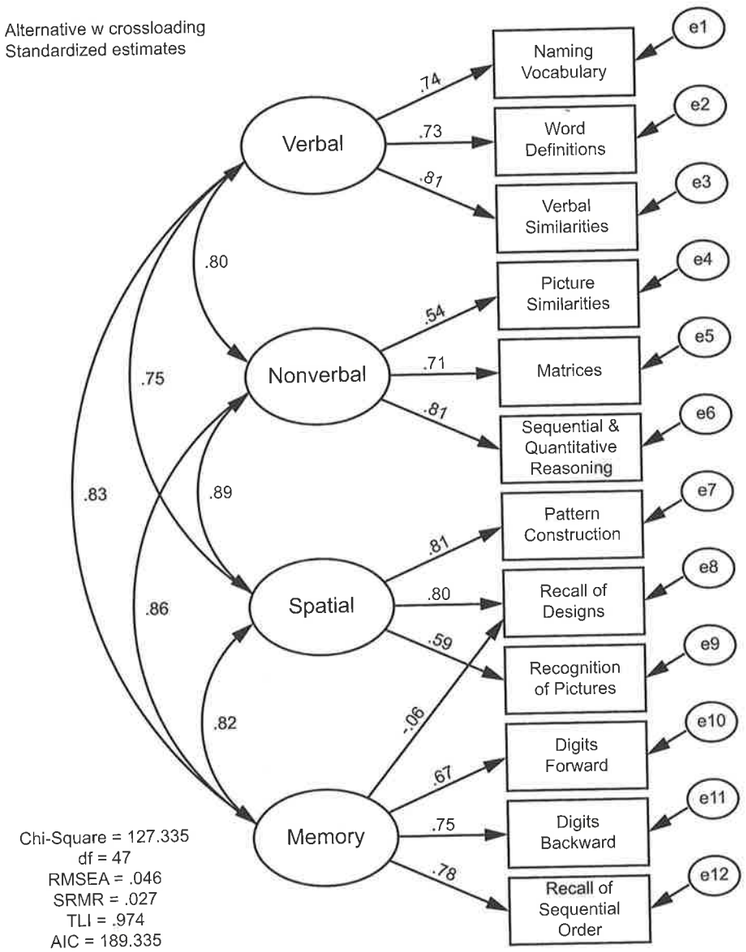


Figure 15.7 Standardized estimates and fit for the cross-loaded model

would reject the hypothesis that is personified by the difference between the two models. In other words, no, the data do not support the cross-loading of Recall of Designs on both the Spatial and the Memory factors; it appears that Recall of Designs indeed measures visual-spatial reasoning skills, not short-term memory.

A Three-Factor Combined Nonverbal Model

Although I have argued that the DAS-II should measure four underlying constructs, we have already noted the very high correlation (.89) between the Nonverbal Reasoning and the Spatial factors. Perhaps these two factors really are equivalent, meaning that we could collapse them into one? We could easily argue that the Spatial and the Nonverbal Reasoning subtests

Table 15.2 Comparison of Fit Indexes for Alternative Models of the Structure of the DAS-II

Model	χ^2	*df*	$\Delta\chi^2$	*df*	*p*	*AIC*	*aBIC*	*RMSEA*	*TLI*	*CFI*	*SRMR*
1. Initial four-factor	127.826	48				187.826	233.023	.046	.974	.981	.027
2. Recall Designs cross-loaded	127.335	47	.491	1	.483	189.335	236.038	.046	.974	.981	.027
3. Three-factor (Figure 15.8)	163.651	51	35.825	3	<.001	217.651	258.328	.053	.966	.974	.029
4. Nonverbal-Spatial correlation = 1	156.698	49	28.872	1	<.001	214.698	258.388	.052	.966	.975	.028
5. Equivalent correlations	163.651	51	6.953	2	.031	217.651	258.328	.053	.966	.974	.029

Note: All models are compared to Model 1 with the exception of Model 5. The $\Delta\chi^2$ for Model 5 is a comparison to the previous model (Model 4).

should be considered as measuring a single underlying ability. After all, most of these tests require some degree of spatial awareness and nonverbal reasoning; why separate the two factors? Thus, we have both a priori logical as well post hoc data-driven reasons for suggesting another plausible model, one that combines these two factors. Figure 15.8 shows such a plausible three-factor model. Although it is not obvious, the model *is* nested with the model in Figures 15.1 through 15.5. This three-factor model is equivalent to the model shown in Figure 15.4 (the standardized model) with the following constraints:

1. Set the Nonverbal Reasoning–Spatial correlation to 1.0 (in the standardized model). This constraint essentially equates the factors.
2. Constrain other factor correlations to be equal to one another across these factors. That is, constrain the Memory–Spatial factor correlation to be equal to the Memory–Nonverbal Reasoning correlation, and then constrain the Verbal–Spatial factor correlation to be equal to the Verbal–Nonverbal correlation. The most direct way to do this in Amos is to constrain the correlations to an alphabetical value (e.g., *a* for the first two correlations and *b* for the second two). The result of this constraint is that the values will be freely estimated, but all values with the same letter will be constrained to be equal. Other SEM programs will have other methods of constraining values to be equal.

Because the models are nested, $\Delta\chi^2$can be used to compare the competing models. This model is more parsimonious that the initial four-factor model. (Make sure you understand why this three-factor model is more parsimonious than the initial model.) Thus, if the two models have an equivalent fit, we will favor the more parsimonious three-factor model with the combined Nonverbal factor.

As shown in Figure 15.8, the three-factor combined Nonverbal model showed a good fit to the data according to most of the stand-alone fit indexes (with the exception of RMSEA), yet the χ^2 also increased substantially for this model. The four-factor model had a χ^2 of 127.826 ($df = 48$) versus 163.651 ($df = 51$) for the three-factor Combined Nonverbal model. Change

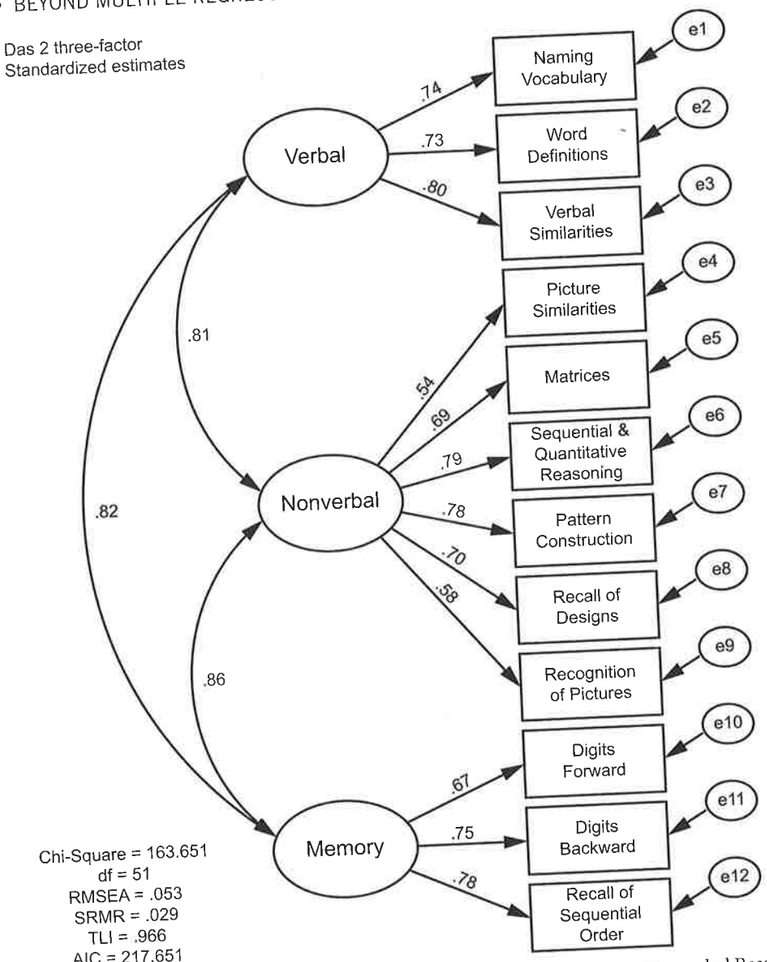


Figure 15.8 Another competing model of the DAS-II. This model combines the Nonverbal Reasoning and Spatial factors into a single Nonverbal factor.

in $\Delta\chi^2 = 35.825$ ($df = 3$), a value that is statistically significant ($p < .001$). This means that the three-factor combined Nonverbal model, although more parsimonious than the four-factor model, does not explain the relations among the DAS-II subtests, the DAS-II structure, as well as does the four-factor model. Said differently, the Nonverbal Reasoning and Spatial factors are indeed statistically distinguishable. The models shown in Figures 15.1 through 15.4 provide a better "theory" for understanding the DAS-II than does the model shown in Figure 15.8. Thus this analysis suggest that the DAS-II tests used in this analysis should be interpreted as measuring four, rather than three, underlying abilities. The fit indexes for this model are also shown in Table 15.2.

Although $\Delta\chi^2$ is our primary method for comparing competing, nested models, it is also worth noting the other fit indexes we discussed as useful for comparing (non-nested) models, the AIC and the aBIC. The rule of thumb for the AIC (and aBIC) is that they favor the model with the lower value; again the four-factor model appears superior if we use the AIC or aBIC to compare models. Again, according to our primary criteria, the four-factor model provides a better fit than does the three-factor model. Table 15.2 also includes fit indexes for the two steps outlined earlier for turning the standardized four-factor model into the three-factor model. (I will not show the analyses or models here, but I encourage you to conduct them.) In the first step, the factor correlation (standardized covariance) between the Nonverbal and Spatial factors was constrained to 1. In the second step, the Nonverbal-Verbal and the Spatial-Verbal factor correlations were constrained to be equal, as were the Nonverbal-Memory and Spatial-Memory factors. Note that the fit statistics associated with this second step are identical to those from the three-factor model as shown in Figure 15.8. As an aside, I don't believe it is necessary to conduct this analysis in two steps, but it does help to understand what is being done.

Before we move to the next topic consider why this approach is equivalent to that in which we simply combined the two factors into one. What we are testing is whether the Nonverbal and Spatial factors should really be considered as the same factor. What would be required for them to be "the same" factor? First and obviously, they should be perfectly correlated with one another. But a perfect correlation is not enough. If the Nonverbal and Spatial factors are really "the same," then they should also have the exact same relation (correlation) with other factors. The second step, constraining factor correlations to be the same value, fulfills this part of the requirement that the factors be "the same factor."

MODEL FIT AND MODEL MODIFICATION

A common response when a model does not fit well is to examine more detailed aspects of fit with an eye toward modifying the model. I won't try to dissuade you from this practice, because it is indeed useful and necessary, but I encourage you to do so sparingly, unless you are primarily involved in model development and exploration (as opposed to testing a priori models). I am not alone in this ambivalence concerning model modification: "As a statistician, I am deeply suspicious of modification indices. As a data analyst, however, I find they are really great" (Dag Sörbom, one of the authors of LISREL, quoted in Wolfle, 2003, p. 32). There are several aspects of the printout that may help in this process.

Modification Indexes

To illustrate the use of the more detailed fit indexes, let's examine the combined three-factor Nonverbal model from Figure 15.8. If we had started with this model—if we had not compared this model with the initial four-factor model (e.g., Figure 15.1), could we have figured out that the four-factor was better? Would the modification indexes or the other detailed fit statistics have led us to what we have concluded was a better model? And are there other changes we need to make in our models?

Figure 15.9 shows the modification indexes from the Amos output for this model. With some programs, all modification indexes are printed; with Amos you can request modification indexes above a certain level. The figure shows the default, indexes greater in magnitude than 4.0 (recall that 3.9, or approximately 4, is the value of $\Delta\chi^2$ that is statistically significant with 1 df). When models do not fit well, you may be able to improve the fit by freeing parameters in the model. Recall that freeing a parameter reduces the degrees of freedom (parsimony) of the model and improves the $\Delta\chi^2$ to some degree. The question we ask with such

Modification Indices

Covariances:

			M.I.	Par Change
e11	<-->	Nonverbal	5.334	2.320
e11	<-->	Verbal	5.116	-3.381
e8	<-->	Verbal	4.521	-3.288
e8	<-->	e7	21.895	7.878
e8	<-->	e9	14.549	8.641
e6	<-->	e8	5.886	-4.243
e5	<-->	e7	6.381	-4.414
e5	<-->	e12	7.027	5.542
e5	<-->	e10	7.128	-6.454
e5	<-->	e8	21.672	-9.739
e5	<-->	e6	21.896	8.493
e4	<-->	Verbal	10.847	5.969
e4	<-->	e6	4.884	-4.551
e2	<-->	Nonverbal	5.909	-2.399
e2	<-->	e7	4.741	-3.500
e2	<-->	e8	15.275	-7.488
e2	<-->	e5	4.273	4.107
e2	<-->	e4	4.568	4.787
e1	<-->	e7	6.716	4.340
e1	<-->	e11	6.197	-4.838
e1	<-->	e8	4.259	4.120
e1	<-->	e6	4.944	-3.870
e1	<-->	e5	4.929	-4.596

Regression Weights:

			M.I.	Par Change
pcss	<---	rdss	10.309	.072
dfss	<---	mass	4.597	-.064
rpss	<---	rdss	6.734	.078
rdss	<---	pcss	7.334	.079
rdss	<---	rpss	9.224	.081
rdss	<---	mass	10.490	-.085
rdss	<---	wdss	9.454	-.086
sqss	<---	mass	10.718	.075
mass	<---	rdss	10.090	-.088
mass	<---	sqss	6.989	.076
psss	<---	wdss	5.161	.074
wdss	<---	rdss	10.836	-.083

Figure 15.9 Modification indexes for the 3-factor combined model.

relaxations in the model is whether the decrease in $\Delta\chi^2$ is worth the reduction in the *df*. The modification indexes estimate the minimum decrease in $\Delta\chi^2$ that will result from freeing the listed parameter. Modification indexes are shown for covariances and for regression weights (the first row, for example, lists a modification index of 5.334). Although the actual output also had a table for variances, there were no modification indexes associated with variances greater than 4.0 so the table was blank and is not included in the figure.

Note the modification index for the covariance between e5 and e6: a value of 21.896. This modification index suggests that $\Delta\chi^2$ can be reduced by at least 21.896 by freeing the covariance between e5 and e6. Although this is a statistically significant decrease in $\Delta\chi^2$ with a *df* of 1, we need to consider whether this change makes theoretical sense. The variables e5 and e6 represent the unique variances of Matrices and Sequential & Quantitative Reasoning. The column marked "Par Change" shows the expected value of this parameter (covariance) if we were to free this constraint, that is, if we were to allow these two unique variances to correlate. Note that the expected parameter change is positive (this shows the expected value of this parameter in the unstandardized solution if it, and it alone, were freed). If we were to free this covariance (correlation), it would suggest that we think the unique variances of the Matrices and Sequential & Quantitative Reasoning subtests are related above and beyond the effect of Nonverbal Reasoning on each subtest. The factors correlate with each other because they are both affected by Nonverbal Reasoning, but could they be correlated for other reasons, as well? Stated differently, do Matrices and Sequential & Quantitative Reasoning measure something in common other than the factor Nonverbal Reasoning? Given our other analyses, it is fairly easy to answer this question: yes, these two subtests likely measure a more narrow Nonverbal Reasoning factor that is separate from Spatial ability. Note also the modification index for the covariance between e7 and e8 (21.895), suggesting that we free the covariance between the unique variances of the Pattern Construction and Recall of Designs subtests. Again, given our knowledge of the four-factor solution, we can say that yes, these two tests indeed do measure something in common, a Spatial factor that is separate from the Nonverbal reasoning factor.

If we had started with this three-factor (combined Nonverbal) model and IF we were skilled in reading the modification indexes, or IF we had some knowledge of the theory underlying the DAS-II, then the modification indexes may have suggested to us to split this factor into two factors. This example also illustrates that the modification indexes are not always easy to interpret!

The other large modification index in Figure 15.9 is between e5 and e8 (21.672). This modification index suggests that freeing the covariance between the unique variance for Matrices subtest and that of the Recall of Designs would results in a $\Delta\chi^2$ of at least 21. This "suggestion" by the modification indexes would seem to be in the opposite direction from the previous ones, because we know from the four-factor model that Matrices and Recall of Designs measure separate abilities. But note also that the expected parameter change is negative. This finding, in turn, suggests that these two tests measure less in common than our three factor model would predict. Again, given our additional knowledge, this finding also suggests the possibility of placing these two subtests on separate factors. The question is whether they would have suggested this possibility if we did not have this additional knowledge!

Here are common rules of thumb for using modification indexes. Examine the larger values of the modification indexes. Note that in actual practice you may have even more modification indexes to examine than those shown for this model. What is large? Modification indexes, like χ^2, are sample-size dependent; if our model fit much worse or if we had a larger sample size, we would have larger modification indexes and more modification indexes greater than 4.0. Thus, you should examine the larger values of the modification indexes relative to the other values. Again, the modification indexes show the expected minimum reduction in χ^2 if the listed parameter is freed, at a cost of 1 *df*. Next, consider whether each change is justifiable through theory and previous research. Make the single change that makes the most theoretical sense and results in the largest improvement in model fit, and then re-estimate the model. You can then repeat the process. Generally we don't use the

modification indexes to make several changes at a time, because with each additional change the modification indexes are likely to differ. I remind you to use the modification indexes cautiously. You will find it is all too easy to justify model modifications *after* examining modification indexes; do so sparingly and with an eye toward theory and previous research. If you see the modification index and smack yourself in the head because you should have thought of that model change a priori, then the model change is probably reasonable. If you find yourself having to do mental gymnastics to justify freeing a parameter, then you probably should not.

One final note on the modifications indexes. None of the MIs for the second table (Regression Weights) were particularly large, but if they had been, and if they were between a subtest and a factor, they would have suggested the possibility of allowing for cross-loadings of tests on other factors.

Residuals

Another aspect of fit to examine to understand why a model does not fit well is the matrix of standardized residuals (Standardized Residual Covariances) shown in Table 15.3 (this matrix is also from the output for results of the model analyzed in Figure 15.8). Recall from Chapter 13 that the various fit statistics examine the consistency between the actual covariance matrix and the covariance matrix implied by the model. The difference between these two matrices is the matrix of residual covariances; the matrix of standardized residual covariances simply puts these residuals on the same standardized scale so that they can be compared. That matrix is shown in Table 15.3.

Standardized Residual Covariances

For this matrix, as well, we are looking for relatively larger values, regardless of sign. One rule of thumb suggests examining standardized residual covariances (commonly referred to as standardized residuals) greater in absolute magnitude than 2.0; but the standardized residuals are also sample-size dependent, so with larger samples you may have many values greater than 2, whereas with smaller samples there may be few or no standardized residuals that

Table 15.3 Standardized Residual Covariances for the Three-Factor Combined Nonverbal Model.

	pcss	*soss*	*dbss*	*dfss*	*rpss*	*rdss*	*sqss*	*mass*	*psss*	*vsss*	*wdss*	*nvss*
pcss	.000											
soss	–.548	.000										
dbss	.192	–.040	.000									
dfss	–.455	.173	–.130	.000								
rpss	–.226	–.383	–.111	–1.014	.000							
rdss	***1.614***	–.411	.264	.353	***1.910***	.000						
sqss	–.196	–.095	.693	–.411	–.125	–.815	.000					
mass	–.891	1.163	1.169	–1.123	–.238	***–1.981***	***1.609***	.000				
psss	–.288	–.680	.050	.322	.377	.533	–.936	–.052	.000			
vsss	.420	–.207	–.166	.447	–.655	–.531	.650	.387	1.291	.000		
wdss	–1.036	.552	–.581	.295	–1.372	***–2.297***	–.316	.610	1.480	.108	.000	
nvss	1.012	.227	–1.052	1.005	–.316	.489	–.542	–.609	1.366	–.348	.365	.000

reach this level. Again, focus on the relatively larger values; these are bolded and italicized in the table. For the present example, the combined Nonverbal DAS-II model, there is only one value greater than 2.0, between the Recall of Designs and the Word Definitions (–2.297).

What does this value mean? Recall how this matrix is created: the implied covariance matrix is subtracted from the actual covariance matrix to create the residuals. The residuals are then standardized to create this matrix. This means that for positive values the *actual* correlation between two measured variables is larger than the *implied* correlation. For negative residuals, just the opposite is the case: the implied correlation is larger than the actual correlation. This means that positive standardized residuals suggest that the model does not adequately account for the observed correlation between two variables, whereas for negative residuals the model more than accounts for the original correlation between variables. Positive residuals are thus generally more informative for purposes of model modification in that they suggest ways the model can be modified to improve the fit.

In the current example, the highest value, –2.297, is between Recall of Designs (rdss) and Word Definitions (wdss). The value is negative, which suggests that the model—in which these subtests load on the Spatial and Verbal factors, factors which correlate .75—more than accounts for the correlation between Word Definitions and Recall of Designs. Given the loadings of these subtests on their factors, and given the correlation between the factors, we would expect these two subtests to be more highly correlated than they are. This standardized residual thus seems to hint a different aspect of local misfit than we saw with the modification indexes, although it is not clear what this means or if there is anything we should do about it.

The other larger standardized residuals (those with values greater than 1.5 are highlighted) tell the same story as did the modification indexes. There are high positive values for Recall of Designs with Pattern Construction and with Recognition of Pictures. The model does not adequately explain the correlations between Recall of Designs and the other two Spatial tests. Likewise, the model does not adequately explain the correlation between the Matrices and the Sequential & Quantitative Reasoning subtests (both measures of Nonverbal Reasoning), but more than accounts for the correlation between Matrices and Recall of Designs (which, in the four-factor model measure two different underlying abilities). Again, if we were skilled and theoretically savvy, we might have taken these as hints that these six subtests should be split into two factors rather than loaded all on one. Or maybe not. The other thing the pattern of higher loadings suggests is that the Recall of Designs subtest is a general source of misfit in this model.

Residual Correlations

Table 15.4 shows a related but potentially useful matrix, the matrix of residual correlations. As noted in chapter 13, this matrix shows the residuals for the actual and implied correlation matrices. The downside is that many SEM programs do not produced this matrix (Amos does not, as least as of this writing). But the matrix is easy to produce; I simply copied and pasted the sample correlation matrix and the matrix implied by the model into Excel and subtracted the latter from the former. Again I have highlighted the higher values in this matrix (here, values greater than .06 in absolute value). Note that the subtests highlighted the same as in the previous table, which should always be the case. This table, however, shows differences in correlations, so the values are readily interpretable. The value for Recall of Designs and Word Defintions (–.088) means the actual correlation between these two subtests is .088 lower than that predicted by the model. If you focus on the model (Figure 15.8), it shows Word Definitions with a standardized loading of .73 on the Verbal factor and Recall of Designs with a loading of .70 on the Nonverbal factor, with these factors correlating .81 with each other. The expected or implied correlation between these two subtests would thus

Table 15.4 Residual Correlations for the Three-Factor Combined Nonverbal Model.

	pcss	*soss*	*dbss*	*dfss*	*rpss*	*rdss*	*sqss*	*mass*	*psss*	*vsss*	*wdss*	*nvss*
pcss	0											
soss	–.022	0										
dbss	.008	–.002	0									
dfss	–.018	.007	–.005	0								
rpss	–.009	–.015	–.004	–.038	0							
rdss	***.065***	–.016	.010	.014	***.073***	0						
sqss	–.008	–.004	.028	–.016	–.005	–.033	0					
mass	–.036	.045	.045	–.043	–.009	***–.078***	***.065***	0				
psss	–.011	–.026	.002	.012	.014	.020	–.036	–.002	0			
vsss	.017	–.008	–.007	.017	–.025	–.021	.026	.015	.048	0		
wdss	–.040	.022	–.023	.011	–.051	***–.088***	–.012	.023	.055	.004	0	
nvss	.040	.009	–.041	.038	–.012	.019	–.021	–.023	.051	–.014	.015	0

be .73 x .70 x .81, or .41 using the tracing rule. In fact, the actual correlation between these two subtests was .32, a difference of –.09 (rounded). Again, this model predicts a higher correlation between these two subtests than was found in the actual data.

For both the standardized residuals and the residual correlations, consider whether the larger positive values share some characteristic in common (you can do the same for the larger negative values, which may suggest additional constraints to the model). Although the residuals are somewhat more difficult to interpret than the modification indexes, they also sometimes show a pattern, and thus may be very useful in suggesting additional paths, correlations, or even minor factors to add to a model.

The residual correlations should highlight the same sources of misfit as the standardized residuals. The advantage is that these residuals are on a scale with which we are familiar, that of a correlation coefficient. As a result, we can devise informal rules of thumb for problematic values. Kline, for example, suggests "correlation residuals" greater than .10 as potentially problematic (2011, p. 202). Kline also suggests examining the residual correlations whenever the χ^2 for the model is statistically significant; I would simply add that this is a good idea when any of the fit indexes suggest a lack of fit.

Adding Model Constraints and *z* Values

You can modify a model by relaxing constraints to the model (estimating a parameter that was previously set to zero), as discussed previous. Model relaxations will always improve χ^2, but will make the model less parsimonious. Sometimes the relaxation of constraints is worth the improvement in fit. Another direction in modifying models is to add constraints, generally by constraining a previously estimated value to zero (or some other value). If, for example, some of the factor loadings had been statistically not significant according to the critical ranges (*z* values), we might have constrained these values to zero (i.e., removed the path) in subsequent models. Adding constraints to the model will always lead to a larger (worse) χ^2, but a more parsimonious model. If the $\Delta\chi^2$ is not statistically significant, the constraint makes

sense. These same rules apply to many other fit indexes, as well: relaxations will improve fit, constraints will degrade fit. The exception to this rule is with fit indexes that take model parsimony into account; these indexes may improve with constraints and degrade with relaxations. Of the indexes we have discussed, the TLI, RMSEA, aBIC, and (commonly) the AIC also take parsimony into account. Indeed, the AIC and related fit indexes (e.g., BIC) are designed to prevent "overfitting," or making small, sample-specific changes solely to improve fit.

Cautions

I again encourage you to be cautious when making model modifications. Extensive model modifications will take you far afield from the supposedly confirmatory, theory-testing nature of SEM and CFA and can even lead to erroneous models (MacCallum, 1986). Some authors make the useful distinction between the use of SEM and CFA in a theory-testing versus a more exploratory matter (Joreskog & Sorbom, 1993). I believe this is a useful distinction, and encourage you to know where you are along this continuum. If you make more than minor changes to your model, you should not think of what you are doing as theory testing unless you have retested the model with new data.

HIERARCHICAL MODELS

Higher-Order Model Justification and Setup

The analyses so far have pointed to the model in Figure 15.1 as a more valid representation of the structure of the DAS-II than the models in Figures 15.7 and 15.8. But the model in 15.1 is not complete, either. In addition to measuring the four abilities shown in Figure 15.1, the DAS-II is also designed to measure overall general intelligence. The model shown in Figure 15.10, then, is probably a more accurate reflection of the intended structure of the DAS-II: rather than simply having the first-order factors correlated, these factors are shown as reflections of second-, or higher-order factor, general intelligence, usually symbolized as *g*, in a hierarchical model. Note that this type of hierarchical model (with higher-order factors) is generally referred to as a higher-order model; another type of hierarchical model (the bifactor model) is also discussed (also see Keith & Reynolds, 2012 or Reynolds & Keith, 2013).

There are several reasons for developing and estimating higher-order models. In the arena of intelligence, higher-order models are more consistent with commonly accepted theories of intelligence (e.g., three-stratum or Cattell–Horn–Carroll theory, Carroll, 1993) than are first-order models and are more consistent with the actual structure of most intelligence tests. Higher-order and other hierarchical models may be equally relevant in many other areas of research. Higher-order models can also lead to a better understanding of the first level of factors. Just as the first level of factors helped us understand what the subtests measured, the second-order factor(s) may help us better understand the first-order factors.

The mechanics of estimating a higher-order CFA also need comment. Note that the scale of the second-order factor (*g*) is set in the same way as the first-order factors, by fixing one path from it to one of the first-order factors to 1.0. We could also set the scale by fixing the variance of *g* to 1.0 (in this case we would still need to set the scale of the first-order factors by setting a path to 1.0, and thus the first-order factor solution will not be standardized). The higher-order model differs from the first-order model in that the first-order factors have small latent variables pointing toward them, labeled uf1 through uf4 (for unique factor variance). These latent variables have the same essential meaning as other disturbances/residuals: they represent all influences on the first-order factors (Verbal, Nonverbal Reasoning, etc.) other than *g*. To put it another way, *any variable—whether measured or latent—that has an*

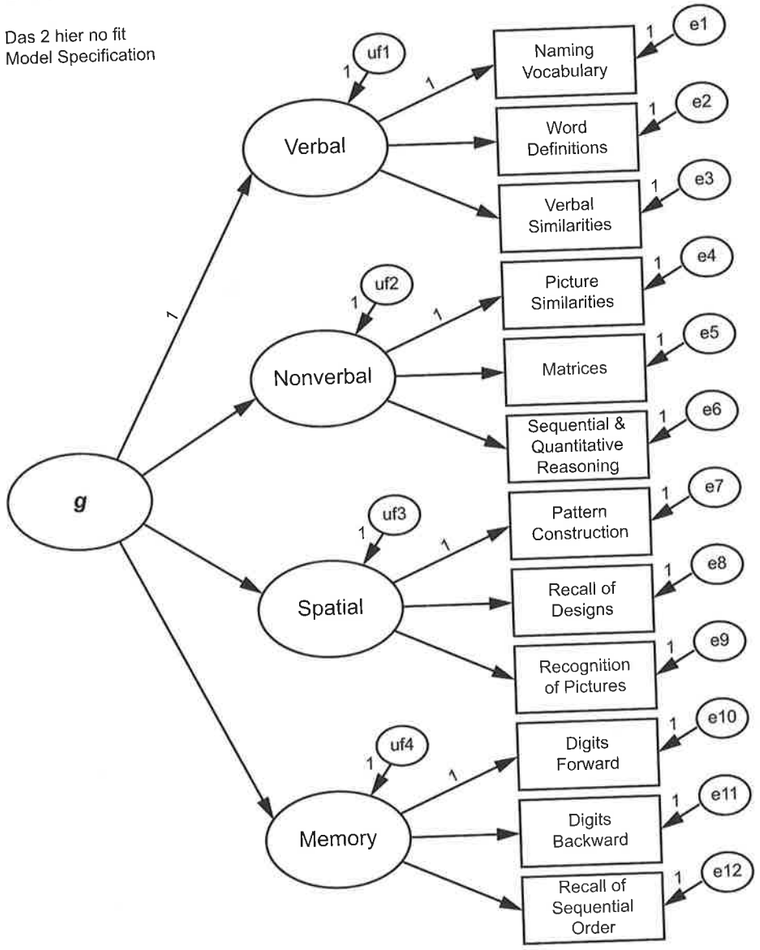


Figure 15.10 Higher-order model of the DAS-II. The model specifies that the DAS-II measures general intelligence in addition to the four broad cognitive ability factors.

arrow pointing to it must also include a latent disturbance/unique variable to represent all other influences on the variable. Finally, the model shown here includes three levels—measured variables, first-order factors, and a second-order factor—but additional levels are possible and are capable of estimation using these same methods.

Higher-Order Model Results

Figure 15.11 shows the fit statistics and standardized estimates for the higher-order analysis. Note that the first-order factor loadings are the same as they were for the initial, first-order analysis (Figure 15.2; these will not always be identical but should be very similar).

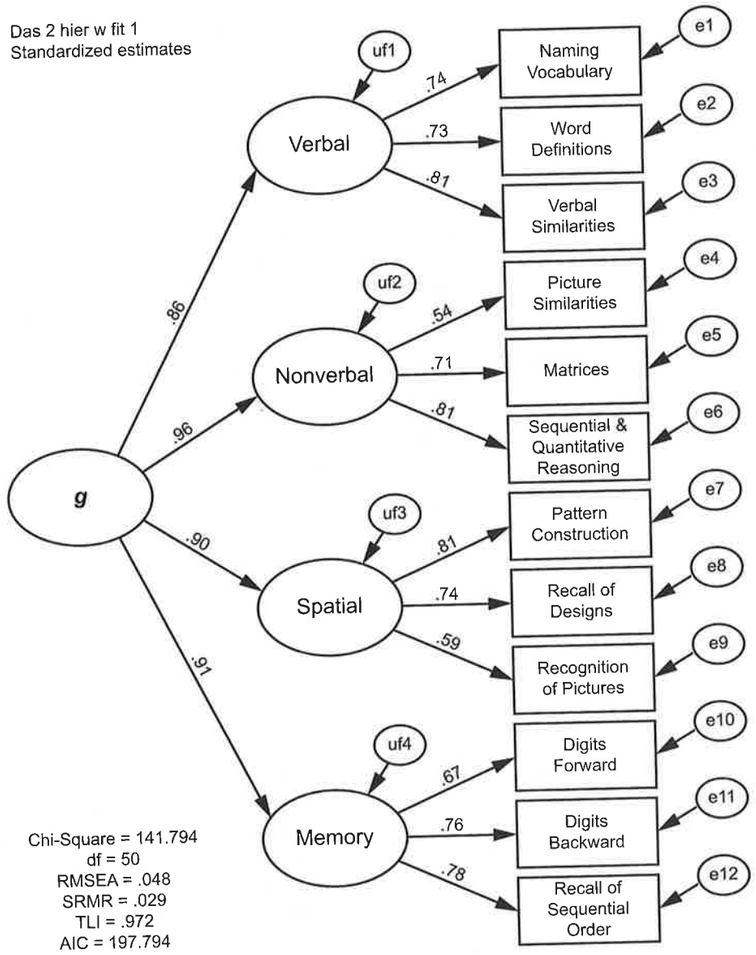


Figure 15.11 Standardized estimates for the higher-order DAS-II model.

The equivalence is because the essential difference between the higher-order and the first-order model is that the higher-order model explains the correlations (covariances) among the first-order factors with a specific structure. The first-order factor model helps explain why the *subtests* correlate with each other: because there are four abilities that partially cause students to perform at a certain level on the eight subtests. The second-order model adds to that a possible explanation of the reason for the correlations among the four *factors*: because there is one general intellectual ability factor that influences, in part, the four more narrow abilities. Conceptually, the factor analysis of latent variables (second-order) is equivalent to the factor analysis of measured variables (first-order).

The fit of the model looks good; with the exception of the statistically significant χ^2, all indexes suggest a good fit of the model to the data. An examination of the modification

indexes, standardized residuals, and correlation residuals show similar results as for the initial four-factor model, and suggest no major problems.

Given that the higher-order model is the same as the initial four-factor model with the addition of paths explaining the correlations among factors, it should be clear, then, that the higher-order model may be considered a more constrained, more parsimonious version of the first-order model (Rindskopf & Rose, 1988). The first-order model places no constraints on the factor correlations, whereas the higher-order model says that these correlations are the product (in this case) of another latent variable, *g*. Given this similarity, you may consider the two models as nested, and thus we could use $\Delta\chi^2$ to compare the two models. If we were to do so, $\Delta\chi^2 = 13.968$ [2], $p = .001$, and we would likely reject the higher-order model as not worth the increase in χ^2 compared to gain in parsimony. Likewise, the AIC and aBIC are worse for the higher-order model compared to the four-factor first-order model. I generally don't compare first-order with higher-order models in this way (as nested models), however. It seems to me that at least in the area of intelligence, such models are justified on purely theoretical grounds, without reliance on fit indexes to compare them to agnostic, non-higher-order models. In addition, theorists recognize the likelihood of there being intermediate factors between the first-order factors and *g*, (Carroll, 1993, chap. 16) and such factors, if accurate, would improve the fit of the higher-order models. With the DAS-II, for example, allowing the unique variances of the Nonverbal and Spatial factors to correlate would lead to a higher-order model that fit as well as the first-order model (Keith et al., 2010). Allowing this "correlated error" is statistically equivalent to specifying that Nonverbal Reasoning and Visual-Spatial skills are reflections of an intermediate factor between them and *g* (can you figure out why this would be the case?). Or perhaps I just have a soft spot in my heart for higher-order models of intelligence.

Let's be sure we understand where these two extra degrees of freedom come from as you look over the model. For the first-order model, there were six covariances among the first-order factors and four first-order factor variances. This is calculated as $\frac{p\times(p+1)}{2}$) "moments" in the variance/covariance matrix where p = the number variables, in this case, first-order factors, and thus $\frac{4\times(5+1)}{2}$) = 10. The higher-order model uses up eight of these free parameters to estimate three of the second-order to first-order factor loadings (recall that one path was set to 1), along with the variance of the *g* factor and the variances of the new disturbances (uf1 through uf4), leaving two extra *df*. This means that if there are only three first-order factors the higher-order portion of the model will be just-identified; the two models will then have identical fit and cannot be compared statistically. If we try to add a higher-order factor to a model with only two first-order factors, the higher-order portion of the model will be underidentified and estimation will be impossible unless we make additional constraints (e.g., constraining the two second-order loadings to be the same). You need to pay attention to the identification status of the higher-order portions of such models (identification was discussed in Chapter 11).

Recall that one reason for investigating higher-order models is to help understand the first-order factors. Indeed, the second-order factor loadings are interesting. The highest loading (near 1.0, unity) was by the Nonverbal Reasoning factor. Nonverbal Reasoning thus appears to be the most intellectually laden of the first-order factors. This finding suggests that the deductive and inductive reasoning that underlies the tasks on this factor is close to the essence of general intelligence.

Total Effects

Psychometric researchers are often also interested in understanding which of the subtests are most highly related to the global general intelligence factor. We can calculate these loadings of the subtests on the second-order factor by multiplying paths (e.g., the loading of Word Definitions on *g* would equal .86 x .73 = .63). If this process sounds familiar, it should; we

Standardized Total Effects, Higher-Order Model

	g	Memory	Spatial	Nonverbal	Verbal
Memory	.913	.000	.000	.000	.000
Spatial	.903	.000	.000	.000	.000
Nonverbal	.955	.000	.000	.000	.000
Verbal	.858	.000	.000	.000	.000
pcss	**.736**	.000	.815	.000	.000
soss	**.709**	.776	.000	.000	.000
dbss	**.693**	.758	.000	.000	.000
dfss	**.608**	.665	.000	.000	.000
rpss	**.531**	.000	.588	.000	.000
rdss	**.667**	.000	.738	.000	.000
sqss	**.770**	.000	.000	.806	.000
mass	**.682**	.000	.000	.714	.000
psss	**.516**	.000	.000	.540	.000
vsss	**.692**	.000	.000	.000	.807
wdss	**.626**	.000	.000	.000	.730
nvss	**.635**	.000	.000	.000	.740

Figure 15.12 Standardized total effects for the higher-order model. The bolded coefficients are the total effects of *g* on the subtests. These may also be considered the loading of the subtests on the higher-order *g* factor.

are simply calculating the indirect effect of *g* on each subtest. Because there are no direct effects (all the effects from g are mediated by the first-order factors), these indirect effects are also the total effects. Figure 15.12 shows the total standardized effects of *g* on the first-order factors and subtests (for some reason, the order of sub-tests in the Figure is almost reversed from the order of the subtests in the figure). The total effects from *g* to subtests are shown in boldface. As shown in the figure, the Sequential and Quantitative Reasoning subtest had the highest total effect from *g* (.770). Thus, this subtest is most closely related to *g*, or *g* has a stronger effect on this subtest than on any of the other subtests.

Bifactor Model Justification and Setup

There is another type of hierarchical model, often known as the bifactor model. You may see this model referred to by other names, as well, including the nested-factors or direct hierarchical model. A bifactor version of the DAS-II is shown in Figure 15.13. This model, like the higher-order one, includes both Verbal, Nonverbal, and the other first-order factors, and it also includes a more general factor, here symbolized as *G*. With the bifactor model, however, both the narrow and the general factor are first-order factors, whereas in the higher-order model the general factor is a higher-order one designed to explain the correlations/covariances among the first-order factors. Because the general factor in a bifactor model is also a first-order factor, it is often symbolized in intelligence models as *G* as opposed to *g* (used for a second- or higher-order factor).

Note several other aspects of the bifactor model. First, note that the more narrow factors (the "broad abilities" in intelligence lingo) are often specified as uncorrelated with one another, and as uncorrelated with the general factor. This is done because if we allowed all of these to be correlated with one another the model would be underidentified and thus we could not estimate it. Because models imply theories, the bifactor model thus says that

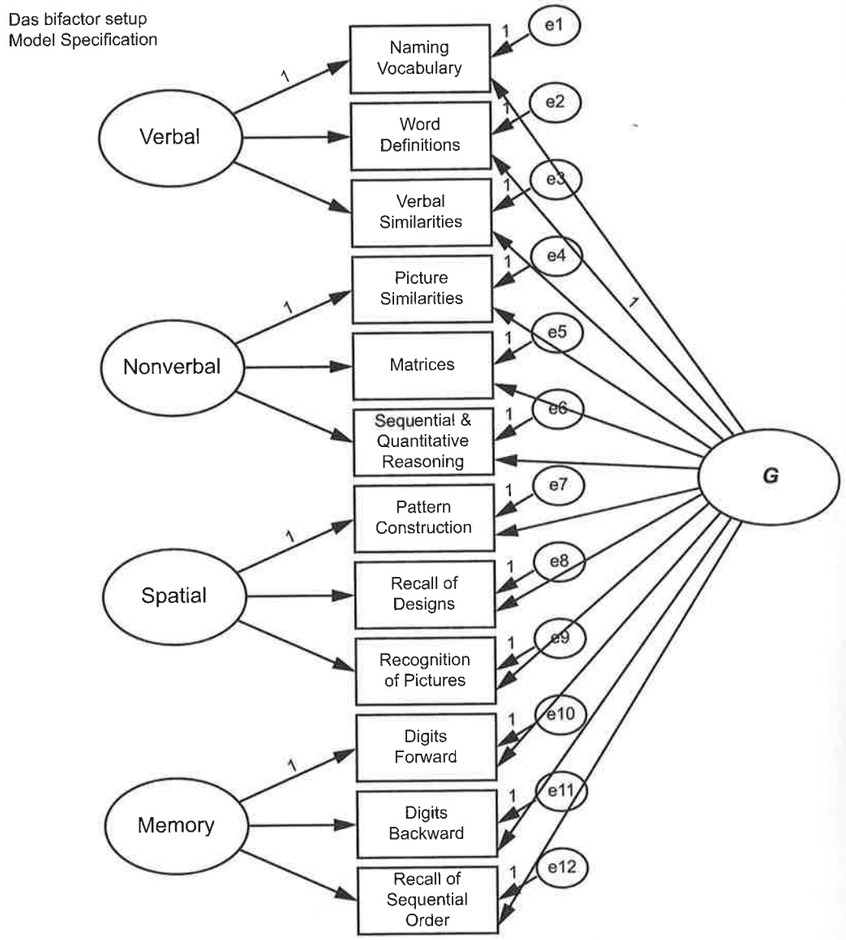


Figure 15.13 A bifactor hierarchical model for the DAS-II. This model has both *G* and the broad abilities as first-order factors.

the broad abilities and *G* are unrelated to one another. This model also says that each of the DAS-II tests measures two things: a general ability shared by all the DAS-II tests, and one of four other underlying constructs. Note also that the scales of both the broad abilities and *G* were set using ULI (unit loading identification). It is also possible to use UVI to set the scale for the broad abilities, or for *G*, or for both.

Bifactor Model Results

The initial analysis of the bifactor model returned the error message that the variance associated with e6 was negative, as shown in Figure 15.14. Variances, which are squared terms (one

The following variances are negative. (Group number 1 - initial model)

	e6
	-35.264

Figure 15.14 Error message for the initial bifactor model. Variances cannot be negative.

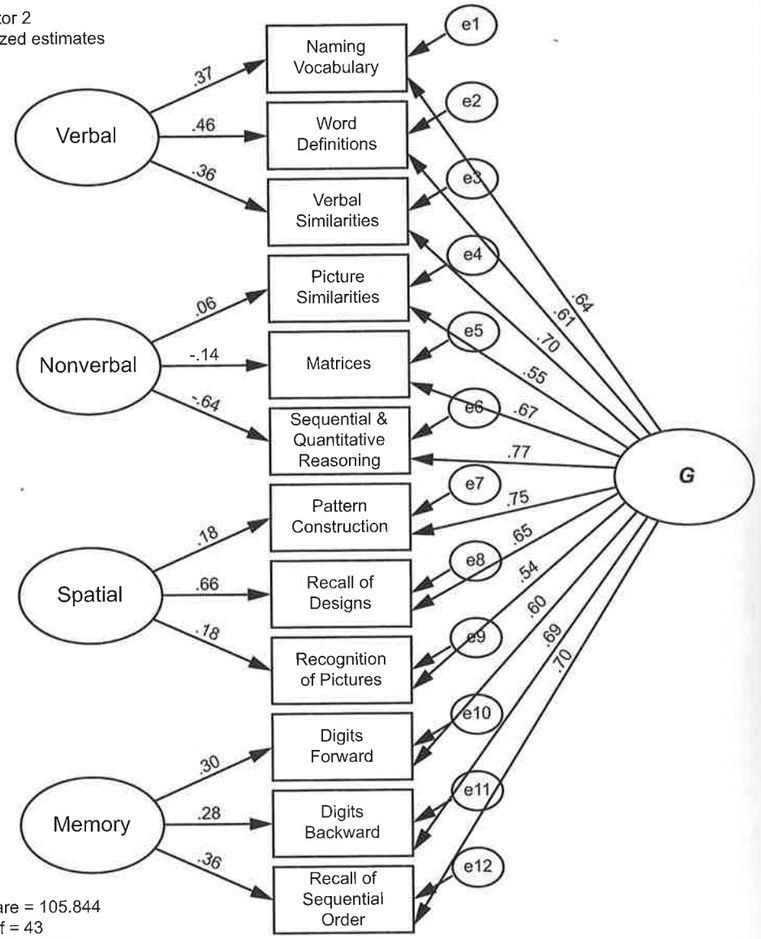
way of thinking of them is that they are the standard deviation squared), cannot be negative, so the model would not run. This problem is common enough in factor analysis that it has a name, a "Heywood case." In CFA, a Heywood case generally shows up as a negative error variance resulting from the path or paths to a variable explaining 100% or more than 100% of its variance. In the present example, the Nonverbal and G factors, together, explain more 100% of the variance in the Sequential and Quantitative Reasoning test. It is worth noting that Heywood cases are not unique to bifactor models; they also show up in higher-order models (one should always check the first-order residual/disturbance variances carefully in higher-order models), and even in first-order models. One common method of dealing with a negative variance is to set the offending value to zero.[1]

Figure 15.15 shows the standardized results for the bifactor analysis with the error variance for Sequential and Quantitative Reasoning (e6) constrained to zero. As shown in the figure, the model fit the data well, with RMSEA = .043, SRMR = .025, and TLI = .977. Indeed, the bifactor model fit better than did the higher-order model (AIC = 175.844 versus 197.794 for the higher-order DAS-II model). We will return to this issue of fit momentarily.

Beyond fit, you probably noticed a few curious aspects of the model results, like the negative loadings of two of the tests on the Nonverbal factor. These exist because the Picture Similarities test was chosen as the reference variable for the unit loading identification. If the loading from Nonverbal to Matrices or to Sequential and Quantitative Reasoning (SQR) had instead been set to one, the Picture Similarities tests would have shown and small negative loading on the Nonverbal factor (–.06) and the Matrices and SQR loadings would have been positive (.14 and .64, respectively).

The loadings of each test on *G* were large and statistically significant. Note also how similar these values are to those shown as the total effects of *g* on the subtests for the higher-order model (Figure 15.12). Although the rank order changes slightly, the subtests that were the best measures of *g* for the higher-order model are also the best measures of *G* for the bifactor model, and the worst for one are also the worst for the other. In contrast, note how much lower are the loadings for the subtests on the four broad factors in the bifactor model compared to all previous models. Indeed, although not shown in the figure (but would be in the detailed output), some of these paths/factor loadings are not statistically significant. Why, you may wonder? The short answer is that these two models imply quite different theories about the nature of intelligence. The higher-order model says that the primary reason that the 12 tests shown correlate with one another is that they measure four underlying cognitive abilities. *g*, in turn, affects these broad cognitive abilities, and *g* affects the specific tests only indirectly. The bifactor model, in contrast, says that there are two reasons for the correlations among these 12 tests: first, they all measure *G*, and second, they all measure some other broad cognitive abilities that are independent from one another. In the bifactor model, *G* has direct effects on the specific tests. In the higher-order model, then, *g* can be understood by the nature of the broad cognitive abilities that underlie it, and those cognitive abilities can be understood as more or less related to *g*. For the bifactor model, the nature of *G* can be referenced to specific tests.

Das bifactor 2
Standardized estimates



Chi-Square = 105.844
df = 43
RMSEA = .043
SRMR = .025
TLI = .977
AIC = 175.844

Figure 15.15 Standardized bifactor model results. The variance associated with residual e6 was set to zero to allow estimation. See the text for the explanation of the negative factor loadings for the Nonverbal broad ability.

Comparing the Hierarchical Models

It is possible to obtain similar (smaller) loadings for the higher-order models as for the bifactor model, and doing so also aids in understanding their differences (or similarities). One way of doing so is illustrated in Figure 15.16. For the previous higher-order models I specified that the paths from the disturbances for the first-order factors were equal to 1 and that

Das hier like bifactor
Model Specification

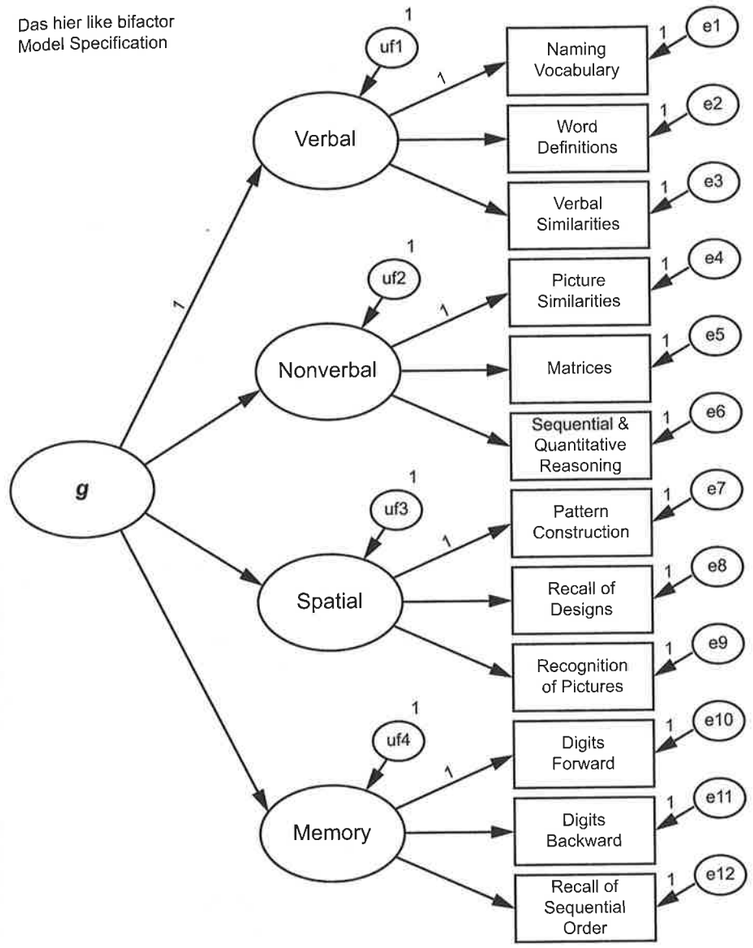


Figure 15.16 Model setup allowing the comparison of the broad ability loadings from the higher-order to the bifactor models.

the unique factor variances were estimated (i.e., ULI specification for uf1 through uf4). In Figure 15.16, in contrast, UVI was used for identification of the disturbances. With this setup it is possible to calculate the indirect effect of uf1 through uf4 on the various DAS-II subtests. These indirect effects are generally quite similar to those for the loading of the subtests on the broad abilities in the bifactor model (they are not identical because the underlying models are different). Consider what this means. uf1 through uf4 represent all other influences on the broad abilities, once *g* is taken into account. These indirect effects, then, represent the unique effects of the broad abilities on the subtests, *once g is removed*. Such estimates may indeed be of interest, and are in fact equivalent to the Schmid-Leiman transformation that is a popular method for interpreting higher-order exploratory factor results.

Although it is common to treat the bifactor and higher-order models as non-nested (as we have done here), it is possible to go from the bifactor model to a model that is equivalent to the higher-order model. Note that because *g* affects the subtests only through the broad abilities in the higher-order model, this places constraints on the relative loadings on the subtests on *g*. It would be possible, then, to go from the bifactor to the higher-order model by adding proportionality constraints (a topic beyond the scope of this text) (Yung, Thissen, & McLeod, 1999). What is important to realize at this stage of understanding is that higher-order model is equivalent to a more constrained version of the bifactor model. Thus, the bifactor model will generally fit as well or better than the more constrained higher-order model.

The bifactor model is popular right now (my colleague Tiffany Whitaker calls it the "little black dress" of CFA), and it does indeed have some advantages over a higher-order model (Chen, West, & Sousa, 2006; Reise, 2012). Chief among these is that it fits as well or better than does a higher-order model (see the previous example). One could consider its lack of specification of a relation between *G* and the broad factors as agnostic (not sure how they are related) rather than well-defined (it actually specifies that they are unrelated). With this change in thinking, the bifactor model would seem to be a good choice for a hierarchical model when there is no theory specifying how general and broad factors are related, or when that theory is undefined on this point. I will note that this is not the case in the area of intelligence, but it may be the case in many other areas where hierarchical CFA is of interest.

The bifactor model also has some disadvantages. As should be obvious by now, models imply theories, and the model you choose should be consistent with the theory you wish to test. Although some researchers treat the higher-order and the bifactor models as interchangeable, even our cursory explanation shows that they imply different theories. If one of these models is more consistent with the theory you wish to test, then that is the one you should use. If one theory says the structure of your construct of interest is one way (e.g., a bifactor-type model), and another theory says it is another way (e.g., a higher-order-type model), then you should compare the two (with knowledge that the bifactor model will fit as well or better). Such comparisons should make reference to the underlying theory being tested.

Another problem with the bifactor model is that it is not always easy to estimate, and the results can be quirky. With the present example you saw that we had to make an additional constraint to one error variance in order for the model to work. The fact that this model specifies that each measure is a reflection of two underlying factors sometimes leads to problems with estimation and convergence of the factor solution. As a result, it is not unusual to have to specify "start values," or initial guesses of what parameters might be (this is easy to do in most SEM programs, although you may have to do some digging to find out how). If the broad ability factors are uncorrelated, they must be referenced by three or more measures or the model will be underidentified.[2] More concerning is the fact that you may get different results for the bifactor model depending on how you go about estimating your model. In the present example, when I constrained the second test on each factor as the reference variable (i.e., Matrices loading set to 1 instead of Picture Similarities), the various standardized loadings showed the same magnitude but a different pattern of nonsignificance (e.g., the unstandardized SQR on Nonverbal loading was nonsignificant for the initial analysis but statistically significant for this one). When a UVI specification was used (factor variances set to 1), there were many fewer nonsignificant factor loadings. Finally, when I analyzed the initial model in Mplus, it suggested a negative variance for e5, whereas Amos suggested a negative variance for e6. All these differences are likely related to the fact that the variances for some of broad abilities were small and, depending on estimation method, nonsignificant. But whatever the reason, finding such differences is disconcerting (cf. Milsap, 2001), and in my experience they are more common with a bifactor as opposed to a higher-order factor model.

A final disadvantage of the bifactor model is that it may lend support to an incorrect model (Maydeu-Olivares & Coffman, 2006; Murray & Johnson, 2013). These simulation studies show that, for example, the bifactor model may fit the data better than a higher-order model, even when a higher-order model is the correct model (Murray & Johnson, 2013).

My current take on the bifactor model, as compared to a higher-order model, is that the bifactor model may indeed be a useful model when one is agnostic or unclear about how the most general factor should relate to the more specific factors. Likewise, if one believes that the structure of the underlying data conform to something like a bifactor model, then it should be used in those cases as well. When the guiding theory specifies a higher-order relation between the most general and more specific factors, however (as with most theories of intelligence), the bifactor model results may be misleading. I am not sure if these tentative conclusions will be supported five years from now, however. Despite the long history of the bifactor model, we are still learning about it! For more detailed comparisons of the two models see some of the references already listed (Chen, West, & Sousa, 2006; Murray & Johnson, 2013; Reise, 2012). Keith and Reynolds (2012) and Reynolds and Keith (2013) also compare these two models with intelligence data, and Mulaik and Quartetti (1997) and Yung and colleagues (1999) show some important statistical comparisons.

ADDITIONAL USES OF MODEL CONSTRAINTS

Occasionally, it is useful to be able to specify single-indicator factors. This may seem impossible, given that we earlier noted that we needed to have multiple measures of each construct to have a latent variable model. As you will see, with single indicators the portion of the measurement model is underidentified, but there are ways of working around this problem.

Pretend for this example that the DAS-II only included a single measure of short-term memory skills, the Digits Forward subtest. Is there some way we could model a Memory factor despite this weakness in the data? There are several ways we could do so. One method is shown in Figure 15.17, which shows a Memory factor with a single indicator, Digits Forward. This sort of model is more difficult to estimate because, without further constraints, this portion of the model is underidentified. We can work around this problem of estimating a single-indicator latent variable in SEM (and CFA) by fixing the value of the unique–error variance to some value; this brings this portion of the model back into a just-identified state.

We could, of course, constrain the value of the unique/error variance to zero. This approach tacitly suggests that we believe the measured variable is measured without error, that the measured variable and the factor are exactly the same. Whether we realized it or not, this is what we were doing when we were analyzing path models (and when we were doing multiple regression): we assumed that a single measure was a perfectly valid and reliable indicator of the constructs we were interested in.

Another approach is to use information about the estimated *reliability* of the measured variable in the model, if we know it or can estimate it. One minus the reliability provides an estimate of the proportion of error in the measured variable; if this value is multiplied by the variance of the variable, the result is the variance in the measured variable that can be attributed to error. Figure 15.17 shows a model that uses this methodology. The estimated (internal consistency) reliability for the Digits Forward test, across ages 5–8, is .91 (Elliott, 2007), and the variance of Digits Forward for the present sample is 121.523 (from the variance/covariance matrix). The estimate used for the error variance for Speed of Processing (u9) is thus 10.94:

$$V_e = (1 - r_{tt})V = (1 - .91) \times 121.523 = 10.937 .$$

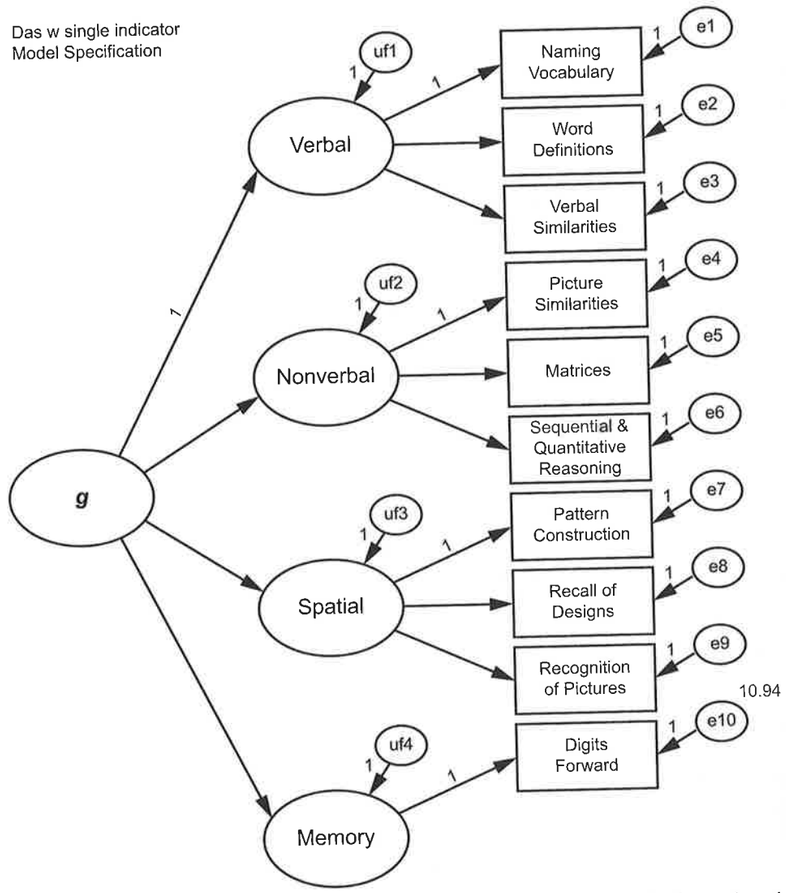


Figure 15.17 Modeling a single-indicator factor. In this model the memory factor has only a single measured variable.

Study this portion of the model. As for all other factors, one path from the latent to the measured variable is set to 1 in order to set the scale. The only difference is that there is only one path from the factor to the measured variable. The path from the unique variance to the subtest is also set to 1, again to set the scale. Recall when we discussed estimating path models via SEM programs we noted we can either estimate the path from the disturbance or estimate the variance of the disturbance. It is the same with the unique and error variances. Normally, we set the path from the unique–error variance to 1 and estimate the unique and error variance. With only a single measured variable, we have to fix the unique–error variance as well as the path to allow model estimation. The value 10.94 beside e10 shows that we have done so, and with this constraint we can estimate the model successfully. Again, this is a common method for dealing with single-indicator latent variables; for more detail, see Hayduk (1987, chap. 4). In fact, this was the method I used to estimate the models showing

the effects of different degrees of error in the previous chapter. It is also possible to use estimates of validity to account for both unreliability and invalidity. The use of reliability probably provides a very conservative (lower-bound) estimate for the unique and error variance (e10). In the complete higher-order model (the model in Figure 15.11), the estimate for Digit Forward's unique and error variance was 67.91 (this information is contained in the text output or the unstandardized estimates, neither of which are shown here). Some writers recommend using a range of values in such single-indicator analyses to make sure the estimates obtained for loadings and paths are reasonable.

The results of this analysis are shown in Figure 15.18. With this approach the Memory factor had a considerably lower loading on the *g* factor than did the other first-order factors (and it was considerably lower than in higher-order model with three memory indicators).

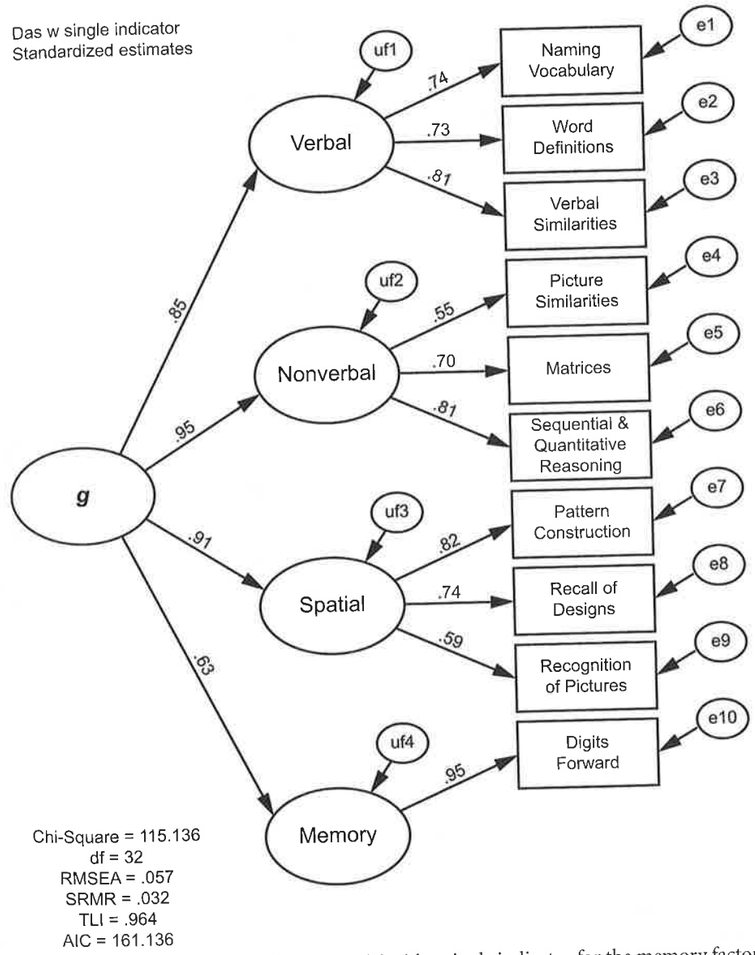


Figure 15.18 Standardized solution for the model with a single indicator for the memory factor.

Although this method allows us to estimate a model with single-indicator factors, it obviously provides less information about these factors than do factors defined by multiple measured variables. For the current example, the model tells us the relative effect of *g* on Memory (with Memory defined as very closely related to the Digits Forward), but it provides little additional information concerning the nature of the Digits Forward subtest or the Memory factor. Although many SEM users regard this method for dealing with single indicators as a trick to allow estimation, Hayduk has argued persuasively for advantages for this approach in path analysis and SEM (1987).

Let's briefly review two alternative methods for dealing with single indicator factors. Figure 15.19 shows the results of a model in which the unique and error variance for Digits

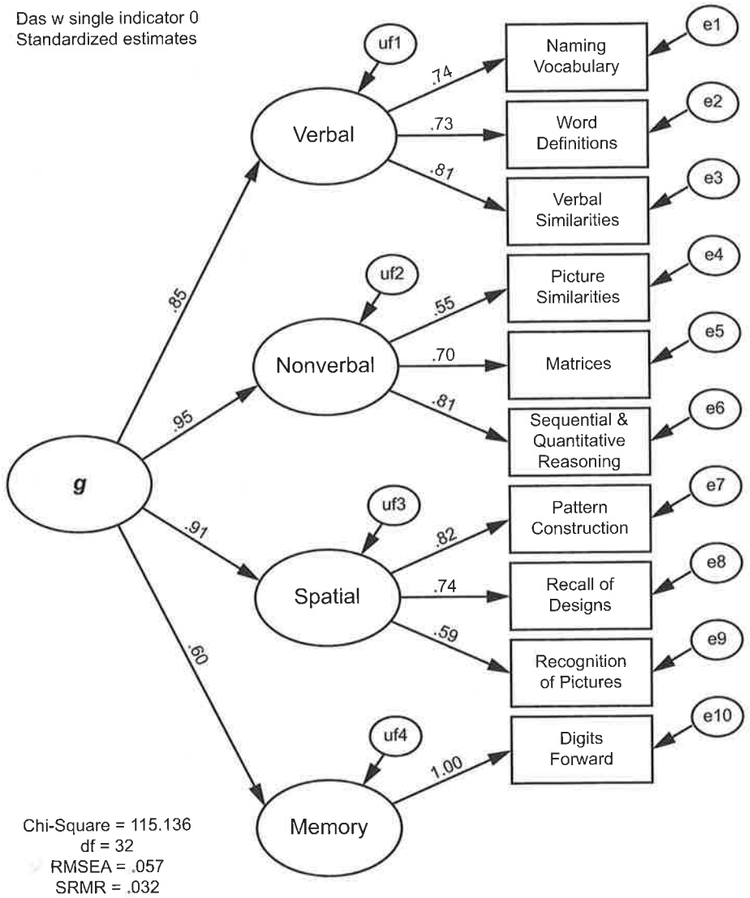


Figure 15.19 An alternative specification with a single indicator. Here, we have constrained the error variance for the Digits Forward test to zero, which essentially says that the subtest is perfectly reliable and that the memory factor and the subtest are equivalent.

Forward (e10) was set to zero. Note that the fit indexes for this model are the same as those shown for Figure 15.18, but that the estimates of the first and second-order factor loadings for Digits Forward and Memory are different. A third possible method is shown in Figure 15.20, in which Digits Forward is loaded directly on the *g* factor; here we essentially say that we don't know what the Digits Forward test measures other than general intelligence. It may not be immediately obvious, but this model is statistically and conceptually equivalent to the previous one. Note that in Figure 15.19 by setting e10 to zero we essentially said that the Memory factor and Digits forward are the same "thing." Note also that the loading of Digits Forward on the second-order *g* factor are identical in the two models (Figures 15.19 and 15.20). Whether you think you will ever use single-indicator latent variables or not, I encourage you to try estimating these three models. You will learn a lot about latent variables

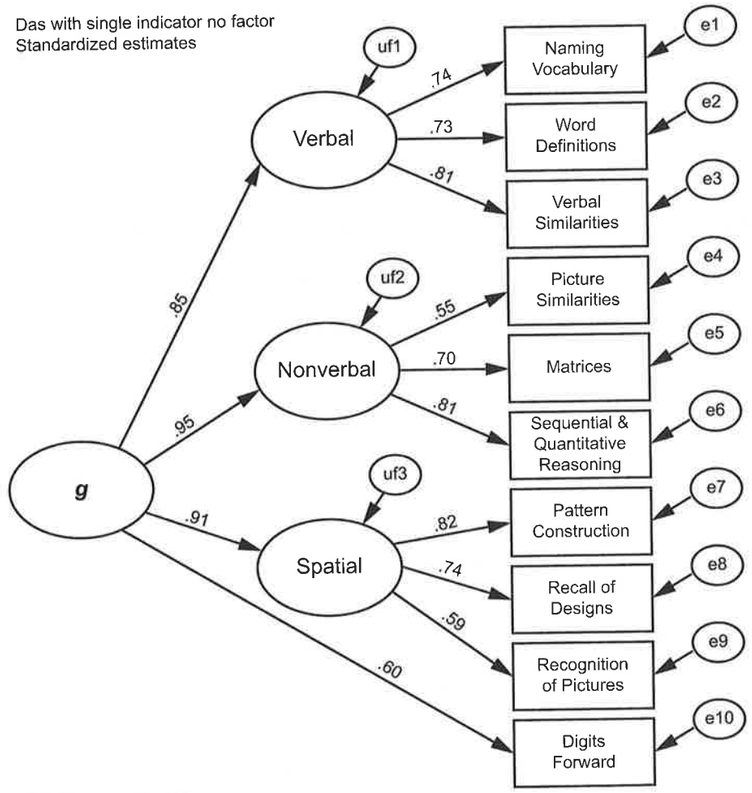


Figure 15.20 Yet another method for dealing with a single-indicator. Although it seems quite different, this model is interchangeable with the previous one.

and alternative models in the process. Make sure you carefully examine the unstandardized estimates in addition to the standardized values shown here.

If the DAS-II indeed only included a single measure of short-term memory another, there is a more powerful method for better understanding the nature of the constructs being measured by the DAS-II Digits Forward test and the Memory factor (in addition to the other factors). This more powerful method would be to factor analyze the DAS-II with another test that includes known measures of short-term along with other related factors. For example, Stone (1992) analyzed the original DAS along with another intelligence test, the Wechsler Intelligence Scale for Children—Revised (Wechsler, 1974) to better understand the constructs measured by both tests.

The examples in this chapter have focused on testing the validity of existing measures. CFA can also be used to test theories. I have mentioned three-stratum theory in the area of intelligence. The DAS-II, it appears, measures several important constructs from three-stratum theory, and thus we can use three-stratum theory to develop a better understanding of what the DAS-II measures. We can turn this process around, as well, to examine the validity of the guiding theory. If we develop multiple measures of the constructs in three-stratum theory, CFA can be used to determine whether a three-stratum-derived model fits the data better than do plausible alternative theories (see Keith & Reynolds, 2012 for more information).

SUMMARY

In the preceding chapter we introduced the full latent variable SEM model. In this chapter we focused on the measurement portion of this model. As it turns out, the measurement model portion of SEM is a useful methodology of its own, generally termed Confirmatory Factor Analysis (CFA). Because the history of factor analysis is so intertwined with the history of intelligence testing, the chapter illustrated CFA through the analysis of a common measure of intelligence, the Differential Ability Scales, Second Edition (DAS-II).

The example used 12 subtests of the DAS-II that supposedly measure four underlying constructs. We drew a model that shows the relations among the factors and subtests (latent and measured variables) (Figure 15.1). The model specifies, with paths drawn from factors to subtests, which subtests load on, or measure, which factors. Consistent with our rules for other path models, each subtest also has a small latent variable pointing to it that represents all other influences on the subtest beyond the four latent factors. With CFA/measurement models, these other influences represent a combination of errors of measurement along with unique or specific influences. With the addition of constraints to set the scales of the latent factors and the unique variances and correlations among the factors—the conceptual model underlying the DAS-II is a testable confirmatory factor model.

We estimated the DAS-II model with data derived from the DAS-II standardization sample. The initial model fit the data well according to the stand-alone fit indexes that we have used in previous chapters (e.g., RMSEA = .046, SRMR = .027), and most of the subtests appeared to measure their corresponding factors strongly. That is, the paths from factors to measured variables, or factor loadings, were generally high. Another way of interpreting these loadings is that the latent constructs (e.g., verbal ability, spatial ability) had strong effects on the corresponding subtests. The factors, or latent constructs, also correlated substantially with each other; all correlations were .75 or larger. This finding suggests that these latent, broad abilities are substantially related to each other.

The common method of setting the scale of latent variables is to set one path from each latent variable to 1, which sets the scale of the variable to be the same as that of the measured variable (the Unit Loading Identification, or ULI, approach, Kline, 2011). An alternative method is to set the variance of the latent variable to 1 (the UVI, or unit variance identification approach). When done with first-order factors, this method turns the factor covariances in the unstandardized solution into factor *correlations*, because a correlation matrix is simply a covariance matrix among standardized variables. This methodology may be useful to test hypotheses about factor correlations.

Just as we can test competing path models using fit statistics, so can we test alternative competing CFA models. We illustrated the testing of competing models by comparing the initial four-factor DAS-II model with a model with a cross-loading and with an alternative three-factor model. In both cases, the initial model fit the data better than did the competing models.

When we wish to use information from the model results to revise the model, several aspects of the SEM program output may be useful. Modification indexes and standardized residual covariances may suggest relaxations in the model that will lead to a better fit. Residual correlations may also be used and have an easier-to-understand metric than standardized residual covariances. Residual correlations are not displayed as output in many SEM programs but are easy to compute. Using these data for model modifications will result in less parsimonious but presumably better fitting models. Using the t (or z) values may lead to values that can be constrained and thus should lead to more parsimonious but equivalent fitting models. You should use such methods to modify models sparingly or else recognize that you are using CFA in an exploratory rather than a theory-testing manner. Model modifications should also be justifiable based on logic, theory, and previous research.

We are often interested in higher-order or other hierarchical models. The field of intelligence is replete with higher-order models, but such models may be relevant in other fields, as well. For the DAS-II example, we hypothesized that a more general factor, often symbolized as g for general intelligence, affects each of the four latent variables, which, in turn, affect the subtests. Said differently, our higher-order model explains that the correlations among the latent factors is a product of their each being affected, in part, by another, more general factor.

An alternative hierarchical model, commonly known as the bifactor model, was also illustrated and tested against the DAS-II data. The bifactor model has shown renewed popularity in recent years and is sometimes considered as a more agnostic version of a hierarchical model. As always, I urge you to consider your underlying theory carefully and allow that theory to guide your model.

It is possible to model latent variables or factors when some of these latent variables include only a single measured variable by constraining the unique–error variance (i.e., e10 in Figure 15.14) to some value. A common method of estimating that unique–error variance uses estimates of the reliability of the measured variable (and thus really only models the error variance, not the specific variance). This may prove a useful method when we only have a single indicator, but we recognize that the variables are not error free. The method can be used in both CFA and SEM models. The chapter ended with a hint of some other uses of CFA.

EXERCISES

1. Conduct the analyses outlined in this chapter. If you have a student version program that only allows a certain number of variables, you may be able to estimate a portion of the models. The initial four-factor model is on the accompanying Web site (www.tzkeith.com) as the file "DAS-II first 1.amw," and the data are in the file labeled "das 2 cov.xls" or "DAS 2 cov.sav."
2. The NELS data include a series of items (BySS44a to BySS44m) designed to assess students' self-esteem and locus of control. Choose several or all of these items that you

believe best measure self-esteem and locus of control and subject them to confirmatory factor analysis. First use SPSS (or another general statistical program) to create a matrix for analysis in Amos (or one of the other programs). Then analyze your model using this matrix. I recommend using the matrix for analysis in order to temporarily avoid dealing with missing data in Amos.

3. The files "DAS 5–8 simulated 6.sav" and "DAS 5–8 simulated 6.xls" include 500 cases of simulated data for the DAS-II.
   a. Conduct the first-order factor analyses from this chapter using the simulated data. Interpret the findings. How do the results compare with those in this chapter (and in Exercise 1)? Would you come to different conclusions following these analyses than we did in the chapter?
   b. Note the fit indexes. Which changed the most from the analyses in the chapter? Why do you think this may be?
   c. As you examine your analyses, are any other hypotheses or models suggested by the findings? If so, conduct these analyses and interpret the findings.

Notes

1 In higher-order intelligence models, Heywood cases often show up in connection with Fluid Reasoning factors (Gf, in the DAS-II represented by the Nonverbal Reasoning factor). When this happens, the *g* to Gf path may approach or exceed 1 and the associated unique factor variance become negative. Note in Figure 15.11 that the *g* to Nonverbal Reasoning loading approached 1. One implication of such a finding is that *g* and Gf factors are not separable. Some researchers use this not-uncommon finding to argue that the Gf factor is redundant with *g*, whereas others argue that this shows that *g* is redundant. As noted, one common method for dealing with negative variances is to set the value to zero. This makes sense if the value is fairly close to zero but is less defensible if it is a large negative value (which likely indicates problems with the model). There are also other possible ways to deal with negative variances, including constraining the value to be positive.

2 Here is an interesting conundrum. When factors are correlated, it is possible (although not desirable) to have factors referenced by only two measured variables each. So, for example, a correlated two-factor, four-measured variable model would have one degree of freedom. But when factors are uncorrelated, each factor requires a minimum of three measured variables for identification, and with three measured variables each factor is just-identified (as in the present bifactor example). That means that if a bifactor model includes fewer than three variables for a factor, the researchers will need to either make additional constraints (e.g., constrain the two factor loadings to be equal) or, counter-intuitively, relax constraints (e.g., allow that factor to be correlated with another factor). As you are reading research using the bifactor model and you notice only two measured variables on a factor, make sure the researchers tell you what they have done to solve this problem! This conundrum of identification also occasionally leads to a phenomenon known as "empirical underidentification" in which a model allows factors to be correlated, but that correlation is small and nonsignificant. If one of the offending factors involves fewer than three measured variables, it will thus be underidentified. The phenomenon of empirical underidentification applies to first-order factor models as well (Kenny, 1979).

16
Putting It All Together
Introduction to Latent Variable SEM

Let's review our progress in our adventures beyond MR. You know how to conduct path analysis using MR. This experience includes the estimation of standardized and unstandardized paths, the calculation of disturbances ($\sqrt{1 - R^2}$), and the calculation and comparison of direct, indirect, and total effects using two different methods. We transitioned into estimating path models using Amos and other SEM programs and focused again on the estimation of both standardized and unstandardized effects and direct, indirect, and total effects. With Amos, we switched from the estimation of the paths from disturbances to estimating the variances of the disturbances, although either is possible. We have defined just-identified, overidentified, and underidentified models, and I suggested that you use a SEM program to estimate overidentified models but use either MR or an SEM program if your models are just-identified. We have examined fit indexes for overidentified models and have highlighted a few that are useful for evaluating a single model and those that are useful for comparing competing models. We briefly focused on equivalent models, nonrecursive models, and longitudinal data. We focused on the effects of measurement error on path analysis, MR, nonexperimental research, and research in general and began considering the use of latent variables as a method of obviating this threat. We expanded our knowledge of latent variables, their meaning, and estimation via confirmatory factor analysis.

PUTTING THE PIECES TOGETHER

In this chapter, we will begin putting all these pieces together in latent variable structural equation modeling. As noted in Chapter 14, you can consider latent variable SEM as a