An Introduction to Structural Equation Modeling[1]

J.J. Hox
University of Amsterdam/Utrecht University
T.M. Bechger
CITO, Arnhem

Abstract *This article presents a short and non-technical introduction to Structural Equation Modeling or SEM. SEM is a powerful technique that can combine complex path models with latent variables (factors). Using SEM, researchers can specify confirmatory factor analysis models, regression models, and complex path models. We present the basic elements of a structural equation model, introduce the estimation technique, which is most often maximum Likelihood (ML), and discuss some problems concerning the assessment and improvement of the model fit, and model extensions to multigroup problems including factor means. Finally, we discuss some of the software, and list useful handbooks and Internet sites.*

What is Structural Equation Modeling?

Structural Equation Modeling, or *SEM*, is a very general statistical modeling technique, which is widely used in the behavioral sciences. It can be viewed as a combination of factor analysis and regression or path analysis. The interest in SEM is often on *theoretical constructs*, which are represented by the latent factors. The relationships between the theoretical constructs are represented by regression or path coefficients between the factors. The structural equation model implies a structure for the covariances between the observed variables, which provides the alternative name *covariance structure modeling*. However, the model can be extended to include means of observed variables or factors in the model, which makes covariance structure modeling a less accurate name. Many researchers will simply think of these models as 'Lisrel-models,' which is also less accurate. LISREL is an abbreviation of LInear Structural RELations, and the name used by Jöreskog for one of the first and most popular SEM programs. Nowadays structural equation models need not be linear, and the possibilities of SEM extend well beyond the original Lisrel program. Browne (1993), for instance, discusses the possibility to fit nonlinear curves.

Structural equation modeling provides a very general and convenient framework for statistical analysis that includes several traditional multivariate procedures, for example factor analysis, regression analysis, discriminant analysis, and canonical correlation, as special cases. Structural equation models are often visualized by a graphical *path diagram*. The statistical model is usually represented in a set of matrix equations. In the early seventies, when this technique was first introduced in social and behavioral research, the software usually required setups that specify the model in terms of these matrices. Thus, researchers had to distill the matrix representation from the path diagram, and provide the software with a series of matrices for the different sets of

---

parameters, such as factor loadings and regression coefficients. A recent development is software that allows the researchers to specify the model directly as a path diagram. This works well with simple problems, but may get tedious with more complicated models. For that reason, current SEM software still supports the command- or matrix-style model specifications too.

This review provides a brief and non-technical review of the basic issues involved in SEM, including issues of estimation, model fit, and statistical assumptions. We include a list of available software, introductory books, and useful Internet resources.

Examples of SEM-Models

In this section, we set the stage by discussing examples of a confirmatory factor analysis, regression analysis, and a general structural equation model with latent variables.

Structural equation modeling has its roots in path analysis, which was invented by the geneticist Sewall Wright (Wright, 1921). It is still customary to start a SEM analysis by drawing a path diagram. A path diagram consists of boxes and circles, which are connected by arrows. In Wright's notation, observed (or measured) variables are represented by a rectangle or square box, and latent (or unmeasured) factors by a circle or ellipse. Single headed arrows or 'paths' are used to define causal relationships in the model, with the variable at the tail of the arrow causing the variable at the point. Double headed arrows indicate covariances or correlations, without a causal interpretation. Statistically, the single headed arrows or paths represent regression coefficients, and double-headed arrows covariances. Extensions of this notation have been developed to represent variances and means (cf. McArdle, 1996). The first example in Figure 1 is a representation of a confirmatory factor analysis model, with six observed variables and two factors.

Confirmatory Factor Analysis
The model in Figure 1 is a confirmatory factor model for data collected by Holzinger and Swineford, extracted from the AMOS manual (Arbucle, 1997, p. 375, see also Jöreskog & Sörbom, 1989, p. 247). The data are the scores of 73 girls on six intelligence tests. There are two hypothesized intelligence factors, a verbal and a spatial ability factor, which are drawn as latent factors which are assumed to cause the variation and covariation between the six observed variables. There is a double-headed arrow between the two factors, which indicates that we assume that the two factors are correlated. The arrows from the factors to the variables represent linear regression coefficients or 'factor loadings'. We do not assume that the latent factors completely explain the observed variation; each observed variable is associated with a residual error term, which is also unmeasured and depicted by a circle.
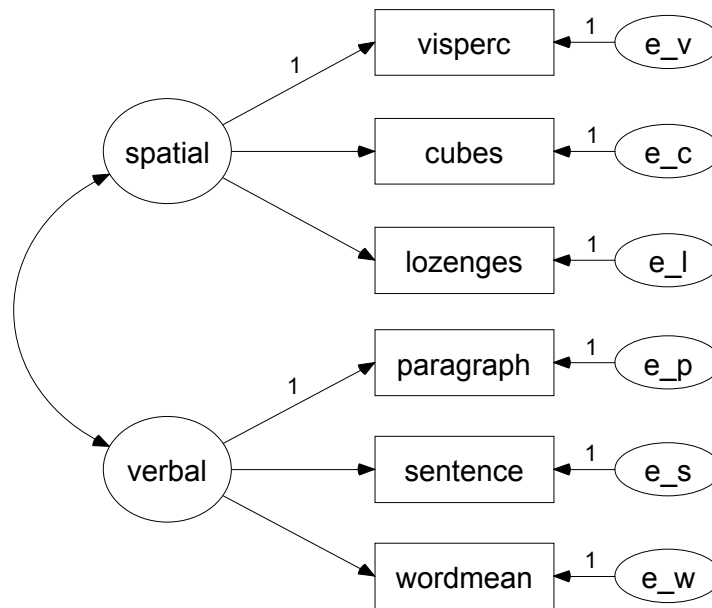
Figure 1. Confirmatory Factor analysis;
Holzinger and Swineford data.

Factor analysis assumes that the covariances between a set of observed variables can be explained by a smaller number of underlying latent factors. In exploratory factor analysis, we proceed as if we have no hypothesis about the number of latent factors and the relations between the latent factors and the observed variables. Statistical procedures are used to estimate the number of underlying factors, and to estimate the factor loadings. In exploratory factor analysis, the model is arbitrary: *all* variables load on *all* factors. Typically, a transformation method such as Varimax rotation is used to improve the interpretation of the results. In contrast, the path diagram in Figure 1 represents a clear hypothesis about the factor structure. Models of this kind are called restricted or confirmatory factor analysis (CFA) models. In structural equation modeling, the confirmatory factor model is imposed on the data. In this case, the purpose of structural equation modeling is twofold. First, it aims to obtain estimates of the *parameters* of the model, i.e. the factor loadings, the variances and covariances of the factor, and the residual error variances of the observed variables. The second purpose is to assess the fit of the model, i.e. to assess whether the model itself provides a good fit to the data. We will deal with these issues in more detail later.

Typically, some of the factor loadings are constrained or *fixed* to be zero. In Figure 1, the absence of arrows going from the verbal factor to 'visperc,' 'cubes' and 'lozenges,' means that the corresponding loadings in the factor matrix are fixed to zero. Similarly, the loadings of 'paragraph,' 'sentence' and 'wordmean' on the spatial factor are also fixed to zero. The factor model in Figure 1 shows a perfect simple structure: each variable loads on one factor only. Confirmative factor analysis can specify such a structure exactly, and test whether it is plausible, while explorative factor analysis can only approximate such simple structures by rotation.

For each factor, we also must fix one loading to one. This is needed to give the latent factor an interpretable scale. If we do not fix one factor loading to one for (or to another number not equal to zero), the scale of the latent factor is undetermined. For each latent factor, we can estimate the loadings given a fixed variance for the latent factor,

which standardizes the scale of the factor to a *Z*-score, or we can estimate the factor variance given at least one fixed loading. Since the loadings are a function of the variance of the latent factor, and the variance of the latent factor is a function of the loadings, we cannot simultaneously estimate unique values for all of these. Thus, one solution here is to fix the variance of all factors to one, and estimate all factor loadings. In SEM, it is more customary to use the other solution, which is to fix one loading for each factor to one, and estimate the factor variances. Only the Lisrel program fixes the variances of the factors by default.

The factors 'e_v' to 'e_w' to the right of Figure 1 represent measurement errors. We expect that the verbal and spatial factor will not perfectly predict the observed variables, and this is modeled by specifying a specific error factor for each observed variable. These measurement errors are often pictured as single arrows. The representation in Figure 1, which is the standard representation in the program AMOS (Arbucle, 1997), makes clear that these errors are also unobserved factors.

In SEM, we must specify a model before we start the analysis. The model specification is usually guided by a combination of theory and empirical results from previous research. Once we have specified a model, we can estimate factor loadings and (co)variances. We can also conduct a statistical chi-square test to assess how well the hypothesized model fits the data. If the chi-square is highly significant, the hypothesized model is rejected, and we could search for a better model. Of course, the chi-square test in SEM shares with all statistical tests the problems of the need for assumptions and the dependence of its power on the sample size. We will address these problems later.

The chi-square for the two-factor model in Figure 1 is 7.9 with 8 degrees of freedom and a p-value of 0.45. Apparently, the two-factor model is a plausible model for these test data. The estimates of the factor loadings are presented in Table 1.

Table 1. Factor loadings Holzinger & Swineford data (loading (s.e.)) and Squared Multiple Correlations (SMC)

	unstandardized loadings		standardized loadings		SMC
	spatial	verbal	spatial	verbal	
visperc	1.00		.70		.49
cubes	.61 (.14)		.65		.43
lozenges	1.20 (.27)		.74		.54
paragraph		1.00		.88	.77
sentence		1.33 (.16)		.83	.68
wordmean		2.23 (.26)		.84	.71

The standard errors that are in parentheses next to the unstandardized loadings can be used to assess their significance. The statistic formed by dividing an estimate by its standard error is called the critical ratio (C.R.). With large samples, the critical ratio can be referred to the standard normal distribution. Thus, a value for the C.R. of 1.96 or higher (and −1.96 and lower), indicates two-sided significance at the customary 5% level. In SEM, for instance in the program EQS, the C.R. test is sometimes referred to as the *Wald* test. In our confirmatory factor analysis, the critical ratio tests indicate that all loadings are significant.

In SEM, it is usual to analyze the covariance matrix and not the correlation matrix, for sound statistical reasons (see Bollen, 1989, or Loehlin, 1998, for details). However, the software also produces standardized estimates, which are generally used for interpretation. The standardized loadings in Table 1 can be interpreted in the same way as

in exploratory factor analysis. The squared multiple correlations (SMC's) in table 1 are in fact the communalities of the variables. Finally, the model estimates the correlation between the two factors as 0.49, which is moderately high. Exploratory factor analysis can also produce factor intercorrelations, but these are strongly influenced by details of the specific rotation method used. The ability of SEM to produce a meaningful identification of the correlations between factors is a key strength.

Multiple Regression Analysis
It is instructive to see how a familiar analysis procedure, such as multiple regression analysis, looks when represented as a path model. Figure 2 is a multiple regression model to predict the perceived burden of having children from the variables 'child care by neighbors,' 'integration in the neighborhood,' 'inability to be alone,' 'relatives in the area,'and 'child care by relatives.' The example is taken from Goldsteen and Ross (1989).
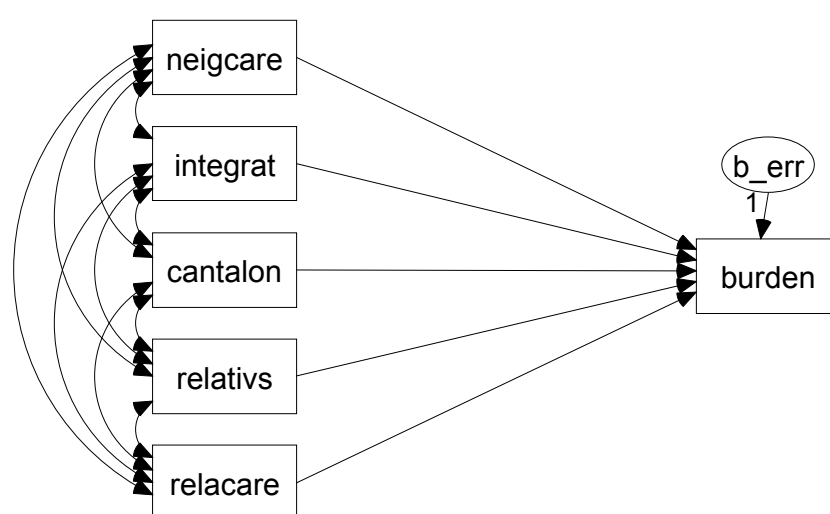

Figure 2. Perceived burden of children, multiple regression.
NOTE: neigcare = child care by neighbors; integrat = integration in neighborhood; cantalon = inability to be alone; relativs = relatives in area; relacare = child care by relatives.

Figure 2 makes two things quite clear. Firstly, in multiple regression analysis, we generally assume that the independent variables are correlated; in Figure 2 we find that assumption as the two-headed arrows between the predictor variables. Secondly, the residual error in multiple regression analysis is actually an unobserved, latent variable. Note that we again must fix the loading of the residual error factor to one, to achieve identification.
        If we estimate the model in Figure 2, we obtain unstandardized and standardized regression weights, a variance estimate for the residual errors, and the squared multiple correlation of the dependent variable 'burden.' In this case, we do not get a chi-square test for the model fit. The reason is that the model in Figure 2 estimates precisely as many parameters as we have data points. Let us count them. We have six variables, which gives us a 6×6 variance/covariance matrix. This matrix has six variances and 15 unique covariances, together 21 data points. The path model in Figure 2 implies estimation of the following parameters: the variances and covariances of the three independent variables

(15), the regression coefficients (5), and the error variance (1). These also sum to 21. Since we have as many parameters as data points, the model is in fact a re-expression of our data. As a consequence, the multiple regression model by itself can not be tested. We may, of course, decide to reject it if the squared multiple correlation is to low to our taste, but this is not strictly a model test.

Using SEM to estimate multiple correlations has few advantages, we use this example only to show how a familiar analysis method looks when cast in the SEM framework. The real strength of SEM is, that we may specify and estimate more complicated path models, with intervening variables between the independent and dependent variables, and latent factor as well. The next section discusses a general path model with latent factors. Figure 3 shows a path model for the variables from the previous example, also from Goldsteen and Ross (1989). Goldsteen and Ross do not ues structural equation modeling on their data; the path diagram in their article is a summary of a series of regression analyses. If we estimate the model in Figure 3, we obtain a chi-square of 20.0 on 8 degrees of freedom, which gives a p-value of 0.01. Statistically, the model is rejected. In a later section, we will discuss ways to improve the model of figure 3, in order to obtain a better fitting model.
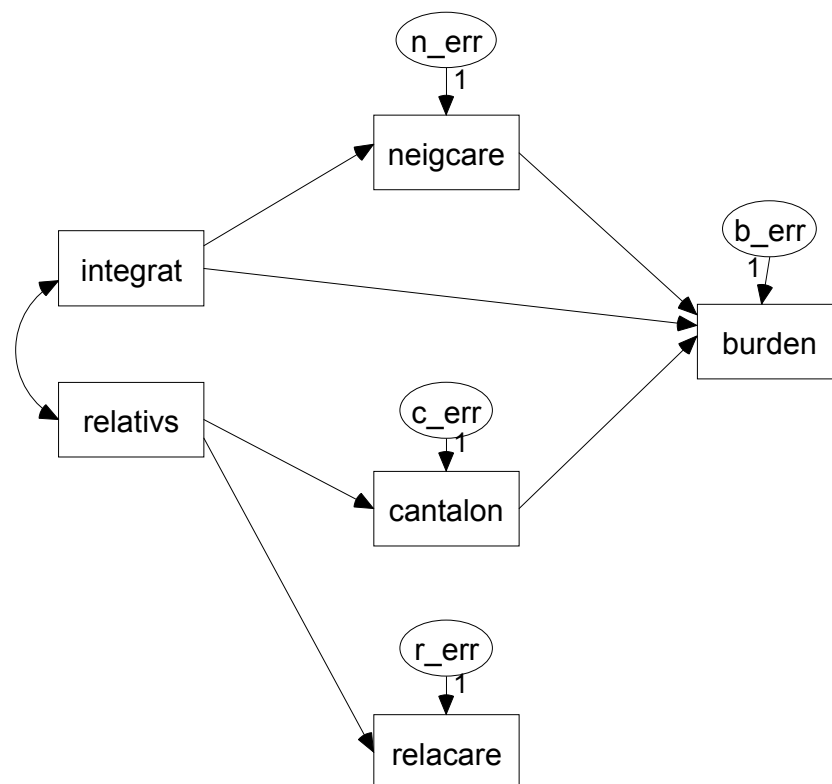

Figure 3. Path model for perceived burden of children.
NOTE: neigcare = child care by neighbors; integrat = integration in neighborhood; cantalon = inability to be alone; relativs = relatives in area; relacare = child care by relatives

Path Model with Latent Factors
Structural equations models can be quite complex, and incorporate both latent factors and observed variables, with either directed or undirected paths among them. Figure 4 is a

diagram for a model on the quality of sleep. It contains two factor models. The first relates the latent factor 'neurotic' to the observed variables 'test attitude,' 'neuroticism,' and 'somatic neuroticism.' The second relates the latent factor 'quality' to the observed variables 'difficult to fall asleep,' 'feel rested,' and 'sleep quality.' There is also a regression model that relates the two latent factors; 'neurotic' is assumed to have an effect on 'quality.'
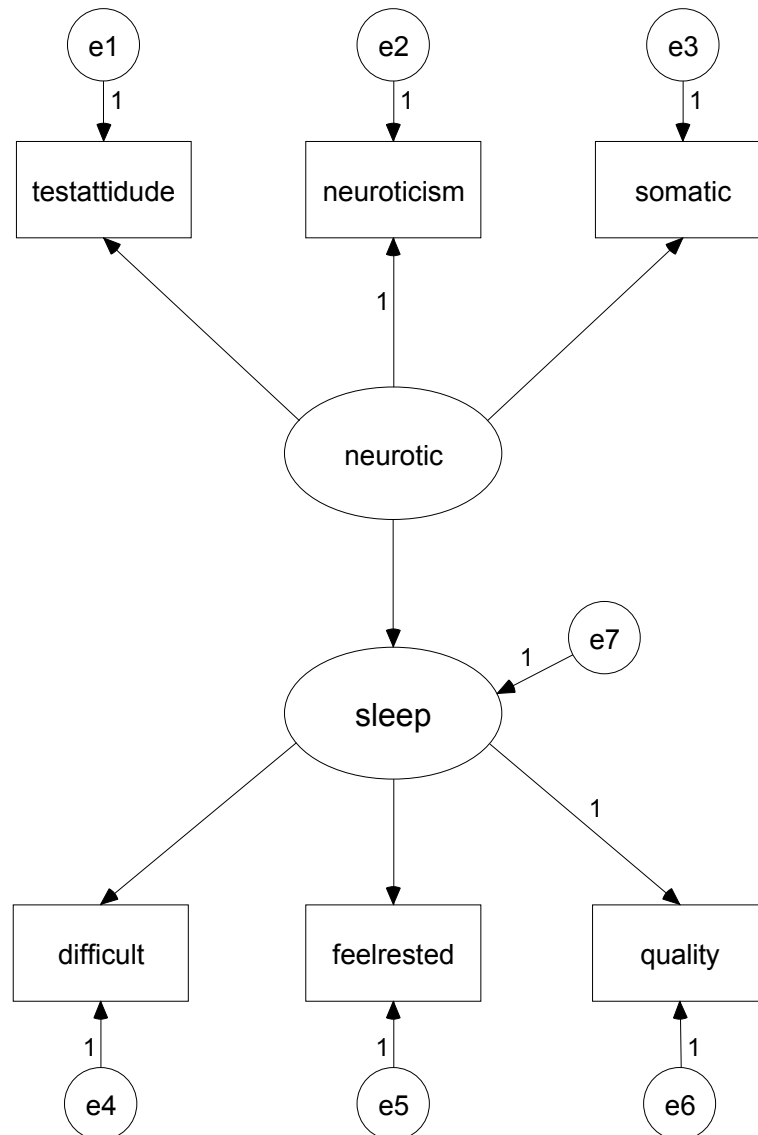

Figure 4. Path model with latent variables for quality of children's sleep.

If we fit the model depicted in figure 4, we obtain a chi-square of 17.2 on 8 degrees of freedom, with a p-value of 0.03. Statistically, the model is rejected. In a later section, we will discuss ways to improve the model.

Estimation and Model Fit

Estimation
Fitting a model to data means solving a set of equations. On the one hand, there is the model with its parameters, whose values we wish to estimate. On the other hand, there are the sample statistics that we 'know' to be good estimates of the corresponding population values. In SEM, it is usually assumed that the sample data follow a multivariate normal distribution, so that the means and covariance matrix contain all the information. The basic model in statistical modeling is: DATA = MODEL + ERROR. SEM software uses complex algorithms that maximize the fit of the model, taking all model restrictions (fixed parameters, equality constraints) into account. The method most widely used for estimation is Maximum Likelihood (ML) estimation, which assumes multivariate normal data and a reasonable sample size, e.g. about 200 observations. There are a variety of estimation procedures that can be used for non-normal continuous data. With non-normal data, the means and covariance matrix does not represent all the information, and usually these alternative estimation methods need raw data.

Simulation research has shown that with a good model and multivariate normal data a reasonable sample size is about 200 cases (cf. Boomsma, 1997), although there are examples in the literature that use smaller samples. If the data are continuous but not normal, an alternative estimation method is Asymptotically Distribution Free (ADF; the LISREL PROGRAM calls this method WLS for Weighted Least Squares). Another alternative is to correct the chi-square statistic for the degree of non-normality. Simulation research (e.g., Chou & Bentler, 1995; Boomsma, 1997) shows that with non-normal data, ADF estimation requires very large samples, typically more than a thousand cases. Maximum likelihood estimation still produces good estimates in most cases, but larger sample sizes are needed, typically at least 400 cases. The so-called Satorra-Bentler chi-square correction (cf. Chou & Bentler, 1995) appears to work well; it is currently the most promising method to accommodate non-normal data.

A different problem is the analysis of ordinal categorical data. It is customary, to consider ordinal categorical measurements as imprecise observations on continuous normally distributed variables. For example, a psychological trait can be thought of as continuous, even when the data consist of categorical responses to test items. Given this assumption, it is possible to compute *polychoric correlations*, which are the estimated correlations among the unobserved normal variables. The analysis then uses ADF to estimate the model parameters from the polychoric correlations. This, again, requires large sample sizes. Another option is to simply ignore the categorical nature of the variables, which appears to work well, provided that the number of categories is not small (at least five categories, cf. Chou and Bentler, 1995), and the data are close to normal.

Model Fit: Goodness-of-Fit Indices
Statistical tests for model fit have the problem that their power varies with the sample size. If we have a very large sample, the statistical test will almost certainly be significant. Thus, with large samples, we will always reject our model, even if the model actually describes the data very well. Conversely, with a very small sample, the model will always be accepted, even if it fits rather badly.

Given the sensitivity of the chi-square statistic for sample size, researchers have proposed a variety of alternative fit indices to assess model fit. All goodness-of-fit measures are some function of the chi-square and the degrees of freedom. Most of these fit indices not only consider the fit of the model, but also its simplicity. A saturated model, that specifies all possible paths between all variables, always fits the data perfectly, but it is just as complex as the observed data. That does not necessarily make it useless. The regression analysis presented earlier involves a saturated model, which is

quite informative. But, if two models have the same degree of fit, we should prefer the simplest of the two. In general, there is a trade-off between the fit of a model and the simplicity of a model. Several goodness-of-fit indices have been proposed to assess simultaneously both the fit and the simplicity of a model. The goal is to produce a goodness-of-fit index that does not depend on the sample size or the distribution of the data. In fact, most goodness-of-fit indices still depend on sample size and distribution, but the dependency is much smaller than that of the routine chi-square test.

Modern SEM software computes a bewildering array of goodness-of-fit indices. All of them are functions of the chi-square statistic, but some include a second function that penalizes complex models. For instance, Akaike's information criterion (AIC), is twice the chi-square statistic minus the degrees of freedom for the model. For an overview and evaluation of a large number of fit indices, including those mentioned here, we refer to Gerbing and Anderson (1993).

Jöreskog and Sörbom (1989) have introduced two goodness-of-fit indices called GFI (Goodness of Fit) and AGFI (Adjusted GFI). The GFI indicates goodness-of-fit, and the AGFI attempts to adjust the GFI for the complexity of the model. Two other well-known measures are the Tucker-Lewis Index TLI (Tucker & Lewis, 1973), better known as the Non-Normed Fit Index or NNFI, and the Normed Fit Index NFI (Bentler & Bonett, 1980). Both the NNFI and the NFI adjust for complexity of the model. Simulation research shows that all these indices still depend somewhat on sample size, while TLI/NNFI shows the best overall performance. If the model fits perfectly, the fit indices should have the value 1. Usually, a value of at least 0.90 is required to accept a model, while a value of at least 0.95 is required to judge the model fit as 'good.' However, these are just rules of thumb.

A relatively modern approach to model fit is to accept that models are only approximations, and that perfect fit may be too much to ask for. Instead, the problem is to assess how well a given model approximates the true model. This view led to the development of an index called RMSEA, for Root Mean Square Error of Approximation. If the approximation is good, the RMSEA should be small. Typically, a RMSEA of less than 0.05 is required, and statistical tests or confidence intervals can be computed to test if the RMSEA is significantly larger than this lower bound.

Modification Indices

If the fit of a model is not adequate, it has become common practice to modify the model, by deleting parameters that are not significant, and adding parameters that improve the fit. To assist in this process, most SEM software can compute *modification indices* for each fixed parameter. (The statistical test used is called a Lagrange multiplier test, and the program EQS calls it by its proper statistical name. However, most programs use the term 'modification index.') The value of a given modification index is the minimum amount that the chi-square statistic is expected to decrease if the corresponding parameter is freed. Researchers often use this information to conduct a sequence of model modifications. At each step a parameter is freed that produces the largest improvement in fit, and this process is continued until an adequate fit is reached. For example, if in a confirmative factor model a loading that is fixed to zero shows a large modification index, we may free this parameter and estimate its value. This will improve the fit of the model, at the cost of one degree of freedom.

Associated with sequential model modification is the danger of capitalization on chance properties of the sample. Generally, the advice is to apply modifications only when there is a theoretical justification for them. The problem here is, of course, that this theoretical justification is evaluated post hoc, and one suspects that researchers are very creative in justifying modifications. Simulation research has show that model modification often fails to find the correct model (Spirtes, Scheines & Glymour, 1991),

and that models so achieved cross-validate badly (Maccallum, 1986; MacCallum, Roznowskei & Necowitz, 1992). Many authors recommend an alternative strategy for the development and evaluation of SEM models: the use of multiple a-priori models and cross-validating on a different sample (Cudeck & Henley, 1991; MacCallum et al., 1992).

In a previous section, we discussed a path model for the perceived burden of children, and a latent factor model for the quality of children's sleep. In both cases the model test by chi-square led to a rejection of the model. The sample size for the perceived burden data is 549, and for the quality of sleep data, the sample size is 449. This is not excessive, and we may conclude that the model rejection is not merely the result of a very large power for the chi-square test. Consequently, it is advisable to see if the model can be improved in a meaningful way.

The modification indices of the perceived burden of children example suggest adding a path from 'relatives' to 'child care by neighbors.' Figure 5 presents the modified model, with standardized estimates. The path coefficient for the added path is negative, which suggests that if there are many relatives living in the area, the amount of child care by neighbors is less. This is logical, and we decide to retain the modification. The resulting model fits, with a chi-square of 12.9 and 7 degrees of freedom, and a p-value of 0.08. The various goodness of fit indices are also acceptable: the GFI is 0.99, the AGFI is 0.98, and the TLI/NNFI and the NFI are both 0.96 while the RMSEA is 0.04.

The quality of sleep model can also be improved by adding one parameter to the model. The modification indices suggest adding various covariances between error terms. One covariance term is theoretically justifiable: the covariance between the residual errors for 'neuroticism' and 'somatic neuroticism.' These are strongly related concepts, and it is feasible that these measures have something specific in common that is lacking in 'test attitude,' the third indicator for the 'neurotic' factor. Adding this covariance results in the model depicted in figure 6, which also contains the standardized estimates. This model also fits, with a chi-square of 12.9 on 7 degrees of freedom, and a p-value of 0.07. The goodness of fit indices GFI, AGFI, TLI/NNFI and NFI are all 0.97 or higher, and RMSEA is 0.04. Thus, we retain the modified model.
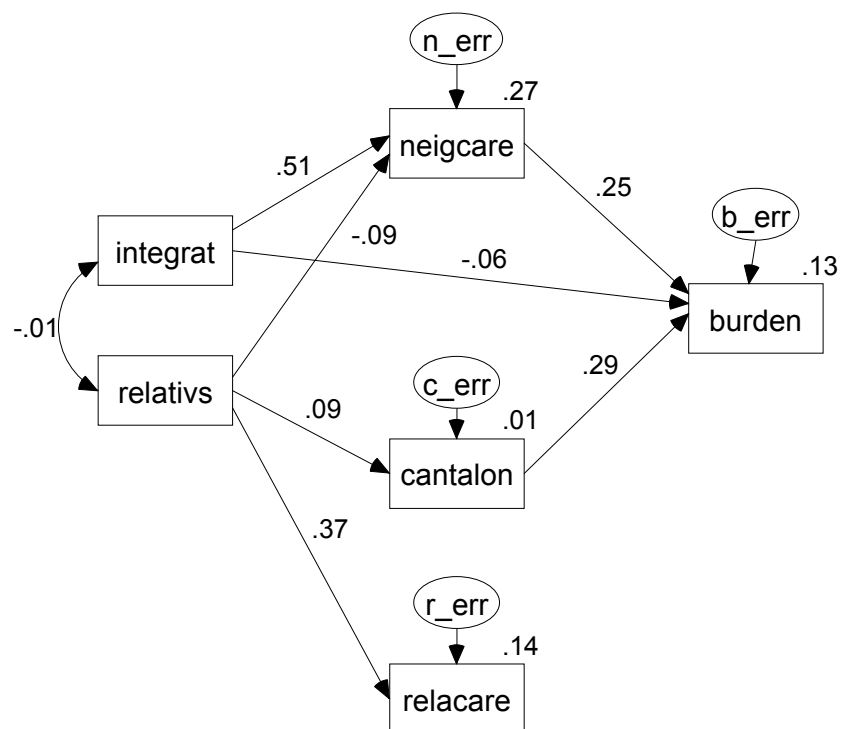
Figure 5. Modified path model for perceived child burden.
NOTE: neigcare = child care by neighbors; integrat = integration in neighborhood; cantalon = inability to be alone; relativs = relatives in area; relacare = child care by relatives. Standardized estimates.
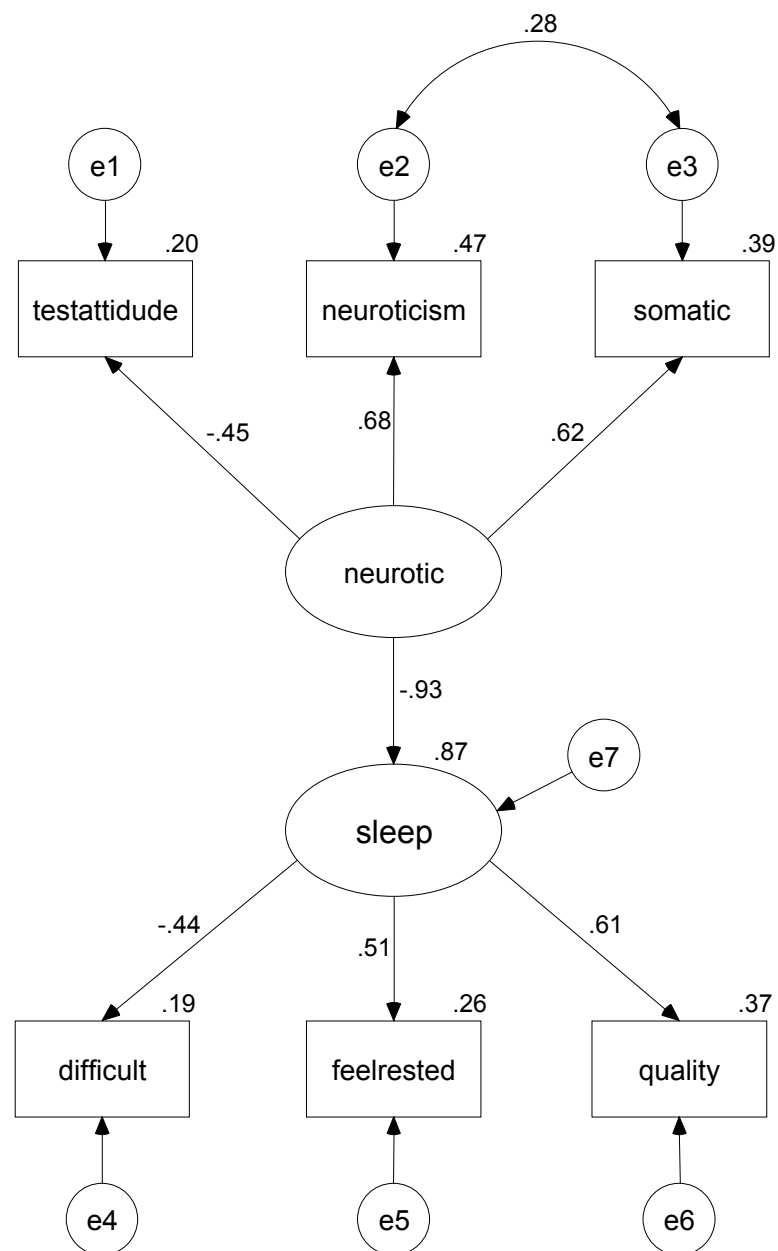
Figure 6. Modified model for children's quality of sleep, standardized estimates.

Extended Models

The basic structural equation model can be extended in various ways. Two useful extensions are multigroup analysis and the inclusion of means. These extensions are especially powerful when they are applied simultaneously, because they offer an alternative to Analysis of (Co)variance for testing multiple groups for different means.

Adding means to a structural equation model is relatively straightforward. In the data, we must include the means with the covariance matrix or provide raw data. In the model, we must include an intercept term in the regressions for the observed variables, and estimate factor means. If we have only one group, this is not very informative, because the factor means are just a function of the observed variables' means. However, in SEM, it is possible to analyze more than one group simultaneously. This is an interesting feature, because it allows us to investigate whether a specific model fits equally well in different groups. We do this by placing *equality constraints* on parameters across groups.

If we place an equality constraint on a parameter across two groups, we in fact command the software to estimate one parameter that is equal in both groups. The result is of course some kind of average of the two estimates we would get without this parameter constraint. If we compare the model with and without the equality constraint, we will see that in the model with the constraint, the chi-square has gone up, and we have gained one degree of freedom. We can formally compare the two models by calculating the difference between the two chi-squares, and testing this difference against the chi-square distribution with one degree of freedom. In general, we can compare a model without constraints with the same model with a number of added constraints, by calculating the difference in chi-squares and in degrees of freedom, and using these in a formal chi-square test.

In a combination of a model with means and multigroup analysis, we can compare an unrestricted model with a model that constrains the means in both groups to be equal. The interesting feature is that we can place equality constraints on the means of the latent factors. This would be comparable to a one-way analysis of variance or a t-test, but we are performing this test on the latent factors. Details on this kind of analysis are given by Byrne, Shavelson and Muthén (1989), Sörbom (1982) and Werts (1979). As an example the confirmatory factor analysis of the Holzinger-Swineford data is extended with a test on the factor means.

The Holzinger-Swineford data consist of six intelligence measures that can be described well by the two-factor model presented in Figure 1. The data actually are a sample of 73 girls and a sample of 72 boys. If we analyze these data using a multigroup model, we can impose a variety of equality constraints. To compare the factor means for the boys and the girls, it is usually considered necessary to constrain the factor loadings to be equal across groups, to guarantee that we are actually measuring the same latent factor in both groups (cf. Bechger, 1997). To compare the means, we also must impose equality constraints on the intercepts of the observed variables across the two groups. This set of restrictions is the minimum required for a valid comparison of the boys' and girls' means. If this model is rejected, we must conclude that the factor structure for the boys differs from that for the girls. In such a situation, comparing factor means makes no sense.

We can investigate if we can impose even stronger constrains on the multigroup model. For instance, it is interesting to test if the error variances and the covariance between the two latent factors are equal across both groups. Finally, we may impose the constraint that the factor means are equal across both groups. If we statistically compare

the model with the equality constraints on the factor means with the model without these constraints, we have the SEM equivalent of an analysis of variance on the factor means.

Table 2 presents the overall fit measures of a series of models, starting with the least restricted model (only intercepts and loadings are constrained to be equal), and ending with the most restricted model (all parameters are equal).

Table 2. Overall fit results for multigroup comparison of boys and girls

Equality restrictions on	chi^2	df	p	TLI	RMSEA	pclose*
intercepts, loadings	22.6	24	.54	1.00	.00	.87
plus error variances	27.1	30	.62	1.00	.00	.93
plus factor (co)variances	30.2	33	.61	1.00	.00	.93
plus factor means	38.4	35	.32	1.00	.03	.79

* pclose is the p-value of a test on RMSEA

The chi-square goodness-of-fit test and the other fit indices indicate that even the most restricted model has a good fit to the data. If we use the chi-square difference test, we may conclude that imposing equality constraints on the factor means, sgnificantly lowers the fit of the model: the chi-square difference is 38.4-30.2=8.2, with 2 degrees of freedom and a p-value of p=0.02. In the most restricted model, the factor means are fixed at zero in both groups. In the next model, the factor means are restricted to zero in the boys group, but left free to be estimated in the girls group. The estimated factor means in the girls group are: spatial -1.09 (s.e. 0.86), verbal 0.95 (s.e. 0.52). The Z-ratio for a test on the mean is −1.26 for the spatial factor, and 1.82 for the verbal factor. The test for the verbal mean is marginally significant, and we conclude that boys and girls do not differ in spatial ability, but girls are higher on the latent verbal factor. We can refine the analysis somewhat by putting an equality restriction on the means of the spatial factor. This leads to a model with a very good fit: the chi-square is 31.8 (df=34, p=0.58), TLI=1.00, RMSEA=0.00, pclose=0.92. In this model the verbal factor mean for the girls is estimated as 1.23, with a standard error of 0.47 and a critical ratio of 2.63. In this, the final model, the difference between boys and girls on the verbal factor is clearly significant. The variance of the verbal factor is estimated as 8.45 in both groups. We can calculate that the difference of 1.23 is 0.42 of the standard deviation, which is a difference Cohen (1982) puts between a small and a medium effect size.

Critiques on SEM

Structural equation modeling contains a variety of powerful analysis techniques. Modern SEM software and inexpensive computers have made it increasingly easy to apply structural equation models to all sort of data. Although these developments can have a positive impact on substantive research in more applied fields such as family research, they also make it more easy to misuse these techniques.

Most critiques that have been raised against the use of SEM center around two issues. One issue is the importance of statistical assumptions and needed sample sizes. There has been much research on the importance of the normality assumption and the sample sizes needed to have confidence in the results. Some of this research is referenced in the handbooks mentioned in then next section. Since this research tends to be technical, applied researchers using SEM are not always aware of all the technicalities. Fortunately, recent handbooks and edited volumes (see next section), try to inform non-statistical users of the errors to avoid, and the choices they have.

The second, and probably more important issue concerning the casual use of SEM, is the issue of causal interpretation. Most applications of SEM are on non-experimental data. Many applications of SEM nevertheless interpret the final model as a causal model. This may be correct, but there is of course nothing in structural equation modeling that magically transforms correlational data into causal conclusions. A classic article by Cliff (Cliff, 1983) treats a number of fallacies that are the result of causal interpretation of correlational data. At the least, SEM users should remind themselves that the fact that a SEM model has been corroborated by the data, does not mean that it has been proven true. It just has not been falsified, but there may be competing models that would not have been falsified either. As usual, explicit comparison of competing models carries more conviction than testing (and failing to reject) just one model.

Software and Handbooks

Software does not make a statistician, but the advent of powerful and user-friendly software for structural equation modeling has had a large impact on the field. There are several programs, of which the best known are probably LISREL, EQS, and AMOS. Several reviews (Waller, 1993, Hox, 1995, Miles, 1998) state that, although each program has its own strengths and weaknesses, for standard analyses one can use any package. Model specification has become much easier with recent versions of the software. Early SEM software required that the model was specified by referring to special parameter matrices labeled by Greek letters (e.g., LISREL, LISCOMP), or by writing out the structural equations (e.g., EQS). Modern versions let the user specify the model in a simplified command language (the SIMPLIS command language in LISREL, Text AMOS), or let the user draw the path diagram for the model specification (Windows versions of LISREL, EQS, and AMOS, a graphical preprocessor for MX). As a result, estimating complex structural equation models has become much easier than it used to be. For example, Dolan et al. (1998) discuss recently developed models that can now be fit using standard SEM programs, although the original papers state that some relevant estimation procedures have not yet been developed.

Of course, researchers still need to understand the modes that they use. The growing popularity of SEM has given rise to the publication of several introductory and more advanced handbooks and readers. A classic, advanced, book about structural equation modeling is Bollen's "Structural equation modeling with latent variables" (Bollen, 1989). Introductory books include Loehlin's "Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis" (Loehlin, 1998), Mueller's "Basic Principles of Structural Equation Modeling" (Mueller, 1996) and Schumacker and Lomax's "A Beginners Guide to Structural Equation Modeling" (Schumacker & Lomax, 1996). Some specialized topics, such as reviews of goodness-of-fit indices, bootstrapping, accommodating non-normal data, are found in Hoyle's "Structural Equation Modeling: Concepts, Issues and Aplications" (Hoyle, 1995) and in Marcoulides and Schumacker's "Advanced Structural Equation Modeling: Issues and Techniques" (Marcoulides & Schumacker, 1996). An interesting collection of introductory articles focussing on SEM models for multigroup and longitudinal data is found in Little, Schnabel and Baumert's "Modeling Longitudinal and Multiple-Group Data" (Little, Schnabel & Baumert, in press).

There is one scientific journal devoted entirely to SEM: "Structural Equation Modeling, an interdisciplinary journal." Methodological journals that often contain SEM related articles include "Multivariate Behavior Research," "Psychological Methods," and "Sociological Methods & Research."

SEM on the Internet

Researchers working with structural equation modeling apparently include a number of Internet aficionados, as there are many interesting SEM resources on the Internet.

For beginners, a good place to start is the SEMNET FAQ (Frequently Asked Questions) page on Ed Rigdon's home page at <http://www.gsu.edu/~mkteer/>. Another good place is the homepage of the Germany-based working group on structural equation models at <http://www.uni-muenster.de/SoWi/struktur/> (note: these capital letters are needed!). These home pages include links to other SEM pages, information about handbooks and software, and links to the home pages of software publishers.

Some software publishers provide a free downloadable demonstration version of their program. The data for the examples given above are included in the demonstration version of AMOS (http://www.smallwaters.com/amos) and LISREL (http://www.ssi.com). There is also a free demonstration version of the program EQS (http://www.mvsoft.com). A special case is the program MX, which is freeware and available for downloading from Mike Neale's home page at <http://griffin.vcu.edu/mx/>.

Joel West maintains a large SEM related home page at <http://students.gsm.uci.edu/~joelwest/SEM/> which contains a very fine (partly annotated) list of SEM related books and articles.

A different Internet resource is the SEMNET discussion list. This discussion list is devoted to SEM-related topics. Members of the list include both novices and several eminent researchers in the field of structural equation modeling. The discussions vary from long diatribes on appropriate SEM strategies, to concise questions and answers about specific software problems. The information on how to subscribe to this list is on Ed Rigdon's home page at <http://www.gsu.edu/~mkteer/>.

References

Arbucle, J. (1997). Amos user's guide. Chicago: Smallwaters.

Bechger, T.M. (1997). Methodological aspects of comparison of educational acievement. Amsterdam: University of Amsterdam, Ph.D. Thesis.

Bentler, P.M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.

Bollen, 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing the equivalence of factor covariance and mean structure: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18, 115-126.

Chou, C.-P. & Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In: Hoyle, R.H. (ed.). *Structural equation modeling: concepts, issues and applications*. Newbury Park, CA: Sage.

Cohen, J. (1982). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cudeck, R. & Henley, S.J. (1991). Model selection in covariance structure analysis and the 'problem' of sample size: a clarification. *Psychological Bulletin*, 109, 512-519.

Gerbing, D.W. & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. Sociological Methods & Research, 21. 132-161

Dolan, C.V., Bechger, T.M. & Molenaar, P.C.M. (1998). Using structural equation modeling to fit models incorporating principal components. *Structural Equation Modeling*, in press.

Goldsteen, K., Ross, C.E. (1989). The perceived burden of children. *Journal of family Issues*, 10, 4, 504-526.

Hoogland, J. & Boomsma, A. (1997). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26, 329-367

Hoyle, R.H. (1995) (ed.). *Structural equation modeling: concepts, issues and applications*. Newbury Park, CA: Sage.

Hox, J.J. (1995). Amos, Eqs and Lisrel for Windows: A comparative review. *Structural Equation Modeling*, 2, 79-91.

Jöreskog, K.G. & D. Sörbom (1989*). Lisrel 7: A guide to the program and applications*. Chicago: SPSS.

Little, T., Schnabel, K.U. & Baumert, J. (1998). *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples*. Hillsdale, NJ: Erlbaum.

Loehlin, J.C. (1997). *Latent variable models: An introduction to factor, path, and structural analysis*. Hillsdale, NJ: Erlbaum.

MacCallum, R.C. (1989). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100, 533-541.

MacCallum, R.C., Roznowskei, M. & Necowitz, L.B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 114, 490-504.

Marcoulides, G.A. & Schumacker, R.A. (1996) (eds.). Advanced structural equation modeling: Issues and techniques. Hillsdale, NJ: Erlbaum.

McArdle, J.J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science*, 5, 11-18.

Miles, J. (1998). Review type: Statistical analysis/structural equation modelling. *Psychology Software News*, 8, 2, 58-65.

Mueller, R.). (1996). *Basic principles of structural equation modeling*. New York: Springer.

Schumacker, R.E. & Lomax, R.G. (1996). *A beginners guide to structural equation modeling*. Hillsdale, NJ: Erlbaum.

Sörbom, D. (1982). Structural equation models with structured means. In: Jöreskog, K.G., & Wold, H. (eds.). *Systems under indirect observation*. Amsterdam: Elsevier.

Spirtes, P., Scheines, R. & Glymour, C. (1991). Simulation studies of the reliability of computer-aided model specification using the Tetrad II, Eqs, and Lisrel. *Sociological Methods & Research*, 19, 3-66.

Tucker, C. & C. Lewis (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.

Waller, N. G. (1993). Software review. Seven CFA programs: EQS, EzPATH, LICS, LISCOMP, LISREL7, SIMPLIS, and CALIS. *Applied Psychological Measurement*, 17, 73-100.

Werts, C.E. (1979). Confirmatory factor analysis applications: missing data problems and comparison of path models between populations. *Multivariate Behavior Research*, 14, 199-213.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.