Papeles del Psicólogo, 2010. Vol. 31(1), pp. 34-45
http://www.cop.es/papeles

STRUCTURAL EQUATION MODELS

Miguel A. Ruiz, Antonio Pardo and Rafael San Martín
Psychology Faculty. Universidad Autónoma de Madrid

In this article Structural Equation Models (SEM) are presented. Structural equation modelling is a multivariate statistical technique used to test models proposing causal relations between their variables. After defining this type of model and presenting a typical example, we discuss the concept of causation and its meaning in the present context. The general model structure, together with the types of variables used and how to represent them in path diagrams are also discussed, and illustrated with an example. We describe the steps needed to build a model and the different types of relationships, and briefly discuss the goodness-of-fit concept. Some typical problems found with these models are also discussed. Finally, we suggest some additional resources.
Key words: *Structural equation models, Latent variables, Observed variables, Path diagrams, Path analysis, Analysis of covariance structures, Confirmatory factor analysis, Goodness of fit, Causal models.*

En este artículo se presentan los modelos de ecuaciones estructurales, una técnica de análisis estadístico multivariante utilizada para contrastar modelos que proponen relaciones causales entre las variables. Tras la definición de este tipo de modelos y la presentación de un ejemplo típico, se discute el concepto de causalidad, para entender su utilización en este contexto. A continuación se discute la estructura general que tiene un modelo, los tipos de variables que se pueden utilizar en ellos y su representación mediante diagramas estructurales, acompañado de la discusión de un ejemplo. Posteriormente se presentan los pasos en la elaboración de un modelo y los tipos de relaciones posibles. También se comentan brevemente el concepto de ajuste y los problemas típicos de estos modelos. Por último se ofrecen algunos recursos adicionales.
Palabras clave: *Modelos de ecuaciones estructurales, Variables latentes, Variables observadas, Diagrama de rutas, Modelos de rutas, Análisis de estructuras de covarianza, Análisis factorial confirmatorio, Bondad de ajuste, Modelos causales.*

Structural equation models are a family of multivariate statistical models that make it possible to estimate the effects and relations between multiple variables. Structural equation models emerged in response to the need for greater flexibility in regression models. They are less restrictive than regression models, since they permit the inclusion of measurement errors in both criterion (dependent) and predictor (independent) variables. They might be thought of as several factor analysis models that permit direct and indirect effects between factors.

Mathematically, these models are more complex to estimate than other multivariate models such as those of Linear Regression and Exploratory Factor Analysis. Hence, they were not widely used until 1973, when the LISREL (Linear Structural Relations; Jöreskog, 1973) analysis software appeared. LISREL, on being refined, gave way to LISREL VI (Jöreskog & Sörbom, 1986), which offered greater variety of estimation methods. EQS (an abbreviation of "Equations"; Bentler, 1985) is the other package traditionally used for this type of analysis. Today, other estimation software programs are also available running in graphical environments, such as AMOS (Analysis of Moment Structures; Arbuckle, 1997). The influence of estimation software has been so determinant in the development of structural equation models that they are often called LISREL models. In the international literature they are generally referred to by their initials, as SEM.

The great advantage of these types of model is that they allow us to propose the type and direction of the relationships expected to be found between the different variables they contain, to then proceed to estimate the parameters specified by the relationships proposed at a theoretical level. For this reason they are also called *confirmatory models*, since the primary interest is to "confirm", through analysis of the sample, the relationships proposed from the explanatory theory chosen to be used as a reference.

As it can be seen in the following example, the theoretical specification of the model permits the proposal of causal structures between the variables contained in it, so that some variables have an effect on other variables, which in turn can transfer these effects to other variables, creating concatenations of variables.

Correspondence: Miguel Ángel Ruiz. *Departamento de Psicología Social y Metodología. Facultad de Psicología. Universidad Autónoma de Madrid. Calle Iván Pavlov 6. 28049 Madrid. España. Email: miguel.ruiz@uam.es*

Figure 1 shows a structural equation model from the health field (González & Landero, 2008). Models such as the one shown here are also called "path analysis" models, and all the variables contained in them are observable, except prediction errors. The purpose of this specific model is to predict the magnitude of a person's psychosomatic symptoms from a set of personal antecedents. The model proposes the existence of three predictor variables (self-esteem, self-efficacy and social support) that influence the individual's stress level. In turn, stress has a direct influence on psychosomatic symptoms, as well as an indirect one, modulated by the emotional exhaustion level. As it can be seen, the model proposed is somewhat more complex than a regression model, since some variables play the role of predictor variable and dependent variable simultaneously.

A rapid interpretation of the magnitude and direction of the studied parameters reveals that the predictor variables have a negative effect on the level of stress, so that the lower the perceived self-efficacy, the lower the self-esteem and the lower the social support, the greater the level of stress. Moreover, perceived self-efficacy is the predictor with the greatest effect, and all the predictors are related to each other. With the predictors employed we can account for 42% of the variability of the stress. Moreover, stress has a direct and positive (0.16) influence on psychosomatic symptoms, but the indirect effect via emotional exhaustion is greater (0.21=0.54*0.39). Overall, 24% of the differences in the subjects' psychosomatic symptoms are accounted for. The meaning of these and other elements of the figure will be explained later.

Structural equation models are so called because it is necessary to use a set of equations to represent the relationships proposed by the theory. To represent the relations in the above example, three regression equations are being used and estimated simultaneously.

There are many types of models with different levels of complexity and with different purposes. All of them are models of a statistical nature. This means that they consider the presence of measurement errors in the observations obtained from reality. They normally include multiple observable variables and multiple unobservable (latent) variables, though some, such as that of the example, only include latent variables corresponding to the prediction errors.

As regards their estimation, structural equation models are based on the correlations between the variables measured cross-sectionally in a sample of subjects. Therefore, in order to make the estimations, it is sufficient to measure a set of subjects at a given moment in time. This makes these models especially attractive. Even so, it should be borne in mind that the variables must permit the calculation of correlations, so that they must be quantitative, and preferably continuous.

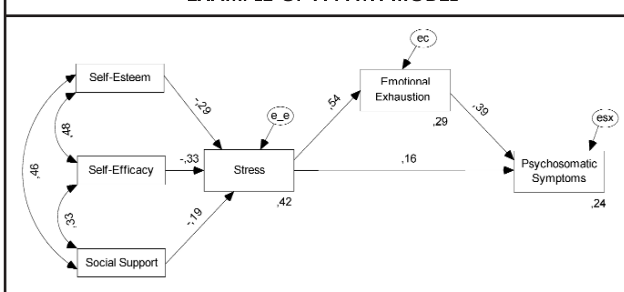
The strong points of these models are: having developed a set of conventions that permit their diagrammatic representation, the possibility of hypothesizing causal effects among the variables, the fact that they permit the concatenation of effects between variables, and the fact that they permit reciprocal relations between variables.

There are many types of models that can be defined with this methodology. The most popular of those mentioned in the statistical literature are: Multiple regression with multicollinearity, Confirmatory factor analysis (see Ferrando & Anguiano, 2010), Second-order factor analysis, Path analysis, Complete causal model with latent variables, Latent curve model (see Bollen & Curran, 2006), Multilevel models (see Skrondal & Rabe-Hesketh, 2004), Multi-group models, Mean-based models (ANOVA, ANCOVA, MANOVA and MANCOVA; see Bagozzi & Yi, 1994) and Mediation analysis (see Preacher et al., 2007).

THE CONCEPT OF CAUSALITY

An interesting feature of these types of models is the possibility they offer of representing the causal effects among their variables. Although it is very attractive to be able to represent diagrammatically the causal influence of one variable on another, and although we are also able to estimate the parameter corresponding to that effect, it should be stressed that the estimation of the parameter does not "demonstrate" the existence of causality. The existence of a causal relation between the variables must

FIGURE 1
EXAMPLE OF A PATH MODEL



Adapted from Gonzalez & Landero (2008).

be supported by the theoretical articulation of the model, and not by its estimation with data of a cross-sectional nature. In order to demonstrate scientifically that there is a causal relation it is necessary to design a controlled experiment with random assignment of subjects to the study conditions (see Pardo, Ruiz & San Martín, 2009, pp. 356-359). It should be borne in mind that structural equation models are used in studies of correlational nature, in which only the magnitude of the variables is observed, and the variables are never manipulated.

The work of Boudon (1965) and Duncan (1966) opened up a new possibility for approaching the problem of causality, distinct from that of experimental manipulation, proposing dependence analysis or path analysis. In this type of analysis a causal theory is studied through the specification of all the variables that are important for that theory. Subsequently, the relations between the causal effects can be derived from the causal theory, to eventually estimate the size of those effects. The generalization of the path analysis model gave rise to structural equation models for the testing of theories – or causal models, which amount to the same. According to the logic of these models, on the basis of the theory underlying the model, it will be possible to derive the covariance measures expected among the variables directly from the causal effects contained in the model. If the theory is correct, the covariance measures derived from the model and the covariance measures obtained from the data should be equal.

STRUCTURE OF A MODEL

A complete structural equation model comprises two basic parts: the measurement model and the structural relations model.

The measurement model contains the form in which each latent construct is measured through its observable indicators, the errors that affect the measurements and the relationship expected to be found among the constructs when they are related to one another. In a complete model there are two measurement models, one for the predictor variables and another for the dependent variables.

The structural relations model is that which we actually want to estimate. It contains the effects and relationships between the constructs, which will typically be latent variables. It is similar to a regression model, but can also contain chained effects and loops among variables. Moreover, it contains the prediction errors (which are distinct from the measurement errors).

There are two exceptional cases in which the model does not contain both parts, and which are used relatively frequently. First of all, confirmatory factor analysis models contain only the measurement model, and the relations between latent variables can only be of a correlational nature. Secondly, path analysis models do not contain latent variables; instead, the observable variables are equated with the latent variables; consequently, the structural relations model is the only one present. As a result of this drawback, measurement errors and prediction errors are combined in a single common term.

TYPES OF VARIABLES

In a structural model two different types of variable are distinguished, according to their role and their measurement.

- ✔ Observed or indicator variable. Variables measured in the subjects, e.g., the items of a questionnaire.
- ✔ Latent variable. Characteristic that it is wished to measure but that cannot be observed and is free of measurement error. For example, a dimension of a questionnaire or a factor in an exploratory factor analysis.
- ✔ Error variable. Represents both the errors associated with the measurement of a variable and the set of variables not considered in the model and which may affect the measurement of an observed variable. These are considered to be latent variables, since they are not directly observable. The error associated with the dependent variable represents the prediction error.
- ✔ Grouping variable. Categorical variables that represent membership of the different subpopulations to be compared. Each code represents a subpopulation.
- ✔ Exogenous variable. Variable that affects another variable and that is not affected by any variable. The independent variables of a regression model are exogenous.
- ✔ Endogenous variable. Variable that is affected by another variable. The dependent variable of a regression model is endogenous. All endogenous variables must be accompanied by an error.

STRUCTURAL DIAGRAMS: CONVENTIONS AND DEFINITIONS

In order to represent a causal model and the relations to be included in it, it is customary to use diagrams similar to flow diagrams. These diagrams are called *causal* *diagrams, path diagrams or structural diagrams*. The structural diagram of a model is its graphic representation, and is of great help in specifying the model and the parameters contained in it. Indeed, today's software makes it possible to define the model in its entirety on representing it in the graphic interface. On the basis of the structural diagram, the software itself derives the equations of the model and reports on the restrictions necessary for it to be completely identified. Structural diagrams follow some particular conventions that one needs to know so as to be able to derive the corresponding equations.

- ✔ The observable variables are shown within rectangles.
- ✔ The unobservable (latent) variables are shown within ovals or circles.
- ✔ The errors (be they of measurement or prediction) are shown without rectangles or circles (though some software programs show them in the same way as latent variables).
- ✔ Bi-directional relations (correlations and covariations) are shown as curved vectors with an arrow at each end.
- ✔ All structural effects are shown as a straight arrow, whose origin is the predictor variable and whose end, at the point of the arrow, is the dependent variable.
- ✔ The model parameters are shown on the corresponding arrow.
- ✔ Any variable that is affected by other variables of the model must also include an error term.
- ✔ Although it is not necessary for the user to specify it, software programs tend to include, together with each variable, its variance, and, if it is a dependent variable, its corresponding proportion of explained variance.

Structural diagrams also serve for the adequate specification of the model with a view to estimation with statistical software. The restrictions are made diagrammatically or by imposing values on the diagram itself. Moreover, statistical software programs make it possible to check the specified model on the basis of the diagram they generate. This helps to avoid overlooking fundamental parameters in the definition of the model, so that the user does not have to write down the model's equations explicitly and trust that the equations are the right ones.

Let us review the model considered previously as an example, but this time defined with more complexity. Figure 2 shows a new version of the model that contains six latent variables: self-esteem, self-efficacy, social support, stress, emotional exhaustion and psychosomatic symptoms. The first three latent variables are exogenous (because they are not directly affected by any other variable) and the last three latent variables are endogenous, because they are affected by other variables. The three endogenous variables have a term that represents their prediction error (e_e, ece and esx).

Each endogenous latent variable is measured by means of three observable variables called indicators. The latent variable psychosomatic symptoms is measured in subjects through three scales called INDI1SX, INDI2SX and INDI3SX. The model assumes that a person with many

FIGURE 2
COMPLETE STRUCTURAL MODEL FOR PREDICTING PSYCHOSOMATIC SYMPTOMS

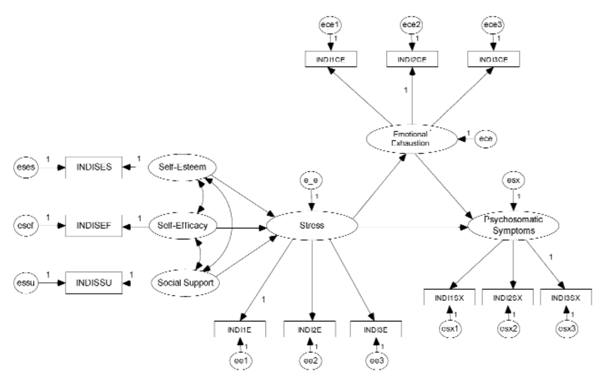
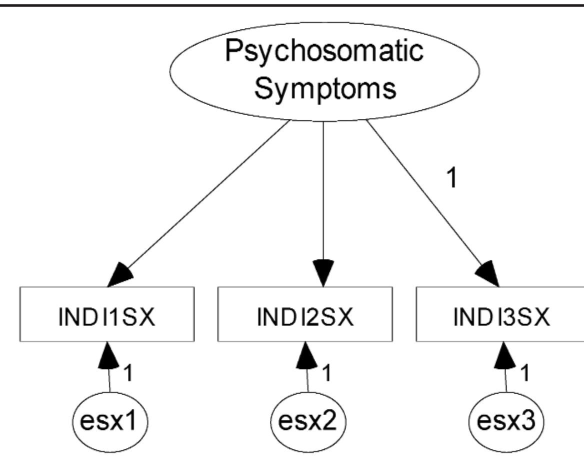


FIGURE 3
MEASUREMENT MODEL OF THE VARIABLE PSYCHOSOMATIC SYMPTOMS



psychosomatic symptoms will score high in all three indicators and a person with few psychosomatic symptoms will score low. The indicators are observable but are not perfect measures of their latent variable. Therefore, each indicator has an associated measurement error. The measurement error of the indicator INDI1SX is the unobservable variable esx1. Figure 3 shows the measurement model of the latent variable psychosomatic symptoms. In the case of exogenous latent variables, each construct is measured by a single indicator, so that this part of the model can be simplified by identifying the latent variable with its indicator, as can be seen in the final estimated model in Figure 5.

Figure 4 shows the structural relations model. This model contains only the latent variables. It is easy to appreciate in it that the exogenous variables can correlate with one another (which would be impossible in an ordinary regression model), and that each endogenous variable has an associated prediction error that accounts for part of its variability (this error is not associated with the measurement errors, which are represented in the measurement model).

FIGURE 4
STRUCTURAL RELATIONS MODEL

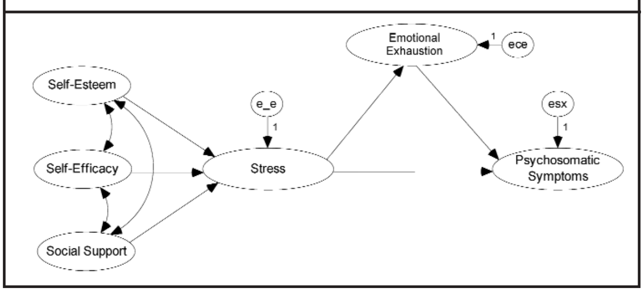


FIGURE 5
FINAL ESTIMATED MODE

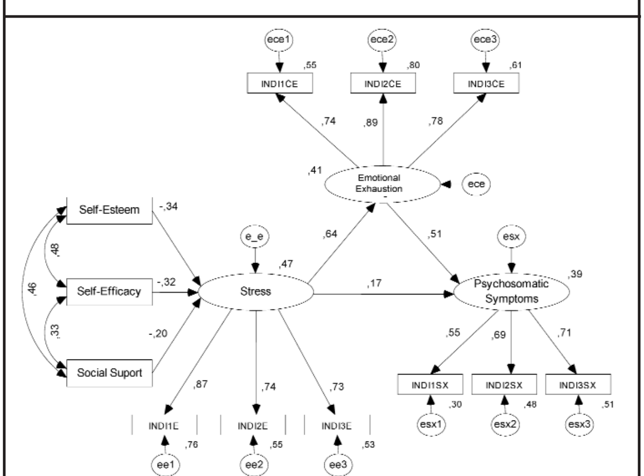


Figure 5 represents the final estimated model, once it has been simplified with respect to the exogenous variables. On the left are the three exogenous variables used for predicting the stress level. The three variables are observable and correlate with one another (they are multicollinear). Their negative effect on stress indicates that lower levels of self-esteem, self-efficacy and social support permit us to predict a higher level of stress. (Although not indicated in the diagram, all the regression weights differ significantly from zero). The combination of the three predictors permits the explanation of 47% of the variance of stress (free of measurement error), which is indicated numerically on the latent variable. The proportion of variance of the stress accounted for by its predictors is inversely proportional to the variance of its prediction error, so that it is not necessary to indicate this second value, while the corresponding error variable is indeed shown (e_e). Each endogenous latent variable is measured by three indicators. Each arrow that proceeds from a latent variable toward its indicator is interpreted in the same way as the loading in factor analysis, and (in the standardized solution) corresponds to the correlation of the indicator with the latent variable it intends to measure. The numerical value shown next to the box of an observed variable is the proportion of variance shared by the indicator and the corresponding latent variable (similar to the communality) and which is not attributable to the measurement error. In the central part of the model are the effects of each latent variable on the others. As it can be seen, stress has a greater direct effect on emotional exhaustion than on psychosomatic symptoms. In turn, the effect of stress on psychosomatic symptoms is lower than the effect received from emotional exhaustion. The figure does not show the total effect of stress on psychosomatic symptoms (0.50), which would be the sum of the direct effect (0.17) and the indirect effect (0.64*0.51=0.33) via emotional exhaustion.

Comparing the complete model with the path model in Figure 1, it can be seen that the effects have increased substantially in some cases and that, moreover, there has been a rise in the proportion of explained variance of the endogenous variables. It can also be appreciated that not all the indicators are equally accurate. Finally, it is to be expected that this equivalent model, though useful, will obtain poorer fit values than the path model for the mere fact of containing a greater number of variables (which affects the model's degrees of freedom and the goodness-of-fit statistics).

Just as there is a set of conventions for representing the models in diagrammatic form, there are also conventions for naming each element of a model, be they variables or parameters, in the mathematical notation. We shall not go into an explanation of this notation here, but it is useful to know that it is customary to use Greek letters (see Ruiz, 2000; Hayduk, 1987).

STEPS IN THE DEVELOPMENT OF A MODEL

The estimation of a model begins with the formulation of the theory that supports it. This theory should be formulated so that it can be tested with real data. Specifically, it must contain the variables considered important and which are to be measured in the subjects. The theoretical model must specify the relations expected to be found between the variables (correlations, direct effects, indirect effects, loops). If a variable is not directly observable, mention should be made of the indicators that permit its measurement. Normally, the model will be formulated diagrammatically, making it easy to identify the equations and parameters.

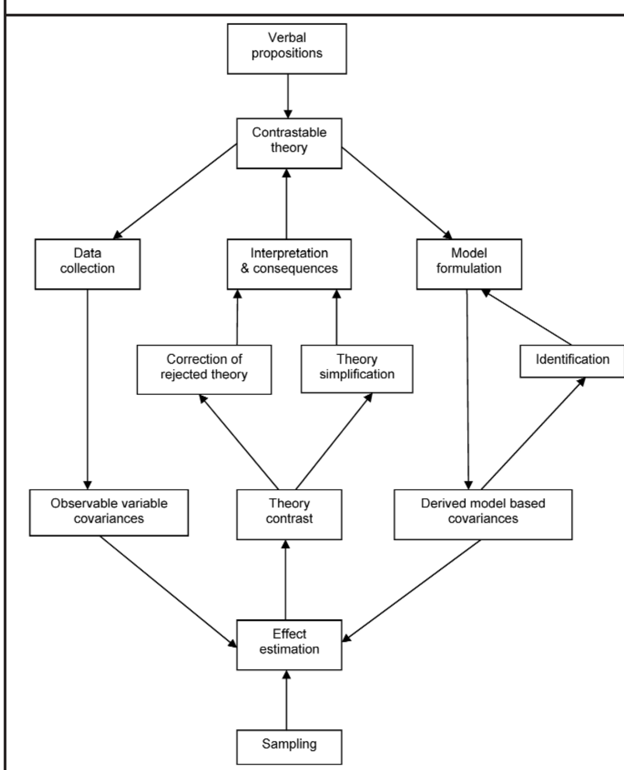
Once the model has been formulated, each parameter must be correctly identified and derivable from the information contained in the variance-covariance matrix. There are strategies for ensuring that all the parameters are identified, such as using at least three indicators per latent variable and making the measurement scale of each latent variable equivalent to one of its indicators (this is achieved by setting the weight of one of the indictors arbitrarily at a value of 1). Even so, it may occur that the model is not completely identified, which will mean that it is being attempted to estimate more parameters than the number of pieces of information contained in the variance-covariance matrix. In that case it will be necessary to impose further restrictions on the model (setting the value of one or more parameters) and reformulate it.

When the variables that form part of the model have been selected, it must be decided how to measure the observable variables. These measurements (generally obtained by means of scales or questionnaires) will permit us to obtain the variances and covariances on which to base the estimation of the parameters of a model correctly formulated and identified (we assume that we are working with a representative and sufficiently large sample of the population under study).

The estimation of the model's parameters is followed, first of all, by an assessment of the model fit. If the estimations obtained do not correctly reproduce the observed data, the model will have to be rejected, and with it the theory that supported it; the model can then be corrected by making additional theoretical assumptions. Secondly, a technical assessment is made of the estimated values for the parameters. Their magnitude must be adequate, the effects must differ significantly from zero, inappropriate estimations (such as negative variances) must not be obtained, and so on. It may be the case that some of the estimations attain a value close to zero; when this occurs it is recommended to simplify the model by eliminating the corresponding effect. Finally, all the parts of the model must be interpreted. If the model has been accepted as a good explanation of the data it will be interesting to validate it with other samples, and quite possibly to use it as an explanation of more complex theories that can be tested. The process is summarized in diagrammatic form in Figure 6.

FIGURE 6

PROCESS OF DEVELOPMENT AND ESTIMATION OF A MODEL



TYPES OF RELATIONSHIP

In multivariate techniques we are accustomed to studying the simultaneous relationship of diverse variables among one another. In these techniques the relationships between dependent and independent variables are all of the same level or the same type. In a structural equation model we can distinguish different types of relationship. Understanding these different types of relationship can be of great help in formulating the models on the basis of ordinary language. Below we discuss these types of relationship, following the scheme proposed by Saris and Stronkhorst (1984).

COVARIATION vs CAUSALITY

We say that two phenomena covary, or are correlated, when on observing a greater quantity of one of the phenomena we also observe a greater quantity of the other (or a lesser quantity if the relationship is negative). Likewise, low levels of the first phenomenon are associated with low levels of the second. Thus, for example, when we say that aptitude and performance correlate with one another, we expect subjects with a higher level of aptitude to show better performance, and vice versa. However, as we have already stressed, covariation and causality are not the same thing. When we observe a strong relationship (covariation) between two variables, we should not interpret this as signifying a causal relationship between them. There may be other variables we have not observed, and which enhance or attenuate the relationship. For example, it may be that motivation and performance are related, and that their relationship is conditioning the relationship between aptitude and performance (enhancing it or attenuating it). A possibly clearer example is that proposed by Saris. If we gather data on the number of vehicles and the number of telephones in different populations, we shall surely find a covariation between the two variables. However, that must not lead us to think that the greater number of vehicles is the cause of the greater number of telephones.

Another level of analysis is causality. If we collect information on the number of smokers in a room and the amount of smoke in that room, we will observe a high covariation between the two variables. It seems reasonable to go one step further in our interpretation of this result and argue, conceptually, that the quantity of smokers causes the amount of smoke, and that changes in the quantity of smokers will cause a change in the amount of smoke.

The change of perspective from the covariation observed to the causality attributed to two variables is made by the researcher, who is the one hypothesizing the causality. It is good practice for statements we make to be worded explicitly with respect to the type of relationship we wish to test between two variables.

The examples we have considered in this section can be represented by means of the diagrams we have developed up to now.

If we are studying the correlation between aptitude and performance it should be represented with a curved arrow between the two variables.

Figure 7. Covariation relationship

In contrast, the causal relation between the number of smokers and the amount of smoke will be represented by a vector from the cause to the effect.

Figure 8. Relationship of a causal nature

SPURIOUS RELATIONSHIP

A basic causal relationship or a covariation relation involves two variables. In a spurious relationship the relation is made up of at least three variables. A spurious relationship is one in which the covariation between two variables is due, totally or partially, to the common relation of the two variables with a third one. This is why the covariation between two variables may be very high, while at the same time their causal relation is null. A typical example of a spurious relationship is that found between height and intelligence in pre-school children. If we measure the two variables in such children it is quite possible that we will find a strong relationship between them; however, nobody would seriously think that height causes intelligence. There is a third variable, the child's development (age) that is the cause of both variables and which leads to this relationship being observed. This can be represented diagrammatically as follows:

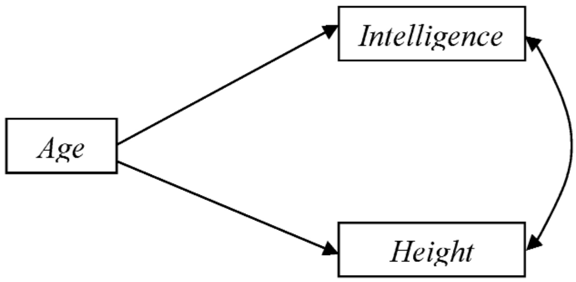


Figure 9. Spurious relationship

To study the presence of this phenomenon we use the partial correlation coefficient, which measures the relationship between two variables after partialling out the effect of a third one (it is also possible to partial out the effect of more than one variable). In our example, the correlation between the three variables will be high and positive, whilst the partial correlation between intelligence and height (partialling out the effect of age) will be practically null.

In general, we can state that the causal relation between two variables implies that the two covary, the rest of the variables remaining constant. But the opposite is not true: covariation between two variables does not necessarily imply that there is a causal relation between them – the relationship may be spurious, false, fictitious (see Pardo, Ruiz & San Martín, 2009, pp. 356-357).

DIRECT AND INDIRECT CAUSAL RELATIONSHIP

Up to now we have mentioned only direct causal relationships. An indirect causal relationship implies the presence of three variables. There is an indirect relationship between two variables when a third variable modulates or mediates the effect between them – that is, when the effect between the first and second variable is via a third one. Variables that mediate in an indirect relationship are also referred to as modulating variables.

Let us consider the relation between aptitude, performance and motivation. We can think of motivation level as a variable that modulates the relation between aptitude and performance. This relationship can be represented thus:

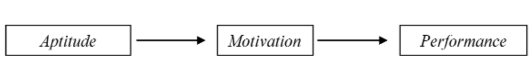


Figure 10. Indirect causal relationship

The model in the figure proposes a direct effect of aptitude on motivation and of motivation on performance. Moreover, there is an indirect effect between aptitude and performance. The indirect effect of the variable aptitude on performance can be enhanced (or attenuated) by the modulating variable motivation.

The existence of an indirect effect between two variables does not preclude the possibility that there will also be a direct effect between them. Thus, the relations proposed in Figure 10 can become more complex, as follows:

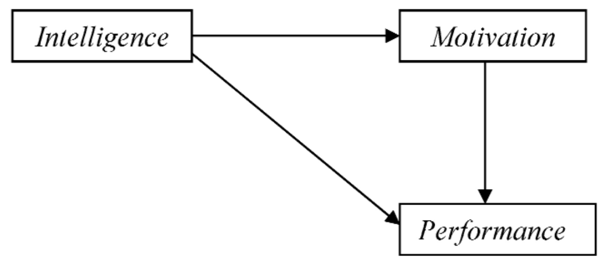


Figure 11. Direct and indirect relationships

Once more, it is the researcher who must make explicit the types of relationship that his or her theory is capable of explaining.

RECIPROCAL CAUSAL RELATIONSHIP

The causal relationship between two variables can be either reciprocal or unidirectional. When the relationship is reciprocal (bidirectional), the cause variable is in turn the effect of the other one. This type of relationship is represented by two separate arrows pointing in opposite directions. A reciprocal relationship is actually a feedback loop between two variables. The reciprocal causal relationship can be direct or indirect, involving other variables before the loop is closed.

The relationship between Anxiety and Performance can be represented as a reciprocal loop: the greater the anxiety, the poorer the performance; and the poorer the performance, the greater the anxiety.

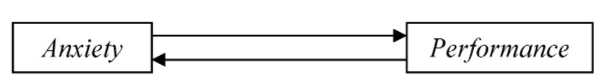


Figure 12. Reciprocal causal relationship

TOTAL EFFECTS

As we have seen, each type of causal relationship is represented by means of a type of effect. There is a final type of effect (or relationship) that we have not mentioned, that of *non-analyzed effects*. In the diagram these would be the arrows that could have been shown but are not. Their absence may be for one of two reasons. On the one hand, it may be that important variables for explaining the covariation in the data have been left out (specification error). On the other hand, it may be because it is assumed that the rest of the variables not considered in the model compensate one another, their effect being incorporated in the model's error terms. The sum of the spurious effects and the non-analyzed effects results in so-called *non-causal effects*. Once the model is defined, the spurious effects appear when the endogenous variables are correlated beyond the estimated effects (covariances emerging between the prediction errors). Non-analyzed effects appear when the observable variables are correlated beyond what the model predicts (covariances emerging between the measurement errors).

Given that an endogenous variable can receive a direct effect from another variable and also an indirect effect from that same variable modulated by other variables, it is customary to add the two types of effect, giving rise to the *total effect*.

THE CONCEPT OF "FIT"

To understand the basis of structural equation models it is necessary to change our perspective on what the concept of a model's *fit* means. In linear regression, when we speak of the parameter estimates we choose those estimations that best fit the model to the data, in the sense that they minimize the prediction errors committed with the model for the set of subjects in the sample (in the least squares method). In contrast, in structural equation models the aim is to fit the covariances between the variables, rather than seeking the fit to the data. Instead of minimizing the difference between the predicted and observed values at an individual level, what is minimized is the difference between the covariances observed in the sample and the covariances predicted by the structural model. This is why these models are also referred to as *covariance structure models* (Long, 1983). Thus, the residuals of the model are the difference between the covariances observed and the covariances reproduced (predicted) by the theoretical structural model.

The fit of a model can be expressed in a basic hypothesis, which proposes that, *if the model is correct* and we know the parameters of the structural model, the populational covariance matrix could be reproduced accurately based on the combination of the model's parameters. This notion of fit is summarized in the following equation

$$H_0:\ \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \qquad (1)$$

where is the populational variance-covariance matrix among the observable variables, is a vector that contains the parameters of the model and ( ) is the variance-covariance matrix derived as a function of the parameters contained in the vector .

We can see the meaning and scope of this hypothesis through an example (Bollen, 1989). Let us consider the model shown in Figure 13:

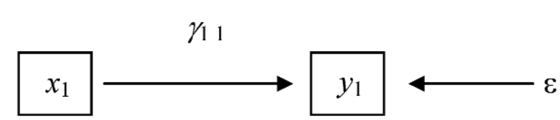


Figure 13. Simple regression model

The regression equation that defines it is as follows (subindices have been removed)

$$y = \gamma x + \varepsilon \qquad (2)$$

Where g is the regression coefficient and is the variable that represents the error term, which is assumed to be independent of *x* and whose expected value is zero. The variance-covariance matrix among the observed variables *x* and *y* is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{VAR}(y) & \mathrm{COV}(x,y) \\ \mathrm{COV}(x,y) & \mathrm{VAR}(x) \end{pmatrix} \qquad (3)$$

This is the matrix we obtain directly on analyzing the data descriptively, and represents the relationships between the variables in the sample. However, the dependent variable *y* is a function of the variables *x* and , and of the parameter g. We can rewrite the elements of the matrix according to equation (2). Operating, it is relatively easy to demonstrate that the variance of the dependent variable is a function of the parameter g and of the error variance:

$$\mathrm{VAR}(y) = \gamma^2\ \mathrm{VAR}(x) + \ \mathrm{VAR}(\varepsilon) \quad (4)$$

It is also possible to demonstrate that the covariance between *x* and *y* is a function of the parameter g and of the variance of the predictor variable:

$$\mathrm{COV}(x,y) = \ \gamma\ \mathrm{VAR}(x) \quad (5)$$

Substituting in equation (3) the derived expressions written according to the parameters of the model, we arrive at the reproduced populational variance-covariance matrix:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \gamma^2\,\mathrm{VAR}(x) + \mathrm{VAR}(\varepsilon) & \gamma\mathrm{VAR}(x) \\ \gamma\mathrm{VAR}(x) & \mathrm{VAR}(x) \end{pmatrix} \quad (6)$$

This matrix is also called an *implied* variance-covariance matrix. We can now substitute in equation (1) and express the basic hypothesis once more as:

$$H_0:\quad \boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{VAR}(y) & \mathrm{COV}(x,y) \\ \mathrm{COV}(x,y) & \mathrm{VAR}(x) \end{pmatrix} =$$

In this expression the elements on the right and those on the left correspond one to one, given the specifications of the model proposed. If the model is correct and we know the values of the parameters on the right of the equation, it will not be difficult to check the equality of the terms. The objective of the estimation is to obtain those values of the parameters (in this case the regression coefficient and the error variance) that permit this equation to be maintained with the sample data.

To be able to estimate the parameters of the model we had to await the development of specialized computer software. In this brief approach to structural equation models it suffices to know that the estimations are made with a view to maximizing the fit of the model. To this end we use some measure summarizing the magnitude of the differences between the variances and covariances observed (left-hand side of the equation) and reproduced (right-hand side of the equation), and we attempt to minimize such differences.

GOODNESS-OF-FIT STATISTICS

Once the model has been estimated its quality must be assessed. To this end, goodness-of-fit statistics are used. There are three types of goodness-of-fit statistics: those of absolute fit (which rate the residuals), those of relative fit (which compare the fit with respect to another model with poorer fit) and those of parsimonious fit (which rate the fit with respect to the number of parameters used). None of them offers all the information necessary for assessing the model, and it is customary to use a set of them, reporting them all together (see Schreiber et al., 2006).

The following table lists those most widely used, together with their standard abbreviations and the reference value that should be attained to indicate good fit. The chi-squared statistic is conceptually the most appealing, since it allows us to test the null hypothesis that all the model's errors are null, a hypothesis which it is pertinent to maintain with the sample used. However, this statistic is highly sensitive to sample size: with large samples (over 100 or 200 cases) it is relatively easy to reject the null hypothesis when the model actually attains a good fit. For this reason, as well as assessing its statistical significance, a comparison is usually made with its degrees of freedom. This statistic is always reported.

TYPICAL PROBLEMS

We should point out various typical problems that tend to be found in the published models, some limitations that should be taken into account and the safety measures that must be taken on using the models.

In the definition of a model it is essential not to exclude important variables from the theoretical point of view.

TABLE 1
GOODNESS-OF-FIT STATISTICS AND REFERENCE CRITERIA

Index	Abbreviation	Criteria
Absolute fit		
Chi-square	χ^2	Significance > 0.05
Chi-square / degrees of freedom	χ^2 / df	Less than 3
Comparative fit		
Comparative goodness of fit index	CFI	≥ 0.95
Tucker-Lewis index	TLI	≥ 0.95
Normed Fit Index	NFI	≥ 0.95
Parsimonious fit		
Parsimony-adjusted NFI	PNFI	Close to 1
Other		
Goodness of fit index	GFI	≥ 0.95
Adjusted goodness of fit	AGFI	≥ 0.95
Root mean square residual	RMR	Close to 0
Root mean square error of approximation	RMSEA	< 0.08

First of all, it should be attempted to measure all the pertinent variables. Secondly, any model in which the conceptually central variables lack significant effect should be called into question.

The fact that a model obtains good fit with a sample does not rule out the possibility of there being other tentative models that might also fit the data well. It is always of interest to test other models that might also be supported by the theory (or by rival theories).

Occasionally, models are published that contain both the effects corresponding to parameters differing from zero and effects which after the estimation can be considered null. Although more space would be required for a full explanation, information should be provided on both the theoretical model with all the parameters and variables proposed and the final model containing only the parameters that differ from zero and the variables with statistical effect.

It is well known that goodness-of-fit statistics deteriorate rapidly as sample size increases, and many researchers report on small samples so as to avoid deterioration of the fit values. Therefore, one should call into question those models estimated with small or relatively unrepresentative samples. It is customary to require sample sizes in excess of 100 subjects, and those of over 200 subjects give good guarantees.

These models admit the use of a small number of variables (10-20). The greater the number of variables, the more difficult it is to correctly reproduce the covariances observed. Moreover, the more variables there are, the greater the sample size should be (a rate of over 10 subjects per observed variable is recommended).

Many studies employing these models abuse the fitting and re-fitting of the possible theoretical relations, including and excluding effects and variables in a tentative manner. With this purpose they use the significance values and the modification indices of the individual parameters (of both the analyzed effects and the excluded effects), which provide information on the problems of fit present in the data. These over-manipulated models tend to be quite unstable, and lose good properties of fit when replicated with other samples. Unfortunately, replication studies are scarce, so that it is recommended to maintain a certain degree of scepticism when a study fails to report in detail on how the data and the model may have been manipulated.

Categorical variables should not be used, since, ideally, all the variables should be quantitative and continuous to justify the use of variance and covariance statistics. As we have seen, it is essential for the sample estimation of the variances and covariances between the observed variables to be accurate, if the process of estimation of the model's parameters is to be successful. However, it is very common to use items in ordinal, Likert-type format to measure subjects, given the fact that they can be answered easily. In such cases it is appropriate to group the individual items to form scales with more continuous measurement (see Finney & DiStefano, 2006).

FINAL CONSIDERATIONS

In spite of the limitations mentioned, structural equation models represent a highly powerful tool for formalizing in an explicit way relatively complex theories; moreover, they make it possible to test those theories and to include complex or hierarchical relations between multiple variables.

They also permit the extension of some traditional models on including, for example, measurement errors in factor analysis models, on estimating directly the loadings and correlations among factors (without using rotation), or on including individual significance tests for the estimated factor loadings.

Furthermore, in these models it is possible to separate measurement errors from prediction errors, attenuating the effect of the measurement errors on the assessment of the model's predictive capacity.

These models, together with canonical regression models, are the only ones to allow the analysis of problems in which there is more than one dependent variable and to analyze those variables simultaneously.

Although the estimation of these models has been greatly simplified with the advent of estimation software which uses a graphic interface, it is important to bear in mind that its use is laborious. Even so, they are undoubtedly of inestimable help in facing the challenge of developing explanatory theories of human behaviour.

ADDITIONAL RESOURCES

Those who wish to explore these models in more depth, while remaining at a relatively basic level, may consult the manuals by Byrne (1994, 1998, 2001, 2006), while those who would prefer an even more elementary introduction might wish to look at the book by Saris and Stronkhorst (1984) and the brief papers by Long (1983a, 1983b, 1990). A good explanation of the development and interpretation of these models can be found in the last three chapters of the manual by Hair et al. (2006), which is highly practical even though it contains scarcely any formulation and lacks demonstrations. The manual by Bollen (1989) is excellent, and quite complete, but requires a good level of previous knowledge in statistics.

Also highly recommended are the manuals of the most widely used estimation software: AMOS (Arbuckle, 1997), LISREL (Jöreskog & Sörbom, 1986; SPSS, 1990, 1993), EQS (Bentler, 1985) and CALIS, which is part of SAS (Hatcher, 2003).

There are also two model estimation software packages that can be used free of charge: HYBALL (http://web.psych.ualberta.ca/~rozeboom/) and TETRAD (http://www.phil.cmu.edu/projects/tetrad/).

REFERENCES

Arbuckle, J. L. (1997). *Amos Users' Guide. Version 3.6.* Chicago: SmallWaters Corporation.

Bagozzi, R. O. & Yi, Y. (1994). Advanced Topics in Structural Equation Models. In: R. P. Bagozzi (Ed.). *Advanced Methods of Marketing Research*. Cambridge: Blackwell Publishers.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Bollen, K. A., Curran, P. J. (2006). *Latent Curve Models. A Structural Equation Perspective*. Wiley.

Boudon, R. (1965). A method of linear causal analysis: Dependence analysis. *American Sociological Review, 30*: 365-373.

Byrne, B. M. (1994). *Structural Equation Modeling with EQS and EQS/WINDOWS: Basic Concepts, Applications, and Programming*. SAGE Publications.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M. (2006). *Structural Equation Modeling With Eqs: Basic Concepts, Applications, and Programming (Multivariate Applications)*. SAGE Publications.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah: Lawrence Erlbaum Associates, Inc.

Duncan, O. D. (1966) Path analysis: Sociological examples. *American Journal of Sociology, 72*: 1-12.

Ferrando, P.J. & Anguiano, C. (2010). El análisis factorial como técnica de investigación en Psicología [Factor analysis as a research technique in Psychology]. *Papeles del Psicólogo, 31(1)*, 18-33.

Finney, A.J. & DiStefano C. (2006). Non-Normal and Categorical Data in Structural Equation Modeling. In *G. R. Hancock & R. O. Mueller (Eds). Structural Equation Modeling: A Second Course, 269-314*. Greenwich: Information Age Publishing

González, M.T. & Landero, R. (2008). Confirmación de un modelo explicativo del estrés y de los síntomas psicosomáticos mediante ecuaciones estructurales [Confirmation of an explanatory model of stress and psychomatic symptoms by means of structural equations]. *Revista Panamericana de Salud Pública, 23 (1)*, 7-18.

Hayduk, L. (1987). *Structural equation modeling with LISREL Essentials and Advances*. Baltimore: The Johns Hopkins University Press.

Hair, J. E., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham R. L. (2006). *Multivariate Data Analysis (6th Edition)*. Upper Saddle River: Pearson-Prentice Hall

Hatcher, L. (2003). *A Step-by-Step Approach to using SAS for Factor Analysis and Structural Equation Modeling*. Cary: SAS Institute Inc.

Jöreskog, K. G. (1973) A general method for estimating a linear structural equation system. In A. S. Goldberger and O. D. Duncan (eds.), *Structural Equation Models in the Social Sciences, pp. 85-112.* New York: Seminar.

Jöreskog, K. G. & Sörbom, D. (1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*. Mooresville, IN: Scientific Software, Inc.

Long, J. S. (1983a). *Confirmatory Factor Analysis: A Preface to LISREL*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 007-033. Newbury Park, CA: Sage.

Long, J. S. (1983b) *Covariance Structure Models: An introduction to LISREL*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-034. Beverly Hills and London: Sage.

Long, J. S. (1990). *Covariance Structure Models: An Introduction to LISREL*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 007-034. Newbury Park, CA: Sage.

Pardo A., Ruiz, M.A. & San Martín, R. (2009). *Análisis de datos en ciencias sociales y de la salud* (volumen I)

[*Data analysis in the social and health sciences* (Vol. I)]. Madrid: Síntesis.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*, 185-227

Ruiz, M.A. (2000). *Introducción a los modelos de ecuaciones estructurales* [*Introduction to structural equation models*]. Madrid, UNED.

Saris, W. E. & Stronkhorst, L. H. (1984). *Causal Modelling in Non-Experimental Research*. Amsterdam: Sociometric Research.

Schreider, J.B., Stage, F.K., King, J., Nora, A., Barlow, E.A. (2006). Reporting structural equation modelling and confirmatory factor analysis results: a review. *The Journal of Education Research, 99 (6)*, 323-337.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman & Hall / CRC.

SPSS (1990). *SPSS® LISREL® 7 and PRELIS®: User's Guide and Reference*. Chicago, IL: SPSS.

SPSS (1993). *SPSS® LISREL® 7 and PRELIS®: User's Guide and Reference*. Chicago, IL: SPSS.