

Latent Variables

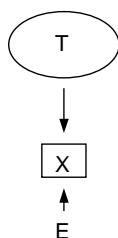
The concept of latent variables is based on classical test theory, which assumes that any measure is a function of two variables: the true score and error variation. This assertion can be written as a formula:

$$X = T + E$$

in which X represents the observed score on the measure, T is the person's true score, and E is error variation.

In the social sciences, we attempt to measure many phenomena that are not directly observable. The real variable or construct of interest is not precisely the one that is measured. A simple example is the measurement of an attitude, say about statistics. A response to a single items such as "Do you like statistics?" is a function of one's true attitude but also a function of other more transient factors such as the specific item wording, the respondents mood, or recent traumatic experiences with statistics. The true score, T , is the actual attitude, the observed score X is the expressed attitude on the question, and E is any factors that impact X other than T .

Notice that the classical test theory formula is also a regression formula. X is predicted by true score with some residual error remaining. In SEM, latent variables are designed to represent true scores. CFA models are visually represented in the following way:



Latent variables are represented by ellipses, and measured variables are represented by square boxes.

To the extent that there is measurement error¹ (i.e., the measure is not perfectly reliable) for a construct we are trying to measure, the remaining unaccounted for variance in X (represented by the E here) will be greater than zero. If we think about things in terms of variance components, that means that the variance of T will be less than the variance of X whenever measurement error is present. Now, think about what effect having a smaller or larger variance of X has on the correlation coefficient or the regression coefficient.

Unstandardized regression coefficient:

$$B_1 = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

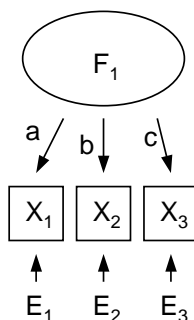
Standardized Regression Coefficient (Simple Regression):

$$r = \beta = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Measurement error will attenuate the correlation or standardized regression coefficient, and it will impact the regression coefficient if there is measurement error in the predictor. So, if we could remove the measurement error and then examine the associations among true(r) scores, we would find stronger bivariate associations.

¹ Although many SEM users equate E with measurement error, there are other sources that can contribute error that can contribute to the value of E , usually referred to as "unique variance." Unique variance may represent a systematic true source of variance rather than random error. Most of us are aware of this distinction but tend to be sloppy in our reference to E as being equal to measurement error.

Deriving Factor Loadings



We can use Wright's tracing rules to derive factor loadings. The correlation between X_1 and X_2 , r_{12} , should be equal to the product of ab , because we trace from X_1 to F_1 and back to X_2 . Similarly, bc will equal the correlation between X_2 and X_3 . To obtain the factor loadings for the above model, there are three equations:

$$r_{12} = ab$$

$$r_{23} = bc$$

$$r_{13} = ac$$

As long as we have values for r_{12} , r_{23} , and r_{13} , we can solve the equations for a , b , and c . Thus, there will be three equations and three unknowns. If we had just two variables loading on one factor, we would have two paths to estimate but only one correlation. That model is unsolvable.

If the number of unknowns is equal to the number of equations, the model is called *just identified*. If the number of unknowns is greater than the number of equations, the model is said to be *underidentified*, and there is no solution possible. An *overidentified* model is one in which there are fewer unknowns than equations. This is preferred.

Generally, the number of correlations among a set of variables can be described as:

$$\# \text{ correlations} = \frac{v(v-1)}{2}$$

where v is the number of variables. One can determine if the model is identified by calculating whether there are more correlation elements than paths to be estimated.² Thus, one formula for degrees of freedom for structural models is:

$$df = \frac{v(v-1)}{2} - p$$

where v is the number of measured variables in the model and p is the number of free parameters that need to be estimated (not including residual errors or disturbances).

Suggested Reading

Bollen, K.A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53, 605-34

² This formula is a convenient formula that works for many situations, but is not always appropriate. Some texts use a different formula, based on a count of the number of variance/covariance elements (i.e., the diagonal variance elements are counted), $df = \frac{v(v+1)}{2} - p$. In this formula, however, the number of parameters, p , must be counted differently by including the variances estimated in the model.