Latent Variables

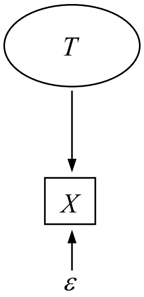
The concept of latent variables can be explained generally in terms of classical test theory,[1] which states that any measure is a function of two components: true score variation and error variation. This assertion can be written as a formula:

$$X = T + \varepsilon$$

in which X represents the observed score on the measure, T is the person's true score, and ε ("epsilon") is error variation.

In the social sciences, we attempt to measure many phenomena that are not directly observable (or, at least, *not directly observed*). The true 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 item like "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 respondent’s mood, recent traumatic experiences with statistics, or even random error. The true score, T, is the actual attitude, the observed score X is the expressed attitude on the question, and ε represents any factors that impact X other than T.

The concept of the latent variable from confirmatory factor analysis and structural equation modeling can be viewed in parallel to the classical test theory formulation. The latent variable is like a true score that is not directly observed, the observed variable is the measurement that is directly observed, and some degree of random measurement error may exist such that the observed score does not perfectly match the true scores. In these terms, we could say that the true score causes a portion of the observed score typically with some unaccounted variance remaining. This causal hypothesis suggests a regression or path model that can be graphically depicted with ellipses that stand for latent variables and square boxes that stand for measured variables.



Effects of Measurement Error

According to the classical test theory framework, when there is measurement error (i.e., the measure is not perfectly reliable) for a construct we are trying to measure, remaining unaccounted for variance in X (i.e., ε) will be greater than zero. If we think about things in terms of variance components, it suggests that the variance of T will be less than the variance of X whenever measurement error is present, evident in the following equation:

$$Var(X) = Var(T + \varepsilon) = Var(T) + Var(\varepsilon)$$

If there is no measurement error (i.e., $Var(\varepsilon) = 0$), then the variance of X and T will be equal. But when there is some measurement error, then the observed score variance, $Var(X)$, will be larger than the true score variance, $Var(T)$.

[1] I use a very informal definition of “latent variable” here and there are many definitions that can be used (see Bollen, 2002, for a more extensive discussion of previously proposed and possible definitions; and an update by Bollen & Hoyle, 2023). My goal is to provide an intuitive definition that points toward the advantages of using latent variables in structural models.

Remember that reliability is the absence of measurement error and can be expressed as a ratio of true score variance to observed variance, $Var(T)/[Var(T) + Var(e)] = Var(T)/Var(X)$. A measure has perfect reliability when this ratio is equal to 1.

Now, think about what effect having a smaller or larger variance of X or Y has on the correlation coefficient or the regression coefficient in the simple regression case.

Unstandardized regression coefficient:

$$B = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

Correlation and standardized regression coefficient:

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

Measurement error will attenuate the correlation or standardized regression coefficient, and it will impact the unstandardized regression coefficient if there is measurement error in the predictor.[2] The unstandardized regression coefficient is unaffected by the presence of measurement error in Y. The covariance, $Cov(X,Y) = \left[\sum(X-\bar{X})(Y-\bar{Y})\right]/(n-1)$, is also unaffected by measurement error in X or Y, because it is not divided by the variance of either variable, so it remains unbiased by measurement error.

Multiple regression. The effects of measurement error in the context of multiple regression is more complex. Assume two predictors in a multiple regression in which X_1 is measured without measurement error and X_2 is measured with measurement error. Because the correlation between X_1 and X_2 and between X_2 and Y will be attenuated due to the measurement error in X_2, the regression coefficient for X_1 will not fully remove the effects of X_2. So, if there is some measurement error in X_1 and X_2, the consequences are less predictable. The direction of the biases will be unknown, but results will be less accurate than if there was no measurement error.

Standard errors. Significance tests are also adversely affected by measurement error via effects on the standard error, which will reduce statistical power. The following formula illustrates that reduction of R^2 (resulting from measurement error on X or Y) and adding measurement error to Y will increase standard errors.

$$SE_B = \sqrt{\frac{1-R^2}{n-k-1}} \cdot \frac{sd_Y}{sd_X} = \sqrt{\frac{1-R^2}{n-k-1}} \cdot \frac{\sqrt{Var(Y)}}{\sqrt{Var(X)}}$$

Means. This is a good place to note that measurement error does not bias the mean, or the expected value of X. Below the expected values is denoted by $\mathrm{E}(.)$

$$\mathrm{E}(X) = \mathrm{E}(T+\varepsilon) = \mathrm{E}(T) + \mathrm{E}(\varepsilon)$$

Because the mean of random error, $\mathrm{E}(\varepsilon)$ is equal to 0, then

$$\mathrm{E}(X) = \mathrm{E}(T) + 0 = \mathrm{E}(T)$$

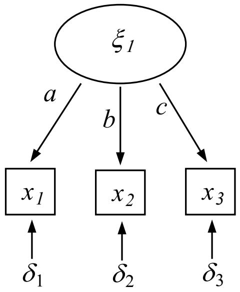
And the expected value of the observed score is the same as the expected value of the true score. Even though the mean is not biased by measurement error, measurement error will lead to lower power on significance tests because standard errors for tests of mean differences are impacted in the same way that for regression coefficient standard errors are impacted.

[2] Incidentally, this is related to the regression assumption is that X is fixed or measured without error.

So, in general, measurement error adversely affects regression or path coefficients. If there was a way to remove measurement error from the path model, then we could adjust for the effects of measurement error and mitigate its undesirable consequences. Structural equation modeling attempts to do just this by estimating path coefficients among latent variables.

Deriving Factor Loadings

In practice, we generally need three or more observed measures to estimate a latent variable.[3] To illustrate conceptually how this is done, we can use the same concepts of path analysis can be applied to estimation of a latent variable. Below is a figure for a latent variable with three measured variables (or sometimes "indicators" or “manifest variables”). The Greek letter ξ ("ksi") is usually used to represent the latent variable and the Greek letter δ ("delta") is usually used to represent the error. I switch from X to x for the measured variable to be consistent with the LISREL notation.



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 ξ_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 (usually) 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 said to be *just identified* (or sometimes, just “determined”). If the number of unknowns is greater than the number of equations, the model is said to be *underidentified*, and there is no single solution possible. An *overidentified* model is one in which there are fewer unknowns than equations. In practice, overidentified models are 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. Thus, one formula for degrees of freedom for structural models is:

[3] It is sometimes possible to estimate latent variables with only two indicators. If there is more than one latent variable in the model, the degrees of freedom from the full model may be positive and loading estimates can be obtained by use of all of the information in the model. This may work in some instances without using any special constraints, particularly with higher intercorrelations among the observed variables and when the sample size is larger. But, in other instances, the researcher may need to assume the two loadings are equal to one another. Although this is potentially a strong assumption, it may be preferable to combining the two items and assuming no measurement error.

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

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).[4]

Measurement Errors

Although many SEM users informally refer to each δ in a latent variable model as "measurement error," there are other sources that can contribute to the residual or unaccounted for variance. So, some authors refer to δ as "unique variance" or "uniqueness", because it represents unaccounted for variance, which may be random or systematic variance that is unique to that variable. In other words, it is variance that is not accounted for by the common factor, but may not necessarily be measurement error. For example, "I think math is difficult" may be an indicator of a latent variable representing a student's attitudes toward statistics, but there may be other factors, such as the number of prior math courses in school that impact only that indicator and not others. This would be systematic variance that is unique to this item if other items do not address math. "Measurement residual" is an alternative term that is perhaps a little less jargony than "uniqueness" perhaps with less connotation that δ is random measurement error. In one sense, unique variance is measurement error in that it represents variance in the observed variable that is not part of the underlying construct. Mathematically, the measurement residual is simply the variance in the observed variable unaccounted for by the factor, or (typically the square root of) $1 - R^2$ (if using standardized values for the residual).

Assumptions

There are several assumptions for latent variables, of which I will only mention a few here. The errors are assumed to be independent (but this can be relaxed later), and, like regression analysis, the latent variable variance is assumed to independent from the measurement residual variance. It is also necessary to make a constraint to have a metric for the latent variable variance. Typically, this is done by setting a loading equal to 1 or setting the factor variance equal to 1. We will talk more about alternative ways to define the latent variable variance later. Additionally, valid latent variable estimation also assumes that the model is correct. For example, if two constructs actually underlie a set of observed variables, then our use of a single latent variable will be incorrect and will lead to inaccurate loadings and other results. Relatedly, a latent variable only represents the domain of the construct as defined by the set observed variables used to estimate it. So, a latent variable’s content validity (breadth or narrowness) is tied to the range of observed variables used to define it.

General Comments

The concept of latent variables is very useful for investigating the internal reliability and the factor structure of a measure. Confirmatory factor analysis is often an end goal of the research, with no immediate goal of adjusting for measurement error in a structural model. The simple example above is a single-factor model, but one can conceptualize and test two, three, or more factors. The ability to estimate latent variables provides an important advantage of structural equation models over regression analysis. Latent variables allow us to estimate measurement error and we can attempt to remove it from the estimates of the relations among variables, which will produce less biased estimates of causal and non-causal relations.

Further Reading

Bollen, K.A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53, 605-634.
Bollen, K.A., & Hoyle, R.H. (2023). Latent Variables in Structural Equation Modeling. In R.H. Hoyle, *Handbook of structural equation modeling*, second edition (pp. 97-109). Guilford.
Brown, T. A. (2014). *Confirmatory factor analysis for applied research, second edition*. New York: Guilford Press.

[4] I use this formula because I am assuming correlations rather than covariances. This formula is more convenient formula and works for many situations. A different formula, based on a count of the number of variance/covariance elements (i.e., the diagonal variance elements are counted), $df = \left[v(v+1)/2\right] - p$ **,** is more commonly used. If using the $v+1$ formula, however, the number of parameters, p, must be counted differently where the variances estimated in the model are included.