Electronic Reprint of:

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In Brian Everitt and David C. Howell, *Encyclopedia of Statistics in Behavioral Science*, (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd.

To find out about the Encyclopedia please see:

http://www.wiley.com//legacy/wileychi/eosbs/

Structural Equation Modeling and Test Validation

BRUNO D. ZUMBO

Volume 4, pp. 1951 – 1958

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling and Test Validation

Ideally, test and measurement validation entails theoretical as well as empirical studies (*see* **Validity Theory and Applications**). Moreover, the term validation implies a process that takes place over time, often in a sequentially articulated fashion. The choice of statistical methods and research methodology for empirical data analyses is of course central to the viability of validation studies. The purpose of this entry is to describe developments in test and measurement validation as well as an important advancement in the statistical methods used in test validation research, structural equation modeling. In particular, a generalized linear **structural equation model** (GLISEM) that is a **latent variable** extension of a **generalized linear model** (GLIM) is introduced and shown to be particularly useful as a statistical methodology for test and measurement validation research.

A Brief Overview of Current Thinking in Test Validation

Measurement or test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals from a given sample and in a given context. The concept, method, and process of validation are central to constructing and evaluating measures used in the social, behavioral, health, and human sciences, for without validation, any inferences made from a measure are potentially meaningless.

The above definition highlights two central features in current thinking about validation. First, it is not the measure per se that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from scores obtained from a scale may appear subtle at first blush but, in fact, it has significant implications for measurement and testing because it highlights that the validity of the inferences one makes from test scores is somewhat bounded by place, time, and use of the scores resulting from a measurement operation.

The second central feature in the above definition is the clear statement that inferences made from all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an 'objective' educational, economic, or health indicator such as number of students finishing grade 12, or a more psychological measure such as a self-report depression measure, one must be concerned with the validity of the inferences.

It is instructive to contrast contemporary thinking in **validity theory** with what is commonly seen in many introductory texts in research methodology in the social, behavioral, health, and human sciences.

The Traditional View of Validity

The traditional view of validity focuses on (a) validity as a property of the measurement tool, (b) a measure is either valid or invalid, various types of validity – usually four – with the test user, evaluator, or researcher typically assuming only one of the four types is needed to have demonstrated validity, (c) validity as defined by a set of statistical methodologies, such as correlation with a gold-standard, and (d) reliability is a necessary, but not sufficient, condition for validity.

The traditional view of validity can be summarized in Table 1.

The process of validation then simply portrayed as picking the most suitable strategy from Table 1 and conducting the statistical analyses. The basis for much validation research is often described as a correlation with the 'gold standard'; this correlation is commonly referred to as a validity coefficient.

The Contemporary View of Validity

Several papers are available that describe important current developments in validity theory [4, 5, 9, 12, 13, 20]. The purpose of the contemporary view of validity, as it has evolved over the last two decades, is to expand upon the conceptual framework and power of the traditional view of validity seen in most introductory methodology texts. In brief, the recent history of validity theory is perhaps best captured by the following observations.

Table 1 The traditional categories of validity

Type of validity	What does one do to show this type of validity?
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure with the criterion measure obtained in the present
B. Predictive	Select a criterion and correlate the measure with the criterion measure obtained in the future
Construct (A. Convergent and B. Discriminant):	Can be done several different ways. Some common ones are (a) correlate to a 'gold standard', (b) factor analysis, (c) multitrait multimethod approaches

1. Validity is no longer a property of the measurement tool but rather of the inferences made from the scores.
2. Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum.
3. Construct validity is the central most important feature of validity.
4. There are no longer various types of validity but rather different sources of evidence that can be gathered to aid in demonstrating the validity of inferences.
5. Validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods.
6. As one can see in Zumbo's [20] volume, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall validity framework for evaluating test interpretation and use came the consideration of the role of ethical and social consequences as validity evidence contributing to score meaning. This movement has been met with some resistance. In the end, Messick [14] made the point most succinctly when he stated that one should not be simply concerned with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of 'normal' test use. The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results affecting educational financing and staffing, as well as health care outcomes and financing in the United States and Canada.
7. Although it was initially set aside in the move to elevate construct validity, content-based evidence is gaining momentum again in part due to the work of Sireci [19].
8. Of all the threats to valid inferences from test scores, test translation is growing in awareness due to the number of international efforts in testing and measurement (see, for example, [3]).
9. And finally, there is debate as to whether reliability is a necessary but not sufficient condition for validity; it seems that this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in their inferences. Specifically, reliability is a question of *data quality,* whereas validity is a question of *inferential quality.* Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter.

In a broad sense, then, validity is about evaluating the inferences made from a measure. All of the methods discussed in this encyclopedia (e.g., factor analysis, **reliability**, **item analysis**, **item response modeling**, **regression**, etc.) are directed at building the evidential basis for establishing valid inferences. There is, however, one class of methods that are particularly central to the validation process, **structural equation models**. These models are particularly important to test validation research because they are a marriage of regression, path analysis, and latent variable modeling (often called **factor analysis**). Given that the use of latent variable structural equation models presents one of the most exciting new developments with implications for validity theory, the next section discusses these models in detail.

Generalized Linear Structural Equation Modeling

In the framework of modern statistical theory, test validation research involves the analysis of covariance matrices among the observed empirical data that arise from a validation study using **covariance structure models**. There are two classes of models that are key to validation research: **confirmatory factor analysis** (CFA) (*see* **Factor Analysis: Confirmatory**) and multiple indicators multiple causes (MIMIC) models. The former have a long and rich history in validation research, whereas the latter are more novel and are representative of the merger of the structural equation modeling and item response theory traditions to what will be referred to as generalized linear structural equation models. Many very good examples and excellent texts describing CFA are widely available (e.g., [1, 2, 10]). MIMIC models are a relatively novel methodology with only heavily statistical descriptions available.

An Example to Motivate the Statistical Problem

Test validation with SEM will be described using the Center for Epidemiologic Studies Depression scale (CES-D) as an example. The CES-D is useful as a demonstration because it is commonly used in the life and social sciences. The CES-D is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. The CES-D prompts the respondent to reflect upon his/her last week and respond to questions such as 'My sleep was restless' using an ordered or Likert response format of 'not even one day', '1 to 2 days', '3 to 4 days', '5 to 7 days' during the last week. The items typically are scored from zero (not even one day) to three (5–7 days). Composite scores, therefore, range from 0 to 60, with higher scores indicating higher levels of depressive symptoms. The data presented herein is a subsample of a larger data set collected in northern British Columbia, Canada. As part of a larger survey, responses were obtained from 600 adults in the general population -290 females with an average age of 42 years with a range of 18 to 87 years, and 310 males with an average age of 46 years and a range of 17 to 82 years.

Of course, the composite scale score is not the phenomenon of depression, per se, but rather is related to depression such that a higher composite scale score reflects higher levels of the latent variable depression. Cast in this way, two central questions of test validation are of interest: (a) Given that the items are combined to created one scale score, do they measure just one latent variable? and (b) Are the age and gender of the respondents predictive of the latent variable score on the CES-D? The former question is motivated by psychometric necessities whereas the latter question is motivated by theoretical predictions.

CFA Models in Test Validation

The first validation question described above is addressed by using CFA. In the typical CFA model, the score obtained on each item is considered to be a linear function of a latent variable and a stochastic error term. Assuming p items and one latent variable, the linear relationship may be represented in matrix notation as

$$y = \Lambda\eta + \varepsilon, \tag{1}$$

where y is a $(p \times 1)$ column vector of continuous scores for person i on the p items, Λ is a $(p \times 1)$ column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, η is the latent variable score for person i, and ε is $(p \times 1)$ column vector of measurement residuals. It is then straightforward to show that for items that measure one latent variable, (1) implies the following equation:

$$\Sigma = \Lambda\Lambda' + \Psi, \tag{2}$$

where Σ is the $(p \times p)$ population covariance matrix among the items and Ψ is a $(p \times p)$ matrix of covariances among the measurement residuals or unique factors, Λ' is the transpose of Λ, and Λ is as defined above. In words, (2) tells us that the goal of CFA, like all factor analyses, is to account for the covariation among the items by some latent variables. In fact, it is this accounting for the observed covariation that is fundamental definition of a latent variable – that is, a latent variable is defined by local or conditional independence.

More generally, CFA models are members of a larger class of general linear structural models for a p-variate vector of variables in which the empirical data to be modeled consist of the $p \times p$ unstructured estimator, the sample covariance

matrix, S, of the population covariance matrix, $\sum$. A confirmatory factor model is specified by a vector of q unknown parameters, θ, which in turn may generate a covariance matrix, $\sum(\theta)$, for the model. Accordingly, there are various estimation methods such as generalized **least-squares** or **maximum likelihood** with their own criterion to yield an estimator $\hat{\theta}$ for the parameters, and a legion of test statistics that indicate the similarity between the estimated model and the population covariance matrix from which a sample has been drawn (i.e., $\sum = \sum(\theta)$). That is, formally, one is trying to ascertain whether the covariance matrix implied by the measurement model is the same as the observed covariance matrix,

$$S \cong \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi} = \Sigma(\hat{\theta}) = \hat{\Sigma}, \tag{3}$$

where the symbols above the Greek letters are meant to imply sample estimates of these population quantities.

As in regression, the goal of CFA is to minimize the error (in this case, the off-diagonal elements of the residual covariance matrix) and maximize the fit between the model and the data. Most current indices of model fit assess how well the model reproduces the observed covariance matrix.

In the example with the CES-D, a CFA model with one latent variable was specified and tested using a recent version of the software LISREL (*see* **Structural Equation Modeling: Software**). Because the CES-D items are ordinal (and hence not continuous) in nature (in our case a four-point response scale) a polychoric covariance matrix was used as input for the analyses. Using a polychoric matrix is an underlying variable approach to modeling ordinal data (as opposed to an item response theory approach). For a polychoric correlation matrix (*see* **Polychoric Correlation**), an underlying continuum for the polytomous scores is assumed and the observed responses are considered manifestations of respondents exceeding a certain number of latent thresholds on that underlying continuum. Conceptually, the idea is to estimate the latent thresholds and model the observed cross-classification of response categories via the underlying latent continuous variables. Formally, for item j with response categories $c = 0, 1, 2, \ldots, C-1$, define the latent variable y* such that

$$y_j = c \text{ if } \tau_c < y_j^* < \tau_{c+1}, \tag{4}$$

where τ_c, τ_{c+1} are the latent thresholds on the underlying latent continuum, which are typically spaced at nonequal intervals and satisfy the constraint $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{C-1} < \tau_C = \infty$. It is worth mentioning at this point that the latent distribution does not necessarily have to be normally distributed, although it commonly is due to its well understood nature and beneficial mathematical properties, and that one should be willing to believe that this model with an underlying latent dimension is actually realistic for the data at hand.

Suffice it to say that an examination of the fit indices for our example data with the CES-D, such as the root mean-squared error of approximation (RMSEA), a measure of model fit, showed that the one latent variable model was considered adequate, RMSEA = 0.069, with a 90% confidence interval for RMSEA of 0.063 to 0.074.

The single population CFA model, as described above, has been generalized to allow one to test the same model simultaneously across several populations. This is a particularly useful statistical strategy if one wants to ascertain whether their measurement instrument is functioning the same away in subpopulations of participants (e.g., if a measure functioning the same for males and females). This multigroup CFA operates with the same statistical engine described above with the exception of taking advantage of the statistical capacity of partitioning a likelihood ratio Chi-square and hence testing a series of nested models for a variety of tests of scale level measurement invariance (see [1], for details).

MIMIC Models in Test Validation

The second validation question described above (i.e., are age and gender predictive of CES-D scale scores?) is often addressed by using ordinary least-squares **regression** by regressing the observed composite score of the CES-D onto age and the dummy coded gender variables. The problem with this approach is that the regression results are biased by the measurement error in the observed composite score. Although widely known among psychometricians and statisticians, this bias is ignored in a lot of day-to-day validation research.

The more optimal statistical analysis than using OLS regression is to use SEM and MIMIC models. MIMIC models were first described by Jöreskog

and Goldberger [7]. MIMIC models, in their essence, posit a model stating that a set of possible observed explanatory variables (sometimes called *predictors or covariates*) affects latent variables, which are themselves indicated by other observed variables. In our example of the CES-D, the age and gender variables are predictors of the CES-D latent variable, which itself is indicated by the 20 CES-D items. Our example highlights an important distinction between the original MIMIC models discussed over the last three decades and the most recent developments in MIMIC methodology – in the original MIMIC work the indicators of the latent variable(s) were all continuous variables. In our case, the indicators for the CES-D latent variables (i.e., the CES-D items) are ordinal or Likert variables. This complicates the MIMIC modeling substantially and, until relatively recently, was a major impediment to using MIMIC models in validation research.

The recent MIMIC model for ordinal indicator variables is, in short, an example of the merging of statistical ideas in generalized linear models (e.g., logit and probit models) and structural equation modeling into a generalized linear structural modeling framework [6, 8, 16, 17, 18]. This new framework builds on the correspondence between factor analytic models and **item response theory** (IRT) models (see, e.g., [11]) and is a very general class of models that allow one to estimate group differences, investigate predictors, easily compute IRT with multiple latent variables (i.e., multidimensional IRT), investigate differential item functioning, and easily model complex data structures involving complex item and test formats such as testlets, item bundles, test method effects, or correlated errors all with relatively short scales, such as the CES-D.

A recent paper by Moustaki, Jöreskog, and Mavridis [15] provides much of the technical detail for the generalized linear structural equation modeling framework discussed in this entry; therefore, I will provide only a sketch of the statistical approach to motivate the example with the CES-D. In this light, it should be noted that these models can be fit with either Mplus or PRELIS-LISREL. I chose to use the PRELIS-LISREL software, and hence my description of the generalized linear structural equation model will use Jöreskog's notation.

To write a general model allowing for predictors of the observed (manifest) and latent variables, one extends (1) with a new matrix that contains the predictors x

$$y^* = \Lambda z + Bx + u, \text{ where}$$
$$z = Dw + \delta, \tag{5}$$

and u is an error term representing a specific factor and measurement error and y^* is an unobserved continuous variable underlying the observed ordinal variable denoted y, z is a vector of latent variables, w is a vector of fixed predictors (also called *covariates*), D is a matrix of regression coefficients and δ is a vector of error terms which follows a $N(0, I)$. Recall that in (1) the variable being modeled is directly observed (and assumed to be continuous), but in (5) it is not.

Note that because the PRELIS-LISREL approach does not specify a model for the complete p-dimensional response pattern observed in the data, one needs to estimate the model in (5) with PRELIS-LISREL one follows two steps. In the first step (the PRELIS step), one models the univariate and bivariate marginal distributions to estimate the thresholds and the joint covariance matrix of y^*, x, and w and their asymptotic covariance matrix. In the PRELIS step there is no latent variable imposed on the estimated joint covariance matrix hence making that matrix an unconstrained covariance matrix that is just like a sample covariance matrix, S, in (3) above for continuous variables. It can therefore be used in LISREL for modeling just as if y^* was directly observed using (robust) maximum likelihood or weighted least-squares estimation methods.

Turning to the CES-D example, the validity researcher is interested in the question of whether age and gender are predictive of CES-D scale scores. Figure 1 is the resulting generalized MIMIC model. One can see in Figure 1 that the correlation of age and gender is, as expected from descriptive statistics of age for each gender, negative. Likewise, if one were to examine the t values in the LISREL output, both the age and gender predictors are statistically significant. Given the female respondents are coded 1 in the binary gender variable, as a group the female respondents scored higher on the latent variable of depression. Likewise, the older respondents tended to have a lower level of depression compared to the younger respondents in this sample, as reflected in the negative regression coefficient in Figure 1. When the predictive relationship of age was investigated separately for males and females via

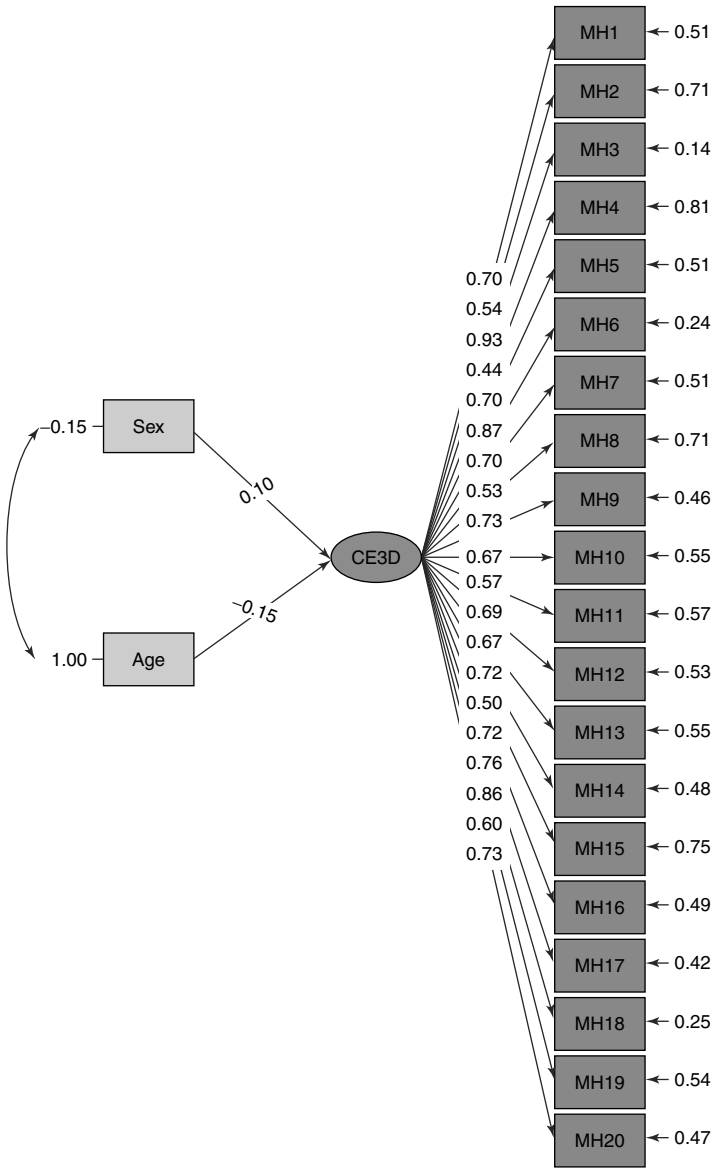

Figure 1 MIMIC model of age and gender for the CES-D (Standardized solution)

this generalized MIMIC model, age was a statistically significant (negative) predictor for the female respondents and age was not a statistically significant for male respondents. Age is unrelated to depression level for men, whereas older women in this sample are less depressed than younger women. This sort of predictive validity information is useful to researchers using the CES-D and hence supports, as described at the beginning of this entry, the inferences made from CES-D test scores.

References

[1] Byrne, B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*, Lawrence Erlbaum, Hillsdale, N.J.

[2] Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, Lawrence Erlbaum, Mahwah.

[3] Hambleton, R.K. & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, Netherlands, pp. 153–171.

[4] Hubley, A.M. & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going, *The Journal of General Psychology* **123**, 207–215.

[5] Johnson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices, *Educational and Psychological Measurement* **58**, 736–753.

[6] Jöreskog, K.G. (2002). Structural equation modeling with ordinal variables using LISREL. Retrieved December 2002 from `http://www.ssicentral.com/lisrel/ordinal.htm`

[7] Jöreskog, K.G. & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable, *Journal of the American Statistical Association* **10**, 631–639.

[8] Jöreskog, K.G. & Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches, *Multivariate Behavioral Research* **36**, 341–387.

[9] Kane, M.T. (2001). Current concerns in validity theory, *Journal of Educational Measurement* **38**, 319–342.

[10] Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, Newbury Park.

[11] Lu, I.R.R., Thomas, D.R. & Zumbo, B.D. (in press). Embedding IRT in structural equation models: A comparison with regression based on IRT scores, *Structural Equation Modeling*.

[12] Messick, S. (1989). Validity, in *Educational Measurement*, R.L. Linn, ed., 3rd Edition, American Council on Education/Macmillan, New York, pp. 13–103.

[13] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *American Psychologist* **50**, 741–749.

[14] Messick, S. (1998). Test validity: A matter of consequence, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 35–44.

[15] Moustaki, I., Jöreskog, K.G. & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: a comparison of LISREL and IRT approaches, *Structural Equation Modeling* **11**, 487–513.

[16] Muthen, B.O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables, *Journal of Educational Statistics* **10**, 121–132.

[17] Muthen, B.O. (1988). Some uses of structural equation modeling in validity studies: extending IRT to external variables, in *Test validity*, H. Wainer & H. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 213–238.

[18] Muthen, B.O. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika* **54**, 551–585.

[19] Sireci, S.G. (1998). The construct of content validity, in *Validity Theory and the Methods used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 83–117.

[20] Zumbo, B.D., ed. (1998). Validity theory and the methods used in validation: perspectives from the social and behavioral sciences, Special issue of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* **45**(1–3), 1–359.

BRUNO D. ZUMBO