Committee II
Theoretical Empiricism: A General
Rationale for Scientific Model-Building

Second Draft --
for Conference Distribution Only

THE BASIC PRINCIPLES OF MODEL BUILDING

by

Jan-Bernd Lohmoeller
Assistant Professor
Center for Social Research
Free University of Berlin
West Germany

The Thirteenth International Conference on the Unity of the Sciences
Washington, D.C. September 2-5, 1984

1/2

Sc.1 Wold"s model for knowledge.

Scientific knowledge, as distinguished from every-day knowledge, is characterized by the clear distinction between theoretical and empirical aspects. Because the distrinction between the two cannnot be part of the theoretical or the empirical, as a third element of the model for scientific knowledge the frame of reference has to be introduced. The frame of reference, formulated (more ore less) in every-day language, contains a mixture of theoretical ("T") and empirical ("E") contents, and is shown as a rectangle in the small graph that is taken from Wold (1969):

Eq.01

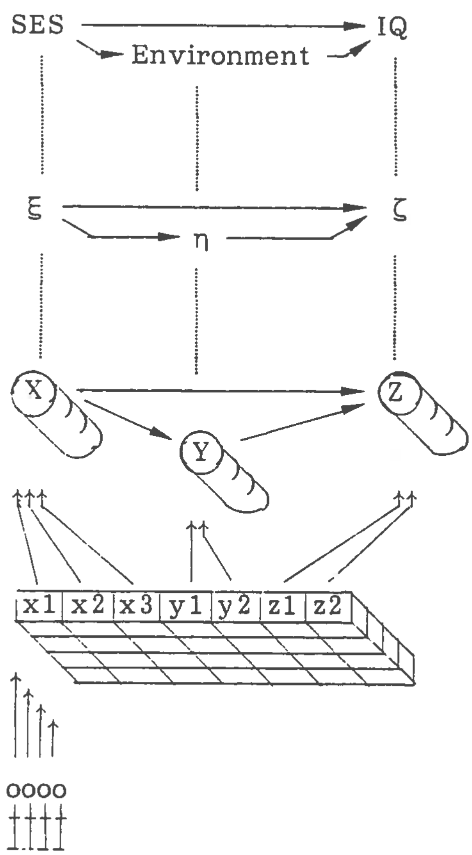
$$\boxed{T \Leftrightarrow E}$$

Within the frame of reference, T and E are kept apart in order to prove that they match. Matching (denoted by the double-headed arrow) means that conclusions can be drawn from T about E (deduction) and from E about T (induction). The process of matching E and T is the model validation an can be understood as a general description of the job of science (Cronbach & Meehl 1955, Wold 1969, Bentler 1978). Matching E and T may involve a reduction of the range of observations E explained by the theory, which is adverse to the basic esthetical qualities of theories: A theory should be as simple, elegant, consistent and general as possible.

Sc.11 Levels of theory and data

The process of matching becomes more transparent and controllable when the rigid distinction between E, T, and frame of reference is relaxed in favour of

Figure 1
Matching between different levels of theorical and empirical knowledge.

	Diagram	Level	Elements	Statements
T1	 SES Environment IQ ξ η ζ X Y Z x1 x2 x3 y1 y2 z1 z2 oooo	substantive theory	concepts, hypotheses	causal, functional, correlational
T2		mathematical model	random variables	functional correlational
E3		aggregated data	compound variables, (numbers)	correlational (functional)
E2		data	numbers	correlational
E1		observations	observational units	verbal

more levels. Five levels T1, T2, E3, E2, E1 may be distinguished, with a matching process between each pair of neighouring levels, as shown in Figure 1. Each level has its own language, with its own primitive elements, syntax and semantics.

T1 is the level of substantive theory. The elements of the theory are concepts (constructs). The syntax which decides which elements are put together to a correct statement is the syntax of a natural language like German or English. The semantics of a statement is determined within the frame of reference of the whole model, in which also the concepts are defined and limited to a meaning which may be different from every-day meaning. What is depicted in the top of Fg.01 is the theoretical statement that the socioeconomic status (SES) of the parents has a direct influence on the intellectual abilities (IQ) of the children, as well as an indirect influence which is meditated by the learning environment at home.

T2 is the level of a mathematical-statistical model. The elements are random variables, here denoted by ξ,η,ζ. The syntax is the algebra of expectations and linear algebra. The meaning of model T2 is given by the correspondence to the substantive theory T1. When limited to path models with latent variables, the elements can be inner and outer variables, inner and outer residuals, and the syntax controls how to formulate correct equations.

E1 is the level of observations. The elements are observational units like individuals, or countries, or time points. The statements on this level are natural-language statements about the behavior of the units. **E2** is the level of data. The elements are real numbers. The mapping of behaviour on numbers is called measurement. **E3** is the level of functions of the data. The elements are real numbers, again, but they are formed as functions of the measured data. This can include compound variables, or estimation functions for parameters.

The five levels (the number five has some arbitrariness) form a chain, and each member has to link up with and to adjust to its neighbour. The mathematical model T2 must represent as completely as possible the substantive theory T1. The statistical function on the level E3 must be determined from the data on E2 so as to be best estimators of the model variables on T2. The data E2 must be gathered from the level E1 so as to be informative for a comprehensive estimation (on E3) of the latent variables (on T2) which stand for the constructs (on level T1).

Residuals. The matching of E and T is never perfect, and there remain unexplained parts on both sides. The deductive specialization of Eq.01 shows the empirical content as being partly a function of the theoretical content and partly unexplainable by the theory. The formal notation is:

Eq.02a $E = E(T) + \varepsilon$,

and the notation in plain words:

Eq.02b data = systematic part + residual part

Eq.02c data = fit + rest

The right-hand side of Eq.02 is often called the model, in a narrower sense. The residual ε may be interpreted as measurement error, prediction error, sample fluctuation, or "systematic" variation which is left out from the systematic part because it related only to small and specific parts of the observation (unique variation, specification error).

The inductive specialization of Eq.01 shows the theoretical content as being partly a function of the observation and partly unobserved:

Eq.03 $T = T(E) + \delta$

The residual δ may be interpreted as theoretical surplus or empirical lack of the model. If, for example, T is the theoretical concept of intelligence, and E is an intelligence test, than T(E) is the IQ and δ is that what the test fails to measure, which can be understood as the lack of validity of the test.

In the process of model building the researcher tries to minimize ε and δ, to extend E, and to simplify T.

The notion of model is found to be used with different ranges. Some scientists set the concept of model identical to what here is called T, or E(T); some call the left-hand side of Eq.02 the model which then is contrasted to the data; some call Eq.02 together with Eq.03 a model. Wold's notion of model goes beyond this in that he includes the frame of reference into the notion of a model.

Sc.12 Causality and latent variables

Time, space and causality are the basic categories for our understanding of the world. The development of science in the past centuries has not undercut the the usage of the concepts of time, space and causality in every-day language, and even scientists use these concepts. Geographers use the concept of space as continuous and three-dimensional, even if physicists talk of bent space, and historian would not change their notion of time, if theoretical physistics would tell them that time runs unsteadily, jerkily, or even backwards. Empirical observations have lead physicist to extend the concepts of time and space.

Causality. The notion of causality has taken a different history. Assaults on the concept of causality came from philosophical consideration, not from empirical findings, and it came in form of an anathema and prohibition sign. Basically, the rejection of the concept of causality came from William of Occam's principle of parsimony: *Pluralitas non est ponenda sine necessicate*. Taking the philosophical objections into consideration, the usage of causal notion has been reestablished in the fifties by Lazarsfeld, Wold and Simon (cf. Bernert 1983), and it was mainly for practical reasons that they introduced causal terminology. A relation X=f(Y) does not become better or richer or more powerful if a causal interpretation is added, but more understandable. Whether the notion of cause-effect relationship can be applied to the relation of latent (LVs) and manifest variables (MVs), will be considered in this chapter.

Cause and effect are two different things. No thing can be the cause of itself. When I "compel myself" to write a paper, I have introduced the distinction between my willing mind and my weak flesh, two different things. Whether mind and body really are different entities, is a question left to philosophers.

By having more than just the two levels of latent and manifest variables, the problems to be addressed in this chapter become more clear. A model with several hierarchical levels of LVs is presented by Noonan (this volume). He distinguishes two types of relationships between LVs, the hierarchical and the causal-predictive relationships. Are the hierarchical relations causal or not, are the variables involved in a hierarchical relationship "different things", i.e. "things" at all?

Theoretical constructs. Variables, whether latent or manifest, are creatures of the human mind, constructions which are found helpful to order the chaos of sensational impressions. Some of these constructions are merely mathematical constructions and can be removed from a model without loss of predictive power. Others can not be removed in this sense, and these variables are called theoretical constructs (Maccorquodale & Meehl 1948, Falter 1977, Falter & Lohmöller 1982). As an example may serve a canonical correlation model for ten manifest predictors $\mathbf{x}$ and ten manifest predictands $\mathbf{y}$, which are transformed into ten latent predictors $\boldsymbol{\xi}$ and ten latent predictands $\boldsymbol{\eta}$. For the sake of argument, assume that two canonical correlation coefficients are of remarkable size, and the other eight correlations are neglegible. Then the eight last dimensions of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ can be retained to ease the algebraic treatment of the model, but can also be removed without loss of predictive power. The first two dimensions, however, can not be removed, and these two dimensions are empirical constructs or even theoretical constructs, if they can be interpreted and named in the framework of a substantive theory. Latent variables which according to this criterion are not theoretical constructs are not thus different from their indicators that they are "different things" and that they can enter a cause-effect relationship between LVs and MVs.

Dispositional terms. The researcher has to decide which one is the "real thing", the MV or the LV, or the LV of which level of a hierarchy. In the social sciences the latent variables often denote dispositions. Examples for dispositions of individuals are intelligence, extraversion, party identification, anomia. The manifest behavior then can be understood as being caused --*inter alii*-- by the disposition, i.e. solving a cognitive task requires intelligence, adressing unkown people requires some sort of extraversion, etc. In case the LVs are theoretical constructs and dispositional terms it seems appropriate to apply the realistic (not a ontological) interpretation to the LVs and to interpret the LVs as causes of the MVs.

No question, if the LVs can be understood as causes of the MVs, they can also be understood as causes of each other, and a causal interpretation can be applied to the inner part of a LV path model. This, however, becomes difficult if a LV is not a theoretical construct, but merely an intervening variable, a transformation or collection of MVs, a basket full of candidates for a cause-effect relationship, a mixed bag of MVs whit suspected explanaotry power, a typical case where PLS mode B weight estimation is advised.

Sc.2 Steps of model building

The process of model building can be described as a three-step procedure, compromising (i) the model specification, (ii) estimation of unknowns, and (iii) model evaluation.

Model specification includes at least two decisions, the first about the empirical phenomena to be explained, and the second about the theoretical form of the explanation. The specification of the theoretical content of the model may be more or less rigorous. We will focus on models which require a statistical treatment, i.e. which are on the one hand specified so far that a formal treatment is possible, but on the other hand are non-deterministic. The elements of a model of this type may include in the theoretical part manifest and latent variables, which are specified with respect to the total distribution or only to the conditional expectations.

Model estimation may require additional assumptions which have no counterpart in the first, substantive-oriented level of the theoretical model part. Estimation methods like least squares and maximum likelihood can be characterized in terms of robustness, availability, and precision of resulting estimates.

Model evaluation, in its classical form, requires a sample of independent observations on a completely specified distribution. The evaluation may involve the model as a whole or single parameters (standard errors). In Sc.3 less demanding evaluation methods will be discussed.

Sc.21 Specification of the data

Observational units. The phenomena to be explained by the model can be the data as observed, say $x_{k,n}$ for k=1..K attributes of n=1..N observational units, or an aggregation of the data, say the covariances s_{kl}, k,l=1..K. This choice includes a decision about the character of the observational units: If the observational units are considered to be genuinely different, with interpretable individual differences, like the children in a classroom (the one child is known to be the *primus* and the other the clown) and the years in an economic time series (the one year is known to be the oil crisis and the other the "1968 cultural revolution"), then the observational units must be specified to be part of the phenomenon under exploration. If, however, the observational units are considered to be replication of one and the same experiment, without any specific distinctions or individual differences, all being identically distributed, like the repeated throw of a die, then one can sum over the observational units without losing any information of substantive interest.

Cases vs. replications. In the first type of data, the observational units are called cases, they are specified by the model builder, and the model must provide certain unkowns or incidental parameters or "factor scores" accounting for the individual differences; in the second type the units are mere replications, anonymous sample points, unspecified. The distinction between "cases" and "replications" introduced in this way points, of course, to the extremes of a dimension with several intermediate steps of more or less specified, less or more randomly chosen observational units.

Applied vs. general science. The distinction between cases and replications corresponds to two types of science, applied and general. A problem of general psychology, for example, is the dependence of school success on intelligence, which is stated so as to apply to specified population and not to a specific individual. The corresponding problem of applied psychology is the prediction of school success of the *primus* and the classroom-clown when their IQs are known to be 120 and 100 points, respectively. The individual-psychology problem presumes that --on the level of general psychology-- it has been established that a relation between intelligence and school success exists. Hence, the individual psychology problem includes as a subproblem the general psychology problem, and not the other way round.

Example. In order to demonstrate the consequences of the distinction between cases and replications, applied and general science, the statistical methods of principal component analysis and factor analysis may serve as examples. The not-completely-specified form of the linear model which is common for both methods is:

Eq.21 $$x_k = \sum_j \pi_{jk} \xi_j + \varepsilon_k \quad ,$$

where the ξ_j denote latent variables, the ε_k residual variables, and the coefficients π_{jk} the so-called loadings. The index n for observational units does not occur in Eq.21 which demonstrates that it represents a general, not an applied model. Additional specifications on the first and second moments of the right--hand variables in Eq.21 lead to the common factor model. The model Eq.21 is re--stated on the individual level:

Eq.22 $$x_{kn} = \sum_j \ddot{\pi}_{jk} \xi_{jn} + \varepsilon_{kn}$$

Now there are two sorts of unknowns on the right-hand side of Eq.22, the parameters π_{jk} as before which account for the structural relations between the variables, and the unknowns which carry the index n, the LV scores ξ_{jn} and the residual scores ε_{kn}. Knowing only the distribution and the moments of these variables instead of the the scores, one would be unable to reconstruct the observed values x_{kn}.

The comparison of Eq.21 and Eq.22 demonstrates that more fundamental than the specification of the model is the specification of the data, the left-hand side of both equations. Indices which happen to appear on the data-side of the equation have also to show up on the right-hand side, the model side. If the observational units are considered to be more than replications, the researcher has to specify them by including the index n on the left-hand side, and consequently he has to put up a systematic model part on the right-hand side which includes also the case index.

The problem of specifying the data becomes more complex -- and perhaps even more clear -- when there are three sets of indices, say units, time points and attributes. Then two index sets may be considered as specified and knwon, and one as random and unspecified. For more details see Sc.4.

Sc.3 Model evaluation

Model evaluation should make use of not more than what as been assumed when specifying the model and the data. The most critical assumption that leads to a parting of the way is related to the distribution of the variables. If the distribution has been fully specified, maximum likelihood estimates of the model parameters can be obtained, and likelihood ration tests can be performed, at least for some simple models. If no distributions are made for the estimation of the model, it is nonsense to intoduce these for the model evaluation. From an array of distribution-free methods we name blindfolding, bootstrapping, jackknifing and perturbation. The methods are different with respect to their assumption and their sensibility. The critical assumptions are independance of observations, identical but unspecified distributions, and known zero point of scale of variables.

Sc.3.1 Distribution-free evaluation methods

Perturbation. Suppose a set of parameters P has been estimated from a data set X, and the question is how strong small changes in X influence the consequent changes in P. Let s(X) and s(P) denotes the Euclidean lengths, $s(X)=\sqrt{\sum_i x_i^2}$, and $e_i \sim N(0,s(X))$. A new "perturbated" data set is created by adding a random error, $x_i^*=x_i+\alpha e_i$, where α is a small number (say one percent) and the new data set has the variation $s(X^*) = (1+\alpha)s(X)$, and a new set of parameters P^* is estimated from X^*. If the ratio $s(P-P^*)/s(P)$ is close to α, this is an indication of stable results. Belsley (1984) reports a small artificial data set where a 1% change in the data produced a 40% change in the parameters, clearly a demonstration of unreliable results. As Belsley points out, the critical assumption is wether the data are treated as raw values or as deviations from their means. If the means are removed, results become smooth.

Bootstrapping implies the assumption that the residuals are independent. If they are independent, they can be exchanged without disturbing the estimates. Let be given a time series y_t, t=1950..1980, and the model $y_t=\beta y_{t-1}+\varepsilon_t$. Now a new time series x_t^* is created by exchanging residuals, say $y_{1961}^* = by_{1960}+e_{1971}$ and $y_{1971}^* = by_{1970}+e_{1961}$, or by a random exchange of residuals. Than the model is reestimated on the new time series y_t^*, giving a new estimate b* and new residuals e_t^*. The redistribution of resiudals and reestimation of the model is continued until a stabil estimate of β is established. Clearly, the bootstrapping method requires the assumption of independence of residuals.

Blindfolding means to omit one part of the data matrix while estimating the parameters from the remaining data, and then to reconstruct the omitted part by the estimated parameters. This procedure of omitting and reconstructing is repeated, until each data point is omitted and reconstructed once. The blindfolding technique provides two types of results, (a) the generalized crossvalidation measures as an evaluation of the model as a whole, and (b) the jackknife standard errors for the single parameter estimates. Both types of results are helpful in deciding on the quality and relevance of a model. The blindfolding technique requires no distributional assumptions, so it fits the PLS technique "like hand in glove" (Wold 1981, Wold and Apel 1982).

The jackknife technique was developed to construct the distribution of parameter estimates without assumptions on the distribution of the variables involved (Quenouille, Tukey). This is done by estimating the parameters N times in a data set with N observations, each time cutting off just one observation. The N estimates for the same parameter, then, are used to compute the mean, the standard deviation, and other distributional characteristics of the parameters.

The generalized crossvalidation measures indicate how well the observed values can be reconstructed by the model and its parameters. Standard crossvalidation utilizes one data set to estimate the parameters and another data set to test the validity of the estimates. For example, regression coefficients and R_1^2 are estimated in the first data set; then the regression coefficients are applied to the second data set, and the squared correlation $R_2^2(y;\hat{y})$ is computed, and usually the shrinkage phenomenon $R_2^2<R_1^2$ is observed (see Stone 1974, and Winteler 1983 for an counterexample). In generalized crossvalidation, however, the blindfolding technique is used to split the data set at hand repeatedly into an estimation set and a test set which may contain a single data point only.

PRE - Proportional reduction of error. A general principle for the construction of measures of predictive power is based on the proportional reduction of error (Guttman 1941, Goodman & Kruskal 1954). This principle implies the comparison of errors made under two prediction rules. The rule1 prediction (often called the trivial prediction) is based on the distribution of the predictand y alone, without any knowledge of the predictors $\mathbf{x}$. The rule1 prediction for continous variables is the mean $\bar{y}$ or the jackknifed mean $\bar{y}_{(-i)}$. The rule2 prediction is based on the joint distribution of $\mathbf{x}$ and y. The definition of the error depends on the scale quality of the variables. If y is a categorical variable, the error is simply the number of misclassifications. If y is continuous, the square sum of the differences between observed and predicted values is an appropriate error term, $\sum_n(y_n-\hat{y}_n)^2$. The standard formula for PRE measures is:

Eq.31 $$\text{PRE} = 1 - (\text{rule2 error}) \,/\, (\text{rule1 error})$$

For the (descriptive) multiple regression model, the rule1 prediction is the mean of y, and the rule1 error is the variance of y; the rule2 prediction is $\hat{y}$, and the rule2 error is the variance of the residual variable e; and the PRE measure is identical to the squared multiple correlation.

Eq.32 $$\text{PRE} = 1 - s_e^2/s_y^2 = (s_y^2 - s_e^2)/s_y^2 = \text{var}(\hat{y})/\text{var}(y) = R^2$$

Because var(y) is the observed variance and var(e) is the error variance in the regression model, the PRE coefficient is noted shortly as 1-E/O.

SG test. If both prediction rules include the blindfolding device, the PRE measures belong to the realm of methods proposed by Stone (1974) and Geisser (1974). The title of Stone's article with translation: Cross-validatory (= using blindfolding) choice (=estimation) and assessment (=testing) of statistical predictions. Stone proposes to use the deletion procedure for both the estimation of the unkowns of a model and the test of the predictive validity of that model. Geisser applies his "Predictive Sample Reuse Methode" only for the estimation, but does it by deleting more than one case at time.

If with rule2 all of the errors made under rule1 can be eliminated, than PRE=1, and the prediction rule (prediction model) is valid. PRE=0 indicates that rule2 has no relevance for the data at hand and that rule2 is no improvement over rule1. If PRE is negative, the non-blindfolded parts of the data matrix are misleading when guessing the blindfolded parts, and in general the rule2 is misleading for the prediction of the data. This can happen when the parameter estimates for the rule2 are unstable or when the data set is not homogenous (i.e. influenced by outliers) with respect to the hypothesized model.

Sc.3.2 Predictive testing

The blindfolding device can be applied in very different ways and can be used for different sorts of inference. This will be demonstrated on a very common model, a multiple regression model. The predictors are at the beginning considered as directly observed, later as weighted aggregates (LVs) of observed variables (MVs). Three aspects of regression will be distinguished here, called description, forecast, and generalization, and corresponding PRE measures will be presented. The three models generate rule2 predictions, and the respective residual sums of squares will be called DRESS, FRESS, and GRESS. Two different rule1 residual sums of squares will be entertained, called RESS0 and RESS1. FRESS, GRESS, and RESS1 are based on different applications of blindfolding, whereas DRESS and RESS0 are the usual OLS residuals such that R^2=1-DRESS/RESS0.

The multiple regression model to be investigated relates the predictor values $\mathbf{X} = [x_{jn}]$ (j=1..J predictor variables, n=1..N cases) to the predictand values $\mathbf{y} = [y_n]$ by the model

Eq.40 $$y_n = \sum_j b_j x_{jn} + e_n ,$$

where $\mathbf{e} = [e_n]$ is the residual variable and $\mathbf{b} = [b_j]$ is the vector of regression coefficients. If the variables are not centered to zero mean, the first variable has to be taken as unit, x_{1n}=1 for all n, and consequently b_1 will be the regression constant.

The different applications of the blindfolding vary along the dimension "Which model parameters and which data points do we take as known and what must be reestimated for each blindfolding sample?" With respect to the data points which are blindfolded or not we distinguish three cases:

(D.) Description: Nothing is blindfolded.

(F.) Forecast: The predictand value y_i is blindfolded.

(G.) Generalization: Both x_{ji} and y_i are blindfolded.

With respect to the model parameters and the moments of the variables we distinguish the cases:

(P0) Only the regression parameters are unknown.

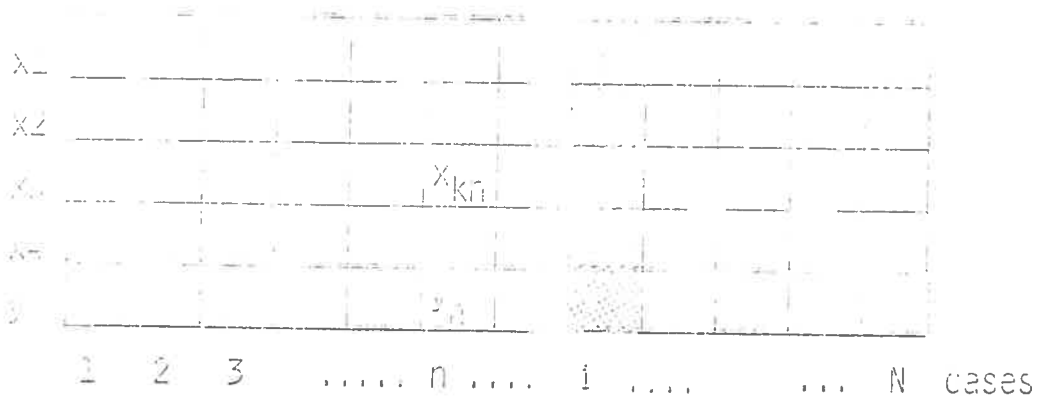
(P1) The mean $\bar{y}$ is known.

(P2) The means $\bar{x}_j$ are known.

(P3) The weights for forming the predictor LVs are known.

Whether a case is omitted totally or only partly from the data matrix depends on the intended conclusions. We will distinguish the description, the forecast, and the generalization approach. With respect to Figure 2, in the forecast approach only the double-shaded data point y_i is blindfolded, whereas in the generalization approach all shaded data points x_{ji} and y_i are omitted.

Figure 2
Blindfolding for multiple regression.



RESS0 and RESS1 - Trivial Prediction error. As rule1 error terms, two residual sums of squares will be defined. RESS0 is the residual sum of squares when the mean $\bar{y}$ of y is taken as predictor of y_i. The residual is

Eq.34 $$e_i = y_i - \bar{y} \quad ,$$

and the residual sum of squares, called RESS0, is identical to the variance of y:

Eq.35 $$\mathrm{RESS0} = (1/N) \textstyle\sum_n e_n^2 \quad .$$

RESS1 is the residual sum of squares when the mean $\bar{y}_{-i}$ of y is taken as predictor of y_i, but when this mean is computed, y_i is omitted. It can be demonstrated that

Eq.36 $$e_{J,i} = y_i - (N/(N-1))\bar{y} \ ,$$

and that RESS0 und RESS1 are related by

Eq.37 $$\mathrm{RESS1} = (N/(N-1))^2\ \mathrm{RESS0} \ .$$

Hence RESS0<RESS1. Notice that RESS0 implies an assumption, namely P1, and this assumption pays off in a smaller error term.

DRESS - Description error. The descriptive regression model makes use of the total data sample ($\mathbf{X}$,$\mathbf{y}$) when estimating the parameters. The residuals are denoted by e_D; the residual sum of squares associated with this model is:

Eq.38 $$\mathrm{DRESS} = (1/N) \textstyle\sum_n e_{D,n}^2 \ .$$

FRESS - Forecast error. In the forecast approach it is assumed that N-1 cases are known totally, and that from an N^{th} case only the predictor but not the predictand values are known. In order to make the most efficient use of the information at hand when estimating the regression parameters, $\mathbf{b}' = \mathbf{m}'_{yx}\mathbf{M}^{-1}_{xx}$, the inverse of $\mathbf{M}_{xx}$=($\mathbf{XX}'$)/N should be based on the total sample, but the predictor--predictand relation can be based only on N-1 cases, $\mathbf{m}_{yx,-i}$, where the subscript -i indicates that the i^{th} case is omitted. When the actual value of y_i becomes known, the residual $e_{F,i}$ are calculated. Notice that the value y_i was not used when the parameters b_{-i} where estimated and the prediction $\hat{y}_{F,i}$ was derived. The real-world forecasting situation can be simulated N times, by blindfolding each value y_i once and estimating the parameters and the prediction. The residual sum of squares is:

Eq.39 $$\text{FRESS} = (1/N) \textstyle\sum_i e^2_{P,i} .$$

GRESS - Generalization error. In the generalization approach it is stipulated that N-1 cases are known and the estimates are to be generalized on an N^th^ case. In order to test this stipulation, each case is omitted once from the data, an both the predictor, and the predictand, values are blindfolded. With case i blindfolded, the regression parameters are estimated by $\mathbf{b}'_{G,i} = \mathbf{m}'_{-i}\mathbf{M}^{-1}_{-i}$. It can be shown that the DRESS-residuals and the GRESS-residuals are related by

Eq.40 $$e_{G,i} = (1/(1-q_i))\, e_{D,i} ,$$

where q_i is the normed Mahalanobis distance, a function of the predictor values of i. The residual sum of squares for the generalization approach is:

Eq.41 $$\text{GRESS} = (1/N) \textstyle\sum_i e^2_{G,-i} .$$

PRE-coefficients Five PRE-coefficients, based on the error sums defined above, are defined now:

Eq.42
$$Q^2(D0) = 1 - \text{DRESS/RESS0}$$
$$Q^2(F1) = 1 - \text{FRESS/RESS1}$$
$$Q^2(F0) = 1 - \text{FRESS/RESS0}$$
$$Q^2(G1) = 1 - \text{GRESS/RESS1}$$
$$Q^2(G0) = 1 - \text{GRESS/RESS0}$$

The first coefficient $Q^2(D0)$ is identical to the squared multiple correlation coefficient in case the variables are centered to zero mean. The ordering of the five coefficients is roughly according to the information used for their estimation, and hence according to the expected order of magnitude.

Examples. The behaviour of Q^2 in different data sets, with different model specification, and with different blindfolding approaches is demonstrated in Table 1. The first data set is analysed under a model with location parameter (row #1) and, centered to zero mean, without location parameter (row #2). If the avarage of y is considered as *a priori* known and exempted from estimation (row #2), then the model has predictive relevance, whereas the model of row #1 is not predictive. A close inspection of the data reveals that the all-cases results are strongly influenced by two single cases. If the tenth case is omitted, R^2 jumps from 0.77 to 0.99, and if the nineth case is omitted, the regression parameters change completely.

Table 1
How predictive validity changes with data, model, and blindfolding approach

Data set and model		D0	F0	G0	F1	G1
	Gaensslen & Schubö (197)					
#1	Raw data	.77	-.04	-.44	.16	-.16
#2	Centered data, no regression constant	.77	.30	.41	.43	.53
	Economic sanctions (Wold 1984)					
#3	4 MV predictors	.78	.07	-.70	.24	-.04
#4	2 MV predictors	.66	.22	.30	.37	.44
#5	2 LV predictors, all-cases w., no constant	.80	.51	.69	.61	.75
#6	2 LV predictors, all-cases w., with const.	.80	.38	.60	.50	.67
#7	2 LV predictors, without-1 w., no constant	-	-.40	-.41	.24	.23
#8	2 LV predictors, without-1 w., with const.	-	-.40	-.41	.23	.23
	Forging force					
#9	Three predictors	.41	-.44	.26	-.32	.20

The second data set is analysed with 4 (row #3) and with 2 predictors (row #4). With 2 predictors the model is predictive, and with 4 it is not: R^2 increases with increasing number of predictors, but Q^2 drops in this case, because the additional parameters can not be estimated reliably. In rows #1 and #3 the rank order of the five coefficients is as expected, G0<F0 and G1<F1, whereas the other models show lower validities for the forecast than for the generalization approach. This is especially so for the third data set (row #9); even if the generalization approach uses less information from the data than the forecast approach, the prediction is better.

LV predictors. The two predictors of the successful model #4 in Table 1 are, in fact, sign-weighted sums of 27 observed predictors; so one can understand this model as one with 2 LV predictors and 27 MV predictors, where the weights w_k were *a priori* chosen as -1 or +1.

Eq.43 $$X_j = \sum_{k_j} w_k x_{k_j}, \quad \forall j$$

The weights were exempted from estimation and blindfolding. Now, in the next two steps on the way from a multiple regression to an LV path model, the weights are not longer treated as known.

All-cases LV weights. Following PLS Mode A technique, the LV weights are taken proportional to the correlation of MV predictors x_k with predictand variable y,

Eq.44 $$w_k \propto \mathrm{cor}(x_k;y) \quad ,$$

and the LVs are scaled to zero mean and unit variance. In this step, we use all cases for the estimation of the LV weights and the scaling of the LVs, and execute the blindfolding procedure only on the regression parameters. As the LVs have zero mean, the constant in the regression model must come out as zero, and could be omitted. However, even if the LVs are standardized over all N cases, they are not standardized in the blindfolded N-1 data set, and results are different for models with and without regression constant, see Table 1, rows #5 and #6. As compared to the MV predictor model (row #4), R^2 is higher because the weights from Eq.44 are "better" than the sign weights. The PRE measures for this model vary between 0.38 and 0.75, depending on the choice of forecasting or generalization approach, RESS0 or RESS1, constant included or not. But under all these variations the PRE is positive, indicating that the model has predictive validity.

Without-one LV weights. Now the LV weights are also subject to blindfolding, and the correlation in Eq.44 is computed with one case omitted. As the predictor LVs change their values due to the different weights, there is not a unique R^2 but N different R^2s. Also the scaling of the LVs becomes ambiguous, and one has to decide whether the LVs are to be standardized over N or over N-1 cases, whether only the predictand or even the predictors are to be rescaled, and whether the regression model should have a constant or not. For all the variations of the data and the model, the PRE measures $Q^2(F1)$ and $Q^2(G1)$ vary from 0.08 to 0.36, and the most resonable results are reported in rows #7,8 of Table 1. As compared to the all-cases weights, there is a sharp drop in the PRE, but nevertheless it is positive, and the additional 27 parameters (the weights) can be estimated reliably from the data. Compared, however, to row #3, the two superfluous predictors do harm to model, and the two additional regression parameters disturb the prediction more than the 27 weights.

Conclusion. It has been demonstrated that the predictive testing by the blindfolding device is a flexible tool for the evaluation of the predictive models. The model to be tested is not presumed to be true, but to have predictive power. Consequently a negative tests result is implies not that the model is wrong, but that the model is useless. Unlike perturbation analysis, the stability of the results is not tested by adding something to the data (and it has to be specified what form and distribution the added error should have), but by omitting given data.

The blindfolding device requires no assumption about independent observations. The assumption that the data set is homogenous with respect to the hypothesized model, i.e. that it has no outliers, can be tested, and the cases which are contradictory to this assumption can be identified. The blindfolding procedure is sensitive to scaling, in two respects: In very small data sets, it makes a difference whether the LVs are standardized over N or over N-1 cases. And in general, the regression constant can turn out to be the most volatile and susceptible parameter.

Sc.4 Three-way data models

Three-way data are ordered by three indices; for example, y_{ptr} may denote the election outcome of a party p at election time point t in a region r, and x_{gtr} the number of employed, and unemployed, workers at time t in region r. If theory is concerned only with the influence of unemployment categories (x_g) on voting outcome (y_p), time points or regions or both may be used as observational units, and we have a two-way ordering of observations. In this case, the indices p and g refer to "specified" coordinates of the data array, and t and r refer to "unspecified" coordinates of replications.

It may, however, turn out that the time dimension taps a causal influence of its own (see Falter). Then the time must be specified theoretically, the model considers the variables x_{gt} and y_{pt} with regions as the only observational units, and we have a two-way ordering of variables. As a general method for modelling variables ordered in two ways and observed in one way (three-way or three-mode data), the three-mode path analysis with latent variables can be used (Lohmöller and Wold 1980, Lohmöller 1984).

References

Apel,H. & Wold,H. (1982) Soft modeling with latent variables in two or more dimensions: PLS estimation and testing for predictive relevance. In K.G. Jöreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (2 vols.). Amsterdam: North Holland (vol.2,p.209--248).

Belsley,D.A. (1984) Demeaning conditioning diagnostics trough centering. American Statistician, 38:73-77,90-93.

Bentler,P.M. (1980) Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 31:419-456.

Bernert,C. (1983) The career of causal analysis in American sociology. british Journal of Sociology, 34:230-254.

Cronbach,L.J. & Meehl,P.E. (1955) Construct validity in psychological testing. Psychological Bulletin, 52:281-302.

Falter,J.W. (1977) Zur Validierung theoretischer Konzepte: Wissenschaftstheoretische Aspekte des Validierungskonzepts. Zeitschrift für Soziologie, 6:349-385.

Falter,J.W. & Lohmöller,J.B. (1982) Manifeste Schwächen im Umgang mit latenten Variablen: Ein Kommentar zu Han-Hermann Hoppes Theologie der LV--Pfadmodell in ZfS Juli 1981. Zeitschrift für Soziologie, 11:69-77.

Geisser,S. (1974) A predictive approach to the random effect model. Biometrika, 61:101-107.

Goodman,L.A. & Kruskal,W.H. (1954) Measures of association for cross classifications. Journal of the American Statistical Association, 49:732-764.

Guttman,L. (1941) An outline of the statistical theory of prediction (p.253-311).

In P. Horst (Ed.), The prediction of personal adjustment; supplementary study B-I (SSRC bulletin No.48). New York: Social Science Research Council.

Lohmöller,J.B. (1984) Path models with latent variables and Partial Least Squares (PLS) estimation. Würzburg: Physica Verlag (in print).

Lohmöller,J.B. & Wold,H. (1980) Three-mode path models with latent variables and Partial Least Squares (PLS) parameter estimation (Paper presented at the European Meeting of the Psychometric Society; University of Groningen, The Netherlands; June 18-21 1980) (Forschungsbericht 80.03 Fachbereich Pädagogik). München: Hochschule der Bundeswehr.

MacCorquodale,K. & Meehl,P.E. (1948) On a distinction between hypothetical constructs and intervening variables. Psychological Review, 55:95-107.

Mueller,J.H., Schuessler,K.F., & Costner,H.L. (1970) Statistical reasoning in sociology (2nd ed.). New York: Hughton Mifflin.

Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, series B, 36:111-147.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiah (Ed.), Multivariate analysis. New York: Academic (p.391-420).

Wold,H. (1969) Mergers of economics and philosophy of science: A cruise in shallow waters and deep seas. Synthese, 20:427-482.

Wold,H. (1981) Systems under indirect observations using soft modelling (Working Paper No. 48). Cleveland,Ohio: Case Western Reserve University, Economics Department.