4

THE TREATMENT OF UNOBSERVABLE VARIABLES IN PATH ANALYSIS

Robert M. Hauser
UNIVERSITY OF WISCONSIN, MADISON
Arthur S. Goldberger
UNIVERSITY OF WISCONSIN, MADISON

This paper was prepared for presentation at the 65th annual meetings of the American Sociological Association, held at Washington, D.C., August 31 through September 3, 1970. At that time, it was distributed as Workshop Paper EME 7030 by the Social Systems Research Institute, University of Wisconsin. Work on this project was in part supported by the Graduate School Research Committee of the University of Wisconsin, by the National Institutes of Health, U.S. Public Health Service (M-6275), and by the Social and Rehabilitation Service, U.S. Department of Health, Education, and Welfare (CRD-314).

Under the rubrics of path analysis (Duncan, 1966; Land, 1969; Heise, 1969) or dependence analysis (Boudon, 1965; 1968), sociologists have recently been introduced to the expression of theories or models as systems of structural equations. In this chapter our purpose is to draw attention to the recurrent problem of estimating the parameters of overidentified path models which feature unobservable variables. Such variables have not been, or perhaps cannot be, measured directly, and the structural coefficients pertaining to their causes and effects must be inferred from the available measurements and the postulated causal structure. While models containing unobservables may be underidentified (Siegel and Hodge, 1968; Duncan, 1969a; Land, 1970) or just-identified (Heise, 1969; Wiley and Wiley, 1970; Land, 1969, pp. 29–33; 1970; Hauser, 1969b, pp. 549–550; and Brewer, Crano, and Campbell, 1970), most frequently they are overidentified (Duncan, Featherman, and Duncan, 1968; Duncan, Haller, and Portes, 1968; Hodge and Treiman, 1968; Duncan, 1969b; Hauser, 1968; 1969a; 1969b; 1970). Overidentification means that alternative estimates of certain parameters can be made (Costner, 1969; Blalock, 1969a; 1970).

Alternative estimates in an overidentified model will not coincide in finite samples even where the model is correct, that is, even where they would coincide in the population. Hence, some means of reconciling the conflicting estimates is required. The theoretical and empirical path-analysis literature tends to slur the sample-population distinction, and it provides little guidance in estimation for overidentified models.[1] Some ad hoc procedures for estimation of overidentified models containing unobservables have been suggested. For example, Blalock (1970) and Land (1970) have advocated that one or more equations be ignored in the estimation process and be introduced only to test for goodness of fit. At least one empirical example using this approach (Hodge and Treiman, 1968; Hauser, 1969b) predates their work. The technique is mentioned in the econometric literature (Christ, 1966, pp. 407–411) but with little enthusiasm. Other analysts have used arbitrary averages of alternative estimators without considering their statistical properties (Duncan, Featherman, and Duncan, 1968; Duncan, Haller, and Portes, 1968; Hauser, 1968; 1969a; 1970).

Principles of estimation imply testing procedures, and the treatment of goodness of fit in the path-analysis literature reflects its casual methods of estimation. In numerous instances we are told only to "see" whether the correlations implied by estimators for an overidentified

[1] The sociological literature is occasionally vague on the reasons for discrepancies among alternative estimates (compare Costner, 1969, pp. 252–253, 259, 262). However, the explanation is no more complicated than that required to account for the fact that, in a random sample from a normal population, the mean and the median will be different, despite their coincidence in the population.

model are "close" to the observed correlations. It is not clear how close is close enough. For example, Costner (1969, pp. 252–262) states,

> Failure of the data to satisfy this [overidentification] equation, at least approximately, indicates that, in some respect, the indicators . . . are not appropriate . . . . [T]he several estimates . . . should all be identical except for random error . . . . It may be reasonably asked what is meant when we say that the consistency criterion is satisfied. Do we mean that the two sides of [the overidentification equation] are exactly identical, that they are approximately identical, or that they should not differ to a degree that is statistically significant at the commonly utilized levels of significance?

The sociological literature on path analysis also conveys the misleading impression that certain problems in estimation and testing have not been analyzed or even that they are not amenable to rigorous analysis. For example, Costner (1969, p. 262) states, "Satisfying the additional consistency criteria in the three-indicator model presents an additional statistical inference problem, the solution to which does not appear to be found in the factor analysis literature." Blalock (1970, p. 103) states,

> If there were absolutely no specification or sampling errors, the data would fit the model exactly, and it would make no difference which equations were treated as redundant. However, in practice, no data will fit the model exactly; therefore, there is a certain arbitrariness in one's selection of the particular equations that will be used for estimation purposes and those that will be treated as excess equations used to test the model. This difficulty . . . would seem to admit of no completely satisfactory solution . . . . Perhaps the issue will reduce to the question of whether one assumes specification errors to be more serious than sampling errors . . . .[2]

In fact, the estimation of overidentified models is not an intractable problem, and it is a central topic in the econometric and psychometric literature. There, standard principles of statistical inference are applied to determine efficient estimates, that is, estimates which have minimum sampling variability. Since path models are linear models of the type considered in econometrics and psychometrics, it should

[2] These remarks call to mind Boudon's (1968, p. 213) assertion, "Of course, when we are dealing with fallible data, these different possible choices will lead to different estimates, and there is obviously no reason to think that one estimate is better than the other."

not be surprising that the efficient estimating procedures developed there can be applied to problems of estimation in path analysis. Standard principles of statistical inference also imply testing procedures, and these, too, are worked out in the econometric and psychometric literature.[3]

We propose to illustrate the utility of econometric and psychometric estimation techniques for path models containing unobservable variables. Our examples are simple and will not do justice to the more elaborate sociological applications. They can be thought of as components of larger models, and they should suffice to document our basic theme. We shall treat two classes of models, those where unobservables appear only as causes of observable variables and those where unobservables appear as both causes and effects of observable variables. These have been recognized as distinct cases in path analysis.

Following Costner (1969), Blalock (1969a, pp. 264–270) discusses multiple indicators of correlated unobservable variables. Then, under the heading of "the instrumental variable approach" (pp. 270–272), he considers the case where causes of the unobservable variables are also observed. He correctly indicates that this gives an econometric flavor to the model, but his emphasis on instrumental variables is somewhat misleading. Land (1970, p. 507) distinguishes "two general cases for which sociologists will be interested in utilizing unmeasured variables in a causal model." "The first of these," he states, "arises in the study of measurement error . . . . The common characteristic of all of these applications of path analysis is that the hypothetical (unmeasured) variables enter the path models only as causes of the observed variables. A second case in which sociologists will utilize unmeasured variables is as variables which intervene between measured variables in a causal model."

We shall see that the first type of path model translates directly into a confirmatory factor-analysis model which psychometricians have studied and for which computer programs are available. The second type of path model translates into a form studied by econometricians and for which computer programs are also available. There are, we should add, cases which require a blending of the factor-analysis and econometric approaches.

As Blalock has implied, efficient estimation of overidentified path models may not be a very important topic. Perhaps it is more important to settle for some reasonable estimate and to concentrate on improving the model than to search diligently for efficient estimators. Still, it is

[3] The idea that path analysis involves a mixture of psychometric and econometric themes is not a new one (see Duncan, 1966, p. 16; Blalock, 1961, pp. 167–169; 1969a; Werts and Linn, 1970; Wright, 1925; 1934; 1954).

worth knowing that there are solutions to some of the puzzles we create, and it is probably worthwhile to build up the stock of tools in advance of the time when rigorous inference becomes the order of the day.

GENERAL CONSIDERATIONS

In most, and perhaps in all, cases the efficient estimators for an overidentified model may be interpreted as appropriately weighted averages of the several conflicting estimators. The weights are chosen to take account of the sampling variability and covariability of the original estimates. Generally, the weights cannot be determined in advance but must be estimated from a sample. For this reason, the computation required to obtain efficient estimates is often more extensive than that for ordinary regression or ad hoc averaging of estimates. For example, the maximum-likelihood principle offers a standard method for efficient estimation of overidentified models (with or without unobservable variables), and the method implicitly involves the construction of weights with which to reconcile the conflicting estimates. Often, as in factor analysis, iterative computation is required.[4]

To avoid elaborate computations, other estimation procedures have been developed for certain classes of models. Generalized least-squares and two- and three-stage least-squares are econometric examples, and the minres criterion for fitting factor-analysis models is an example from psychometrics. Under some conditions estimates produced by such procedures have the same efficiency as maximum-likelihood estimates.

A simple example may clarify the logic of efficient estimation. Suppose we have two independent, unbiased estimates, m_1 and m_2, of a parameter μ, which have variances σ_{11} and σ_{22}, respectively. We wish to construct m, the minimum-variance unbiased estimator of μ. That is, we want to find weights, a_1 and a_2, such that $E(m) = \mu$, where $m = a_1m_1 + a_2m_2$ and $\mathrm{Var}(m)$ is minimized. Clearly, we will choose $a_1 + a_2 = 1$, since

$$\begin{aligned} E(m) &= E(a_1m_1 + a_2m_2) = E(a_1m_1) + E(a_2m_2) = a_1E(m_1) + a_2E(m_2) \\ &= a_1\mu + a_2\mu = (a_1 + a_2)\mu \qquad (1) \end{aligned}$$

[4] The fully recursive model with uncorrelated errors provides an important exception to this pattern. For that model Boudon (1965; 1968) developed an averaging scheme, with weights determined by the sample. However, his estimates are less efficient than those produced by ordinary least-squares regression, which gives all the weight to the regression estimate and ignores the conflicting instrumental-variable estimates (Goldberger, 1970b).

Subject to this, we wish to choose the a_1 and a_2 which will minimize the sampling variance of m, namely,

$$\mathrm{Var}(m) = \mathrm{Var}(a_1m_1 + a_2m_2) = a_1^2\sigma_{11} + a_2^2\sigma_{22} \tag{2}$$

Multiplying equation (2) by $(m_2 - m_1)^2$, we obtain

$$\begin{aligned} a_1^2\sigma_{11}(m_2 - m_1)^2 &+ a_2^2\sigma_{22}(m_1 - m_2)^2 \\ &= \sigma_{11}(m_2a_1 - a_1m_1)^2 + \sigma_{22}(m_1a_2 - a_2m_2)^2 \end{aligned} \tag{3}$$

Noting that $a_1 = 1 - a_2$ and $a_2 = 1 - a_1$, we rewrite this as

$$\begin{aligned} \sigma_{11}(m_2(1 - a_2) &- a_1m_1)^2 + \sigma_{22}(m_1(1 - a_1) - a_2m_2)^2 \\ &= \sigma_{11}(m_2 - a_1m_1 - a_2m_2)^2 + \sigma_{22}(m_1 - a_1m_1 - a_2m_2)^2 \\ &= \sigma_{11}(m_2 - m)^2 + \sigma_{22}(m_1 - m)^2 \end{aligned} \tag{4}$$

Finally, dividing equation (4) by $\sigma_{11}\sigma_{22}$ and cancelling like terms in the numerator and denominator, we have

$$\sigma^{11}(m_1 - m)^2 + \sigma^{22}(m_2 - m)^2 \tag{5}$$

where $\sigma^{11} = 1/\sigma_{11}$ and $\sigma^{22} = 1/\sigma_{22}$, which we can proceed directly to minimize with respect to m. Setting the derivative of expression (5) with respect to m equal to zero, we find that

$$m = \frac{\sigma^{11}}{\sigma^{11} + \sigma^{22}}\, m_1 + \frac{\sigma^{22}}{\sigma^{11} + \sigma^{22}}\, m_2 \tag{6}$$

whence

$$a_1 = \frac{\sigma^{11}}{\sigma^{11} + \sigma^{22}} \quad \text{and} \quad a_2 = \frac{\sigma^{22}}{\sigma^{11} + \sigma^{22}} \tag{7}$$

are the desired weights.

While we could have minimized equation (2) directly, we chose the more tortuous path in order to point out that expression (5) gives the function minimized by the principle of generalized least-squares, namely, the weighted sum of squared deviations of the conflicting estimates from the desired estimate, where each weight is inverse to the variance of the corresponding estimate.

If the estimates were not independent but had covariance $\sigma_{12} \neq 0$, then the GLS principle says that the estimate m should be chosen to minimize

$$\sigma^{11}(m_1 - m)^2 + \sigma^{22}(m_2 - m)^2 + 2\sigma^{12}(m_1 - m)(m_2 - m) \tag{8}$$

where the σ's with superscripts are elements in the inverse matrix

$$\begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{12} & \sigma^{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \tag{9}$$

The resulting estimate, which has the property of minimum-variance unbiasedness, is

$$m = \frac{\sigma^{11} + \sigma^{12}}{\sigma^{11} + \sigma^{22} + 2\sigma^{12}} m_1 + \frac{\sigma^{22} + \sigma^{12}}{\sigma^{11} + \sigma^{22} + 2\sigma^{12}} m_2 \tag{10}$$

which is again a weighted average of m_1 and m_2, with the weights now taking account of sampling covariability as well as variability.

In practice, the σ's are unknown and must be replaced by estimates of them. When this is done, we refer to the procedure as modified generalized least-squares (MGLS). Under quite general conditions the MGLS procedure produces estimates which are as efficient as those produced by the maximum-likelihood principle.

MULTIPLE INDICATORS OF CAUSALLY RELATED UNOBSERVABLE VARIABLES

Specification of the Model

Consider a model in which we observe multiple indicators of two causally related unobservable variables, as shown in Figure 1. In algebraic form the system consists of:

$$y^* = \alpha z^* + \epsilon \tag{11}$$

a linear equation expressing the determination of the unobservable variable y^* by the unobservable variable z^* and an unobservable disturbance ϵ;

$$\begin{aligned} z_1 &= \beta_1 z^* + v_1 \\ &\vdots \quad \vdots \\ z_k &= \beta_k z^* + v_k \\ &\vdots \quad \vdots \\ z_K &= \beta_K z^* + v_K \end{aligned} \tag{12}$$

a set of K linear equations expressing each observable indicator z_k of z^* in terms of z^* and an unobservable disturbance $v_k (k = 1, \ldots, K)$; and

$$\begin{aligned} y_1 &= \gamma_1 y^* + w_1 \\ &\vdots \quad \vdots \\ y_m &= \gamma_m y^* + w_m \\ &\vdots \quad \vdots \\ y_M &= \gamma_M y^* + w_M \end{aligned} \tag{13}$$

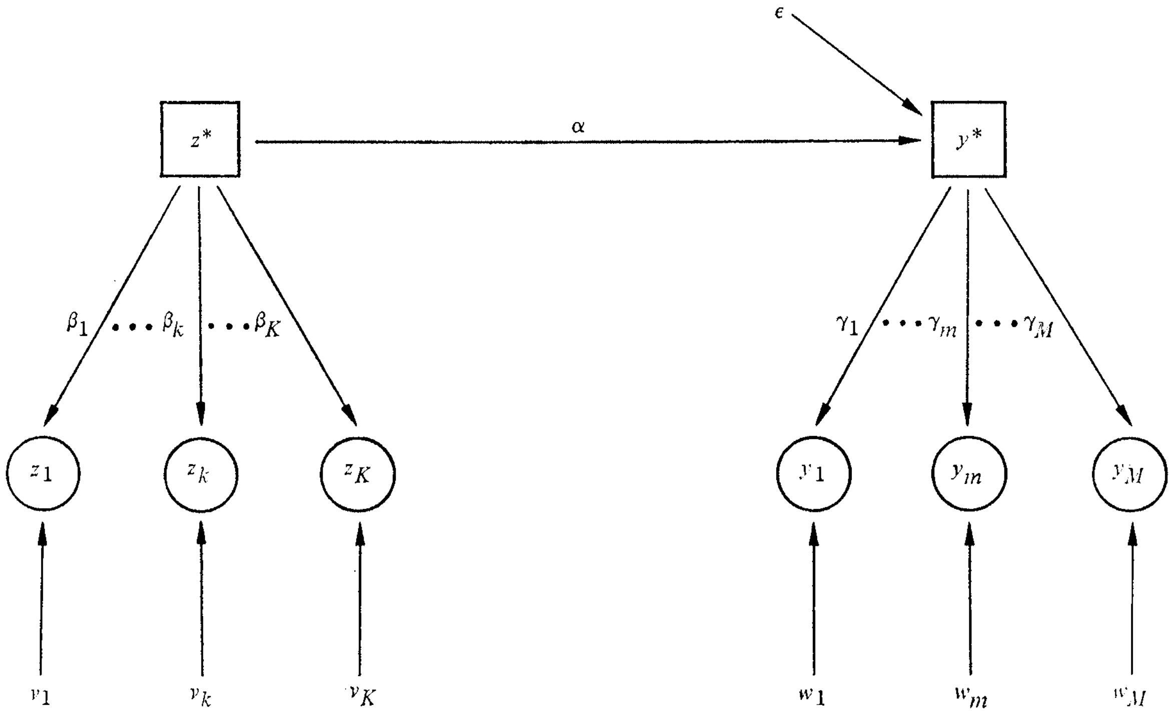

Figure 1. Multiple indicators of two correlated unobservable variables.

a set of M linear equations expressing each observable indicator y_m of y^* in terms of y^* and an unobservable disturbance $w_m (m = 1, \ldots, M)$. It is assumed that the disturbances are independent of z^* and y^* and are mutually independent as well.

The α, β's, and γ's are path coefficients to be estimated along with the disturbance variances, $\sigma_\epsilon^2, \sigma_{v_1}^2, \ldots, \sigma_{v_K}^2, \sigma_{w_1}^2, \ldots, \sigma_{w_M}^2$. We are following the econometric and psychometric convention which leaves the disturbances unstandardized; our variances are just the squares of the residual paths which would appear if the path-analysis convention (in which the disturbances are standardized) had been followed. The unobservables, z^* and y^*, are standardized as are the observables (although the latter is not at all essential).

Costner (1969) and Blalock (1969a) considered at length systems of this type without making it clear that such models have already been thoroughly investigated in the psychometric literature. To clarify the situation, a matrix formulation is convenient.

We introduce the vectors

$$\begin{aligned} x' &= (z_1, \ldots, z_K, y_1, \ldots, y_M) \\ f' &= (z^*, y^*) \\ u' &= (v_1, \ldots, v_K, w_1, \ldots, w_M) \end{aligned}$$

and the matrix

$$\Delta = \begin{pmatrix} \beta_1 & 0 \\ \vdots & \vdots \\ \beta_K & 0 \\ 0 & \gamma_1 \\ \vdots & \vdots \\ 0 & \gamma_M \end{pmatrix} \tag{14}$$

and write equations (12) and (13) compactly as

$$x = \Delta f + u \tag{15}$$

The population variance-covariance matrix of the observable indicators is then

$$\begin{aligned} \Omega &= E(xx') = E(\Delta f + u)(\Delta f + u)' = \Delta E(ff')\Delta' + E(uu') \\ &= \Delta\Phi\Delta' + \Sigma \end{aligned} \tag{16}$$

where we have introduced $\Phi = E(ff')$ as the variance-covariance matrix of the two unobservable variables, $\Sigma = E(uu')$ as the variance-covariance matrix of the disturbances, and used $E(fu') = 0$. It follows from equation (11) that

$$\Phi = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \tag{17}$$

and from the assumptions on the independence of the disturbances that

$$\Sigma = \begin{pmatrix} \sigma^2_{v_1} & & & & & 0 \\ & \ddots & & & & \\ & & \sigma^2_{v_K} & & & \\ & & & \sigma^2_{w_1} & & \\ & & & & \ddots & \\ 0 & & & & & \sigma^2_{w_M} \end{pmatrix} \tag{18}$$

It should now be apparent that what we have is a factor-analysis model (compare Harman, 1967, Chapter 2; Morrison, 1967, Chapter 8). More specifically, the unobservables z^* and y^* represent oblique factors, and the absence of direct paths from z^* to the y's and from y^* to the z's represents certain zero factor loadings.

If we examine the population correlation matrix of the indicators given in equation (16), we see that the model is typically overidentified. The $\frac{1}{2}(K + M)(K + M + 1)$ distinct elements of the symmetric matrix Ω are expressible in terms of only $1 + 2(K + M)$ parameters, $\alpha, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_M, \sigma^2_{v_1}, \ldots, \sigma^2_{v_K}, \sigma^2_{w_1}, \ldots, \sigma^2_{w_M}$; the variance σ^2_ϵ being determined by the standardization of y^*. For example, if $M = 2 = K$, we have $\frac{1}{2}(K + M)(K + M + 1) = 10$ and $1 + 2(K + M) = 9$, so that there is one overidentifying restriction.

The maximum-likelihood principle offers a straightforward approach to estimation of the parameters of the model. It must be emphasized that, having specified certain zero factor loadings in advance, we are concerned with confirmatory factor analysis. In the more traditional exploratory factor analysis, various rotations are used to obtain approximate zero factor loadings in the desired places, but these would not do justice to the present model. On the important distinction between confirmatory and exploratory factor analysis, see Jöreskog and Lawley (1968). Computational procedures for confirmatory maximum-likelihood estimation are spelled out in Lawley and Maxwell (1963, Chapters 2, 6) and Jöreskog (1969a).

Path Analysis Approach to Estimation

Before describing the efficient estimation procedure, we pause to review the path-analysis approach to fitting the model. For the sake of concreteness, we take $K = M = 2$ as in Costner (1969, Figure 4). By inspection of the path diagram, or from equations (11) through (13), the following "estimating equations" are produced:

$$
\begin{aligned}
r_{z1z2} &= \beta_1\beta_2 \qquad & r_{z1y1} &= \beta_1\alpha\gamma_1 \qquad & r_{z1y2} &= \beta_1\alpha\gamma_2 \\
& & r_{z2y1} &= \beta_2\alpha\gamma_1 \qquad & r_{z2y2} &= \beta_2\alpha\gamma_2 \\
& & & & r_{y1y2} &= \gamma_1\gamma_2
\end{aligned} \tag{19}
$$

(compare Blalock, 1969a, p. 251, equations (1) through (6)). In equation (19) there are six equations from which to estimate the five parameters α, β_1, β_2, γ_1, γ_2; the disturbance variances being estimable subsequently. Clearly, the system is overdetermined, there being one excess equation.

In particular, we can estimate α as $a^{(1)}$, the square root of $(r_{z1y1}r_{z2y2})/(r_{z1z2}r_{y1y2})$; and then, with this value in hand, we can go on to solve for estimates of β_1, β_2, γ_1, γ_2, say $b_1^{(1)}$, $b_2^{(1)}$, $c_1^{(1)}$, $c_2^{(1)}$. Alternatively, we can estimate α as $a^{(2)}$, the square root of $(r_{z2y1}r_{z1y2})/(r_{z1z2}r_{y1y2})$; and with that value in hand, we can go on to solve for estimates of β_1, β_2, γ_1, γ_2, say $b_1^{(2)}$, $b_2^{(2)}$, $c_1^{(2)}$, $c_2^{(2)}$.

Even if the model is correct in the population, the distinct estimates of the same parameters will fail to coincide in any sample. Sometimes, the advice given is to average them; thus, Blalock (1969a, p. 266) in effect suggests taking $\frac{1}{2}(a_1^{(1)} + a_2^{(2)})$ as the estimate of α and, with this value in hand, going on to solve for estimates of the β_1, β_2, γ_1, γ_2. Such averaging procedures are obviously arbitrary. In a sense, they put equal weight on conflicting estimates in forming the average. However, the several conflicting estimates are unlikely to have the same sampling variability, and an efficient estimation procedure should take this into account.

EFFICIENT ESTIMATION OF THE MULTIPLE-INDICATOR MODEL

Derivation of Procedure

If we consult the factor-analysis literature, we find that the maximum-likelihood principle calls for the estimates to be chosen as the values of Δ, Φ, Σ, which minimize

$$\log \det(\Omega) + \operatorname{tr}(\Omega^{-1}W) \tag{20}$$

where $\Omega = \Delta\Phi\Delta' + \Sigma$, W is the sample variance-covariance matrix of the indicators, log stands for natural logarithm, det stands for determinant, and tr stands for trace. Specifically, with $M = 2 = K$, we have from equations (14), (17), and (18),

$$\Delta\Phi\Delta' = \begin{pmatrix} \beta_1 & 0 \\ \beta_2 & 0 \\ 0 & \gamma_1 \\ 0 & \gamma_2 \end{pmatrix} \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & 0 & 0 \\ 0 & 0 & \gamma_1 & \gamma_2 \end{pmatrix}$$

so

$$\Omega = \begin{pmatrix} \beta_1^2 + \sigma_{v_1}^2 & \beta_1\beta_2 & \beta_1\alpha\gamma_1 & \beta_1\alpha\gamma_2 \\ & \beta_2^2 + \sigma_{v_2}^2 & \beta_2\alpha\gamma_1 & \beta_2\alpha\gamma_2 \\ & & \gamma_1^2 + \sigma_{w_1}^2 & \gamma_1\gamma_2 \\ & & & \gamma_2^2 + \sigma_{w_2}^2 \end{pmatrix}$$

Also,

$$W = \begin{pmatrix} 1 & r_{z1z2} & r_{z1y1} & r_{z1y2} \\ & 1 & r_{z2y1} & r_{z2y2} \\ & & 1 & r_{y1y2} \\ & & & 1 \end{pmatrix}$$

(Here and throughout the paper we omit the subdiagonal elements of symmetric matrices.)

The equations for minimizing expression (20) and an iterative procedure for solving them can be found in Lawley and Maxwell (1963, pp. 79–81).

The heuristic interpretation is that the maximum-likelihood method seeks parameter values (elements of Ω) which reproduce the observed correlations (elements of W) as closely as possible. The over-identifying restriction prevents perfect reproduction, of course.

The estimates produced by the maximum-likelihood method are guaranteed to be efficient, that is, to have minimum sampling variability.

Numerical Illustration

To illustrate the efficient estimation procedure we draw on Hauser's (1969a) study of schools and the stratification process. The sample consists of some 17,000 white public-school students enrolled in grades 7 through 12. The observed variables (original symbols follow in parentheses) are z_1 = arithmetic mark (A), z_2 = English mark (E), y_1 = educational aspiration (T), and y_2 = occupational aspiration (J). The within-school correlations given above the diagonal in Hauser (1969a, Table 3) are presented here in Table 1.

In our model, shown in Figure 2, the marks z_1 and z_2 are assumed to be indicators of an unobservable variable z^* = academic performance, which determines an unobservable variable y^* = ambition, for which y_1 and y_2 serve as indicators. The maximum-likelihood estimates are

Table 1
Correlations of Academic Performance and Ambition Indicators

W = Observed Correlations

	z_1	z_2	y_1	y_2
z_1	1.000	0.630	0.202	0.238
z_2		1.000	0.272	0.292
y_1			1.000	0.456
y_2				1.000

reported in Table 2. Our estimates of α, β_1, β_2, γ_1, γ_2 appear as elements of $\hat{\Phi}$ and $\hat{\Delta}$; the residual paths to the indicators are the square roots of the elements in $\hat{\Sigma}$; and the standardization of y^* gives the residual path of y^* as the square root of $1 - a^2$, where a is the estimate of α. Also reported in Table 2 is our implied correlation matrix $\hat{\Omega} = \hat{\Delta}\hat{\Phi}\hat{\Delta}' + \hat{\Sigma}$.[5]

Our implied correlations in $\hat{\Omega}$ naturally differ from the observed correlations in W; after all, the latter did not satisfy the overidentifying restriction. The differences are rather small, but the sample size is very large. To translate such remarks into a formal test of the causal model, one can simply draw on the likelihood-ratio test of factor analysis (Lawley and Maxwell, 1963, pp. 84–86). The relevant statistic is $T \log [\det(\hat{\Omega})/\det(W)]$, where T is the sample size. On the null hypothesis that the overidentifying restrictions are correct, this statistic is distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions. In our illustration we have one restriction, a sample size of 17,000, and the determinants are $\det(\hat{\Omega}) = 0.434$ and $\det(W) = 0.423$. This gives a test statistic of 442, which is significant even at the 1

Table 2
Efficient Estimates for Causal Model of Figure 2

$\hat{\Delta}$ = Factor Loadings

	z^*	y^*
z_1	0.77	0
z_2	0.82	0
y_1	0	0.66
y_2	0	0.69

$\hat{\Sigma}$ = Unique Variances

	z_1	z_2	y_1	y_2
	0.64	0	0	0
		0.58	0	0
			0.75	0
				0.72

$\hat{\Phi}$ = Factor Correlations

z^*	1.00	0.47
y^*		1.00

$\hat{\Omega}$ = Implied Correlations

z_1	1.00	0.63	0.24	0.25
z_2		1.00	0.26	0.27
y_1			1.00	0.46
y_2				1.00

[5] Starting with Hauser's correlation matrix, we found the efficient estimation required about three hours on a desk calculator. Computer programs are in fact available (see Jöreskog, 1967).

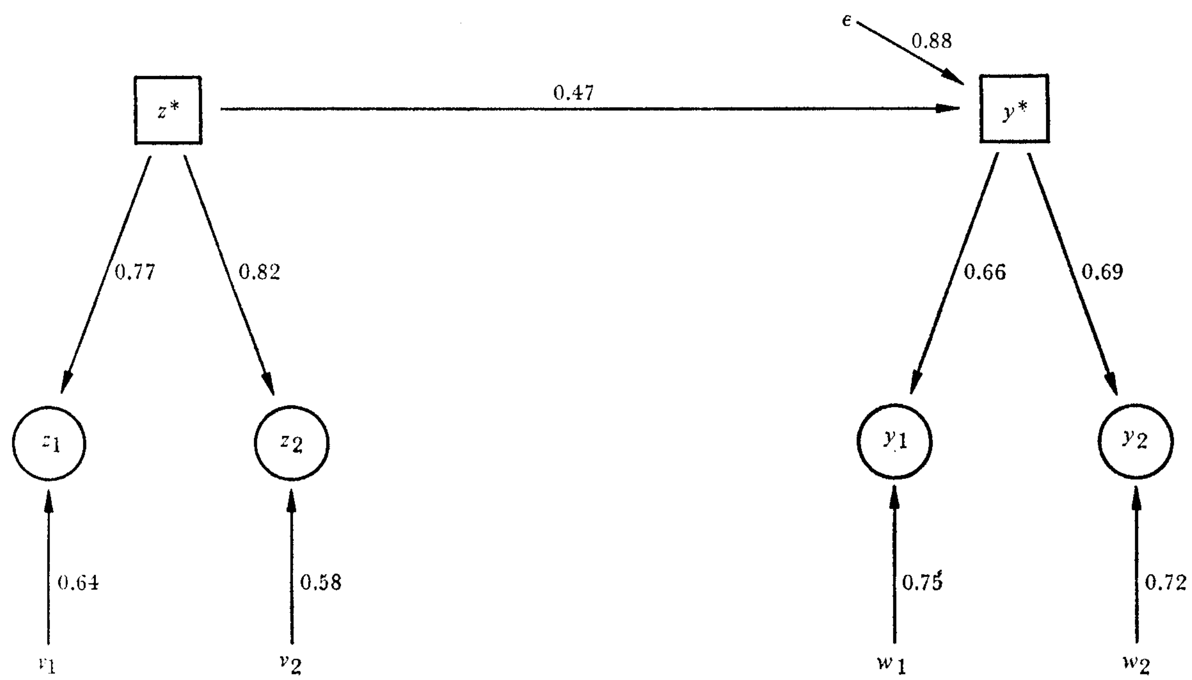

Figure 2. Causal model of academic performance and ambition.

per cent level. Routine procedures of statistical inference, therefore, would lead us to reject the causal model.

Comments

In Table 3 we present the two conflicting sets of parameter estimates obtained by the path-analysis approach sketched above in the section "Path-Analysis Approach to Estimation" along with our efficient estimates. In this case, the efficient estimates of the individual parameters do not all lie within the range of the two conflicting estimates.

Table 3
Conflicting and Efficient Estimates of α, β_1, β_2, γ_1, γ_2

j	$a^{(j)}$	$b_1^{(j)}$	$b_2^{(j)}$	$c_1^{(j)}$	$c_2^{(j)}$
1	0.45	0.72	0.88	0.65	0.70
2	0.48	0.68	0.92	0.62	0.73
Efficient	$a = 0.47$	$b_1 = 0.77$	$b_2 = 0.82$	$c_1 = 0.66$	$c_2 = 0.69$

We have argued that models with multiple indicators of causally related unobservable variables fall directly under the scope of factor analysis, but our illustration was confined to the case of two unobservable variables. Our argument, in fact, requires qualification when there are more than two unobservables bound together in a recursive model. If all direct paths are present in the recursive model, no difficulty arises, there being a one-to-one correspondence between the factor correlations and the paths connecting the unobservables. Estimates of Φ can be converted directly into estimates of the path coefficients in the main model. But, if some direct paths are ruled out of the recursive system, as in Costner (1969, Figure 6), the one-to-one correspondence breaks down. The structuring of the Φ matrix will also have to be taken into account in efficient estimation. This can be done by formulating a "second-order factor analytic model" as in Jöreskog's (1970) general method for analysis of covariance structures. The specific device is spelled out in Jöreskog (1969b, pp. 13–18).[6]

MULTIPLE CAUSES AND MULTIPLE INDICATORS OF AN UNOBSERVABLE VARIABLE

Specification of the Model

We now turn to a model in which we observe multiple causes and

[6] His approach also allows for one to specify that certain direct paths be equal as in Blalock (1970, pp. 106–110).

multiple indicators of a single unobservable variable as shown in Figure 3. In algebraic form the model consists of

$$y^* = \alpha_1 x_1 + \cdots + \alpha_k x_k + \cdots + \alpha_K x_K + \epsilon \tag{21}$$

a linear equation expressing the unobservable variable y^* in terms of its observable causes $x_1, \ldots, x_k, \ldots, x_K$ and an unobservable disturbance ϵ, and

$$\begin{aligned} y_1 &= \beta_1 y^* + u_1 \\ &\vdots \\ y_m &= \beta_m y^* + u_m \\ &\vdots \\ y_M &= \beta_M y^* + u_M \end{aligned} \tag{22}$$

a set of M linear equations expressing each observable indicator y_m in terms of y^* and an unobservable disturbance $u_m (m = 1, \ldots, M)$. It is assumed that the disturbances are independent of the x's and are mutually independent as well.

The α's and β's are path coefficients to be estimated along with the variances of the disturbances $\sigma_{\epsilon\epsilon}, \sigma_{11}, \ldots, \sigma_{MM}$; we are again following the unstandardized-disturbance convention. The unobservable y^* is standardized, as are the observables (although the latter is not at all essential).

We can solve the model into its reduced form by inserting (21) into (22), thus expressing each indicator in terms of the causes and disturbances. The reduced-form equations are

$$\begin{aligned} y_m &= \beta_m \alpha_1 x_1 + \cdots + \beta_m \alpha_K x_K + \beta_m \epsilon + u_m \\ &= \pi_{m1} x_1 + \cdots + \pi_{mK} x_K + v_m \end{aligned} \tag{23}$$

say. Here the reduced-form coefficients are

$$\pi_{mk} = \beta_m \alpha_k \qquad (m = 1, \ldots, M; k = 1, \ldots, K) \tag{24}$$

and the reduced-form disturbances are

$$v_m = \beta_m \epsilon + u_m \qquad (m = 1, \ldots, M) \tag{25}$$

The variances and covariances of the v_m are

$$\omega_{mm} = E(v_m^2) = \beta_m^2 \sigma_{\epsilon\epsilon} + \sigma_{mm} \qquad (m = 1, \ldots, M) \tag{26}$$

and

$$\omega_{mn} = E(v_m v_n) = \beta_m \beta_n \sigma_{\epsilon\epsilon} \quad (m, n = 1, \ldots, M; m \neq n) \tag{27}$$

Note that the v's are not independent of each other since they all have the disturbance ϵ in common.

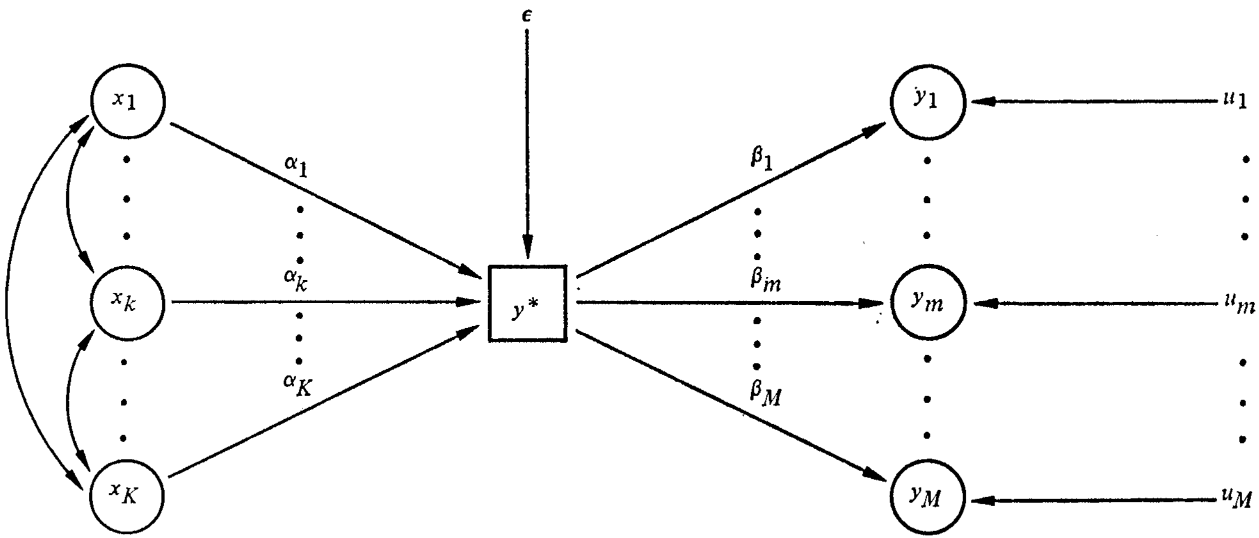

Figure 3. Multiple causes and indicators of an unobservable variable.

A matrix formulation is convenient. We introduce the vectors

$$\begin{aligned} x' &= (x_1, \ldots, x_K) & y' &= (y_1, \ldots, y_M) \\ \alpha' &= (\alpha_1, \ldots, \alpha_K) & \beta' &= (\beta_1, \ldots, \beta_M) \\ & & u' &= (u_1, \ldots, u_M) \end{aligned}$$

and write equations (21) and (22) compactly as

$$y^* = \alpha' x + \epsilon \tag{28}$$

$$y = \beta y^* + u \tag{29}$$

with

$$\Sigma = E(uu') = \begin{pmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{MM} \end{pmatrix}$$

The reduced form is now

$$\begin{aligned} y &= \beta(\alpha' x + \epsilon) + u = \beta\alpha' x + \beta\epsilon + u \\ &= \Pi' x + v \end{aligned} \tag{30}$$

where

$$\Pi = \begin{pmatrix} \pi_{11} & \cdots & \pi_{M1} \\ \vdots & & \vdots \\ \pi_{1K} & \cdots & \pi_{MK} \end{pmatrix} = \begin{pmatrix} \alpha_1\beta_1 & \cdots & \alpha_1\beta_M \\ \vdots & & \vdots \\ \alpha_K\beta_1 & \cdots & \alpha_K\beta_M \end{pmatrix} = \alpha\beta'$$

and

$$v' = (v_1, \ldots, v_M)$$

with

$$\begin{aligned} \Omega &= E(vv') = E(\beta\,\epsilon + u)(\beta\,\epsilon + u)' = \sigma_{\epsilon\epsilon}\beta\beta' + \Sigma \\ &= \begin{pmatrix} \omega_{11} & \cdots & \omega_{1M} \\ \vdots & & \vdots \\ \omega_{M1} & \cdots & \omega_{MM} \end{pmatrix} = \begin{pmatrix} \beta_1^2\sigma_{\epsilon\epsilon} + \sigma_{11} & \cdots & \beta_1\beta_M\sigma_{\epsilon\epsilon} \\ \vdots & & \vdots \\ \beta_M\beta_1\sigma_{\epsilon\epsilon} & \cdots & \beta_M^2\sigma_{\epsilon\epsilon} + \sigma_{MM} \end{pmatrix} \end{aligned} \tag{31}$$

Examination of the reduced form reveals that the model incorporates two sorts of overidentification: (1) The $K \times M$ regression-coefficient matrix Π is expressible as the product of a $K \times 1$ vector α and a $1 \times M$ vector β'. In other words, the $K \times M$ parameters π_{mk} are expressible in terms of only $K + M$ parameters $\alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_M$.

(2) The $\frac{1}{2}M(M + 1)$ distinct elements of the symmetric variance-covariance matrix Ω are expressible in terms of the 1×1 scalar $\sigma_{\epsilon\epsilon}$, the $M \times 1$ vector β, and the M nonzero elements of the diagonal

matrix Σ. In other words, the $\frac{1}{2}M(M+1)$ distinct parameters ω_{mn} are expressible in terms of only $1 + 2M$ parameters $\sigma_{\epsilon\epsilon}, \beta_1, \ldots, \beta_M, \sigma_{11}, \ldots, \sigma_{MM}$. Furthermore, the β's here are the same as under (1).

Taking the two sorts together and allowing for the standardization of y^*, we find that the $(K \times M) + \frac{1}{2}M(M+1)$ distinct reduced-form parameters are expressible in terms of only $2M + K$ distinct structural parameters. This means that the model will typically be overidentified. For example, with $K = 3 = M$ we have $(K \times M) + \frac{1}{2}M(M+1) = 15$ and $2M + K = 9$ so that there are six overidentifying restrictions.

The efficient procedure for estimating the model will have to take account of all these restrictions. Nevertheless, it is instructive to consider the two sorts of overidentification separately.

The first sort of overidentification is of the type dealt with in econometrics (where reduced-form coefficients are combinations of structural coefficients), whereas the second sort of overidentification is of the type dealt with in factor analysis (where covariance matrices are built up from factor loadings, factor variance, and unique variances). Note that ϵ plays the role of a common factor, β the role of factor loadings, and u the role of a unique factor in the expression $v = \beta\epsilon + u$.[7]

The maximum-likelihood principle offers a straightforward approach to efficient estimation of the model, since it takes into account both sorts of overidentifying restrictions. The computation can be performed by Jöreskog's (1970) general method for the analysis of covariance structures.

For present purposes, however, we will be content to consider only the first sort of overidentification. To do so, we simply drop the assumption that the indicator disturbances are mutually independent. In some contexts, no doubt, this is substantively justified and not merely done for the sake of analytical convenience. For example, we might expect to find positively correlated errors among multiple indicators of a single underlying attitude when the indicators are ascertained consecu-

[7] This mixture of econometric and psychometric themes is presumably what Blalock (1969a, pp. 270–272) had in mind in asserting that "Once the basic ideas of each approach [instrumental variables and multiple indicators] have become generally familiar, however, it should become possible to apply them in various combinations to a wide variety of causal models." But, as we have seen in the section "Multiple Indicators of Causally Related Unobservable Variables," the multiple-indicator approach essentially is a factor analysis model. Further, the instrumental-variable approach is simply a particular method for estimation of econometric models. Blalock gives no advice for reconciling the alternative instrumental-variable estimates, nor for reconciling the alternative multiple-indicator estimates, let alone for reconciling both sets. The issues involved are sketched in Goldberger (1971).

tively in a survey interview. If the u's are freely correlated, then the correlations of the v's are no longer patterned as they were in equations (26) and (27) or in equation (31), and the factor-analytic considerations disappear.

In that event, we may as well rewrite the model to make the unobservable variable an *exact* function of its causes, absorbing the disturbance ϵ into the u's and relabelling the latter directly as v's, as in Figure 4. In algebraic form the structural model now reads

$$y^* = \alpha_1 x_1 + \cdot \cdot \cdot + \alpha_K x_K \tag{32}$$

$$y_m = \beta_m y^* + v_m \qquad (m = 1, \ldots, M) \tag{33}$$

with

$$E(v_m^2) = \omega_{mm} \qquad E(v_m v_n) = \omega_{mn} \tag{34}$$

the ω's being unrestricted. Forcing an unobservable variable to be an exact function of its observable causes may seem strange. But, once the disturbances in the indicator equations are allowed to be correlated freely, nothing is gained by retaining a disturbance in the causal equation. Partial correlation among the indicators, controlling on the observable causes, is already present. To put it another way, it would be impossible to distinguish empirically whether the partial correlation was attributable to the common disturbance ϵ or to inherent correlation among the disturbances u. We may as well adopt the latter formulation.[8]

On the understanding that the disturbance variances and covariances are unrestricted, we see the reduced-form system (23) or (30) is just a particularly simple example of the reduced forms which arise in the simultaneous equation models of econometrics (Johnston, 1963, Chapter 7). Indeed, examples of this type have been explicitly

[8] In the two-indicator situation ($M = 2$) our case may be made more strongly. Even if one made the assumption that u_1 and u_2 were uncorrelated, nothing would be lost by dropping ϵ and permitting the v's to be freely correlated; for only two indicators the factor-analysis model is empty. Then, if one insisted on presenting a disturbance in the y^* equation and uncorrelated indicator disturbances, the estimates of the three variances, $\sigma_{\epsilon\epsilon}$, σ_{11}, σ_{22}, could be recovered from unrestricted estimates of the three (co)variances, ω_{11}, ω_{22}, and ω_{12}.

Strictly speaking, even in the $M = 2$ case, we find an exception to our argument that a disturbed y^* equation with uncorrelated indicator disturbances is operationally equivalent to an exact y^* equation with correlated indicator disturbances. The former precludes a correlation between v_m and v_n opposite in sign from the product of β_m and β_n, as can be seen from equation (27); the latter has no such restriction. This exception is a version of the "Heywood case" of factor analysis; compare Harman (1967, pp. 117–118).

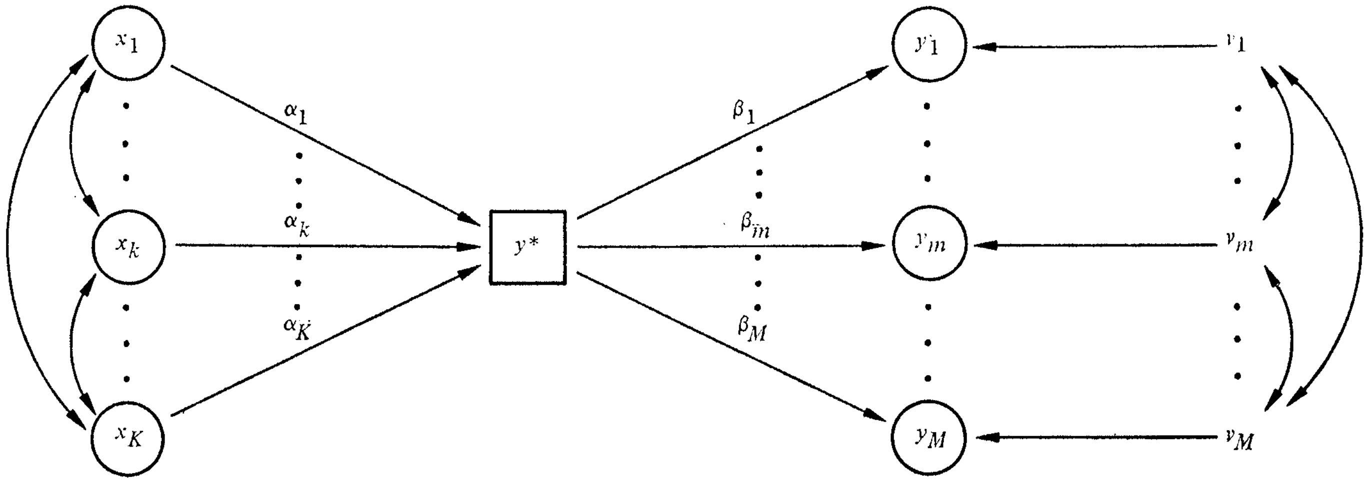

Figure 4. Expression of an unobservable variable as an exact function.

analyzed by Zellner (1970) and Goldberger (1970a). In developing an efficient estimation procedure, we can draw on that literature.

Path Analysis Approach to Estimation

We pause to sketch how a path-analysis approach to fitting the model might proceed. For the sake of concreteness, we take $M = 2$. By inspection of the path diagram, or from equations (32) and (33), the following "estimating equations" are produced:

$$r_{j*} = \sum_{k=1}^{K} \alpha_k r_{jk} \qquad (j = 1, \ldots, K) \tag{35}$$

$$r_{**} = \sum_{j=1}^{K} \sum_{k=1}^{K} \alpha_j \alpha_k r_{jk} = 1 \tag{36}$$

$$r_{jm} = \beta_m r_{j*} \qquad (j = 1, \ldots, K;\ m = 1, 2) \tag{37}$$

$$r_{mn} = \beta_m \beta_n r_{**} + \omega_{mn} \qquad (m, n = 1, 2) \tag{38}$$

Here r_{j*} denotes the correlation of x_j and y^*, r_{jk} the correlation of x_j and x_k, r_{**} the correlation of y^* with itself, r_{jm} the correlation of x_j and y_m, and r_{mn} the correlation of y_m and y_n.

In equations (35) through (38) there are $3K + 4$ equations from which to estimate the $2K + 5$ unknowns, $r_{1*}, \ldots, r_{K*}, \alpha_1, \ldots, \alpha_K, \beta_1, \beta_2, \omega_{11}, \omega_{22}, \omega_{12}$. After solving out the r_{j*} and ω_{mn} via equations (35) and (38) respectively, we find there remain $1 + 2K$ equations from which to estimate the $K + 2$ parameters, $\alpha_1, \ldots, \alpha_K, \beta_1, \beta_2$. Clearly the system is overidentified, there being $K - 1$ excess equations.

In particular, for given r_{j*} there are K distinct estimates of β_1 provided by equation (37), namely

$$b_1^{(j)} = \frac{r_{j1}}{r_{j*}} \qquad (j = 1, \ldots, K) \tag{39}$$

and similarly, there are K distinct estimates of β_2, namely

$$b_2^{(j)} = \frac{r_{j2}}{r_{j*}} \qquad (j = 1, \ldots, K) \tag{40}$$

Even if the model is correct in the population, the distinct estimates will fail to coincide in any sample. One might arbitrarily discard excess equations until a just-determined system obtains which is then solvable for unique estimates. Or, an ad hoc averaging procedure could be adopted (compare Hauser, 1968, pp. 280–287). Thus, equation (37) implies $\beta_1/\beta_2 = r_{j1}/r_{j2}$; so one might estimate β_1/β_2 by

$$\frac{\hat{\beta}_1}{\hat{\beta}_2} = \frac{\sum_{j=1}^{K} r_{j1}}{\sum_{j=1}^{K} r_{j2}} = \frac{\sum_{j=1}^{K} b_1^{(j)}}{\sum_{j=1}^{K} b_2^{(j)}} \tag{41}$$

Then, for given $\hat{\beta}_1$, $\hat{\beta}_2$, there are two distinct estimates of each r_{j*} provided by equation (37), namely

$$r_{j*}^m = \frac{r_{jm}}{\beta_m} \qquad (m = 1,2) \tag{42}$$

which can be averaged into

$$r_{j*} = \frac{r_{j1} + r_{j2}}{\beta_1 + \beta_2} \tag{43}$$

With values of the r_{j*} in hand, the normal equations (35) are then solved for estimates of the α's.[9]

Such averaging procedures are obviously arbitrary since they, in a sense, put equal weight on conflicting estimates. An efficient estimation procedure should take into account the differences in the sampling variabilities of the conflicting estimates.

EFFICIENT ESTIMATION OF MULTIPLE-CAUSE AND MULTIPLE-INDICATOR MODEL

Derivation of the Procedure

Adopting the econometric approach and proceeding to the general M-indicator case, we consider first the estimates of the reduced-form equations obtained by regressing each of the M indicators on all of the observable causes. The normal equations for the typical reduced-form regression equation are

[9] In this description, we have skipped a step in going from $\hat{\beta}_1/\hat{\beta}_2$ to $\hat{\beta}_1$ and $\hat{\beta}_2$. This step is a bit awkward in the present formulation in which the y^* disturbance has been absorbed into the v's. Still, from equation (38) we have $r_{12} = \beta_1\beta_2 + \omega_{12}$ whence

$$\hat{\beta}_1\hat{\beta}_2 = \phi r_{12} \tag{i}$$

where $\phi = 1 - (\omega_{12}/r_{12})$ is temporarily unknown. Combining equations (41) and (i), we have $\hat{\beta}_1$ and $\hat{\beta}_2$ up to a factor of proportionality; then equation (42) gives the $\hat{r}_{j*}$ up to a factor of proportionality. The solution to the normal equations (35) will then estimate the α's up to a factor of proportionality. Finally, the factor of proportionality is determined by equation (36). A more conventional treatment would have $\omega_{12} = 0$, whence $\phi = 1$, giving $\hat{\beta}_1$ and $\hat{\beta}_2$ separately, and estimating $\sigma_{\epsilon\epsilon}$ from $\sigma_{\epsilon\epsilon} = 1 - \Sigma_j\Sigma_k\alpha_j\alpha_k r_{jk}$. The distinction is computational rather than substantive.

$$r_{jm} = \sum_{k=1}^{K} r_{jk}p_{mk} \qquad (j = 1, \ldots, K) \tag{44}$$

where the p_{mk} denote the least-squares regression coefficients. These p's are estimates of the π's but will not satisfy the overidentifying restrictions. According to the model, $\pi_{mk} = \beta_m\alpha_k$; but except by a remarkable coincidence, there will be no set of numbers $b_1, \ldots, b_M, a_1, \ldots, a_K$ such that $p_{mk} = b_m a_k$. Put somewhat differently, the model implies "consistency criteria" such as $\pi_{11}/\pi_{21} = \cdots = \pi_{1k}/\pi_{2k} = \cdots = \pi_{1K}/\pi_{2K}$ (each of these ratios being equal to β_1/β_2), but it will not be true that $p_{11}/p_{21} = \cdots = p_{1k}/p_{2k} = \cdots = p_{1K}/p_{2K}$.

Goldberger (1970a) shows that in the present context maximum-likelihood estimation is identical to modified generalized least-squares estimation. The problem can therefore be posed as follows: Each p_{mk} is an estimate of $\beta_m\alpha_k$; how can we combine them to come up with efficient estimates of the $\beta_1, \ldots, \beta_M, \alpha_1, \ldots, \alpha_K$? In multivariate linear regression models it is well-known that the variances and covariances of the p_{mk} are given by

$$\mathrm{Cov}(p_{mk}, p_{nj}) = \frac{1}{T}\,\omega_{mn}r^{kj}$$

where T is the sample size, ω_{mn} are the elements of Ω, and the r^{kj} are the elements of the matrix inverse to the correlation matrix of the x's. It is also well-known that the ω_{mn} are estimable as the residual variances and covariances s_{mn} from the least-squares regressions. (On these matters, compare Anderson (1958, pp. 178–183) or Goldberger (1964, pp. 207–209).) In view of this, the MGLS procedure calls for estimates of the α's and β's to be obtained as follows: choose the values $\alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_M$ which minimize

$$\frac{1}{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{M}\sum_{m=1}^{M} s^{mn}r_{kj}(p_{mk} - \beta_m\alpha_k)(p_{nj} - \beta_n\alpha_j) \tag{45}$$

where the s^{mn} are the elements of the matrix inverse to the matrix of the s_{mn}. In expression (45) the weight attached to the term involving p_{mk} and p_{nj} is inverse to the estimated covariance of p_{mk} and p_{nj} as called for by the MGLS procedure.

In carrying out the minimization, we find a matrix formulation convenient. Let $X'X$ denote the $K \times K$ matrix of the r_{jk}, $X'Y$ the $K \times M$ matrix of the r_{jm}, and $Y'Y$ the $M \times M$ matrix of the r_{mn}. Further, let

$$P = \begin{pmatrix} p_{11} & \cdots & p_{M1} \\ \vdots & & \vdots \\ p_{1K} & \cdots & p_{MK} \end{pmatrix}$$

be the $K \times M$ matrix of the p_{mk}. The normal equations (44) are compactly expressed as

$$X'XP = X'Y$$

and their solution as

$$P = (X'X)^{-1}X'Y \tag{46}$$

Further, let

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1M} \\ \vdots & & \vdots \\ s_{M1} & \cdots & s_{MM} \end{pmatrix}$$

be the $M \times M$ matrix of the s_{mn}; then

$$S = (Y - XP)'(Y - XP) = Y'Y - Y'X(X'X)^{-1}X'Y \tag{47}$$

The formidable expression (45) can now be compactly written as

$$\frac{1}{T}\,\mathrm{tr}[S^{-1}(P - \alpha\beta')'X'X(P - \alpha\beta')] \tag{48}$$

A simple manipulation shows that the trace of the $M \times M$ matrix in brackets is, apart from an irrelevant constant, equal to the scalar

$$(\alpha'X'X\alpha)(\beta'S^{-1}\beta) - 2\alpha'X'YS^{-1}\beta \tag{49}$$

The MGLS principle thus chooses α and β to minimize expression (49), or rather, if we recall the standardization of y^* as in equation (36), to minimize expression (49) subject to

$$\alpha'X'X\alpha = 1 \tag{50}$$

To minimize expression (49) subject to equation (50), one first forms the expression

$$f = (\alpha'X'X\alpha)(\beta'S^{-1}\beta) - 2\alpha'X'YS^{-1}\beta + \lambda(\alpha'X'X\alpha - 1) \tag{51}$$

where λ is a Lagrangean multiplier and then differentiates with respect to β and α to find

$$\frac{\partial f}{\partial \beta} = (\alpha'X'X\alpha)2S^{-1}\beta - 2S^{-1}Y'X\alpha \tag{52}$$

$$\frac{\partial f}{\partial \alpha} = (\beta'S^{-1}\beta)2X'X\alpha - 2X'YS^{-1}\beta + 2\lambda X'X\alpha \tag{53}$$

Setting equation (52) at zero, using equation (50), and introducing

$$a' = (a_1, \ldots, a_K) \qquad b' = (b_1, \ldots, b_M)$$

as the symbols for the estimates of α and β, we find

$$b = Y'Xa \tag{54}$$

Setting equation (53) at zero and using equations (50) and (54), we find that $\lambda = 0$ and, thus, that

$$a = (b'S^{-1}b)^{-1}PS^{-1}b \tag{55}$$

Then, inserting equation (55) into equation (54), we find

$$b = (b'S^{-1}b)^{-1}Y'XPS^{-1}b = (b'S^{-1}b)^{-1}QS^{-1}b \tag{56}$$

where

$$Q = Y'XP = Y'X(X'X)^{-1}X'Y = P'X'XP = Y'Y - S$$

is the matrix of regression moments. What equation (56) says is that

$$(QS^{-1} - \mu I)b = 0 \tag{57}$$

where $\mu = b'S^{-1}b$. In other words, b is a characteristic vector of the matrix QS^{-1}. It is not hard to show that b should be a vector corresponding to the *largest* characteristic root μ (in order to *minimize* the trace) and that it should be normalized by $b'S^{-1}b = \mu$ (in order to ensure $a'X'Xa = 1$).[10] With this value for b in hand, the value for a follows from (55).

The efficient estimates for α and β can, in short, be obtained by solving a characteristic root-characteristic vector problem of a type which is prevalent throughout multivariate statistical analysis. Standard computer programs can be adapted for this purpose; a desk calculator will suffice if M is no larger than three or four, once the output of least-squares regressions is available. As shown in the Appendix, the computations are intimately related to those of canonical correlation.

[10] When b satisfies equation (57), then premultiplication by $b'S^{-1}$ shows that $b'S^{-1}QS^{-1}b = \mu b'S^{-1}b$; so that when a is computed from equation (55) as

$$a = (b'S^{-1}b)^{-1}PS^{-1}b,$$

we will have $a'X'YS^{-1}b = (b'S^{-1}b)^{-1}b'S^{-1}P'X'YS^{-1}b = (b'S^{-1}b)^{-1}b'S^{-1}QS^{-1}b = \mu$, and $a'X'Xa = a'X'X(b'S^{-1}b)^{-1}PS^{-1}b = (b'S^{-1}b)^{-1}a'X'YS^{-1}b = (b'S^{-1}b)^{-1}\mu$. To make $a'X'Xa = 1$, therefore, we must normalize b according to $(b'S^{-1}b) = \mu$. With these values inserted, expression (49) becomes $(a'X'Xa)(b'S^{-1}b) - 2a'X'YS^{-1}b = 1\mu - 2\mu = -\mu$; since we're minimizing this expression, the desired root is the largest one.

Numerical Illustration

To illustrate the efficient estimation procedure, we draw on Hodge and Treiman's (1968) study of social participation and social status. The sample consists of approximately 530 adult female residents of a Washington, D.C. suburban county. The observed variables are (original symbols follow in parentheses): x_1 = family income (I), x_2 = main earner's occupation (O), x_3 = respondent's education (E), y_1 = frequency of church attendance (C), y_2 = number of voluntary organization memberships (V), and y_3 = number of friends seen (F). The observed correlations given in Hodge and Treiman (1968, Table 2) are presented here in Table 4.

Table 4
Correlations of Status and Participation Variables

$$\begin{pmatrix} X'X & X'Y \\ & Y'Y \end{pmatrix}$$

	x_1	x_2	x_3	y_1	y_2	y_3
x_1	1.0000	0.3040	0.3049	0.1000	0.2835	0.1762
x_2		1.0000	0.3444	0.1561	0.1925	0.1357
x_3			1.0000	0.1580	0.3235	0.2255
y_1				1.0000	0.3601	0.2099
y_2					1.0000	0.2654
y_3						1.0000

The results of unconstrained multiple regression are presented in Table 5. This, in effect, is the estimated model displayed in Hodge and Treiman (1968, Figure 1b); the elements in our P will be found there as paths from causes to indicators, while the elements in our S, converted

Table 5
Results of Unconstrained Multiple Regressions

P = Regression Coefficients

	y_1	y_2	y_3
x_1	0.0335	0.1932	0.1094
x_2	0.1078	0.0484	0.0411
x_3	0.1107	0.2479	0.1780

	$Q = P'X'XP$ = Regression Moments			$S = Y'Y - Q$ = Residual Moments		
	y_1	y_2	y_3			
y_1	0.0377	0.0660	0.0455	0.9623	0.2941	0.1644
y_2		0.1443	0.0965		0.8557	0.1689
y_3			0.0650			0.9350

into standard deviations and correlations, will be found there as residual paths and correlations.

In our model, shown in Figure 5, the influence of status on participation is assumed to be transmitted through a single unobservable variable, y^* = socioeconomic status. The MGLS estimates a and b are reported in Table 6, along with $\hat{\Pi} = ab'$, which is our implied estimate of the compound paths from causes to indicators, and

$$\hat{\Omega} = (Y - X\hat{\Pi})'(Y - X\hat{\Pi}),$$

the matrix of residual moments from the constrained regressions. Converting the elements of $\hat{\Omega}$ into standard deviations and correlations gives the residual paths and correlations displayed in Figure 5.[11]

Our implied estimates in $\hat{\Pi}$ naturally differ from the unconstrained estimates in P; the latter, after all, did not satisfy the overidentifying restrictions. The differences, however, are generally small, which suggests that the unobservable-variable model may be appropriate. (Equivalently, one could compare $X'X\hat{\Pi}$ with $X'Y$ to see how closely our model reproduces the correlations between the x's and the y's.) More to the point is the fact that the diagonal elements of $\hat{\Omega}$ are only slightly larger than the corresponding diagonal elements of S, which suggests that the fit does not deteriorate much when the overidentifying restrictions are imposed.

Table 6
Estimates for Causal Model of Figure 5

	b	a	$\hat{\Pi} = ab'$ = Constrained Regression Coefficient Matrix	y_1	y_2	y_3
x_1	0.1761	0.4815	x_1	0.0848	0.1827	0.1226
x_2	0.3795	0.1476	x_2	0.0260	0.0560	0.0376
x_3	0.2546	0.6638	x_3	0.1169	0.2519	0.1690

$\hat{\Omega} = (Y - X\hat{\Pi})'(Y - X\hat{\Pi})$ = Implied Residual Moments

y_1	0.9690	0.2933	0.1651
y_2		0.8560	0.1688
y_3			0.9352

[11] Starting with the information in Hodge and Treiman (1968), our efficient estimation required about two hours on a desk calculator. The largest root of QS^{-1}, along with the suitably normalized characteristic vector b, was extracted by a standard iterative procedure (compare Morrison, 1967, pp. 234–248). The largest root is $\mu = 0.205$, so that the first canonical correlation between the y's and the x's is $0.41 = \sqrt{0.205/1.205}$ (compare the Appendix). When equations (54) through (57) are taken into account, it turns out that $\hat{\Omega} = (Y - X\hat{\Pi})'(Y - X\hat{\Pi}) = Y'Y - bb'$, which facilitates calculations.

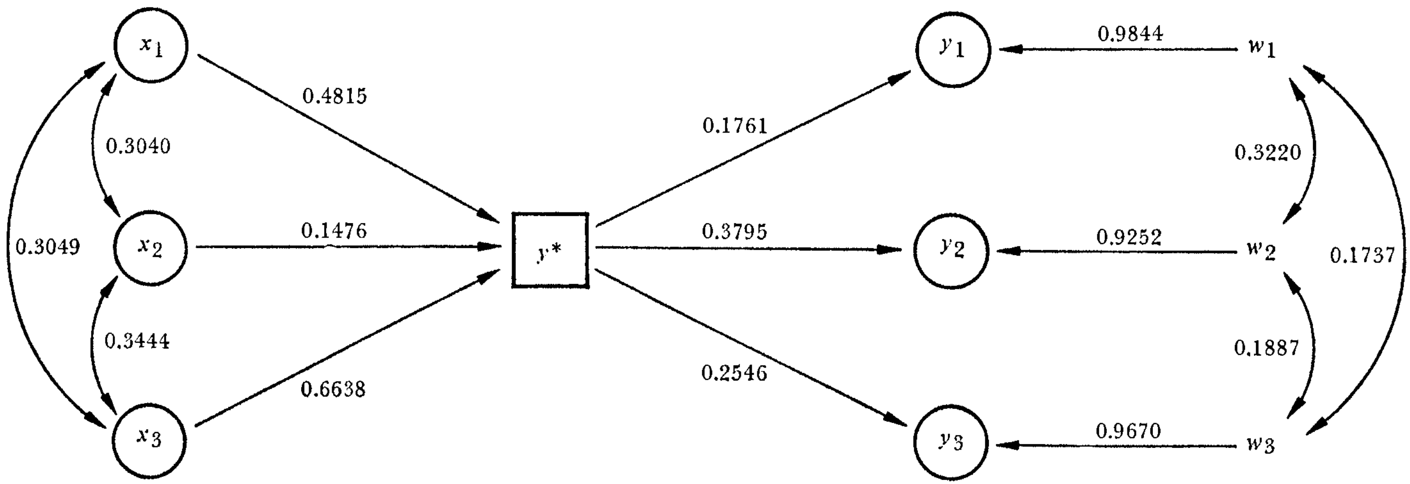

Figure 5. Causal model of socioeconomic status and participation.

To translate such remarks into a formal test of the causal model, we simply draw on the likelihood-ratio test of multivariate analysis (compare Anderson, 1958, Chapter 8). The relevant statistic is $T \log [\det(\hat{\Omega})/\det(S)]$. On the null hypothesis that the overidentifying restrictions are correct, this statistic is distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions. In our illustration we have four restrictions $[4 = (K \times M) - (K + M - 1)]$, a sample size of 530, and the determinants are $\det(\hat{\Omega}) = 0.6549$ and $\det(S) = 0.6607$. This gives a test statistic of 4.5, which is not significant at the 10 per cent level (nor even at the 30 per cent level). Routine procedures of statistical inference, therefore, would not lead us to reject the causal model.

Comments

We can sketch an interpretation of the efficient parameter estimates in terms of averages of conflicting estimates. Taking for example our efficient estimate of β_1, we find from equation (54) that

$$b_1 = \sum_{j=1}^{K} r_{j1} a_j$$

where the a's are our efficient estimates of the α's. Defining

$$\hat{r}_{j*} = \sum_{k=1}^{K} a_k r_{jk} \tag{58}$$

as our efficient estimates of the correlations between x_j and y^* [compare equation (35)] we can rewrite b_1 as

$$b_1 = \sum_{j=1}^{K} \left(\frac{r_{j1}}{\hat{r}_{j*}}\right) a_j \hat{r}_{j*}$$

Recalling equation (39), we write

$$\hat{b}_1^{(j)} = \frac{r_{j1}}{\hat{r}_{j*}} \qquad (j = 1, \ldots, K)$$

which are K conflicting estimates of β_1. Multiplying each r_{j*} in equation (58) by a_j and summing gives

$$\sum_{j=1}^{K} a_j \hat{r}_{j*} = \sum_{j=1}^{K} a_j \sum_{k=1}^{K} a_k r_{jk} = \sum_{j=1}^{K} \sum_{k=1}^{K} a_j a_k r_{jk} = 1$$

in view of $a'X'Xa = 1$ [compare equation (36)]. It follows that

$$b_1 = \sum_{j=1}^{K} b_1^{(j)} w^{(j)}$$

where $w^{(j)} = a_j \hat{r}_{j*}$ and $\Sigma_{j=1}^{K} w^{(j)} = 1$. Thus, b_1 is indeed a weighted average of the conflicting estimates $b_1^{(j)}$.

More generally,

$$b_m = \sum_{j=1}^{K} \hat{b}_m^{(j)} w^{(j)} \qquad (m = 1, \ldots, M)$$

where

$$\hat{b}_m^{(j)} = \frac{r_{jm}}{\hat{r}_{j*}} \qquad (m = 1, \ldots, M)$$

It is important to note that the weights $w^{(j)}$ are not determined in advance. Rather, like the r_{j*}, they involve the estimates of the α's and, hence, fall out as an incidental part of the efficient estimation computation.

To illustrate the interpretation, Table 7 presents the r_{j*} (elements of $X'Xa$), the $b_m^{(j)}$, and the efficient estimates b_m. A similar weighted-average interpretation can be made for the efficient estimates of the α's but will not be demonstrated here.

Table 7
Conflicting and Efficient Estimates of β_1, β_2, β_3

j	$\hat{r}_{j*}$	$b_1^{(j)}$	$b_2^{(j)}$	$b_3^{(j)}$
1	0.7287	0.1372	0.3890	0.2418
2	0.5226	0.2987	0.3683	0.2597
3	0.8614	0.1834	0.3756	0.2618
		$b_1 = 0.1761$	$b_2 = 0.3795$	$b_3 = 0.2546$

The present model might be extended by introducing additional observable variables $z_1, \ldots, z_I$ as direct causes of the observable indicators (compare Zellner, 1970, p. 442). That is, equation (33) might be replaced by

$$y_m = \beta_m y^* + \gamma_{m1} z_1 + \cdots + \gamma_{mI} z_I + v_m \qquad (m = 1, \ldots, M)$$

The reduced-form equations would then express the indicators in terms of the x's, z's, and v's. It is not hard to see that only a portion of the reduced-form coefficient matrix would be restricted by the structural model. The restrictions would be precisely of the form that arises in the econometrician's "limited-information" analysis of a single structural equation of a simultaneous equation model (compare Johnston, 1963, pp. 254–258 or Goldberger, 1964, pp. 338–345). As shown by Goldberger and Olkin (1971), the maximum-likelihood and modified generalized least-squares procedures again yield identical parameter estimates in such situations.

CONCLUSION

In this attempt to spell out procedures for efficient estimation of overidentified unobservable-variable models, we have considered only two simple models in detail. Clearly we have not provided a comprehensive guidebook for the treatment of path models containing unobservable variables. But we think that we have gone far enough to indicate that such a guidebook is feasible. All the models of path analysis are, after all, subsumed under the general linear model of statistics, so the standard principles of statistical inference and the multivariate estimation and testing methods which they entail are relevant. There is no need for a special path-analytic theory of fitting models.

APPENDIX

The estimation procedure in our numerical illustration of the model with multiple causes and multiple indicators has an interpretation in terms of canonical correlation, suggested to us by O. D. Duncan and by H. W. Watts. Blalock (1969b, pp. 42–43) has also discussed the structure of the proportionally constrained regression model with multiple indicators of the dependent variable and recognized its similarity to canonical correlation. Given a set of variables $y_1, \ldots, y_M$ and a set of variables $x_1, \ldots, x_K$, canonical correlation analysis yields the linear combination of the y's, say $\bar{y} = \Sigma_{m=1}^{M} d_m y_m$, and the linear combination of the x's, say $\bar{x} = \Sigma_{k=1}^{K} c_k x_k$, which are most highly correlated with one another (compare Morrison, 1967, Chapter 6). Without loss of generality $\bar{y}$ and $\bar{x}$ are taken to be standardized. If

$$d' = (d_1, \ldots, d_M) \qquad c' = (c_1, \ldots, c_K)$$

it can be shown that d is chosen to maximize

$$\frac{d'Y'X(X'X)^{-1}X'Yd}{d'Y'Yd} = \frac{d'Qd}{d'(Q+S)d}$$

This leads to the characteristic equation

$$(Q - \lambda S)d = 0 \tag{59}$$

with the largest root λ being the required one and with the standardization $d'Y'Yd = 1$ being imposed. Now, equation (59) is equivalent to

$$(QS^{-1} - \lambda I)Sd = 0 \tag{60}$$

and $d'Y'Yd = 1$ is equivalent to $d'Sd = 1/(1 + \lambda)$. Comparing equation (60) with equation (57), recalling that $b'S^{-1}b = \mu$, and recognizing that $\mu = \lambda$, we conclude that

$$b = \sqrt{\lambda(1 + \lambda)}\, Sd \tag{61}$$

Furthermore, in canonical-correlation analysis it is shown that

$$c = \sqrt{(1 + \lambda)/\lambda}\, Pd \tag{62}$$

Comparing equation (62) with equation (55) and using equation (61), we conclude that

$$a = c \tag{63}$$

Thus, our efficient estimates of α and β can be obtained from the c and d of canonical correlation.

If we pursue the point, it follows from equation (63) that the canonical "independent" variable, $\tilde{x} = \Sigma_k c_k x_k$, is identical with the constructed unobservable variable, $\hat{y}^* = \Sigma_k a_k x_k$, implied by our estimates of the α's. Further, it follows from equations (59) through (62) that

$$\begin{aligned} Y'\tilde{x} &= Y'Xc = \sqrt{(1 + \lambda)/\lambda}\, Y'XPd = \sqrt{(1 + \lambda)/\lambda}\, Qd \\ &= \sqrt{(1 + \lambda)/\lambda}\, \lambda Sd = \sqrt{\lambda(1 + \lambda)}\, Sd \\ &= b \end{aligned}$$

which means that the correlation of each indicator with the canonical "independent" variable (that is, with the constructed $\hat{y}^*$) gives our estimate of the path from y^* to that indicator. Alternatively, it can be shown that

$$b = \sqrt{\lambda/(1 + \lambda)}\, Y'\tilde{y}$$

Our estimated paths are proportional to the correlations of indicators with the canonical "dependent" variable $\tilde{y} = Yd$. The factor of proportionality arises from the fact that $\tilde{y}$ and $\tilde{x}$ are not identical; their correlation is just

$$\tilde{y}'\tilde{x} = d'X'Yc = d'b = \sqrt{\lambda(1 + \lambda)}\, d'Sd = \sqrt{\lambda/(1 + \lambda)}$$

which is the so-called first canonical correlation coefficient.

In summary, a canonical-correlation computer program can be adapted to calculate the parameter estimates for the model in the section "efficient estimation of multiple-cause and multiple-indicator model." In more elaborate unobservable-variable models, however, there is no presumption that the efficient estimates can be deduced from the output of canonical correlation.

REFERENCES

ANDERSON, T. W.

1958 *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

BLALOCK, H. M., JR.

1961 *Causal Inferences in Nonexperimental Research.* Chapel Hill: University of North Carolina Press.

1969a "Multiple indicators and the causal approach to measurement error." *American Journal of Sociology* 75: 264–272.

1969b *Theory Construction: From Verbal to Mathematical Formulations.* Englewood Cliffs, N.J.: Prentice-Hall.

1970 "Estimating measurement error using multiple indicators and several points in time." *American Sociological Review* 35: 101–111.

BOUDON, R.

1965 "A method of linear causal analysis: Dependence analysis." *American Sociological Review* 30: 365–374.

1968 "A new look at correlation analysis." Pp. 199–235 in H. M. Blalock, Jr. and A. B. Blalock (Eds.), *Methodology in Social Research.* New York: McGraw-Hill.

BREWER, M. B., CRANO, W. D., AND CAMPBELL, D. T.

1970 "Testing a single-factor model as an alternative to the misuse of partial correlations in hypothesis-testing research." *Sociometry* 33: 1–11.

CHRIST, C. F.

1966 *Econometric Models and Methods.* New York: Wiley.

COSTNER, H. L.

1969 "Theory, deduction and rules of correspondence." *American Journal of Sociology* 75: 245–263.

DUNCAN, O. D.

1966 "Path analysis: Sociological examples." *American Journal of Sociology* 72: 3–16.

1969a "Some linear models for two-wave, two-variable panel analysis." *Psychological Bulletin* 72: 177–182.

1969b "Contingencies in constructing causal models." Pp. 74–112 in E. F. Borgatta (Ed.), *Sociological Methodology: 1969*. San Francisco: Jossey-Bass.

DUNCAN, O. D., FEATHERMAN, D. L., AND DUNCAN, B.

1968 *Socioeconomic Background and Occupational Achievement: Extensions of a Basic Model.* Final Report No. 5-0074 (EO-191), Contract No. OE-5-85-072. Washington, D.C.: Bureau of Research, Office of Education, Department of Health, Education, and Welfare.

DUNCAN, O. D., HALLER, A. O., AND PORTES, A.

1968 "Peer influences on aspirations: A reinterpretation." *American Journal of Sociology* 74: 119–137.

GOLDBERGER, A. S.

1964 *Econometric Theory.* New York: Wiley.

1970a "Criteria and constraints in multivariate regression." Workshop Paper EME 7026. Social Systems Research Institute, Madison: University of Wisconsin.

1970b "On Boudon's method of linear causal analysis." *American Sociological Review* 35: 97–101.

1971 "Econometrics and psychometrics: A survey of communalities." *Psychometrika* 36: 83–107.

GOLDBERGER, A. S. AND OLKIN, I.

1971 "A minimum-distance interpretation of limited-information estimation." *Econometrica* 39: 635–639.

HARMAN, H. H.

1967 *Modern Factor Analysis.* (2nd ed.). Chicago: University of Chicago Press.

HAUSER, R. M.

1968 Family, School and Neighborhood Factors in Educational Performances in a Metropolitan School System. University of Michigan (unpublished doctoral dissertation).

1969a "Schools and the stratification process." *American Journal of Sociology* 74: 587–611.

1969b "On 'social participation and social status.'" *American Sociological Review* 34: 549–554.

1970 "Educational stratification in the United States." *Sociological Inquiry* 40: 102–129.

HEISE, D. R.

1969 "Problems in path analysis and causal inference." Pp. 38–73 in E. F. Borgatta (Ed.), *Sociological Methodology: 1969*. San Francisco: Jossey-Bass.

HODGE, R. W. AND TREIMAN, D. J.
1968 "Social participation and social status." *American Sociological Review* 33: 723–740.

JOHNSTON, J.
1963 *Econometric Methods.* New York: McGraw-Hill.

JÖRESKOG, K. G.
1967 "RMLFA: A computer program for restricted maximum likelihood factor analysis." Research Memorandum RM-67-21. Princeton, N.J.: Educational Testing Service.
1969a "A general approach to confirmatory maximum likelihood factor analysis." *Psychometrika* 34: 183–202.
1969b "Factoring the multitest-multioccasion correlation matrix." Research Bulletin RB-69-62. Princeton, N.J.: Educational Testing Service.
1970 "A general method for analysis of covariance structures." *Biometrika* 57: 239–251.

JÖRESKOG, K. G. AND LAWLEY, D. N.
1968 "New methods in maximum likelihood factor analysis." *British Journal of Mathematical and Statistical Psychology* 21: 85–96.

LAND, K. C.
1969 "Principles of path analysis." Pp. 3–37 in E. F. Borgatta (Ed.), *Sociological Methodology: 1969.* San Francisco: Jossey-Bass.
1970 "On the estimation of path coefficients for unmeasured variables from correlations among observed variables." *Social Forces* 48: 506–511.

LAWLEY, D. N. AND MAXWELL, A. E.
1963 *Factor Analysis as a Statistical Method.* London: Butterworths.

MORRISON, D. F.
1967 *Multivariate Statistical Methods.* New York: McGraw-Hill.

SIEGEL, P. M. AND HODGE, R. W.
1968 "A causal approach to the study of measurement error." Pp. 28–59 in H. M. Blalock, Jr. and A. B. Blalock (Eds.), *Methodology in Social Research.* New York: McGraw-Hill.

WERTS, C. E. AND LINN, R. L.
1970 "Path analysis: Psychological examples." *Psychological Bulletin* 74: 193–212.

WILEY, D. E. AND WILEY, J. A.
1970 "The estimation of measurement error in panel data." *American Sociological Review* 35: 112–117.

WRIGHT, S.
1925 "Corn and hog correlations." Department Bulletin No. 1300. Washington, D.C.: Department of Agriculture.
1934 "The method of path coefficients." *The Annals of Mathematical Statistics* 5: 161–215.
1954 "The interpretation of multivariate systems." Pp. 11–33 in O. Kempthorne, *et al.* (Eds.), *Statistics and Mathematics in Biology.* Ames, Iowa: Iowa State College Press.

ZELLNER, A.
1970 "Estimation of regression relationships containing unobservable variables." *International Economic Review* 11: 441–454.