DOCUMENT RESUME

ED 166 204 TM 007 714

AUTHOR	Wolfle, Lee M.
TITLE	Strategies of Path Analysis.
PUB DATE	Mar 78
NOTE	37p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Correlation; *Critical Path Method; Data Analysis; *Mathematical Models; *Multiple Regression Analysis; Predictor Variables; *Research Design; Research Methodology; Suppressor Variables; Validity

ABSTRACT

The purpose of this paper is to illuminate the advantages of path analysis for the exposition of results in data analytic papers. Probably the greatest advantage is that it provides a means by which the nature of the problem may be handily summarized. The method of path analysis, although conceived over sixty years ago by Sewell Wright, has only recently been introduced into educational literature. Its application in substantive analyses in education has also been infrequent. Path analysis requires the researcher to think about systems of intercausal connections and requires a degree of explicitness desirable in scientific writing. Path models may be classified into four categories: recursive, block, block-recursive, and nonrecursive. Each of these requires a somewhat different strategy of analysis. The strategy for each kind of model is illustrated with substantive examples. The more difficult task of path analysis involves the construction of models that are consistent with sound theory. (Author/CTM)

STRATEGIES OF PATH ANALYSIS

Lee M. Wolfe

Office of Research and Evaluation
College of Education
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

*Prepared for the meetings of the American Educational Research Association, Toronto, Ontario, March 1978. Printed in U.S.A.

Session 25.12

STRATEGIES OF PATH ANALYSIS

The method of path analysis, although conceived over sixty years ago by Sewell Wright (1921, 1925), has only recently been introduced into the educational literature (Anderson and Evans, 1974; Wolfle, 1977). Its application in substantive analyses in education has also been infrequent, although sociologists have been using the method for over a decade to analyze educational attainment and its consequences (Blau and Duncan, 1967; Hauser, 1971; Duncan, Featherman and Duncan, 1972; Sewell and Hauser, 1975; Sewell, Hauser and Featherman, 1976). Several papers have also summarized its methodological aspects; recent reviews include Goldberger (1972), Goldberger and Duncan (1973), Duncan (1975), Heise (1975), and Bielby and Hauser (1977). These reviews obviate the need for an extensive review of the principles of path analysis.[1] This paper assumes the reader's general familiarity with multiple regression analysis, and its similarity to path analysis. Its purpose is to illuminate the advantages of path analysis for the exposition of results in data analytic papers.

A data analytic paper typically begins with a discussion that leads to an unresolved problem. Probably the greatest advantage of path analysis is that it provides a means by which the nature of the problem may be handily summarized. It requires the researcher to think about cause, particularly systems of intercausal connections, and provides an explicit link between *a priori* theoretical notions of causal connections and estimates of causal impact. The formulation of the model requires a degree of explicitness desirable in scientific writing, while requiring the researcher to defend the proposed model.

Once the model is formulated and defended, it provides a systematic guide for the exposition of results. The analysis proceeds in accordance with the logic of the model. First, quantitative estimates are attached to causal effects thought to exist in the proposed model. But what the researcher can do next in the analytic process depends on the kind of model originally formulated. Path models may be classified into four categories: recursive, block, block-recursive, and nonrecursive. Each of these requires a somewhat different strategy of analysis. The remainder of this paper presents the strategy of analysis for each kind of model, with substantive examples.

RECURSIVE EQUATION MODELS

Recursive equation models have no feedback loops, either directly or indirectly; that is, the causal flow in the model is unidirectional. The quantitative estimates attached to the causal links may be either standardized or metric regression coefficients. In many cases the latter are preferred (see Duncan, 1975:51-66; Kim and Mueller, 1976), but for the instructive purposes of this paper, standardized coefficients will be used. Zero-order associations in models using standardized coefficients are measured by correlation coefficients. These may be decomposed into several components using the fundamental theorem of path analysis, which may be written (Duncan, 1966:5):

$$r_{ij} = \sum_q p_{iq} r_{jq} \qquad (1),$$

where i and j denote two variables in the model, and the index q runs over all variables from which direct paths lead to X_i. This decomposition provides the basis for interpreting the results, and the kind of model dictates the interpretations that can be attached to the various components of the decomposition.

The strategy of analysis for recursive models is basically twofold. First, the researcher will want to obtain estimates of the extent to which intervening variables account for relationships among variables. These may be interpreted as indirect causal effects. Second, the researcher will want to obtain estimates of the extent to which antecedent variables account for relationships between other variables. These may be interpreted as spurious effects.

The best way to illustrate a strategy of analysis is to analyze a substantive example. Consider the following set of structural equations:

$$\begin{aligned} x_1 &= p_{12}\,x_2 + p_{13}\,x_3 + p_{15}\,x_5 + p_{16}\,x_6 + p_{17}\,x_7 + p_{1u}\,u \\ x_2 &= p_{23}\,x_3 + p_{24}\,x_4 + p_{25}\,x_5 + p_{2v}\,v \\ x_3 &= p_{34}\,x_4 + p_{35}\,x_5 + p_{36}\,x_6 + p_{37}\,x_7 + p_{38}\,x_8 + p_{3w}\,w \qquad (2). \end{aligned}$$

Under the usual assumptions, the variables on the left-hand side of the equality are dependent upon those to the right. The p_{ij}'s are standardized regression coefficients, known as path coefficients, in which the subscript i denotes the dependent variable. The residuals have expected values of zero, and expected covariations of zero with the other explicitly measured independent variables in the equation, and with each other. The three equations may be diagramed, as shown in Figure 1.

In Figure 1 the straight arrows represent my hypothesis of causal effects; the arrowheads point toward the influenced variables. The arrow from X_4 to X_3, for example, represents the verbal statement, "childhood intelligence is a cause of educational attainment," or "a change in childhood intelligence produces a change in educational attainment." A double-headed arrow represents a correlation between two variables. No causal interpretation is attached to a correlation. In this case, the five variables on the left-hand side of the diagram are called exogenous

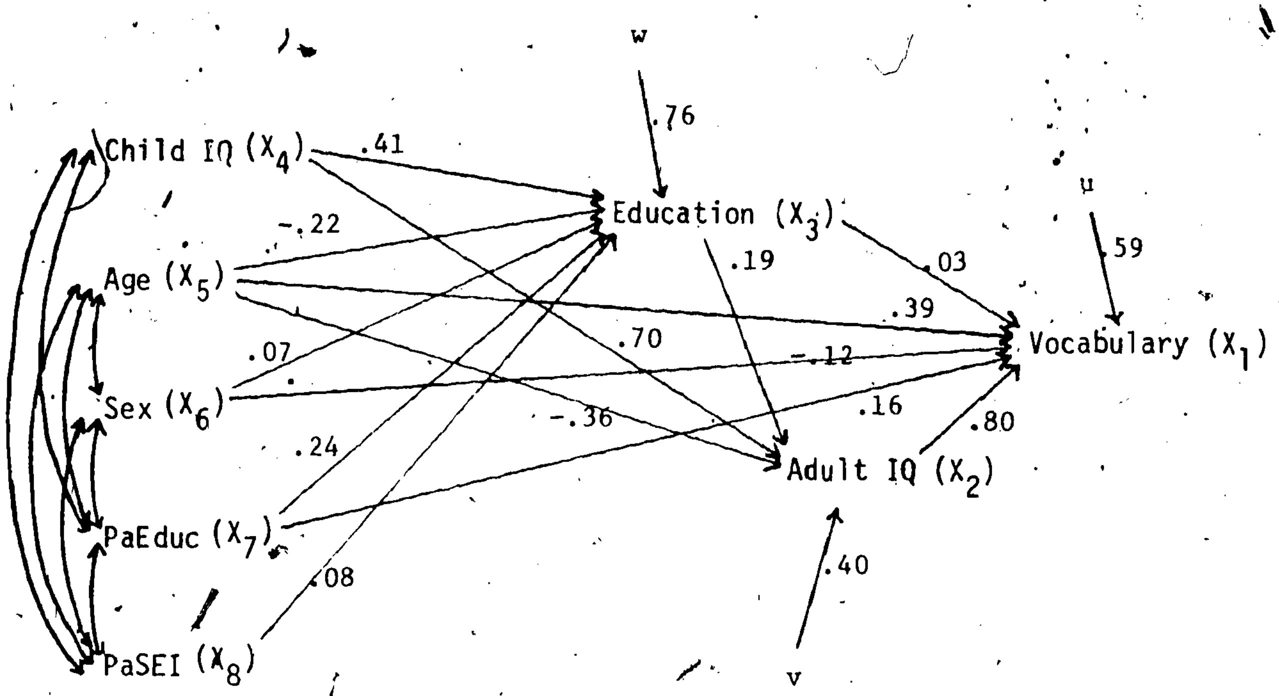


Figure 1. Recursive Path Model of Vocabulary Recognition, U.S. White Civilian Population, Age 25 to 72

variables, because the causes of these variables, whatever they are, come from outside the model. These variables, it is recognized, may be correlated, but for unknown reasons. However, two correlations are specified to be zero *a priori*, namely those between childhood IQ and sex, and childhood IQ and current age.

A fully recursive model is one in which every antecedent variable is allowed to affect every other variable subsequent to it in the model. The model presented in Figure 1 is not fully recursive, because some of the possible causal connections have been eliminated from the model. Some of these were eliminated for *a priori* theoretical reasons. For example, the path from childhood IQ to vocabulary recognition was assumed to be nonexistent. If IQ affects vocabulary it must do so at the time of the test; a person's past IQ has no direct bearing. The paths from father's educational attainment and father's socioeconomic index to adult intelligence were similarly eliminated; the effects were assumed to exist in childhood, but not directly for adult IQ. Other paths were eliminated because the coefficients estimated from sample data failed tests of statistical significance. I therefore inferred that the direct causal path did not exist in the population, removed the path from the model, and recomputed the coefficients of the remaining paths. The missing path from sex to adult IQ, for example, was eliminated because its coefficient was insignificant.

The numbers along each path are standardized regression coefficients obtained by regressing each criterion variable on the variables thought to be causes of it. The data were obtained by combining pieces of evidence from diverse sources. Pending the completion of comprehensive and long-term longitudinal studies of truly representative samples, piecing together correlations from a variety of sources was the only way to obtain estimates for this model. Most of the correlations were obtained from the National Opinion Research Center's 1976 general social survey. The correlations of variables with the two IQ measures were pieced together.[2]

Remember that the strategy of analysis for recursive models is, first of all, to obtain estimates of the extent to which intervening variables account for relationships among variables. The relationship between age (X_5) and vocabulary (X_1) will be used for illustration. Applying the fundamental theorem of path analysis, it can be seen that:

$$r_{15} = p_{12}\, r_{25} + p_{13}\, r_{35} + p_{15}\, r_{55} + p_{16}\, r_{56} + p_{17}\, r_{57} \qquad (3).$$

The correlation of a variable with itself is unity ($r_{55} = 1.0$), and the correlations r_{56} and r_{57} are given as exogenous. The correlations r_{25} and r_{35} may, however, in their turn be decomposed. Thus,

$$r_{25} = p_{23}\, r_{35} + p_{24}\, r_{45} + p_{25}\, r_{55} \qquad (4),$$

$$r_{35} = p_{34}\, r_{45} + p_{35}\, r_{55} + p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58} \qquad (5).$$

Equations 4 and 5 may be simplified by remembering that $r_{55} = 1.0$ by definition, and $r_{45} = 0$ by theoretical assumption.[3] Thus,

$$r_{25} = p_{23}\, r_{35} + p_{25} \qquad (6),$$

$$r_{35} = p_{35} + p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58} \qquad (7).$$

The right-hand portion of equation 7 may be substituted into equation 6 to yield:

$$r_{25} = p_{23}\,(p_{35} + p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) + p_{25} \qquad (8).$$

The right-hand portions of equations 7 and 8 may now be substituted into equation 3 to yield:

$$r_{15} = p_{12}\,[p_{23}\,(p_{35} + p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) + p_{25}]$$
$$+ p_{13}\,[p_{35} + p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}]$$

$$+ p_{15} + p_{16}\, r_{56} + p_{17}\, r_{57} \qquad (9).$$

Notice that equation 9 expresses the correlation between X_1 and X_5 solely in terms of path coefficients and correlations among exogenous variables. Rearranging equation 9 will yield the so-called reduced form equation:

$$\begin{aligned} r_{15} = {} & p_{15} + p_{12}\, p_{25} + p_{13}\, p_{35} + p_{12}\, p_{23}\, p_{35} \\ & + p_{16}\, r_{56} + p_{17}\, r_{57} + p_{13}\, (p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) \\ & + p_{12}\, p_{23}\, (p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) \qquad (10). \end{aligned}$$

This equation provides the means by which the first strategy of analysis may be realized, for each element of the equation has substantive meaning. The first element in equation 10, p_{15}, is the direct effect of age on vocabulary recognition. The coefficient is the number of standard deviations X_1 changes when X_5 increases one standard deviation, *ceteris paribus*.

The second, third and fourth elements of equation 10 are indirect causal effects. For example, $p_{12}\, p_{25}$ is the effect of age on vocabulary recognition through adult intelligence. In other words, age not only has a direct effect on vocabulary recognition, it also has an indirect effect because it affects adult intelligence which is itself a cause of vocabulary recognition. The product, $p_{13}\, p_{35}$, measures the indirect effect through educational attainment, and $p_{12}\, p_{23}\, p_{35}$ measures the indirect effect through *both* education and adult intelligence.

All of the remaining components in equation 10 contain in them a correlation between exogenous variables. Previously I said that no causal interpretation is attached to a correlation; therefore no causal interpretation is attached to a component of the reduced-form equation which contains a correlation. These components exist simply because age is correlated for unknown reasons with other causes of vocabulary recognition. To provide a collective label for these components, I will call them "joint associations."

The results of the decomposition may be presented in tabular form. Table 1 represents one possibility. The total association between age and vocabulary is simply the zero-order correlation coefficient. The total effect is the sum of the direct effect and all indirect effects; that is, the total effect of age on vocabulary is interpreted as the sum of all effects whether they occur directly or through intervening variables. The other components, except for spurious effects which do not apply in this case, have been discussed above.

A substantive interpretation of these results is interesting, but somewhat problematic. In many situations the components are either all positive or all negative, and the components may be interpreted as a percentage of the total association. However, in this case age seems to operate as a suppressor variable, and all of the components are larger in value than the total. I would interpret these results to indicate that age has a strong direct effect on vocabulary--the older a person is, the longer his or her exposure to the language, and thus the greater is his or her vocabulary. At the same time, intelligence decreases with age (Wechsler, 1958), but intelligence has a positive effect on vocabulary recognition. As a result, age has a negative indirect effect through intelligence. Age also has a negative influence on educational attainment; this is probably a cohort effect whereas the other relationships are effects of aging. In sum, the zero-order relationship of age and vocabulary is negligible, but is caused by

Table 1. Decomposition of Association for Age and Vocabulary Recognition

Variables	Type of Effect	Decomposition
Age and Vocabulary	Total Association	-.01
	Total Effect	.06
	Direct Effect	.39
	Indirect Effects	-.33
	1) through X_2	-.29
	2) through X_3	-.01
	3) through X_2 and X_3 jointly	-.03
	Spurious Effects	not applicable
	Joint Association	-.07

strong counterbalancing multivariate effects.

The second strategy of analysis for recursive models is to obtain estimates of the extent to which antecedent variables account for relationships between other variables. These will be interpreted as spurious effects. The relationship between education (X_3) and vocabulary (X_1) will be used for illustration. Once again, the fundamental theorem is used to decompose the correlation between the two variables of interest. By substitution, the reduced-form equation expresses the decomposition in terms of path coefficients and correlations between exogenous variables. The intermediate steps are left for interested readers. The reduced-form equation is:

$$
\begin{aligned}
r_{13} = {} & p_{13} + p_{12}\, p_{23} + p_{15}\, p_{35} + p_{16}\, p_{36} + p_{17}\, p_{37} + p_{12}\, p_{24}\, p_{34} \\
& + p_{12}\, p_{25}\, p_{35} + p_{12}\, p_{24}\, (p_{37}\, r_{47} + p_{38}\, r_{48}) \\
& + p_{12}\, p_{25}\, (p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) \\
& + p_{15}\, (p_{36}\, r_{56} + p_{37}\, r_{57} + p_{38}\, r_{58}) \\
& + p_{16}\, (p_{35}\, r_{56} + p_{37}\, r_{67} + p_{38}\, r_{68}) \\
& + p_{17}\, (p_{34}\, r_{47} + p_{35}\, r_{57} + p_{36}\, r_{67} + p_{38}\, r_{78}) \qquad (11).
\end{aligned}
$$

The first component, p_{13}, of equation 11 is the direct effect of education on vocabulary recognition. The product, $p_{12}\, p_{23}$, is the indirect effect of education on vocabulary through adult IQ. The product, $p_{15}\, p_{35}$, is the portion of the correlation between education and vocabulary that is due to the fact that both of these variables are directly caused by the antecedent variable, age. In common terms, part of the association between education and vocabulary is said to be spuriously caused by age. Actually, the product, $p_{12}\, p_{25}\, p_{35}$, is also a spurious measure of age as it directly affects education and indirectly affects vocabulary through adult IQ. The product, $p_{16}\ p_{36}$, is the spurious effect of sex on education and vocabulary, and $p_{17}\ p_{37}$ is the spurious effect due to father's education; $p_{12}\ p_{24}\ p_{34}$ is the portion of the association between education and vocabulary spuriously due to childhood IQ as it directly affects education and indirectly affects vocabulary through adult IQ. All of the remaining components of equation 11 are joint associations.

Table 2 presents the decomposition of the association between educational attainment and vocabulary recognition. The direct effect of education on vocabulary (p_{13} = .03) is surprisingly small given the amount of attention this effect has received (Bowen, 1977; Hyman, Wright and Reed, 1975; Härnqvist, 1977). Hyman, Wright and Reed's (1975) examination of the enduring effects of education concluded that education had strong effects on knowledge net of social origins. Their analysis showed that the partial correlation of education and a vocabulary item similar to the one used herein, controlling for social origins, was about 80 percent of the zero-order coefficient (1975: 157). The data being analyzed here indicate that the partial correlation of education and vocabulary controlling for childhood IQ, age, sex, and father's education and socioeconomic index was .22. This value is only 43 percent of the zero-order coefficient. The difference between Hyman, Wright and Reed's results and mine probably lies in the ability of the present data to control for early IQ. At any rate, although the direct effect of education is small, the total effect is considerably larger by virtue of education's effect on adult intelligence which in turn strongly affects vocabulary recognition. The spurious effects of antecedent variables are

Table 2. Decomposition of Association for Educational Attainment and Vocabulary Recognition

Variables	Type of Effect	Decomposition	Percentage
Education			
and Vocabulary	Total Association	.51	100.0%
	Total Effect	.18	35.3
	Direct Effect	.03	
	Indirect Effect	.15	
	Spurious Effects	.24	47.1
	1) by X_4	.23	
	2) by X_5	-.02	
	3) by X_6	-.01	
	4) by X_7	.04	
	Joint Association	.09	17.6

particularly noteworthy. The present analysis points to the deficiency of previous studies in this area because they were not able to control for childhood IQ--a measure of the propensity for learning. Forty-five percent of the association between education and vocabulary may be said to be spuriously caused by childhood intelligence. Finally, a portion of the zero-order association may not be decomposed into components worthy of a causal interpretation. About 18 percent of the association results from the correlation of exogenous variables with themselves for unknown reasons.

In explicating this model, I have for purposes of illustration decomposed only two associations of the many that could have been considered. Which associations should the researcher decompose by means of the reduced-form equations? Alwin and Hauser (1975) show how the reduced-form coefficients may be obtained by manipulating the equations in a computer. However, this begs the question of which associations are important enough to discuss, particularly when competition for journal space is considered. The answer lies in which associations are theoretically important. Decompose associations that illuminate the unresolved problem which initiated the exercise.

BLOCK EQUATION MODELS

A block equation model is one in which each of a set of dependent variables is regressed on the same set of independent variables. For example,

$$y_1 = p_{14}\,x_4 + p_{15}\,x_5 + p_{16}\,x_6 + p_{17}\,x_7 + p_{1u}\,u$$

$$y_2 = p_{24}\,x_4 + p_{25}\,x_5 + p_{26}\,x_6 + p_{27}\,x_7 + p_{2v}\,v$$

$$y_3 = p_{34}\,x_4 + p_{35}\,x_5 + p_{36}\,x_6 + p_{37}\,x_7 + p_{3w}\,w \qquad (12).$$

These equations may be diagrammed, and are shown in Figure 2. The data were taken from Sewell and Hauser (1975: 93), and pertain to Wisconsin male high school seniors of 1957 of nonfarm backgrounds. The three endogenous variables are the respondent's best friend's plans to attend college (a dummy variable: = 1 if yes, = 0 if no), parental encouragement to attend college, and teacher's encouragement to attend college (also dummy variables). The four exogenous variables are father's and mother's educational attainment, respondent's Henmon-Nelson IQ score recorded during their senior year, and the respondent's high school grades, recorded as their rank in high school class.

The analysis goals for this model are (1) to compare the partial path coefficients with their corresponding zero-order coefficients in order to determine how much of the latter may be considered a direct effect, and how much a joint association (note that this model does not yield components interpretable as indirect or spurious effects); and (2) to examine the residuals for correlated errors. That is, the researcher employing a block equation model will want to determine how good a job the exogenous variables do of accounting for the correlations among the endogenous variables. To do so, the correlations of residuals are compared with the zero-order correlations of the corresponding endogenous variables.

To effect the analysis of residuals, the correlations between endogenous variables are decomposed by employing the reduced-form equations while allowing for the correlation of their residuals. For example, using the fundamental theorem of path analysis, the correlation between Y_1 and Y_2 may be decomposed:

$$r_{12} = p_{14} r_{24} + p_{15} r_{25} + p_{16} r_{26} + p_{17} r_{27} + p_{1u} r_{2u} \quad (13).$$

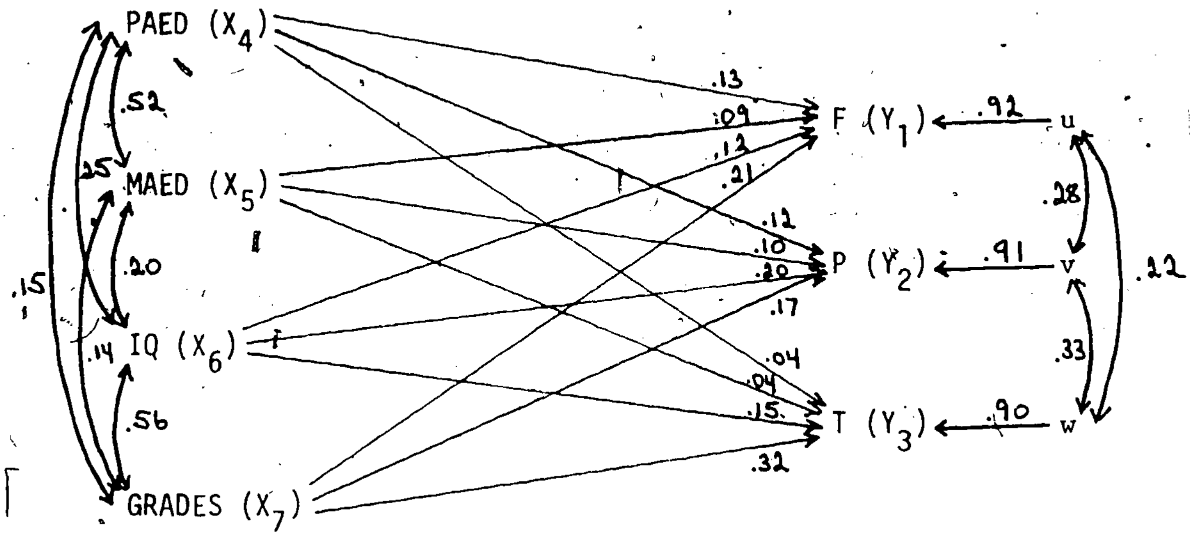


Figure 2. Block Equation Model of Encouragement to Attend College, Wisconsin Male High School Seniors, 1957 (Source: Sewell and Hauser, 1975: 93)

Similarly, the correlation of Y_2 with u may be written:

$$r_{2u} = p_{2v}\, r_{vu} \,. \tag{14}$$[4]

Substituting equation 14 into equation 13 yields the reduced-form equation,

$$r_{12} = p_{14}\, r_{24} + p_{15}\, r_{25} + p_{16}\, r_{26} + p_{17}\, r_{27} + p_{1u}\, p_{2v}\, r_{vu} \quad (15),$$

and

$$r_{vu} = (r_{12} - [p_{14}\, r_{24} + p_{15}\, r_{25} + p_{16}\, r_{26} + p_{17}\, r_{27}])/p_{1u}\, p_{2v} \quad (16).$$

In similar fashion, the correlations of r_{uw} and r_{vw} could be obtained. But notice that $r_{uv} = r_{12.4567}$; that is, the correlation of residuals v and u equals the fourth-order partial correlation of Y_1 and Y_2, controlling for X_4, X_5, X_6, and X_7. With modern computers, the values for the correlated residuals may be obtained directly from a partial correlation routine.

Table 3 presents the decomposition of the associations between the exogenous and endogenous variables. Because there are no intervening variables, there can be no indirect effects. Because there are no antecedent variables, there can be no spurious effects. The direct effects are measured in the usual way; that is, they are standardized regression coefficients. Joint associations could be calculated by using the reduced-form equations, but in this special case are more easily obtained by subtracting the path coefficient from its associated correlation coefficient.

Examining the effects upon parental encouragement to attend college, notice that the respondent's characteristics are the more important cause of parental encouragement, but parental characteristics are by no means unimportant. Moreover, for each of these variables about one-half of the association is a direct effect, while the remainder is due to the inter-

Table 3. Decomposition of Associations for Variables in Block-Equation Model

Variables	Total Association	Direct Effect	Joint Association
PAED & F	.24 (100%)	.13 (54%)	.11 (46%)
MAED & F	.21 (100%)	.09 (43%)	.12 (57%)
IQ & F	.29 (100%)	.12 (41%)	.17 (59%)
Grades & F	.31 (100%)	.21 (68%)	.10 (32%)
PAED & P	.25 (100%)	.12 (48%)	.13 (52%)
MAED & P	.23 (100%)	.10 (43%)	.13 (57%)
IQ & P	.34 (100%)	.20 (59%)	.14 (41%)
Grades & P	.32 (100%)	.17 (53%)	.15 (47%)
PAED & T	.15 (100%)	.04 (27%)	.11 (73%)
MAED & T	.14 (100%)	.04 (29%)	.10 (71%)
IQ & T	.35 (100%)	.15 (43%)	.20 (57%)
Grades & T	.42 (100%)	.32 (76%)	.10 (24%)

correlation (for unknown reasons) of the exogenous variables. My overall impression is that parents who encourage their twelfth-grade children to attend college are nearly equally influenced by a variety of effects. On the other hand, teachers who encourage students to attend college are strongly influenced by the student's performance regardless of social background. To the extent that social background is associated with teacher's encouragement, it occurs primarily through correlations with other effects. For example, the total association of father's education and teacher's encouragement may be decomposed:

$$\begin{aligned} r_{34} &= p_{34} + p_{35}\, r_{45} + p_{36}\, r_{46} + p_{37}\, r_{47} \\ &= .04 + (.04)\,(.52) + (.15)\,(.25) + (.32)\,(.15) \\ &= .04 + .02 + .04 + .05 \\ &= .15 \end{aligned} \qquad (17).$$

In plain English, most of the association of father's education and teacher's encouragement is due to father's education being correlated with student's IQ and school performance, which are themselves effects of teacher's encouragement.

The second strategy of analysis for block equation models is to examine the residuals for correlated errors. We should on substantive grounds expect students who are being encouraged to attend college by their parents also to be encouraged by teachers, and to have best friends whose plans include college. The correlation coefficients lend support to these expectations:

$$\begin{aligned} r_{FP} &= .40 \\ r_{FT} &= .34 \\ r_{PT} &= .44 \end{aligned} \qquad (18).$$

Yet, once again on substantive grounds, one wonders whether these correlations result merely from the fact that they have common antecedent causes. That is, the correlation of parental and teacher's encouragement could be due to similar patterns by which parents and teachers assess the student's social background, ability, and performance. To see if this is so, we examine the correlation of the residuals, equivalent to the fourth-order partial correlation coefficient. If the residuals are uncorrelated, one would conclude that the exogenous variables jointly account for the zero-order association. If the residual correlation coefficients are not zero, one would conclude that the model has not revealed all sources of covariation between the endogenous variables. Indeed, that is what the correlation of errors would lead us to believe, for:

$$r_{12.4567} = .28$$

$$r_{13.4567} = .22$$

$$r_{23.4567} = .33 \qquad (19).$$

Consider the zero-order correlation between parental encouragement and teacher's encouragement ($r_{PT} = .44$). Once again, decomposition permits causal interpretations to be attached to the components of the reduced-form equation:

$$r_{23} = p_{24}\, r_{34} + p_{25}\, r_{35} + p_{26}\, r_{36} + p_{27}\, r_{37} + p_{2v}\, r_{vw}\, p_{3w} \qquad (20).$$

and $r_{23} = .44 = .17 + .27$ (21).

where the first portion (.17) is the sum of products of direct effects and correlations, and the second portion (.27) is due to the effects of correlated errors. In other words, of the total association of teacher's and parental encouragement, 39 percent can be attributed to correlated antecedent variables, and 61 percent to correlated errors. In conclusion, it is clear that the correlation between teacher's and parental encouragements do not result solely from common antecedent causes; most of the association is due to the correlation of causes that have not been explicitly measured in the model.

BLOCK-RECURSIVE MODELS

A block-recursive equation model is, as the name implies, a model that combines both recursive and block equations. The analysis goals are, therefore, defined by what one can do with block and recursive equation models. First, the researcher will want to obtain estimates of the extent to which intervening variables account for relationships among variables. Second, one will want to obtain estimates of the extent to which antecedent variables account for relationships between other variables. Third, the researcher will want to examine where appropriate, residuals for correlated errors. Finally, the researcher will want to determine the extent to which zero-order associations are accounted for by contemporaneous, endogenous variables. Let me emphasize here that, as always, the underlying goal of analysis is the explication of the unresolved problem which motivated the analysis. The analysis thus proceeds in accordance with the logic of the model. In the example to follow, each of the four analytic goals will be discussed, but the reader should remember that in substantive models the strategies of analysis make sense only in relation to the theoretical purpose of the model.

Consider the following block-recursive equation model:

$$x_1 = p_{12}\, x_2 + p_{13}\, x_3 + p_{14}\, x_4 + p_{15}\, x_5 + p_{16}\, x_6 + p_{1u}\, u$$

$$x_2 = p_{24}\, x_4 + p_{25}\, x_5 + p_{26}\, x_6 + p_{2v}\, v$$

$$x_3 = p_{34}\, x_4 + p_{35}\, x_5 + p_{36}\, x_6 + p_{3w}\, w \qquad (22),$$

in which $E(uv) = E(uw) = 0$, by assumption, but no assumption is required about the association of the residuals of X_2 and X_3. The model is diagrammed in Figure 3. The data are fictitious; they were published in Kerlinger and Pedazur's (1973:331-3) text, and are used here for purposes of illustration.

The dependent variable is verbal achievement, and the exogenous variables are race, mental ability, and school quality. Two variables are thought to be causally intervening between verbal achievement and the exogenous variables. These are the sociopsychological variables, self-concept and level of aspiration. While it makes good sense to consider self-concept and level of aspiration dependent upon race, IQ, and school quality, it is not clear to me that causality can be unambiguously established between the two sociopsychological variables. For my purposes, it is not necessary to establish the causal order, because my interest is primarily in the extent to which they account for the associations between verbal achievement and the three exogenous variables. Therefore, I specified no causal link between self-concept and aspiration, but do consider the possibility of correlated errors. Of course, if the causal association between these variables was of theoretical interest, the model would require reformulation.

To illustrate the first strategy of analysis for block-recursive models, consider the association between IQ and verbal achievement. The fundamental theorem reveals:

$$r_{15} = p_{12}\, r_{25} + p_{13}\, r_{35} + p_{14}\, r_{45} + p_{15}\, r_{55} + p_{16}\, r_{56} \qquad (23),$$

and by substitution for r_{25} and r_{35}, the reduced-form equation is:

$$r_{15} = p_{15} + p_{12}\, p_{25} + p_{13}\, p_{35} + p_{14}\, r_{45} + p_{16}\, r_{56}$$
$$+ p_{12}\,(p_{24}\, r_{45} + p_{26}\, r_{56}) + p_{13}\,(p_{34}\, r_{45} + p_{36}\, r_{56}) \qquad (24).$$

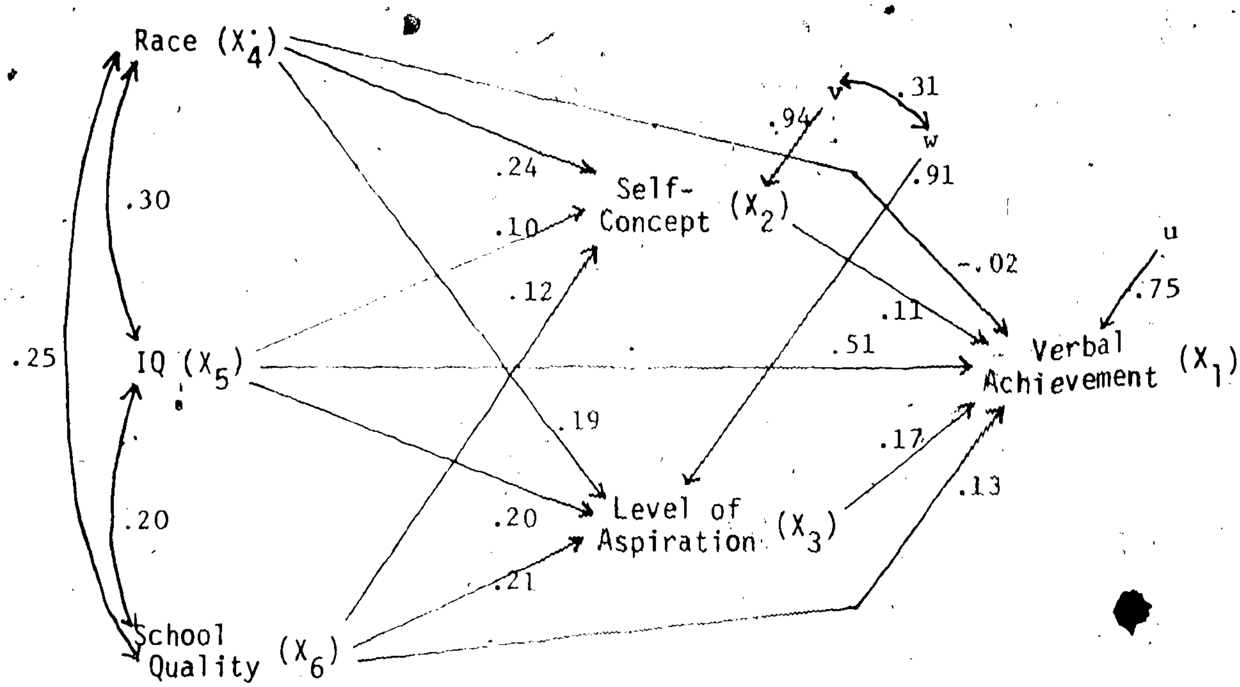


Figure 3. Block-Recursive Equation Model of Verbal Achievement
(Source: Kerlinger and Pedhazur, 1973: 331)

In numeric terms,

$$r_{15} = .60 = .51 + .01 + .03 + .05 \qquad (25).$$

Thus, 85 percent of the association between IQ and verbal achievement was a direct effect, and only seven percent may be said to have occurred through the two intervening variables, self-concept and level of aspiration.

The second strategy of analysis is to determine the extent to which antecedent variables account for relationships between other variables. Consider the association between self-concept and verbal achievement. The decomposition of the correlation is:

$$r_{12} = p_{12}\, r_{22} + p_{13}\, r_{23} + p_{14}\, r_{24} + p_{15}\, r_{25} + p_{16}\, r_{26} \qquad (26).$$

The correlation $r_{22} = 1.0$ by definition, but each of the remaining correlations may themselves be decomposed by means of the fundamental theorem. Decompositions of the three correlations, r_{24}, r_{25}, and r_{26}, are straightforward (the interested readers may insert these intermediate decompositions). The correlation of X_2 and X_3 may be written:

$$r_{23} = p_{24}\, r_{34} + p_{25}\, r_{35} + p_{26}\, r_{36} + p_{2v}\, r_{v3} \qquad (27),$$

and

$$\begin{aligned} r_{23} = {} & p_{24}\, p_{34} + p_{25}\, p_{35} + p_{26}\, p_{36} + p_{24}\, (p_{35}\, r_{45} + p_{36}\, r_{46}) \\ & + p_{25}\, (p_{34}\, r_{45} + p_{36}\, r_{56}) + p_{26}\, (p_{34}\, r_{46} + p_{35}\, r_{56}) \\ & + p_{2v}\, p_{3w}\, r_{vw} \qquad (28). \end{aligned}$$

Therefore, the reduced-form equation for the correlation of X_1 and X_2 is:

$$\begin{aligned} r_{12} = {} & p_{12} + p_{13}\, p_{24}\, p_{34} + p_{13}\, p_{25}\, p_{35} + p_{13}\, p_{26}\, p_{36} + p_{14}\, p_{24} \\ & + p_{15}\, p_{25} + p_{16}\, p_{26} + p_{13}\, p_{24}\, (p_{35}\, r_{45} + p_{36}\, r_{46}) \\ & + p_{13}\, p_{25}\, (p_{34}\, r_{45} + p_{36}\, r_{56}) + p_{13}\, p_{26}\, (p_{34}\, r_{46} + p_{35}\, r_{56}) \\ & + p_{14}\, (p_{25}\, r_{45} + p_{26}\, r_{46}) + p_{15}\, (p_{24}\, r_{45} + p_{26}\, r_{56}) \\ & + p_{16}\, (p_{24}\, r_{46} + p_{25}\, r_{56}) + p_{13}\, p_{2v}\, p_{3w}\, r_{vw} \qquad (29). \end{aligned}$$

The direct effect is given by p_{12} = .11 (which is 37 percent of r_{12} = .30). The next six components on the right-hand side of equation 29 measure the extent to which the antecedent variables X_4, X_5, and X_6 account for the relationship between X_1 and X_2. Thus, a portion of the relationship between X_1 and X_2 is spuriously due to three exogenous variables as they directly affect self-concept and verbal achievement, and indirectly affect verbal achievement through level of aspiration (X_3). In numeric terms, the sum of these six components is .08, or 26 percent of the total, which answers the question addressed in this second strategy of analysis.

The third strategy of analysis for block-recursive models is to examine the residuals for correlated errors. In specifying the model, no assumption was made about the correlation of residuals for X_2 and X_3. It is now appropriate to compare the correlation of residuals to the zero-order correlation; we do so to determine the extent to which race, IQ, and school quality determine the association of self-concept and level of aspiration. The reduced-form equation for the zero-order correlation of X_2 and X_3 has already been shown in equation 28. The zero-order coefficient was .40, and equation 28 suggests that race, IQ, and school quality directly and jointly account for 35 percent (.14) of the zero-order association, and the correlated errors account for the remaining 65 percent ($p_{2v}\ p_{3w}\ r_{vw}$ = .27). In other words, most of the association between these two sociopsychological variables is caused by variables not explicitly included in the model.

The final strategy of analysis is to determine the extent to which zero-order associations are accounted for by contemporaneous, endogenous variables. For example, focus on self-concept (X_2), and on the zero-order relationship between self-concept (X_2) and verbal achievement (X_1);

how much of this relationship is accounted for by the contemporaneous variable, level of aspiration (X_3)? That is, once we admit that X_2 and X_3 have correlated errors, we must also admit the possibility that the zero-order correlation of X_1 and X_2 includes an effect attributable to the correlated errors of X_2 and X_3. This effect is shown in equation 29 as the product, $p_{13}\ p_{2v}\ p_{3w}\ r_{vw}$, which equals the value, .045, or 15 percent of r_{12} = .30.

NONRECURSIVE MODELS

The models considered above were causally unidirectional. However, there are times when a researcher's formulation of a model dictates that one variable thought to be a cause of another, is also thought to be caused by it, either directly or indirectly. In other words, an independent variable in one equation is, in another equation, regressed on the dependent variable in the first equation. In such situations, estimates are no longer obtainable through ordinary least squares procedures.

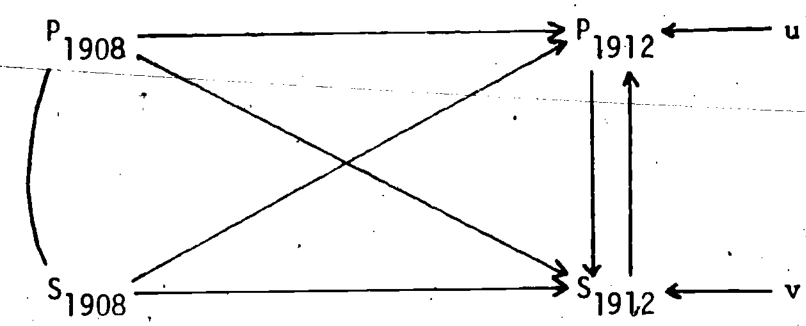
In the area of education, nonrecursive, reciprocal effects were hypothesized by Duncan, Haller, and Portes (1968) between respondent's plans to attend college and the plans of respondent's best friend (see Hout and Morgan, 1975). For purposes of illustration, however, I would prefer to consider the simplest possible nonrecursive model of four variables. However, substantive examples of such models are rare. Erlanger and Winsborough (1976) used such a model to explain the subculture of violence thesis (Wolfgang, 1958). But rather than repeat their already cogent example, I prefer to construct another. It is outside the area of education, and somewhat contrived, yet I hope instructive.

About the turn of the century the Socialist Party of America was a potentially viable political party (Wolfle, 1976). There was strong support for the party among miners, one-crop farmers, lumbermen, some immigrant groups, and others. The party's presidential ticket was headed by Eugene V. Debs, an attractive vote-getter for the Socialists. Yet one is eventually led to ask whether the party's electoral fortunes were due to the popular presidential candidate, or whether the presidential candidate merely gained support from voters attracted to the party's basic philosophy.

To effect this analysis, consider the model diagrammed in Figure 4(a). The four variables are the Illinois presidential SPA vote in 1908 and 1912, and the Illinois SPA vote for the State Treasurer in the same elections. The latter are taken as proxies for structural support for the SPA; the SPA State Treasurer candidate was seldom a factor in the vote cast for the party. The model indicates that the presidential and state treasurer vote in 1908 are correlated for unknown reasons. The curved line in the diagram has no arrowheads to caution us that the decompositions employed in previous models do not apply to nonrecursive models. The presidential vote in 1912 is hypothesized to depend on the candidate's vote-gathering ability in the last election, and the party's structural support in both the current election and the last. At the same time, the party's structural support is hypothesized to depend upon the party's basic support four years earlier, and also the presidential vote in both 1908 and 1912. The equations implied by the model are:

$$x_1 = p_{12}\, x_2 + p_{13}\, x_3 + p_{14}\, x_4 + p_{1u}\, u$$
$$x_2 = p_{21}\, x_1 + p_{23}\, x_3 + p_{24}\, x_4 + p_{2v}\, v \qquad (30).$$

(A)



(B)

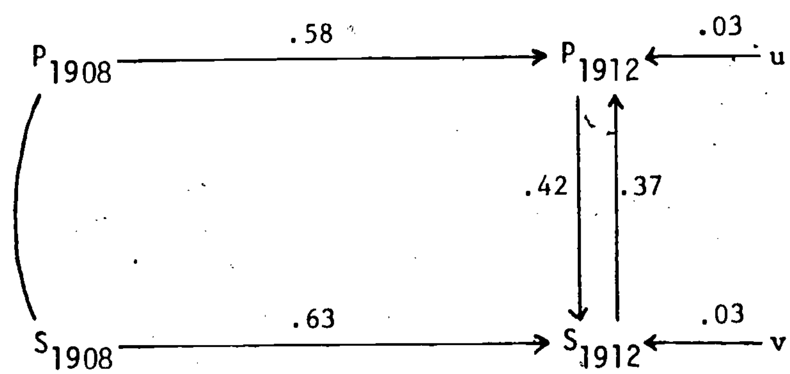


Figure 4. Nonrecursive Equation Models of Socialist Voting in Illinois

Without belaboring the point, this set of equations has no solution. Oversimplifying a bit, there are only six intercorrelations supplied by the data, yet the equations have eight unknowns. The model is said to be unidentified (see Duncan, 1975).

In order to obtain estimates for the model, some prior restraints must be placed on it. This is accomplished by constraining some of the paths to be zero. Which ones depend upon the substantive questions being asked of the data. If the researcher is primarily interested in the cross-lagged effects, then the reciprocal effects are specified zero, and the model may be analyzed as a set of block equations. If the researcher is primarily interested in the reciprocal effects, then the cross-lagged effects are specified zero, such as shown in Figure 4(b). Thus, the strategy of analysis for nonrecursive models depends upon the substantive questions that motivated the analysis. The researcher must choose between models by going beyond the statistical information available. If the reciprocal effects are of theoretical interest, then the strategy of analysis is to compare the size of the reciprocal effects to each other, and to other independent variables.

The model to be estimated is:

$$x_1 = p_{12}\, x_2 + p_{13}\, x_3 + p_{1u}\, u$$

$$x_2 = p_{21}\, x_1 + p_{24}\, x_4 + p_{2v}\, v \qquad (31).$$

To estimate the parameters of this model ordinary least squares procedures are not appropriate. A one-step method has been known for some time (Goldberger, 1964), but most social science computer software packages do not include these two-stage least squares procedures. However, with certain precautions ordinary least squares regression may be employed in two stages. I estimated the coefficients for this model by using SPSS (Nie, et al., 1975), and adjusting the second stage output for inherent computational errors (Hout, 1977).

The results are shown in Figure 4(b). I am led to conclude that the presidential vote in 1908 was relatively more important than the philosophy-based support in 1912 (.58 versus .42) for the presidential vote in 1912. For the state treasurer vote in 1912, the philosophy-based support in 1908 was relatively more important than the presidential vote in 1912 (.63 versus .37). In answer to the question, which was the more important source or electoral support in 1912, the presidential candidate's popularity of the party's basic philosophy, the answer seems to be that these effects were about equal. There did, indeed, seem to be reciprocal effects--the party received votes on the coattails of Eugene Debs, and Debs received votes from people attracted to the party, not necessarily its candidates.

CONCLUSION

Duncan reminded us that "the study of structural equation models can be divided into two parts: the easy part and the hard part" (1975: 149). This paper has been about an easy part--how a researcher can manipulate algebric equations to assist in the interpretation of an existing model. The hard part is constructing models that are consistent with sound theory. To incorporate sound theoretical ideas in structural equation models requires a measure of creativity that goes far beyond the bounds of this paper. As Duncan before me suggested, if I knew how to construct creative, sound models, I would have already done so.

The models that I have presented in this paper to exemplify strategies of analysis are therefore no better than the ideas that went into them. If the models were based on unsound ideas, the models are unsound. Yet I find path models extremely useful, because they force me to formulate my ideas in explicit form, and allow me to read the ideas of others in explicit form. I see this as the most important advantage of path analysis.

Readers who would like to gain further insight into the construction and analysis of models may begin with the collections of Blalock (1971), Goldberger and Duncan (1973), and Sewell, Hauser and Featherman (1976). The *American Sociological Review* often includes articles that incorporate in them structural equation models, and readers may find these instructive. Yet ideas and models are often flawed, and the comment section of the journal contains very instructive dialogs between authors and critics. I recommend these discussions to you, and you might begin with, for example, Havens and Tully (1972), Halaby (1973), Featherman (1973), Hannan, Freeman and Meyers (1976), Alexander and Griffin (1976), or Bohrnstedt (1977).

FOOTNOTES

[1]For those who are unfamiliar with path analysis there are several good introductions. Land's (1969) article is good, but like many early papers on the subject glosses over the distinction between sample estimates and population parameters. Moreover, Land (as well as Kerlinger and Pedhazur [1973:316]) ill-advisedly replicated Duncan's (1966) defective conception of measuring indirect effects. A more proper consideration of indirect effects may be found in Duncan (1971) or Finney (1972). Another introduction to path analysis may be found in Wolfle (1977), and Duncan's (1975) textbook is the best in the field.

[2]A complete description of these data may be found in Wolfle (forthcoming).

[3]In fact it is the population correlation coefficient, ρ_{45}, which is assumed equal to zero. When dealing with sample data the assumed equality will not hold exactly. Setting r_{45} and r_{46} to zero is therefore a practical determination to be distinguished from the formal tenets of statistics.

[4]The complete equation would be:

$$r_{2u} = p_{2v}\, r_{vu} + p_{24}\, r_{4u} + p_{25}\, r_{5u} + p_{26}\, r_{6u} + p_{27}\, r_{7u}$$

but $r_{4u} = r_{5u} = r_{6u} = r_{7u} = 0$ by assumption, i.e., $E(X_i u) = 0$, and the full equation reduces to the form expressed in the text.

REFERENCES

Alexander, Karl L., and Larry J. Griffin
1976 "School District Effects on Academic Achievement: A Reconsideration," American Sociological Review 41:144-52.

Alwin, Duane F., and Robert M. Hauser
1975 "The Decomposition of Effects in Path Analysis." American Sociological Review 40:37-47.

Anderson, James G., and Francis B. Evans
1974 "Causal Models in Educational Research: Recursive Models." American Educational Research Journal 11:29-39.

Bielby, William T., and Robert M. Hauser
1977 "Structural Equation Models." Pp. 137-61 in Alex Inkeles (ed.), Annual Review of Sociology, Vol. 3. Palo Alto: Annual Reviews, Inc.

Blalock, H. M., Jr.
1971 Causal Models in the Social Sciences. New York: Aldine.

Blau, Peter M., and Otis Dudley Duncan
1967 The American Occupational Structure. New York: John Wiley.

Bohrnstedt, George W.
1977 "Use of the Multiple Indicators - Multiple Causes (MIMIC) Model." American Sociological Review 42:656-63.

Bowen, Howard R.
1977 Investment in Learning: The Individual and Social Value of American Higher Education. San Francisco: Jossey-Bass.

Duncan, Otis Dudley
1966 "Path Analysis: Sociological Examples." American Journal of Sociology 72:1-16.

1971 "Path Analysis: Sociological Examples (Addenda)." Pp. 136-8 in H. M. Blalock, Jr. (ed.), Causal Models in the Social Sciences. New York: Aldine.

1975 Introduction to Structural Equation Models. New York: Academic Press.

Duncan, Otis Dudley, David L. Featherman, and Beverly Duncan
1972 Socioeconomic Background and Achievement. New York: Seminar Press.

Duncan, Otis Dudley, Archibald O. Haller, and Alejandro Portes
1968 "Influences on Aspirations: A Reinterpretation." American Journal of Sociology 74:119-37.

Erlanger, Howard S., and Halliman H. Winsborough
1976 "The Subculture of Violence Thesis: An Example of a Simultaneous Equation Model in Sociology." Sociological Methods & Research 5:231-46.

Featherman, David L.
1973 "Comments on Models for the Socioeconomic Career." American Sociological Review 38:785-91.

Finney, John M.
1972 "Indirect Effects in Path Analysis." Sociological Methods & Research 1:175-86.

Goldberger, Arthur S.
1964 Econometric Theory. New York: John Wiley.

1972 "Structural Equation Methods in the Social Sciences." Econometrica 40:979-1001.

Goldberger, Arthur S., and Otis Dudley Duncan
1973 Structural Equation Models in the Social Sciences. New York: Seminar Press.

Halaby, Charles N.
1973 "Hardship and Collective Violence in France: A Comment." American Sociological Review 38:495-500.

Hannan, Michael T., John H. Freeman, and John W. Meyer
1976 "Specification of Models for Organizational Effectiveness." American Sociological Review 41:136-43.

Härnqvist, Kjell
1977 "Enduring Effects of Schooling--A Neglected Area in Educational Research." Educational Researcher 6(November):5-11.

Hauser, Robert M.
1971 Socioeconomic Background and Educational Performance. Rose Monograph Series. Washington, D.C.: American Sociological Association.

Havens, Elizabeth M., and Judy Corder Tully
1972 "Female Intergenerational Occupational Mobility: Comparisons of Patterns?" American Sociological Review 37:774-77.

Heise, David R.
1975 Causal Analysis. New York: John Wiley.

Hout, Michael
1977 "A Cautionary Note on the Use of Two-Stage Least Squares." Sociological Methods & Research 5:335-46.

Hout, Michael, and William R. Morgan
1975 "Race and Sex Variations in the Causes of Expected Attainments." American Journal of Sociology 81:364-94.

Hyman, Herbert H., Charles R. Wright, and John Shelton Reed
1975 The Enduring Effects of Education. Chicago: University of Chicago Press.

Kerlinger, Fred N., and Elazar J. Pedhazur
1973 Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston.

Kim, Jae-On, and Charles W. Mueller
1976 "Standardized and Unstandardized Coefficients in Causal Analysis: An Expository Note." Sociological Methods & Research 4:423-38.

Land, Kenneth C.
1969 "Principles of Path Analysis." Pp. 3-37 in Edgar F. Borgatta (ed.), Sociological Methodology 1969. San Francisco: Jossey-Bass.

Nie, Norman, C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner, and Dale H. Bent
1975 Statistical Package for the Social Sciences, Second edition. New York: McGraw-Hill.

Sewell, William H., and Robert M. Hauser
1975 Education, Occupation, and Earnings. New York: Academic Press.

Sewell, William H., Robert M. Hauser, and David L. Featherman
1976 Schooling and Achievement in American Society. New York: Academic Press.

Wechsler, David
1958 The Measurement and Appraisal of Adult Intelligence, Fourth edition. Baltimore: Williams and Wilkins.

Wolfgang, Marvin E.
1958 Patterns of Criminal Homocide. Philadelphia: University of Pennsylvania Press.

Wolfle, Lee M.
1976 Radical Third-Party Voting Among Coal Miners, 1896-1940. Unpublished Ph.D. Dissertation, University of Michigan.

1977 "An Introduction to Path Analysis." Multiple Linear Regression Viewpoints 8:36-61.

(forthcoming) "Long Term Effects of Education."

Wright, Sewell
1921 "Correlation and Causation." Journal of Agricultural Research 20:557-85.

———1925 Corn and Hog Correlations. Washington, D.C.: U.S. Department of Agriculture Bulletin 1300.