INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Structural Equation Modelling in Educational Research

Dr. Henry A. Ojating
Lecturer, Faculty of Education, Cross River University of Technology, Calabar, Nigeria
Dr. Bassey A. Bassey
Senior Lecturer, Faculty of Education, University of Calabar, Calabar, Nigeria

Abstract:
Research in education, like in the behavioural sciences, involves a wide range of variables that are largely interrelated. The reality of variable inter-connection which explains the utility of multivariate statistics is what essentially gives credence to the concept of structural equation modeling, otherwise known as, path analysis. Structural equation modeling is a statistical approach for decomposing variables to obtain the direct and indirect effect of those hypothesized as "causes" on others hypothesized as "effects". Relationships among variables are being discerned on the basis of knowledge, experience or theory. This paper attempts to shed light on the applicability of structural equation modeling in education research.

Keywords: *Structural equation, modeling, educational research, path analysis*

1. Introduction

Structural equation model(SEM) or path analysis is a technique that provides a graphical framework of measures or magnitudes of causal connections between variables. Such causal connections or relationships between variables are generated by the researcher hypothetically. At best, this is done on the basis of theory, knowledge or experience. The concept of Path Analysis was developed by geneticist Sewall Wright and dates back to 1921.Its application became widespread in the field of behavioural sciences research courtesy of Duncan (1966) and Blalock (1971). Path Analysis is a technique that examines the direct and indirect effect of variables presumed as 'causes' on others presumed as 'effects'.

2. Theoretical Path Model

The theoretical or hypothesized causal model simply describes the researcher's design which prepares him for the actual analysis – model testing. As stated earlier, the researcher draws up the theoretical model based on theory, previousknowledge or experience. This can be best explained by the path model in fig. 1.

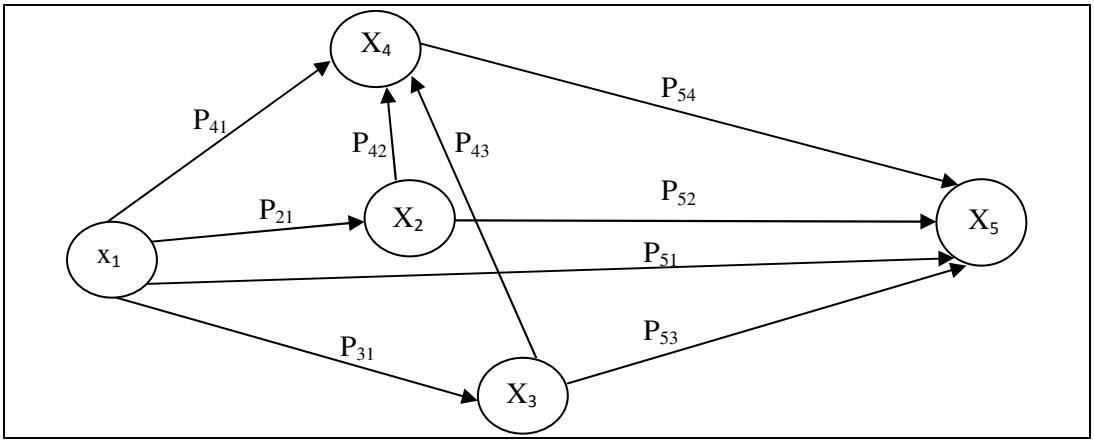

Figure 1: A hypothesized causal model of the four-social environment variables and attitude to school (adapted from Ojating, Bassey and Ayang, 2014)

Key:
X_1 = Socio-economic status (SES)
X_2 = Parental academic stimulation (PAS)

X_3 = Peer group influence (PGI)
X_4 = Teacher-student relationship (TSR)
X_5 = Attitude to school (ATT)

Figure 1 shows a five-variable hypothesized recursive path model which addressed the linkages among variables in a study on social environment variables and students' attitude to school carried out by Ojating, Bassey and Ayang (2014). The model denotes a one way causal flow between variables and the relations among the variables are linear, additive and causal in line with the basic assumptions of path analysis procedures suggested by Kerlinger (1980) and Pedhazur (1984). The hypothesized model comprised one exogenous (x1) and four endogenous variables (x2, x3, x4, x5). Exogenous variables are those caused by factors outside the model, while endogenous variables are those explained by factors within the model. A residual or error (e)estimate is correlated with each endogenous variable. The 'e' term represents variations in an endogenous variable that are due to error.

3. Structural Equations

The hypothesized causal model can be represented either pictorially (i.e. with a path diagram) or in equation form. A theoretical model designed in equation form is usually referred to as a structural equation. Only standardized measures are used to estimate variables in structural equations. The five variables defined in the model are all stated in z-scores (or standard deviation units). In structural equations, according to Adegoke (2013), the direct causal effect of each variable on another is represented by a path coefficient or structural coefficient. And that the coefficient is the same with the standardized coefficient, B derived from the multiple regression analysis.

The path coefficient is denoted by the symbol p with two subscripts. Thus, P_{21} in figure 1 and in equation 2 indicates the direct effect of variable x_1 on variable x_2, while e_2 represents the residual term of variable x_2. The five-variable model shown in figure 1 (with variables expressed in standard scores) can be represented with a system of structural equations:

$$X_1 = e_1 \quad \text{Eqn.1}$$
$$X_2 = P_{21}x_1 + e_2 \quad \text{Eqn.2}$$
$$X_3 = P_{31}x_1 + e_3 \quad \text{Eqn.3}$$
$$X_4 = P_{42}x_2 + P_{41}x_1 + e_4 \quad \text{Eqn.4}$$
$$X_5 = P_{54}x_4 + P_{53}x_3 + P_{52}x_2 + P_{51}x_1 + e_5 \quad \text{Eqn.5}$$

Variable x_1 is exogenous and therefore stands alone in the system. In a structural model, indirect effects are depicted by two or more recursive paths - that is, arrows pointing to the same direction. Indirect path coefficients are obtained by finding the product of all the paths in the chain. As earlier stated, a path coefficient is represented by the direct causal effect of one variable on another.

Multiple regression analysis yields the standardized or unbiased estimates of path coefficients. Repeated regression runs are carried out on each structural equation to obtain the path coefficients. For example, to obtain P_{21}, variable x_2 (parental academic stimulation) is regressed on variable x_1 (socioeconomic status). To estimate $P_{42,}$ variable x_4 (teacher-student relationship) is regressed on variable x_2 (parental academic stimulation) while to obtain $P_{41,}$ variable x_4 is regressed on variable x_1 (socioeconomic status). The process continues for the rest of the equations. The hypothesized model is represented with estimates of the path coefficients in Fig. 2.

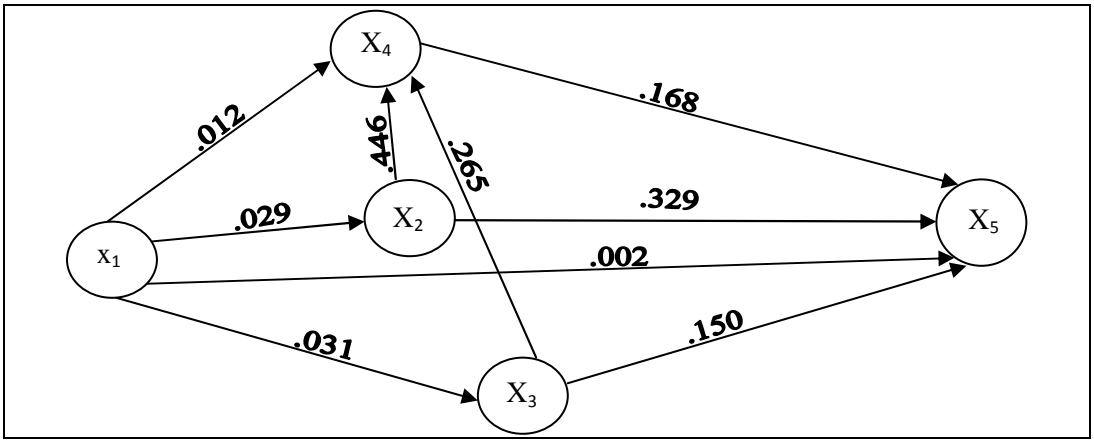

Figure 2: A hypothesized causal model of the four social environment variables and attitude to school (adapted from Ojating, Bassey and Ayang, 2014)

4. Validation of a Path Model

Path coefficients obtained by regression analyses will be useless in determining causal effects among variables if the hypothesized model from which the coefficients were computed is invalid. Kline (2005) noted that before the obtained estimates of path coefficients can be used to explain the causal effects among the variables, one should determine whether or not the model is consistent with the observed empirical correlations among the variables. This is accomplished by determining the reproduced correlations, which are then compared with the empirical correlations. Pedhazur (1982) maintained that, if the observed and the reproduced correlations are

reasonably close (say, within .05 of each other), it can be assumed that the model is consistent with the empirical data. Where discrepancies are beyond the acceptable limits, the model does not fit the data, it should be revised. The overall process of reproducing the bivariate correlations cannot be undertaken by any known statistical package, like SPSS, it can only be done manually or with the use of hand calculators.

5. Deletion of Paths

Paths in the researcher's self-designed or theoretical model that are found to be weak are deleted leaving those that eventually constitute a more parsimonious and meaningful model. Deletion of paths in a causal model is done based on the follow guidelines:

(i) Theory of the researcher and associated literature
(ii) Calculation of path coefficients.

The second guideline (ii) simply involves first obtaining all the path coefficients in the model and then deleting weak ones based on the criteria of statistical significance and meaningfulness. In the first criterion, paths whose beta B values are not significant at the chosen significance level are deleted. This criterion is, however, error prone because with large samples, even weak paths may be found to be statistically significant (Kerlinger&Pedhazur, 1973; Land, 1969). To overcome this problem, Land (1969) recommended the option of meaningfulness where path coefficients with values less than .05 may be treated as not meaningful.

After calculating the zero order correlations of the variables in the researcher's model, the outcomes are compared with the original correlations obtained based on observed data. If discrepancies between the original and reproduced correlations are as small as (less than .05) for virtually all cases, then it may be concluded that the parsimonious model generating the new R matrix is tenable or valid. But where discrepancies are many, the model should be revised.

Practical steps in conducting path analysis would be examined with recourse to the five-variable causal model developed by Ojating, Bassey and Ayang (2014). The following research questions are listed for this purpose:

(i) What is the most meaningful causal model describing the causal effects among the variables x_1, x_2, x_3, x_4 and x_5?
(ii) What are the direct, indirect and total causal effects of variables x_1, x_2, x_3, x_4 on variable x_5?

6. Steps in Conducting Path Analysis

- Step 1: Find the zero-order correlations or obtained correlations among the five variables. This is shown in Table 1

1.	SES (x_1)	1.000			
2.	PAS (x_2)	.029	1.000		
3.	PGI (x_3)	.021	.314	1.000	
4.	TSR (x_4)	.025	.445	.265	1.000	
5.	ATT (x_5)	.018	.451	.298	.354	1.000

Table 1: Original Correlation Matrix

- Step 2: Conduct multiple regression analysis. In this example, four multiple regression analyses were conducted: These are:

One: x_2 was regressed on x_1
Two: x_3 was regressed on x_1
Three: $x_{4\ was}$ regressed on x_2 and x_1
Four: x_5 was regressed on x_4, x_3, x_2, and x_1.

Variables x_2, x_3, x_4, and x_5 were dependent variables in the regression runs conducted while the variables on which they were regressed were independent variables.

- Step 3: The path coefficients or B weights obtained from the regression runs in step 2 should be presented in a Table. They should further be reflected on the hypothesized model in fig. 1. Use the .05 criterion to delete the paths that are not significant. Table 2 presents the path coefficients. On the basis of the criterion of statistical significance of .05, delete paths p_{21}, p_{31}, p_{41}, and p_{51}. Draw up a new model reflecting only the paths whose beta weights are significant. The new model so produced as shown in figure 3, becomes the more parsimonious and meaningful model.

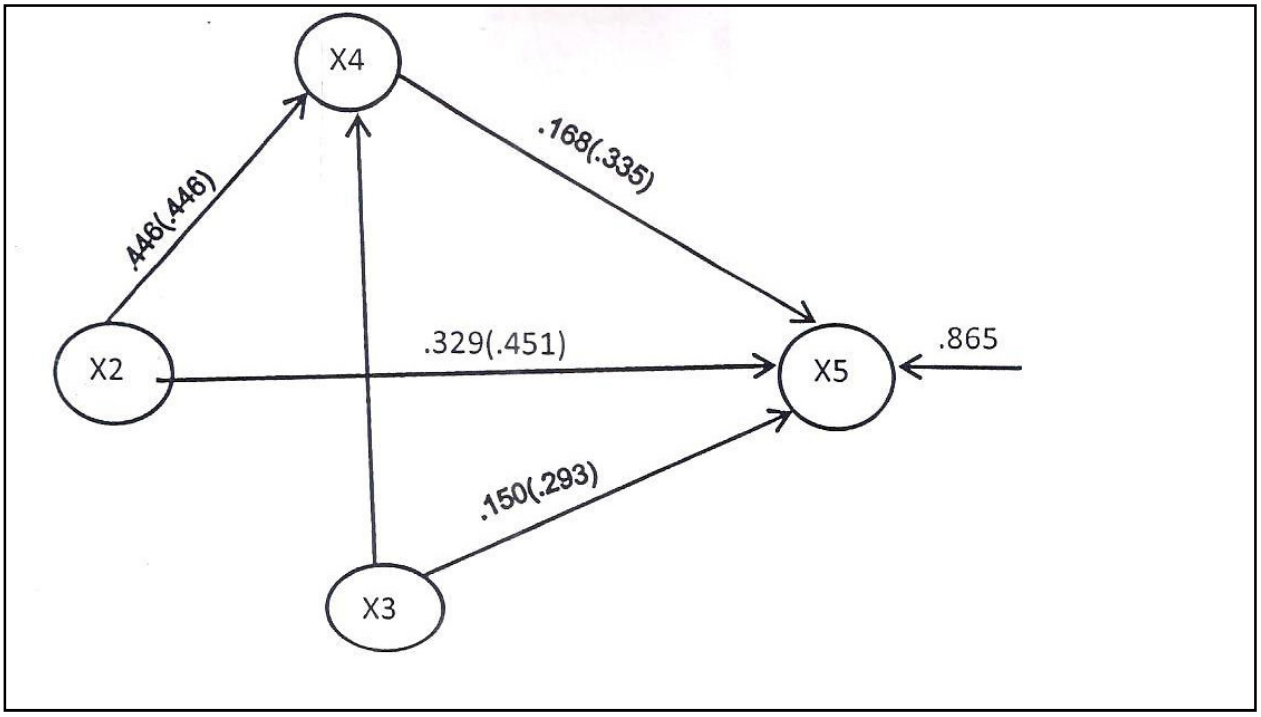

Figure 3: New Causal Models (Trimmed)

- Step 4: Find out how well the parsimonious model is consistent with empirical data. This is done by obtaining the reproduced correlations and comparing them with the original correlations on the basis of the criterion of .05. Reproduced correlations are obtained by tracing the paths in the researcher's presumed model. Path tracing, otherwise referred to as path decomposition, is a process that yields the correlation coefficient for each path, which is equal to the product of all coefficients in the path (Adegoke, 2013). The formula for calculating the reproduced correlation coefficient indicating the effect of variable x on y is given by: $r_{xy} = \frac{1}{N}\sum z_x z_y$ eqn. 6

To obtain $r_{12,}$ for example, equation 6 is stated $r_{12} = \frac{1}{N}\sum z_1 z_2$. The value ($r_{12}$) is the calculated or reproduced correlation. It is derived using the appropriate structural equation (see eqn. 2). Recall that all 'x' values have been transformed to 'z'. Thus, eqn.2 can be written $z_2 = p_{12}z_1 + e_2$. To calculate the first reproduced correlation coefficient – r_{12}, still relying on data adapted from Ojating, Bassey and Ayang (2013), $z_2$is substituted in equation 8 as follows:

$$r_{12} = \frac{1}{N}\sum z_1 z_2.$$
$$= \frac{1}{N}\sum z_1(p_{21}z_1 + e_2)$$
$$= \frac{1}{N}\sum z_1 z_1 p_{21} + z_1 e_2$$
$$= 1 + p_{21} + 0$$
$$= p_{21}$$

Statistically, the variance of standard score $\frac{1}{N}\sum z_1^2$ is equal to 1 and $z_1 e_2 = 0$.

Therefore, $r_{12} = p_{21} = .029$. To obtain r_{13} the same procedure is followed:

$$r_{13} = \frac{1}{N}\sum z_1 z_3.$$

Recall that $z_3 = p_{31}z_1 + e_3$

$$r_{13} = \frac{1}{N}\sum z_1(p_{31}z_1 + e_3)$$
$$= \frac{1}{N}\sum z_1 z_1 p_{31} + z_1 e_3$$
$$= 1 + p_{31} + 0$$
$$= p_{31} = .021$$

The procedure for obtaining the reproduced bivariate correlation coefficients as used here applies for the remaining cases: r_{14}, r_{15}, r_{23}, r_{24}, r_{25}, r_{34}, r_{35}, and r_{45}. The reproduced correlation matrix is presented in Table 2.

1. SES (x_1)	1.000				
2. PAS (x_2)	.029	1.000			
3. PGI (x_3)	.021	.001	1.000		
4. TSR (x_4)	.025	.446	.140	1.000	
5. ATT (x_5)	.019	.451	.298	.355	1.000

Table 2: Reproduced Correlation Matrix

Combining Tables1and 2, we have the obtained and the reproduced at the top and bottom of the diagonal respectively in Table 3.

S/N	Variable	SES(x_1)	PAS(x_2)	PGI(x_3)	TSR(x_4)	ATT(x_5)
1	SES (x_1)	1.000	.029	.021	.025	.018
2	PAS(x_2)	.029	1.000	.314	.445	.451
3	PGI(x_3)	.021	.001	1.000	.265	.298
4	TSR(x_4)	.025	.446	.140	1.000	.354
5	ATT(x_5)	.019	.451	.298	.355	1.000

Table 3: The original and reproduced correlation matrix of social environment variables and students attitude to school

Notice in Table3 that it was only in two cases (z_3z_2 and z_4z_3) that discrepancies in correlations exceeded .05, all others were less than .05. This suggests the hypothesized model fits the empirical data.

- Step five

Obtain the total, direct and indirect effects of the predictor variables (z_1, z_2, z_3, z_4) on the criterion variable (z_5). The total effects are the values of the zero-order correlation between the predictor variables and the criterion. The direct effects are the standardized regression weights obtained from the structural equations. Therefore, the difference between the Pearson's (or zero-order) correlation coefficients and the weights yields the indirect effects as summarized in Table 4.

Variables	Total effect	Direct effect	Indirect effect
X_1(SES)	.018	.002	.016
X_2(PAS)	.415	.329	.122
X_3(PGI)	.298	.150	.148
X_4(TSR)	.354	.168	.186

Table 4: Total effects of the predictors of $x_5$that are direct and indirect.

From Table 4, notice that the total effect of z_1(socioeconomic status) on z_5 (attitude to school) is the least which parental academic stimulation has the highest total effect on attitude to school the predictor variable with the highest direct effect on attitude to school is teacher-student relationship (.049).

7. Conclusion

Path analysis is essentially about the decomposition of causal connections among variables into direct and indirect effects. It attempts to fill the gap of univariate statistics that examines variables in very strict isolation, without considering the complex inter-relationships which subsists among variables especially in the behavioural sciences and education.

8. References

i. Adegoke, B. A. (2013). Multivariate statistical methods for behavioural and social sciences research. Ibadan: Esthom Graphic Prints.
ii. Blalock, H. M. (1971). Causal models in the social sciences. New York: Norton.
iii. Duncan, O. D. (1996). Path analysis: Sociological examples. American journal of Sociology, 72, 1-16.
iv. Kline, R. B. (2005). Principles and practices of structural equation modeling(2nd ed.). New York: Guildford press.
v. Kerlinger, F. N. & Pedhazur, E. (1973). Multiple regression analysis in behavioural research. New York: Holt Rinehart and Winston.
vi. Kerlinger, F. N. (1980). Foundation of behavioural research (2nd ed.). New York: Holt Rinehart Winston Inc.
vii. Land, K. C. (1969). Principles of path analysis. In E. F. Borgatta (Ed.). Sociological methodology. San Francisco: Jossey – Bass.
viii. Ojating, H. A., Bassey, S. W., Ayang, (2014). Causal modeling of social environment variables as determinants of senior secondary school students' attitudes to school. Journal of Teacher Perspective, 8(3),
ix. Pedhazur, E. (1982). Multiple regression in behavioural research. New York: Holt, Rinehart and Winston.
x. Wright, S. (1921). Correlation and causation. Journal of Agricultural research,20, 557-583.