KM 19(1985)
pag 73 - 87

PATH ANALYSIS FOR MIXED QUALITATIVE AND QUANTITATIVE VARIABLES

by

Abby Z. Israëls*)

Abstract

In this report path analysis models are considered for mixed qualitative/ quantitative variables. Only endogenous variables that are dependent in all its relations are supposed to be quantitative, but this restriction can easily be dropped. Qualitative variables are handled using a dummy-variable for each category. Parameters are estimated by ordinary least squares. The method allows decomposition of total effects in direct and indirect effects, which makes interpretation easier.

*) Netherlands Central Bureau of Statistics, Department for Statistical Methods, P.O. Box 959, 2270 AZ VOORBURG. Tel. 070-694341. The views expressed in this paper are those of the author and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics. The author wishes to thank an anonymous reviewer for several valuable suggestions.

1. Introduction

Path analysis may be considered as a system of regression analyses for quantitative variables. The technique was introduced by Wright (1921), however in terms of decomposition of correlation coefficients. A description of the technique can be found for instance in Duncan (1975), Kendall and O'Muircheartaigh (1977) and Saris and Stronkhorst (1984). Various articles have been written about path analysis for discrete variables. In Heise et al. (1975) some of them are compiled. However, most of those articles handle only binary variables, which can be incorporated as a numerical variable in a regular path analysis by assigning the values zero and one to their categories. Also transformations like a logit transformation are considered for binary dependent and intervening variables. (In this paper three types of variables are distinguished: 'dependent variables', which are dependent in all its relations, 'exogenous variables', which are independent in all its relations, and 'intervening variables', which are dependent in some but not in all of its relations.) Goodman (1973) developed a form of path analysis in which all variables may have more than two categories. Its estimation is based on theory of maximum likelihood for multi-way tables. Here, we assume that there is one quantitative dependent variable, although this demand is not strictly necessary. All other variables in the path model may be either qualitative or quantitative. In our procedure one dummy-variable is defined for each category of each qualitative variable. Under this operationalization, path analysis may be applied to the set of all quantitative and dummy-variables. Since dependencies between parameters arise when considering all dummies, there is an opportunity to choose a parametrization that gives a good interpretation. Boyle (1970) already suggested to include dummy variables into a path diagram and Lyons and Carter (1972) elaborated and corrected his work. The main difference with their way of introducing dummy variables into the path diagramn is that they used K-1 dummy variables for each K-categorical variable in order to prevent dependencies between the dummy variables. However, as will be shown here, an analysis of variance like parametrization gives results that are much better interpretable in the case of nominal variables.

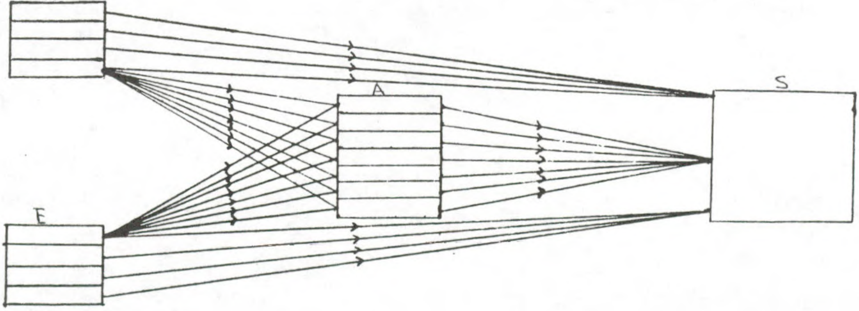
There are other ways of including qualitative variables into a path diagram. Muthén (1984) deals with discrete ordinal variables that are intrinsically continuous, but are measured in a discrete way only by lack of a good measurement instrument. He considers underlying latent variables which are multinormally distributed. Here the qualitative variables are (considered to be) discrete by nature, either nominal or ordinal. Moreover, it is assumed that there is no measurement error in the sense of no errors of placement into categories (Werts and Linn, 1972).

Section 2 discusses path analysis for qualitative exogenous and intervening variables, while in section 3 an example is given. In section 4 we discuss analysis with mixed qualitative/quantitative variables. Section 5 concludes.

2. Path analysis for qualitative variables

In figure 1 a simple example of a path diagram is given. It contains four variables: the dependent variable S, the intervening variable A and the exogenous variables M and E. Variable S is assumed to be quantitative, and A, M and E to be qualitative with seven, four and four categories respectively. In the diagram each category is presented by one layer. There are arrows from each category of M and E to each category of A (only those from the last category of M and the first category of E are given in figure 1), and arrows from each category of M, E and A to S.

Figure 1. Path diagram of variables S, A, M, and E



We distinguish three types of regression analysis (analysis of variance):

1. regression of S on M+E+A ('+' means that interactions are excluded);
2. regression of S on M+E;
3. regression of each category of A on M+E.

Strictly speaking, figure 1 is a combination of only types 1 and 3.

For notation, we introduce the set of dummy-variables $x_a^{(A)}$ (a=1,...,7), $x_m^{(M)}$ (m=1,...,4) and $x_e^{(E)}$ (e=1,...,4) for the variables A, M and E. (A dummy-variable has scores one and zero, according to yes or no belonging to the corresponding category.) Further, $p_a^{(A)}$, $p_m^{(M)}$ and $p_e^{(E)}$ are the fractions of individuals in each category of A, M and E. Replacing S by y, the three regression formulae are respectively

$$y = \mu + \sum_m M_m x_m^{(M)} + \sum_e E_e x_e^{(E)} + \sum_a A_a x_a^{(A)} + \varepsilon \ , \tag{2.1}$$

$$y = \mu + \sum_m M_m^* x_m^{(M)} + \sum_e E_e^* x_e^{(E)} + \varepsilon^* \ , \tag{2.2}$$

$$x_a^{(A)} = p_a^{(A)} + \sum_m M_{am} x_m^{(M)} + \sum_e E_{ae} x_e^{(E)} + \varepsilon_a \qquad (a=1,\ldots,7) \ . \tag{2.3}$$

Parameters for all nine regressions are estimated by ordinary least squares. As restrictions on the parameters, in order to make them unique, we choose

$$\sum_m p_m^{(M)} M_m = \sum_e p_e^{(E)} E_e = \sum_a p_a^{(A)} A_a = 0 \ , \tag{2.4}$$

$$\sum_m p_m^{(M)} M_m^* = \sum_e p_e^{(E)} E_e^* = 0 \ , \tag{2.5}$$

$$\frac{1}{p_a^{(A)}} \sum_m p_{am}^{(AM)} M_{am} = \frac{1}{p_a^{(A)}} \sum_e p_{ae}^{(AE)} E_{ae} = 0 \qquad (a=1,\ldots,7) \ . \tag{2.6}$$

Here $p_{am}^{(AM)}$ is the fraction of people belonging to category a of A and category m of M, and $p_{ae}^{(AE)}$ is defined analogously. So, for each explanatory variable regression coefficients have weighted mean zero in each equation, using the fractions of individuals as weights. As a result, μ is the mean score on S (or y). There is a dependence among the seven regression equations of (2.3), which is easily seen when summing these equations over a. One might therefore drop one of the categories of A, but this will make interpretation much more difficult.

Regression (2.3) is a binary regression for each a. In such cases one usually performs an initial transformation on the dependent variable in order to get the predicted values within the interval (0,1); examples are logit and probit transformations. These kinds of regressions, however, are not compatible with the other two regressions (2.1) and (2.2) in our causal system. They will only be compatible when using the ordinary regression of $x_a^{(A)}$ on M+E for each a. (Compatibility, here, means the possibility of decomposition of the total effect for each category of M and E on S into the direct effect and the indirect effects via the categories of A.) As mentioned in Keller and Verbeek (1984) for the case of regressions like (2.3) "the estimated proportions may be out of the 0-1-range, and we have also sacrified some statistical efficiency (variance of the estimators) to a gain in ease of use." See their work for a further discussion on this subject.

Formula (2.1) gives the direct effect for some category m of M on S by the coefficient M_m; this coefficient is adjusted for the effects of E and A on S, and for the mean score on S. Formula (2.2) gives the total effect for category m of M on S by the coefficient M_m^*; here there is only adjustment for the effect of E on S and for the mean score on S. The indirect effect of category m of M via category a of A is equal to A_aM_{am}. These indirect effects can be found by substituting (2.3) into (2.1), which gives

$$y = [\mu + \sum_a p_a^{(A)} A_a] + \sum_m [M_m + \sum_a A_a M_{am}] x_m^{(M)} + \sum_e [E_e + \sum_a A_a E_{ae}] +$$

$$+ [\varepsilon + \sum_a A_a \varepsilon_a] \ . \qquad (2.7)$$

The first term on the right hand side of (2.7) is equal to μ, because of (2.4). In the second term M_m is the direct effect for category m of M, A_aM_{am} the indirect effect via category a of A and $\sum_a A_aM_{am}$ the total indirect effect via A. Comparing (2.7) to (2.2) we notice that

$$M_m^* = M_m + \sum_a A_a M_{am} \qquad (m=1,\ldots,4) \ , \qquad (2.8)$$

$$E_e^* = E_e + \sum_a A_a E_{ae} \qquad (e=1,\ldots,4) \ , \qquad (2.9)$$

$$\varepsilon^* = \varepsilon + \sum_a A_a \varepsilon_a \ . \qquad (2.10)$$

Formulae (2.8) and (2.9) state the decomposition of total effect in direct and indirect effect. The reason that these identities hold is the assumption of independence between the disturbances $\varepsilon, \varepsilon_1, \ldots, \varepsilon_7$, which underlies the independent treatment of all regressions.

Usually in path analysis, the total effect is simply defined as the sum of direct and indirect effects. For a saturated model like in figure 1, it was possible as well to define the total effects M_m^* and E_e^* as coefficients of a separate regression equation; this made decomposition of effects identical to decomposition of regression coefficients. In a non-saturated model a decomposition of regression coefficients like (2.8) and (2.9) is not always possible. For instance, if the arrows between E and A would be dropped, the parameters M_m from (2.1) and M_m^* from (2.2) would not change, but the parameters M_{am} from (2.3) would, unless the parameters E_{ae} in the saturated model would all be equal to zero. Of course, one only would drop arrows, if they represent no or small effects.

3. Example

The example is taken from the Dutch Life Situation Survey 1977 (CBS, 1978). We have four variables: Satisfaction (S), Activity (A), Marital status (M) and Education (E), see figure 1. Satisfaction has got category numbers 1,2,...,5 as scores for its categories 'not too satisfied', 'rather satisfied', 'satisfied', 'very satisfied', and 'extremely satisfied'. So, Satisfaction is handled as if it were a numerical variable. Other scales than this five points scale are possible, but beyond the scope of this paper.

The descriptions of the categories of the other three variables are given in table 1. In this table the results of regressions (2.1) and (2.2) are given.

Table 1. Regressions of S on M+E+A and on M+E (standard deviations in parentheses)

Explanatory variable	Category	Coefficients for regression on M+E+A	Coefficients for regression on M+E	Number of people
Marital status	Married	.0422 (.0109)	.0406 (.0096)	2 940
	Widowed	-.3460 (.0598)	-.3075 (.0582)	268
	Divorced	-.7753 (.1316)	-.8597 (.1333)	53
	Single	.0116 (.0364)	.0101 (.0301)	847
Education	Low	.0080 (.0125)	-.0022 (.0124)	2 492
	Medium	.0089 (.0263)	.0229 (.0265)	1 027
	High	-.0109 (.0465)	.0135 (.0466)	401
	Unknown	-.1315 (.0686)	-.1246 (.0696)	188
Activity	Employed	.0277 (.0160)		1 987
	Unemployed	-.6921 (.1192)		65
	Not able to work	-.8652 (.0877)		118
	Retired	.1552 (.0537)		314
	Student	.0631 (.0609)		336
	Housewife	.0152 (.0253)		1 203
	Unknown	.0446 (.1035)		85
General effect		3.1804	3.1804	4 108
R^2		.0508	.0184	

The regression of Satisfaction on M+E+A shows the regression coefficients M_m, E_e, and A_a from (2.1), i.e. the direct effects. The regression of Satisfaction on M+E gives the total effects M_m^* and E_e^*. Differences in coefficients between both columns for M and E are due to indirect effects via Activity. So, for divorced people there is an indirect effect of -.8597 - (-.7753) = -.0844 via the categories of Activity. In our case indirect effects are small, but still interpretable, as we will see later on, when considering the regressions (2.3). Most obvious effects in table 1 are the negative effects of being widowed, divorced, unemployed or not able to work. However, total influence of the three characteristics on Satisfaction is small (R^2=.0508 for the first regression). Especially Education has little influence.

In table 2 the seven regressions of the Activity categories on M+E are given, i.e. the results of formula (2.3). Standard deviations are estimated using the method of Keller and Verbeek (1984). General effects are equal to fractions of people for each activity ($p^{(A)}$). The values of R^2 show that being a student or not is strongly related to Marital status and Education; single people with a high level of education have a high proportion of students. Despite these rather strong relations, indirect effects via the category 'student' are not too large because of its rather weak relation with Satisfaction (A_5=.0631). Indirect effects via some Activity category a are equal to $A_a M_{am}$ and $A_a E_{ae}$ for categories of M and E, adjusted for each other. These values are given in table 3.

Table 2. Regression of Activity categories on M+E (standard deviations in parentheses)

Explanatory variable	Category	Regression of Activity categories[a)]						
		1	2	3	4	5	6	7
Marital status	Married	.0376 (.005)	-.0060 (.001)	.0013 (.002)	-.0108 (.003)	-.0742 (.003)	.0545 (.004)	-.0024 (.001)
	Widowed	-.3470 (.030)	-.0125 (.007)	-.0105 (.010)	.1953 (.016)	-.0749 (.016)	.2150 (.027)	.0347 (.008)
	Divorced	-.0981 (.068)	.0595 (.017)	.0450 (.023)	.0152 (.036)	-.0609 (.037)	.0601 (.062)	-.0207 (.019)
	Single	-.0144 (.015)	.0210 (.004)	-.0039 (.005)	-.0252 (.008)	.2850 (.008)	-.2611 (.014)	-.0014 (.004)
Education	Low	-.0590 (.006)	.0008 (.002)	.0127 (.002)	.0197 (.003)	-.0096 (.003)	.0349 (.006)	.0005 (.002)
	Medium	.0879 (.014)	-.0004 (.003)	-.0192 (.005)	-.0297 (.007)	-.0033 (.007)	-.0325 (.012)	-.0027 (.004)
	High	.1073 (.024)	-.0017 (.006)	-.0260 (.008)	-.0324 (.013)	.0744 (.013)	-.1174 (.022)	-.0041 (.007)
	Unknown	.0732 (.036)	-.0052 (.009)	-.0075 (.012)	-.0302 (.019)	-.0131 (.020)	-.0343 (.032)	.0171 (.010)
General effect		.4837 (.008)	.0158 (.002)	.0287 (.003)	.0764 (.004)	.0818 (.004)	.2928 (.007)	.0207 (.002)
R^2		.0638	.0110	.0105	.0518	.2991	.1141	.0053

a) 1 = employed; 2 = unemployed; 3 = not able to work; 4 = retired; 5 = student; 6 = housewife; 7 = unknown.

In agreement with table 1, the category 'divorced' has the highest total indirect effect (in absolute value). Being divorced has not only a negative direct effect on Satisfaction, it also has a negative indirect effect via the categories 'unemployed' and 'not able to work'. From table 2 it appears that being divorced has a 'positive' influence on the proportion of people that is unemployed or not able to work, and from table 1 that belonging to these activity categories leads to less satisfaction. Both influences together account for the rather strong indirect effects. Indirect effects of .01 and higher (in absolute value) are underlined in table 3. The effect for single people via 'student' has already been mentioned. Further, we see an impact of being not able to work on the total effect of Education on Satisfaction. These people have a low education, on the average; it is this indirect effect which is responsible for the negative influence of having a low education on Satisfaction (adjusted for Marital status). Finally, we see that apart from the direct effect of dissatisfaction for widowed people, there is a positive indirect effect on Satisfaction via the category 'retired'.

Table 3. Effects of categories of Marital status and Education on Satisfaction; indirect, direct and total effects

Type of effect	Category[a)]							
	Marital status				Education			
	1	2	3	4	1	2	3	4
via Employed	.0010	-.0096	-.0027	-.0004	-.0016	.0024	.0030	.0020
via Unemployed	.0041	.0086	-.0411	-.0145	-.0006	.0003	.0012	.0036
via Not able to work	-.0011	.0091	-.0389	.0034	-.0110	.0166	.0225	.0065
via Retired	-.0016	.0303	.0024	-.0039	.0031	-.0046	-.0050	-.0047
via Student	-.0047	-.0047	-.0038	.0180	-.0006	-.0002	.0047	-.0008
via Housewife	.0008	.0033	.0009	-.0040	.0005	-.0005	-.0018	-.0005
via Unknown	-.0001	.0015	-.0009	-.0001	.0000	-.0001	-.0002	.0008
Total indirect	-.0015	.0385	-.0844	-.0015	-.0102	.0139	.0244	.0069
Direct	.0422	-.3460	-.7753	.0116	.0080	.0089	-.0109	-.1315
Total	.0406	-.3075	-.8597	.0101	-.0022	.0229	.0135	-.1246

a) Marital status: 1 = Married; 2 = Widowed; 3 = Divorced; 4 = Single.
Education: 1 = Low; 2 = Medium; 3 = High; 4 = Unknown.

All regression coefficients corresponding to (2.1) and (2.3) can be reproduced in the path diagram. However, with as many categories as we have, the figure will be very unclear. Therefore, in figure 2, we left out the coefficients M_{am} and E_{ae} which are not needed for finding indirect effects $A_a M_{am}$ and $A_a E_{ae}$ greater than .01 (in absolute value), and reordered the Activity categories. Alternatively also values of M_{am} and E_{ae} greater than .1 (in absolute value) might be reproduced or coefficients which lead to values of $A_a M_{am}$ or $A_a E_{ae}$ greater than .005. In general, if many variables and categories are used in a path diagram as well as many arrows, the pictures may become inconvenient. In that case the picture may be reproduced by several sub-pictures.

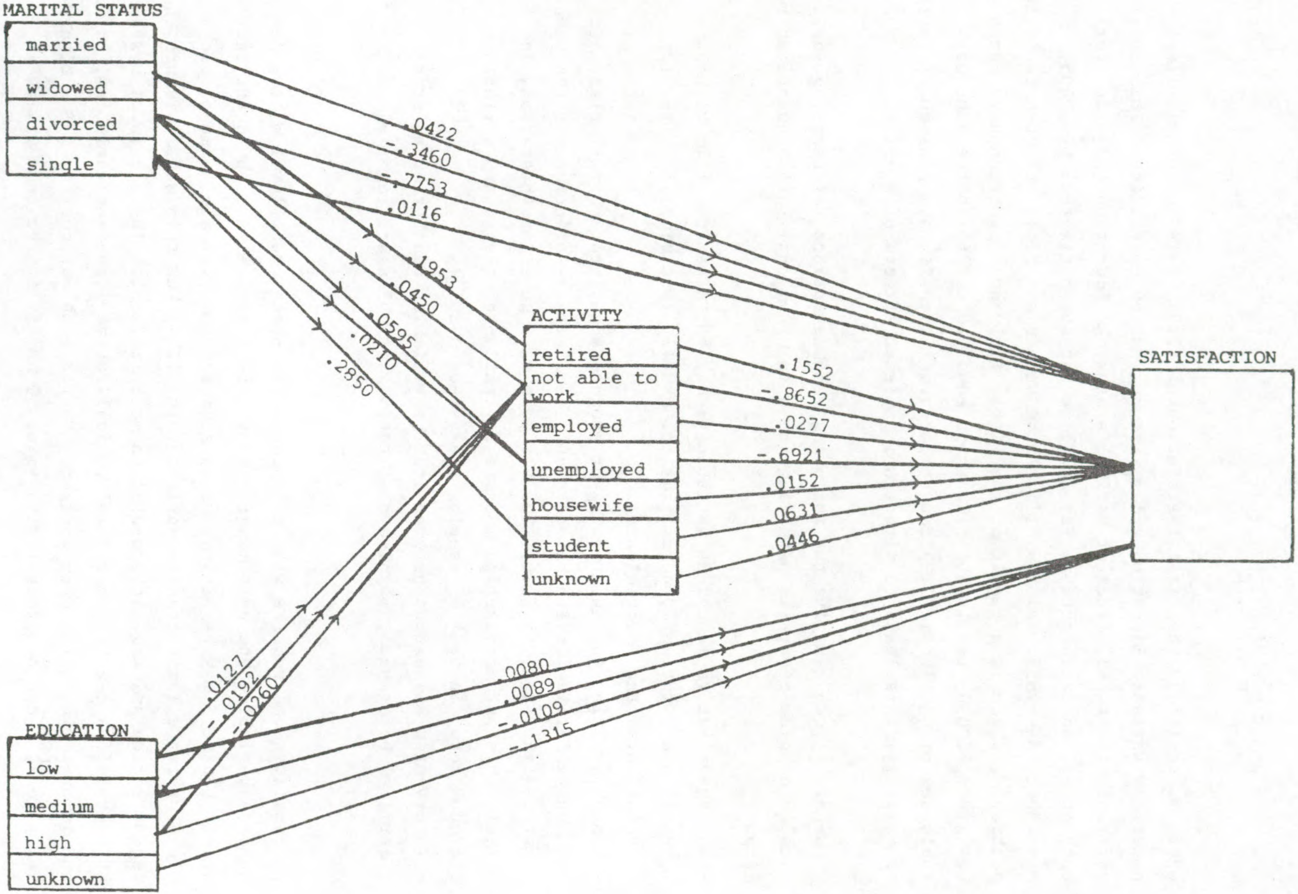
It is not recommended to standardize the dummy-variables like is common practice for quantitative variables in causal models. Coefficients are more difficult to interpret then, especially those of (2.3). Moreover, the sets of restrictions (2.4) and (2.6) would become more cumbersome. See Kim and Ferree (1981) for a discussion on this subject.

4. Analysis with mixed qualitative/quantitative variables

If in figure 1 variable A would be a quantitative variable, the situation would be identical to a usual path analysis; the fact that M and E are qualitative variables does not hinder the analysis when introducing dummy-variables. If A and M are qualitative and E quantitative the situation is much like the one handled in this paper. If one takes E in deviation of its mean, the quantitativeness of E may be seen as a restriction on the parameters E_e, E^*_e and E_{ae}, which must not only fulfil relations (2.4) to (2.6) but must also be a linear transformation of the category scale values.

As an example we may give to the Education categories the values 0 (low), 1 (medium), 2 (high) and 1 (unknown). Standardizing this quantitative variable Education gives the values -.7359 (low), .7629 (medium, unknown) and 2.2617 (high). For the regression of Satisfaction on Marital status + Education we now get the formula (compare to (2.2))

Figure 2. Path diagram of Satisfaction, Activity, Marital status and Education



$$y = \mu + \sum_m M_m^* x_m^{(M)} + \beta_E^* E + \varepsilon^* , \qquad (4.1)$$

where M_m^* satisfies the first restriction of (2.5), E is the standardized numerical variable Education and β_E^* its regression coefficient. This model refers to the total effect of Marital status and Education (adjusted for each other) on Satisfaction. For (4.1) we get the estimates: M_1^*=.0405, M_2^*=-.3092, M_3^*=-.8592 and M_4^*=.0111. These values are close to those found in section 2, which is a result of the little influence Education has. Further we have β_E^*=.0039, μ=3.1804, while R^2 is equal to .0176. In the same way analogues of (2.1) and (2.3) can be performed with Education as quantitative variable; this leads to direct and indirect effects of M and E.

Generally, we can give the following recommendations for handling variables in complex recursive path diagrams with mixed qualitative/quantitative data.

- The quantitative dependent variable(s) may be in natural form without a specific centration or normalization. Usually this simplifies the interpretation of the coefficients.
- Dummy-variables for intervening or exogenous qualitative variables must be unstandardized, with values 0 and 1 and with restrictions on the parameters like (2.4) to (2.6), for the sake of a good interpretation. In fact, this parametrization means a centration of the binary variables around zero, instead of keeping the values 0 and 1.
- Intervening and exogenous quantitative variables can better be standardized beforehand, if also qualitative variables are regarded in the model.

Finally, some remarks will be made on the measurement level of the dependent variable. The path diagram in figure 1 assumed S to be quantitative. Of course S may be binary, using scores 0 and 1. Again, some people might prefer a logit transformation or the like, but this would disturb the decomposition and make interpretation more difficult. The recursive system of regressions can be generalized to qualitative dependent variables, using one dummy-variable for each category. So for each category s of S a binary variable $x_s^{(S)}$ can be created with values 0 and 1, in the same way as is done for A (Activity); one of these binary variables can be found from the other, which leads to a dependence among a set of regressions of $x_s^{(S)}$. For ordinal qualitative variables there are other ways for creating dummy-variables. For instance, one may consider all dichotomizations of S with groupings in a high- and a low-rank group, as considered by Boyle (1970). Notwithstanding the possibility of regarding Satisfaction as an ordinal variable, we used it as a quantitative variable. Main reason was the gain in simplicity of the path diagram. The position of Activity as a qualitative intervening variable is more stressed now.

5. Conclusions

This paper showed that including qualitative variables in a path analysis model does not give any theoretical problems. Apart from the multitude of information, results are very well interpretable.

Sometimes, it may be difficult to decide whether some variable will be included in the model as a qualitative or as a quantitative variable. This is not only a question for ordinal variables, but also for real numerical variables, since handling them as quanti ative variables in a path model assumes linearity. The question may be solved by performing the analysis twice, one time by handling the variable in question as quantitative and one time as qualitative. If relations are much stronger in the latter case, one will decide to publish results only for the model with the qualitative form for that variable. Decisions can also be taken from the qualitative form itself, by looking to the category quantifications.

Literature

Boyle, P.B., 1970, Path analysis and ordinal data. American Journal of Sociology 75, pp. 461-480.

CBS (Netherlands Central Bureau of Statistics), 1978, De leefsituatie van de Nederlandse bevolking 1977, deel 1: Kerncijfers (Well-being of the population in the Netherlands 1977, part 1: Key figures) (Staatsuitgeverij, 's-Gravenhage).

Duncan, O.D., 1975, Introduction to structural equation models (Academic Press, New York).

Goodman, L.A., 1973, Causal analysis of data from panel studies and other kinds of surveys. American Journal of Sociology 78, pp. 1135-1191.

Heise, D.R. (ed.), 1975, Sociological Methodology 1976 (Jossey-Bass, San Francisco).

Keller, W.J. and A. Verbeek, 1984, ANOTA: analysis of tables. Kwantitatieve methoden 15, pp. 28-44.

Kendall, M.G. and C.A. O'Muircheartaigh, 1977, Path analysis and model building. Technical Bulletins no. 2/Tech414 (World Fertility Survey, London).

Kim, J., and G.D. Ferree, 1981, Standardization in causal analysis. Sociological Methods and Research 10, pp. 187-210.

Lyons, M. and T.M. Carter, 1972, Further comments on Boyle's 'Path analysis and ordinal data'. American Journal of Sociology 77, pp. 1112-1132.

Muthén, B., 1984, A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 44, pp. 443-460.

Saris, W.E. and L.H. Stronkhorst, 1984, Causal modelling in non-experimental research; Introduction to the LISREL approach. Series on structural equation models no. 3 (Sociometric research foundation, Amsterdam)

Werts, C.E. and R.L. Linn, 1972, Comments on Boyle's 'Path analysis and ordinal data'. American Journal of Sociology 77, 1109-1112.

Wright, S., 1921, Correlation and causation. Journal of Agricultural Research 20, pp. 557-585.