A CAUSAL MODEL TO EXPLAIN SOURCES OF ERRORS IN DEMOGRAPHIC DATA**

VICTOR JESUDASON*

Social science researchers who collect data to explain and predict human behaviour are aware that their data contain errors. The extent and magnitude of errors depend on many factors such as the nature of data collected, the care with which the study was planned and executed, the population from which the data are gathered etc. Often researchers ignore such errors and draw inferences from data as if there are no errors. Some others assume that the magnitude of such errors may be small and/or random and so ignore them. Very rarely do researchers take pains to examine the magnitude of such errors and to trace their possible sources. Even text books on research methods and sampling theory have often allotted only a few pages to non-sampling errors (for example, Moser and Kalton, 1973; Cochran, 1953). Recently, however, some studies have consciously started examining validity of data (Jesudason, 1975 and references cited therein: Mukherjee 1974 a, b, 1975; Hayness *et al,* 1973; Elder, 1973; Sen Gupta, 1954; Mehree, 1968).

The *purpose* of this paper is to develop a conceptual model for measurement errors, and to examine some possible sources of such errors. It is hoped that such an effort may help in organising data and to facilitate their interpretation in a systematic manner so that corrective measures may be taken to reduce errors.

A Conceptual Model

Concepts in social sciences can be broadly classified into two categories. Some are very close to the operational level and therefore can be measured directly, e.g. sex, age, etc. For some others the operational definitions are only approximations and so they are measured indirectly, e.g. income, need achievement etc. For the purpose of this paper our attention will be confined to the former. Following Blalock (1968) the relationship between true value and measured value may be represented by a causal diagram where X represents the true value, X' the measured value and *e* indicates the

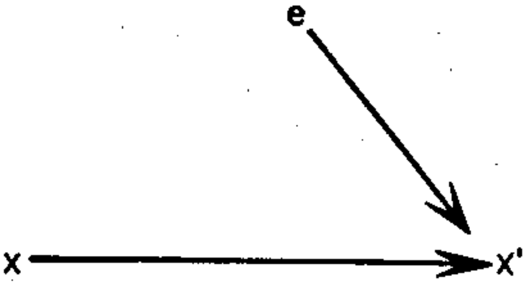


FIG. 1: Causal diagram of relationship between true (X) and measured (X') values.

* Dr. Victor Jesudason is Senior Research Fellow, Council for Social Development, New Delhi.

** This paper grew out of a large research project entitled 'An Experimental Non-Formal Education Project for Rural Women' undertaken by the Council for Social Development, New Delhi. The study was sponsored by the Government of India and funded by the UNICEF. It was under the overall direction of T. A. Koshy, and Prodipto Roy had the overall responsibility for research. Thanks are due to C. Rajalakshmi and K. Satyanadam for computational assistance, and to Vimal Shah and Anita Dighe for critically reading an earlier version of the paper. This study was originally suggested by Prodipto Roy.

source of errors. In the theory of measurement error, the measured value of a variable is represented as the sum of its true value and an error term (Cochran, 1953: 374). Thus,

$$M = T + e \tag{1}$$

where T is the true value, M is the measure of it and *e* an error term. Under this condition, the variance of the measured variable, Var [M] may be expressed as

$$\mathrm{Var}[M] = \mathrm{Var}\ [T] + \mathrm{Var}\ [e] + 2\ \mathrm{Cov}\ [T, e] \tag{2}$$

Ordinarily, the errors are assumed to be uncorrected with the true scores ($r_{Te} = 0$) so that Cov [T, e] is equal to zero. But the variance of the errors is always a non-negative number. So, under the assumption of random measurement error, the variance of the measured score will exceed the variance of the true score by an amount exactly equal to Var [e]. (For a proof of this, see: Gulliksen, 1950; Siegel and Hodge, 1968).

Most social science researchers stop with the assumption that the *e* is a random error. Also the other assumption usually made is that it is randomly distributed with the mean of 0 and standard deviation of 1. In addition, in regression equations, it is also assumed that the errors of the dependent variables (regressand) are uncorrelated with the independent variables (regressors) which are measured without error.

In the measurement of most demographic variables, both floor and ceiling effects operate. For example, some of those with the true value of zero children, can report higher values only. Similarly some of those with high true value of children can only report a lower number. In other words, the errors are not random.

In an earlier paper (Jesudason, 1975) demographic data collected independently from both husband and wife were analysed under various assumptions. Based on that and as an extension to it, it is hypothesised

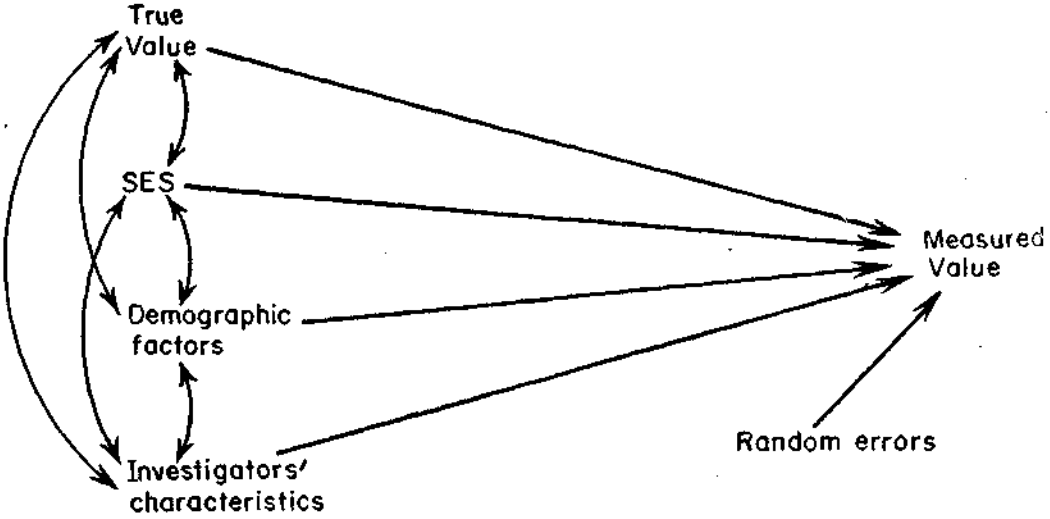


FIG. 2: A Conceptual Model for Measurement Errors.

that the measured values may be caused by (in addition to true value) socio-economic status, demographic factors, and investigator's characteristics and of course random errors. Figure 2 (page 52) displays the relationship in a diagramatic manner.

Data

A subset of data collected for a large study[1] was utilized to test this model. The larger study was designed as a field experiment. As a part of the experimental manipulation, Maternal Child Care (MCC) Centres were established. Free medical care and supplementary foods were provided to selected women through the MCC. The physician (a lady doctor) in charge of MCC visited the MCC Centres once a fortnight and carried out medical check ups.

As part of the bench mark survey, the social science — trained female investigators (eight in number) gathered retrospective pregnancy history[2] from the women who were to participate in the MCC. After about one year of operation of MCC, the physician in charge of MCC once again gathered retrospective pregnancy history from some of the women who participated in the MCC (N = 172). Taking the date of the bench mark survey as the reference point, those children born or dead after that date were eliminated from the latter's data. (See: Hayness, *et al;* 1973 for a similar approach). The data on pregnancy history gathered by social science — trained investigators (SI) and the lady doctor in-charge of MCC, hereafter referred to as Clinical Investigator (CI), were analysed for this paper.

Adequacy of data

The SI's had been working in this area for about nine months before collecting data for this study, and almost all of them had prior survey research experience. As such, they were familiar with the local Telengana dialect used in the villages. For an earlier study (Phase I), the same pregnancy matrix had been used. The SI'S were trained twice in the use of pregnancy matrix — once for Phase I and once for Phase II. They were, therefore, familiar with the matrix format, question sequence and codes. The data which were analysed in this paper were collected from villages where MCC Centres were to be established and where the respondents had expressed willingness to participate in the MCC before the survey was undertaken. Therefore, the investigators were able to build rapport with the respondents very quickly. In terms of question sequence, information on pregnancy history was collected about the middle of the interview. In short, it was an ideal situation for the social science-trained investigators, and we expected minimum of errors in the data collected.

The validation data were collected by a lady doctor. She had been working in this area for about two years, and with the respondents of this study for about a year. Through the MCC she had been meeting these respondents (and their children) at least once a fortnight. As such, she had an excellent rapport with them. If at all these rural women would open up and tell the truth about their pregnancies and their outcomes, it will be to her. It was assumed that the data collected by CI would be more accurate than the data collected by SIs. Whenever discrepancies occurred

[1] For a description of the study, see CSD (1975).

[2] See Shirur (1975) for a more complete analysis of these data.

between the two sets of data, the schedules were manually checked to rule out the possibility of coding or key punching errors.

FINDINGS

Univariate Comparisons

Panel A of Table 1 gives the reliability coefficients for selected aspects of pregnancy. The coefficients, are all above 0.9. This shows that the information on these variables can be collected reliably. Or, it may be that the errors were common for both sets of data. Panel B shows the reliability coefficients for four measures of mortality. In contrast to the earlier panel, the coefficients are small in this panel ranging from 0.3 to 0.7. The coefficient for Total pregnancy wastage (defined as the sum of abortions, miscarriages and still-birth) is the smallest

TABLE 1

RELIABILITY COEFFICIENTS, MEANS AND STANDARD DEVIATIONS OF SELECTED DEMOGRAPHIC DATA GATHERED BY CLINICAL INVESTIGATOR (CI) AND SOCIAL SCIENCE TRAINED INVESTIGATORS (SI). N=172.

Demographic variables	Reliability coefficients	Means and (Std. Dev.)*	
		Data collected by CI	Data collected by SIs
A. *Aspects of Pregnancy*			
1. Total number of pregnancies	.941	4.26 (2.29)	4.23 (2.31)
2. Total number of children born	.935	4.04 (2.20)	4.10 (2.25)
3. Total number of children alive	.924	2.83 (1.78)	2.98 (1.69)
B. *Measures of mortality*			
1. Total pregnancy wastage	.346	0.22 (0.57)	0.20 (0.55)
2. Total perinatal deaths	.591	0.26 (0.59)	0.29 (0.64)
3. Total neonatal deaths	.488	0.09 (0.31)	0.10 (0.36)
4. Total infant mortality	.695	0.59 (0.95)	0.59 (0.87)

* None of the differences between the means are significantly different from zero at .05 probability level.

(0.34). This shows that the data collected through retrospective pregnancy histories on mortality of children have low reliability.

Reading down the columns of means of Table 2 (page 57) it can be seen that the means and standard deviations for both CI and SI are fairly similar. It may be recalled from (2) that the variance of the measured scores will exceed that of the true scores. This study is based on the assumption that the data collected by CI are more accurate than that of the data collected by SIs. Given this, it can be stated that the data collected by CI and SIs are imperfect indicators of the true value, and the latter more so than the former. It follows that the variances of data collected by SI should be larger than that of data collected by CI. Out of seven possible comparisons given in Table 1, this expectation is true for four comparisons only. This shows that, although it is not unambiguous, there is some justification to regard CI's data as more accurate than that of SI's data. It may be recalled that the above formulation holds good under the assumption of random error. The observation that none of the differences between the means are statistically significant indicates that the errors may not be random.

An empirical model

Each of the concepts in the conceptual model (given in Fig. 2) could be measured by many variables. If all those variables are included in an empirical model, the model would become too complicated and its empirical estimation may be difficult. Since the attempt in this paper is first of its kind, the model should not be too complicated. Further, as the term "model" implies, there is a purposeful selection and manipulation of data. In the words of Duncan, *et al.*, (1968:9) the purpose of a model "is not to construct a faithful portrait of reality, but rather to exhibit and rationalize some of the connections between aspects of reality".

An empirical model with a few selected variables is tested in this paper. For such a model, the concept socio-economic status was indexed by the type of house in which the respondent was living.[3] The demographic characteristics of the respondent was indexed by her age (in years). The characteristic of investigators[4] is indexed by age. It was coded as: 30 or above = 1; and less than 30 = 0.

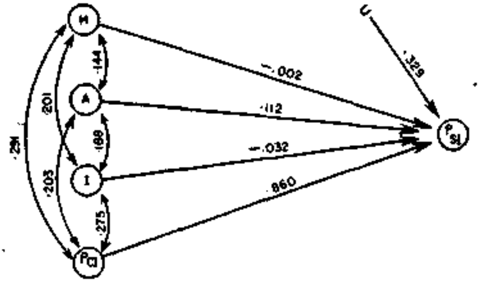
The conceptual model shows that the true value of a measured variable is one of the predetermined variables. By definition the true value is unknown and unknowable. The study was designed with the assumption that the data gathered by the Clinical Investigator, CI, will be closest to the true value. We explicitly assume that the data gathered by the Social

[3] It was coded as: mud hut = 1: kutcha house = 2; house with cement or mortar plastering = 3; and pucca house = 4.

[4] Although many text books on sampling (e.g. Sukhatme and Sukhatme, 1954: 390-391) specify investigators as a source of non sampling error, and an early study (United Nations, 1961) has documented inter-interviewer variations, in the Indian context no serious attempts have been made to examine the characteristics of investigators who may gather inaccurate data. The first such attempt known to me was by Choudhury (1975b) who showed a relationship between marital status and educational level of the investigators with under enumeration of pregnant women.

Science trained investigators, SI, was the dependent or measured variable.

Figure 3 presents the empirical model in a diagramatic form for one dependent variable.



NOTE: variables are: H — Type of house; A — Age of respondents, I — Investigators' age, PCI — Number of pregnancies enumerated by CI, PSI — Number of pregnancies enumerated by SI.

FIG. 3: Empirical causal model for total pregnancies.

The same model can be expressed by a structural equation as:

$$S = \beta SH + \beta SA + \beta SI + \beta SC + \beta Se \quad (3)$$

where the β's are betas or b* in the notation of Walker and Lev (1965). As in the tradition of path analysis (See: Jesudason, 1974) the first subscript indicates the dependent variable, the second the pre-determined variable, and the subscripts referring to the variables held constant are omitted. The βSe refers to the influence of all variables not included in the model and random errors. The model assumes linearity of variables and additivity among the relationships. The model is estimated by least squares procedure.

RESULTS

1. *Number of pregnancies / children*

Table 2 presents the regression coefficients for the various dependent variables. The last line of the table shows that about 90 per cent of the variations in the three dependent variables (namely, total pregnancies, total children born and total children alive) were explained by the model. This indicates that fairly reliable data with regard to these variables can be gathered by social science trained investigators.

The standardized regression coefficients (Panel B) show that age of the respondent was the second most important variable for determining the data collected by SI. This was followed by investigator's age. The coefficients of this variable for two of the dependent variables (namely total pregnancies and total children born) were negative. This shows that the younger investigators (on the average) arrived at higher number of pregnancies and children born than the older investigators. (For the third variable the coefficient, although positive, was of negligible magnitude). In view of the oft reported finding about under reporting of pregnancies and children born, one would assume that the higher numbers were closer to true values than smaller numbers. One could speculate that the older investigators would have had more rapport with the respondents and so would arrive at data which were closer to true values. But the data showed that such a speculation was not justified. It may be that the younger investigators were more diligent in their work than the older investigators.

The regression coefficients (Panel A) show that for each pregnancy enumerated by the CI, the SIs enumeration (on the average) was 0.87. In other words, the SI's enumeration was about 15 per cent less than that of the CI, when other factors were held constant. Similar results can be seen for the other two variables also.

TABLE 2.

Regression Coefficients (and their *t* values) for an Empirical Model to explain Sources of Errors in Selected Demographic data.

Predetermined variables	Dependent variables						
	Total Number of						
	Pregnancies SI	Children born SI	Children alive SI	Pregnancy wastage SI	Perinatal deaths SI	Neonatal deaths SI	Infant mortality SI
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. *Regression Coefficients*:							
1. Type of house	—.004 (0.07)	—.045 (0.91)	—.037 (0.91)	.035 (1.03)	.031 (0.96)	—.022 (1.08)	—.005 (0.12)
2. Age of R	.038 (2.98)	.046 (3.63)	.022 (2.20)	.004 (0.65)	.016 (2.70)	.002 (0.49)	.020 (2.73)
3. Investigator's Age	—.151 (1.24)	—.138 (1.12)	.047 (0.46)	.035 (0.43)	—.029 (0.36)	—.033 (0.66)	—.145 (1.49)
4. Total pregnancies — CI	.866 (23.28)	—	—	—	—	—	—
5. Total children born — CI	—	.855 (22.12)	—	—	—	—	—
6. Total children alive — CI	—	—	.823 (22.05)	—	—	—	—
7. Total pregnancy wastage — CI	—	—	—	.320 (4.47)	—	—	—
8. Total perinatal deaths — CI	—	—	—	—	.609 (9.12)	—	—
9. Total neonatal deaths — CI	—	—	—	—	—	.559 (7.17)	—
10. Total infant mortality — CI	—	—		—	—	—	.602 (11.86)
Constant	—.390	—.376	.168	—.091	—.372	.087	—.218
B. *Standardized regression coefficients*:							
1. Type of house	—.002	—.025	—.027	.077	.060	—.075	—.007
2. Age of R	.112	.140	.088	.048	.169	.034	.155
3. Investigator's age	—.032	—.030	.014	.031	.022	—.045	—.082
4. Total pregnancies — CI	.860	—	—	—	—	—	—
5. Total children born — CI	—	.837	—	—	—	—	—
6. Total children alive — CI	—	—	.869	—	—	—	—
7. Total pregnancy wastage — CI	—	—	—	.330	—	—	—
8. Total perinatal deaths — CI	—	—	—	—	.562	—	—
9. Total neonatal deaths — CI	—	—	—	—	—	.483	—
10. Total infant mortality — CI	—	—	—	—	—	—	.659
Coefficient of determination	.892	.883	.859	.129	.384	.245	.509

Usually in demographic surveys, the data are collected by social science-trained investigators. If the model is properly specified, based on this data it can be estimated that the under enumeration may be about 15 per cent in such surveys when other factors are held constant.[5]

Aspects of infant mortality

Table 2 reports data for four aspects of infant mortality. The last line of the table shows that the per cent of the variation in the data collected by SI explained by the model ranged from 13 per cent for pregnancy wastage to 51 per cent for infant mortality. This shows that if the model is properly specified, the data gathered by SIs were hopelessly inadequate. It should be pointed out that the perinatal and neonatal mortality were classified and recorded on the basis of the age of the child (in days) at the time of death. Since the data collection depended on the respondents' ability to recall the specific age at which death occurred, this could have produced a misclassification. Among the four measures infant mortality had the least error.

Panel B shows that the standardized regression co-efficients for age of respondent (in general) were larger than the coefficients for type of house. This shows that for this sample (i.e. rural illiterate women), the socio-economic status of the respondent had very negligible effect on the data gathered by SIs.

For these set of variables also the investigators' age was negatively related. This shows that the older investigators were enumerating less number of neonatal and perinatal deaths and infant mortality.

Summary

Arguing that it is important for researchers to examine their data for accuracy, a causal model based on prior research was developed to explain sources of errors in surveys. Retrospective pregnancy histories gathered by Social Science trained investigators and by a clinical investigator were used to empirically test the model. It was found that fairly adequate data were gathered by social science-trained investigators with regard to total number of pregnancies, total number of children born and total number of children alive. With regard to other demographic data like pregnancy wastage, perinatal deaths, neo-natal deaths and infant mortality the magnitude of errors was high.

[5] Based on a household survey and a more detailed fertility survey, the *Mysore Population Study* (United Nations, 1961: 222) reported under reporting of births and deaths by 14 per cent and 18 per cent respectively for births and deaths that occurred 15 months or more earlier than the survey date. Based on a resurvey Hayness *et al.*, (1973) reported under reporting of number of children died by 13 per cent.

REFERENCES

Blalock. Hubert M., Jr. 1968 "The Measurement Problem: A Gap Between the Languages of Theory and Research". Pp. 5-27 in HM Blalock and AB Blalock (eds), *Methodology in Social Research,* New York: McGraw Hill.

Chowdhury, Sunanda 1975a "Sample Design" Pp. 121-155 in CSD, *Non-Formal Education for Rural Women.* Final Report submitted to UNICEF, India (Memeo).

1975b An Estimate of Incomplete Enumeration While Developing a Sampling Frame. New Delhi: Council for Social Development (Type script).

Cochran, William 1953 Sampling Techniques. New Delhi: Wiley Eastern.

Duncan, Otis Dudley; David L. Featherman and Beverly Duncan 1968 *Socioeconomic Background and Occupational Achievement.* Ann Arbor: University of Michigan.

Elder, Joseph 1973 "Problems of Cross Cultural Methodology", Pp. 119-143 in M. Armer and AD Grimshaw (eds), *Comparative Social Research,* New York: Wiley.

Gulliksen, Harold 1950 *Theory of Mental Tests,* New York: Wiley.

Hayness, M. Alfred; Aleyamma George; and R. Ramkumar 1973 "Experimental Error in Retrospective Survey", *The Indian Journal of Social Work,* 34: 113-117.

Jesudason, Victor 1974 "An Analysis of the Process of Adoption of Conventional Methods of Family Planning by Male Industrial Workers in Two Factories in India", *The Indian Journal of Social Work,* 35 (October): 205-220.

1975 "Discrepancies Between Husband's and wife's report of Selected Demographic Characteristics". *The Indian Journal of Social Work* 36: 49-60.

Mehree, Jahina 1968 Discrepancies in the Response of Husbands and Wives Regarding the Couples' Practice of Family Planning. *Five Years of Research in Family Planning,* Bombay: Demography Training and Research Centre.

Moser, C. A.; and G. Kalton 1973 *Survey Methods in Social Investigation,* New Delhi: Heinemann Educational Books, Ltd.

Mukherjee, Biswa Nath 1974a "A Factor-analytic Study of Respondent Variability in Demographic Data", *Demography India* 3(2): 375-396.

1974b "Some Correlates of Response Inconsistency in Demographic Data", *Journal of Population Research,* 1 (July-December): 24-43.

1975 "Reliability Estimates of Some Survey Data on Family Planning", *Population Studies,* 29 (March).

Sen Gupta, J. M. 1954 "On the Validity of Fertility Data Collected Through Interview", Calcutta: The Indian Statistical Institute. (Mimeo).

Shirur, Rajani 1975 "Retrospective Pregnancy History", Pp. 186-216 in CSD, *Non-formal Education for Rural Women.* Final Report submitted to UNICEF, India, New Delhi, Council for Social Development (Mimeo),

Sukhatme, P. V.;
and B. V. Sukhatme
1954
Sampling Theory of Surveys with Applications, Bombay, Asia Publishing House.

United Nations
1961
The Mysore Population Study, Report of a Field Survey Carried out in Selected Areas of Mysore State, India, New York: Department of Economic and Social Affairs. The United Nations. Population Studies No. 34 (ST/SOA/Series A/34).

Walker, Helen,
and Joseph Lev.
1965
Statistical Inference, Delhi: Oxford and IBH Publishing Company (Indian Edition).

To explain civil war, we must explain why various and often conflicting micro-level motives combine to produce political violence with the characteristics that we attribute to civil war. If we cannot understand why we get civil war instead of other forms of organized political violence, then we do not understand civil war. a.Â The existence of differences in prediction systems involving test scores across demographic groups continues to be a thorny and unresolved scientific, professional, and societal concern.Â The second is a sequential model that suggests a causal sequence among the dimensions of innovation strategy that may lead to higher performance. We used data from a sample of 149 manufacturing companies to test the models. In an explanatory modeling scenario, however, inference after algorithmic model selection is not a viable option. An intense variable selection process is expected to bias coefficient p-values and deprive us from using the modelâ€™s asymptotic properties. There are various ways to overcome this phenomenon but, in order to explain causal effects, we must rely on domain knowledge to isolate the variables that we consider impactful. For example, an obvious selection of variables would be those related to a patientâ€™s age, gender, country of origin, date on which symptoms started, date on which hospi âˆ' A causal factor. X Such as exposure to benzene at work. 5.Â Â„ Have a standardized protocol for data collection Â„ Make sure sources and methods of data collection are similar. for all study groups Â„ Make sure interviewers and study personnel are unaware of. exposure/disease status Â„ Adapt a strategy to assess potential information bias.Â Â„ Bias is a systematic error in a study and cannot be fixed Â„ Confounding may lead to errors in the conclusion of a study, but, when confounding variables are known, the effect may be fixed. 30. Here, we applied structural equation models to the double-blind randomized controlled trial of simvastatin in secondary progressive multiple sclerosis to investigate causal associations that underlie treatment effects. Our results suggest that beneficial effects of simvastatin on reducing the rate of brain atrophy and slowing the deterioration of disability are independent of serum cholesterol reduction.Â When we deconstructed the total treatment effect into indirect effects, which were mediated by brain atrophy, and direct effects, simvastatin had a direct effect (independent of serum cholesterol) on both the EDSS, which explained 69% of the overall treatment effect on Such covariation-based models of causation are the product of the Humean tradition of radical empiri-cism (Hume, 1739/1978), which posits that humans and other animals rely primarily on observable empirical cues to understand and explain causal sequences. Cur-. Copyright 2003 Psychonomic Society, Inc.Â Taken together, these findings suggest that knowl-edge about a causal mechanism plays an important role in testing causal hypotheses and may even take priority over covariation-based data. A dual-process model of causal reasoning 803.Â The data that are available suggest that these sources of in-formation contribute interactively, as opposed to addi-tively, to causal judgments (Fugelsang & Thompson, 2000).