On the Use of a Conceptual Model in the Empirical Research Setting

I.M. Wilson[1]
C. Timmerman [2]
A. De Muynck[2]
J.B. Levin[1]
I. Beghin[3]
P. van der Stuyft[2]

[1] Statistical Services Centre, University of Reading, United Kingdom.
[2] Unit. of Epidemiology and Biostatistics, Institute of Tropical Medicine, Antwerp, Belgium.
[3] Unit. of Nutrition, Institute of Tropical Medicine, Antwerp, Belgium

Acknowledgements

This work was supported in part by a grant from the CEE (Erasmus programme No. UCP 87–0026B), from the Flemish Ministry of Health, and from the NFWO (PvS): D/E 193/1988

SUMMARY

This paper considers the situation where the process being studied has been analysed and represented by a conceptual model (CM). Operationalising the concepts in the conceptual model is necessary if the CM is to be used to guide the development of empirical research instruments.

This process is illustrated by, and some lessons are drawn from, a conceptual model for ante–natal care take–up by migrant women in Belgium and from questionnaire–based research carried out on a sample of Turkish women in several Belgian cities. However, the general ideas in the paper should be accessible to any researchers interested in CMs.

We consider the targets and limitations of the CM, ways of augmenting its presentation, the selection and validation of measurable variables representative of the ideas in the boxes of the CM, and checking the coverage and content of the proposed set of empirical measurements.

Consideration is given to the statistical analysis of data collected from such empirical research instruments. It is argued that the regression techniques are the essential tool for investigating the relationships represented by the lines of the CM. Particular consideration needs to be given to the design of the research instrument, so that an effective analysis can be carried out. Similarly, the analysis may have implications for the structure of the CM itself.

ACKNOWLEDGEMENT

We are grateful to Dr. David Rogers of the University of Oxford, U.K., and to Matti Lammers and Herwig Mentens of I.T.M., for valuable comments on drafts of this paper.

1. INTRODUCTION : LINKING QUALITATIVE & QUANTITATIVE RESEARCH

This paper is concerned with the development of a useful relationship between different sets of procedures and ideas which are often kept separate in the study of social and behavioural phenomena. It considers how a qualitative approach, aimed at producing an intellectually persuasive case, can be brought together with the quantitative analytical method with its emphasis on data collection and analysis methods.

To support a case that research conclusions are cogent and convincing it is certainly desirable to have both sets of evidence, because the qualitative study and the quantitative will contribute differently to the research results. Frequently the consideration of validity in qualitative research focusses on arguing the case that a particular model is a plausible description of the phenomenon studied. External acceptance may best be seen with the emergence of a consensus view amongst the experts. On the other hand, strictly quantitative research work tends to validate its conclusions more internally, e.g. by reference to internal checks on the consistency of measurements.

Of course there is a range of approaches between the purely conceptual and the totally statistical, and furthermore validity arguments are not wholly determined by the researchers' standpoint on the range from cerebral to mechanical. Below we consider a statistical augmentation of a qualitative model, and the contribution which this broadening may make to the usefulness of the research.

It is intended that the comments made below should be quite widely applicable and generally accessible. They are developed in the context of a specific case-study, but this paper is intended to stand on its own and readers do not require close familiarity with the case-study, which is described in other publications in some detail: see da Silveira et al. (1988), and Levin et al. (1989). Relevant details of the case-study are included below.

2. THE CASE–STUDY :

A CONCEPTUAL MODEL FOR ANTE–NATAL CARE TAKEUP BY MIGRANTS

The general topic of the research was an investigation of factors affecting the take up of antenatal services by Turkish migrant women in Belgium. The first stage of this research was the development of a conceptual model which relates together a number of factors believed to be relevant. The method of constructing the conceptual model follows Beghin (1986) and Beghin et al. (1988). A multidisciplinary team held a series of meetings, "brainstorming sessions", at which factors were identified and fitted into a simplified theoretical representation shown here as Fig. 1. This can perhaps be thought of as a sketch map of the intellectual area in which the research is taking place. Like any map of a territory which is not fully explored, changes are to be expected when research resources are deployed. Features of the map may be somewhat rearranged, new information will be added, and new forms of detailed description will be accumulated which are not easily summarised in the map.

Beghin (1986) makes it quite clear that the interpretation and use of the conceptual model are primarily intellectual rather than mechanical or statistical. One of the manifestations of this is the use of the neutral terms box and line to describe the components of Fig. 1. A conscious process of translation is needed if quantitative models are to be built from these components.

In the same way, the words in the boxes are not usually measurable variables, but rather encoded descriptions of substantial sets of ideas which the brainstorming panel has agreed to bundle together. The lines linking these boxes imply that certain forms of association between closely–linked boxes are hypothesised to be important and meaningful. Figure 1 is therefore not a research endpoint, but rather a framework for research thinking.

An idealised target would be to explain the concept of primary interest, *UTILIZATION OF THE ANC CONSULTATION BY MIGRANT WOMEN*, fully and explicitly in terms of unambiguously defined measurable variables, each corresponding to just one box of the conceptual model. This does motivate the structure of the conceptual model, although a proven form of explanation is some way off. In a broad sense, the immediate or proximate determinants of utilisation are close in the conceptual model to the utilisation box. Logically, historically, psychologically, economically or socially more distant elements of explanation are further away in the diagram and linked through proximate determinants to utilisation.

3. BASES OF CRITICISM OF THE CONCEPTUAL MODEL

(a) A critique is clearly possible which suggests changing the selection of boxes and lines on a priori grounds: this is best left to independent outsiders with relevant research experience of the subject–matter.

(b) Some criticisms of the relationships between content and form in the conceptual model which have been made are that (i) it presents an apparently static view of a situation which is in some instances clearly dynamic, (ii) it does not encompass feedback loops, (iii) it does not illustrate direct connections between boxes separated by one or more other intervening boxes.

Such criticisms are evidence of a misconception that Figure 1 is an analogue of a model from the exact sciences such as an electrical circuit diagram. The conceptual model was always intended to operate at a higher level, so that if for example an electrical engineer set out to design a computer–controlled mensuration system with many electrical circuits, the conceptual model analogous to ours would not be a representation of the circuitry, but rather a guide to the processes of thought and imagination which the engineer should use, in an organised and structured way, to produce a novel, imaginative and powerful mensuration system.

Secondly, the conceptual model is constructed at an early stage of research development, when a detailed explanation may be premature. The criticisms (i) – (iii) above are rather like the complaint that the soils map of a new agricultural area does not explicitly explain its geological history.

Thirdly and most importantly, the conceptual model concerns, in some of its parts, ideas which are intrinsically hard to define or express, mediated through the perceptions of researchers and subjects, and which frequently concern time-limited, sometimes evanescent, phenomena such as *PERSONAL FEELINGS TOWARDS CURRENT PREGNANCY.*

The last 3 paragraphs mean (i) that Figure 1 is necessarily on a consciously-decided plane of simplification, which allows it to be useful in interdisciplinary discourse at the research strategy level, and (ii) that some boxes are duplicated at different places in the model.

It should be noted that the conceptual model illustrates major factors which are persuasively argued to be causally prior to the primary phenomenon investigated – in this case *UTILISATION OF ANTE-NATAL CARE PROVISION.* There is no reason why longitudinal information, retrospective or prospective, should not be included in some boxes of the conceptual model, with the immediate implication of a longitudinal, and dynamic, form of modelling being part of the analysis.

(c) Further development, rather than criticism of the conceptual model, may well result from the use of the model and the discovery from empirical data that the concept descriptions in the diagram are misleadingly vague, over-complicated or otherwise damaging to the use of the instrument in research. Some moves in that direction are made in the report by Levin et al. (unpublished) of findings from the first empirical study of Turkish women's take-up of ante-natal care.

4. AUGMENTING THE CONCEPTUAL MODEL DIAGRAM

Even within the limitations of a two-dimensional "map", there are various possibilities of augmenting the information conveyed in the model diagram. Some of these may be useful to conceptual modellers in encouraging systematic consideration of the characterisation of the words used in boxes.

In the context of preparing the conceptual model, Beghin et al. (1988, Annex I, p.48) recommend restricting the form of model to that shown in Fig. 1. The remarks below do not contradict that. The following comments are concerned with the use of a completed conceptual model, specifically at the stage where it is being scrutinised in the process of preparing an empirical research instrument.

The diagram can be thought of like a hanging mobile: just as this can change in perceived shape, so can the boxes of the diagram be moved relative to one another to bring similar items close together. Closeness or distance in the vertical scale could be manipulated, as statisticians do in cluster analysis "dendrograms", to convey an impression of "intellectual distance" if this were meaningful and useful. Shape, size and colour of the boxes could be used to describe some of the attributes of their intellectual content, data sources, or other typologies, e.g. factual vs. attitudinal. In the same way a greater richness of description could be expressed in the lines between boxes e.g. using line thickness, line pattern and colour hue and intensity, to typify the sorts of connection understood by the brainstormers between the boxes of the diagram. All of these ideas might serve to systematise the description in the conceptual model, to illustrate why some parts are easy and some difficult for data collection, and to focus thinking during further content analysis of the boxes and lines.

There is a clear need to develop the characterisation aspect of the conceptual model and its depiction in the diagram, in the light of data and other evidence about the modelled situation. Even though Beghin's modelling approach explicitly demands starting with a well–circumscribed problem having clearly–defined boundaries of place and time, the model and diagram are still likely to be used in a variety of studies for disparate purposes within the circumscribed boundaries. In the example of our case–study, an investigation of women's attitudes, and a factual questionnaire about their backgrounds and behaviour were two cognate, but distinct, studies in the framework of the model. It has to be recognised that any modification or elaboration of the intellectual structure through one study will not necessarily be suitable or helpful in another.

Similar remarks apply to the physical presentation of the model. There is some danger that commitment to graphics which are difficult, slow or expensive to reproduce will help to "fossilise" the conceptual model. Another reason for care is that since these presentational techniques would be intended to have an aesthetic and perceptual effect on users of the diagram, the graphical presentation itself should be validated to ensure its effects are those anticipated by its designers!

Despite these reservations the process of producing empirical research instruments from the conceptual model might be made more systematic with efforts to superimpose this sort of structure on the original model diagram.

5. CREATING EMPIRICAL RESEARCH TOOLS

It is certainly true that the care and effort which goes into developing a conceptual model needs to be matched with a thoughtful operationalisation of the ideas when they motivate empirical research. Writing about path analysis, Kendall and O'Muircheartaigh (1977) observe that the models "can be very helpful in disentangling a complex set of relationships and, used with care, can add considerably to our knowledge of the mechanisms at work in the population. Not the least of the advantages is the fact that the use of such models forces the researcher to be explicit about his theorising and permits criticism and evaluation of the assumptions built into the models".

Of itself, the conceptual model does not define precisely the variates which are to be measured in empirical study. That step needs to be taken if important parts of the conceptual model are to be tested against actuality by quantitative means. An example where this process is made quite explicit is to be found in Annex 2 of the Guide to Nutritional Assessment by Beghin et al (1988). There is a substantial list of about 60 well–recognised and commonly–used indicators, which relate quite clearly to a causal model. Part of the model and a short segment of the list of indicators are reproduced here as Figures 2 and 3. The indicators refer to the starred boxes in the model.

It is in the operationalisation phase that the conceptual model needs to be made the basis of something more explicit, and any empirical study using it should at least examine the content of the boxes relevant to the study, and develop more explicit sub–models. This content analysis is the basis of face validation of an empirical data collection instrument.

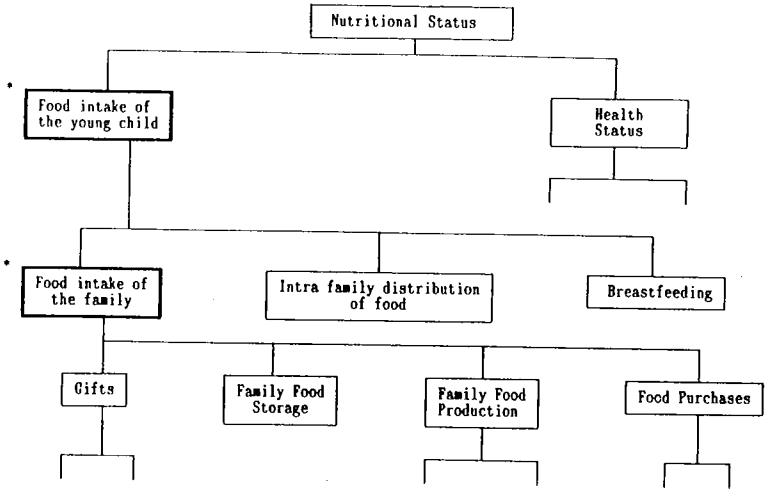

Figure 2

Category of factor	Factor	Data to be collected	Indicator and suggested cut off points
Food intake	Breast–feeding	Weaning age	Average age at weaning (age at which 50% of the infants stop receiving breastmilk).
			% of children still breast fed at 3, 6, 9, or 12 months.
	Food intake of the young child	Daily calorie and protein intake	% of children with calorie intake below recommended daily allowance.
			% of children with protein intake below recommended daily allowance.
	Food intake of the household	Daily calorie and protein intake	% of families eating on average less than the "family basket".
	Biological value of food	Protein quality	Average NPU (Net Protein Utilisation rate) of average diet.
			% of calories from protein origin, average in the group.

Figure 3

(a) VALIDATION

If data collection is based on a survey, each question should be validated as representing an integral part of the observable profile of the concept bundle which is coded in a box, i.e. the response to the question should be informative about an aspect of one of the concepts shown in the model. Equally, where a set of questions is intended to explore aspects of a concept, the overall content of the entire set needs to be assessed in relation to the spectrum of underlying ideas implied in that box of the conceptual model, and consideration of face validity for the set of questions should demand an argument that the "information content profile" of the set is well–balanced in its coverage of aspects of the concept.

Derived measures resulting from the combination of items of the survey should similarly be examined. Both behavioural indices and measures of broad attitudes can be expected to derive from more than one questionnaire item and the form of any proposed compound measurement should also be justifiable in that it illuminates one model concept or in that it clarifies the association between two.

(b) CHECKING THE RELATIONSHIPS OF CONCEPTS AND QUESTIONS

It has been found useful to cross–tabulate concepts under study in an empirical research project with the actual questions and derived measures to be used in survey data collection. If the concepts are labelled C_1, ... C_m and the combined collection of questions and derived measures is denoted Q_1, ... Q_n, a well–organised study should have a suitable array of entries, some appearing in every row and every column of a table like Fig. 4.

A question not connected to any concept may be redundant and a concept not supported by any question is not being measured. A question connected to two or more concepts indicates a failure of the conceptual modelling process, or an ambiguous question.

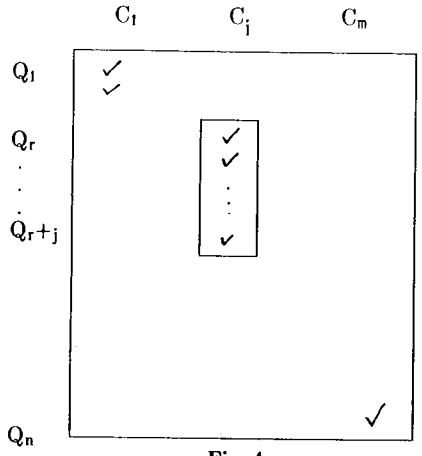

Fig. 4

When considering the batch of Q's related to one concept from the model, it may also be desirable to express the relevant "sub–concepts" or attributes A_1.....A_k as a list like C_1....C_m and to examine this in the same way. See Fig. 5. The aim of any such tabular presentation is to confirm the completeness and balance of the expert consensus, e.g. that Q_r to Q_{r+j} in Fig. 3 "cover" C_i.

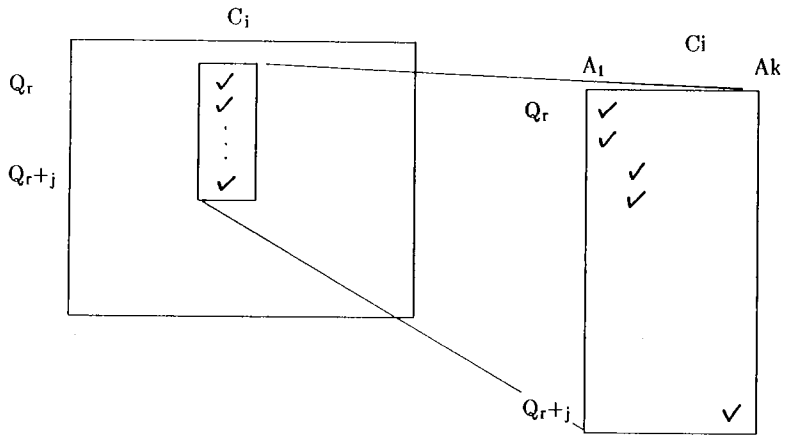

Fig. 5

(c) GENERAL PRINCIPLES OF EMPIRICAL RESEARCH

The "content analysis" of concepts is explored below in an illustration drawn from the case study. Since this content analysis is focussed on the creation of empirical research instruments the illustration is delayed until after some general points about such data collection.

All normal requirements of social scientific research practice remain valid when empirical work is developed out of a theoretical model. For example, questionnaire design should meet normal standards of layout, and data–quality checking must be vigorous e.g. when pregnancy histories are taken, there is merit in incorporating procedures for date–checking, gap–filling and memory–prompting for older respondents, as recommended for example in W.F.S.(1975). A few specific relevant points are developed below.

(i) Definition of complex derived variables

Life history data for migrant women is of course multi–facetted information and systematic attention is necessary to elucidating exactly which attributes and measures could be sensibly used, and how. For example, if a girl is brought up in a Turkish village, then in the space of one year marries, moves to Belgium and has her first child, there is a series of substantial adaptations in her life and limited opportunity to adapt in a leisurely way. If the effects of this process are being investigated, the researcher needs to define measures of *adaptation* and *opportunity to adapt* and to set out hypotheses about their relationships to other variates. Here an apparently straightforward piece of empirical data–collection turns out to require something rather specific from the conceptual model.

(ii) **The difficulty of retrospective research on attitude formation**

The last point refers to a limitation unavoidable with cross–sectional data collection. Individuals' attitudes change with age, societal exposure and personal experience, but not even the most self–aware person will be able to recall the details of the development of their own attitudes. Cross–sectional data may be ascertainable which will profile the current attitudinal stance of an individual, but causal dependencies or even just temporal sequencing in attitude formation can only be guessed and the discovery of statistical associations between life history events and current attitudes generally makes only a limited and informal contribution to reinforcing such guesses. The dependency on conceptual modelling must be substantial in interpreting such situations.

6. AN EXAMPLE OF CONTENT ANALYSIS OF CONCEPTS

This section is necessarily specific to our case–study and can be omitted by readers uninterested in this specialised research field.

Figure 1 includes *AGE OF THE WOMAN, NUMBER AND SEX OF OTHER CHILDREN,* and *PREVIOUS EXPERIENCES WITH PAST PREGNANCIES.* Implicit in the simple–sounding *AGE OF THE WOMAN* are other aspects e.g. *age at migration to Belgium, age at marriage.* It is, however, plausible to characterise *AGE OF THE WOMAN* as a marker in the sense of da Silveira et al. that it is not susceptible to manmade intervention, being a record of historical fact. The same description fits *Number and Sex of Children* and *Previous Experience with Past Pregnancies* so both are also markers.

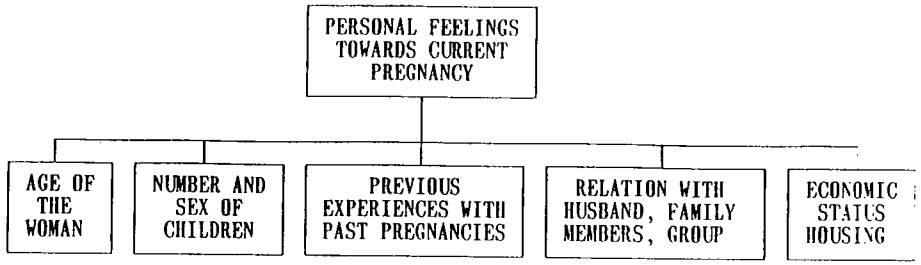

Figure 6

But when potential analyses are considered there is another distinction or characterisation. "Number and Sex of Children" could subsume a measure such as whether the woman perceives herself to have had "enough sons" for her to have achieved the maximum available status in the eyes of her family. Such a variate could be directly used in "explaining" the attitudes expressed in relation to the box above in the conceptual model, *PERSONAL FEELINGS TOWARDS CURRENT PREGNANCY.* On the other hand *AGE OF THE WOMAN* taken on its own does not have the same explanatory role (unless it is used as a crude surrogate for others of the co–determinants of *PERSONAL FEELINGS TOWARDS CURRENT PREGNANCY*). Essentially *AGE OF THE WOMAN* conditions the interpretation of data on other concepts in the same set: this box could be characterised as providing "background", "subsidiary" or "indirect" information – probably used through standardisation or stratification exercises.

In the above example of "explaining" the woman's feelings towards her current pregnancy in terms of whether she considers she already has enough sons, both quantities are attitudinal, but there is no logical difficulty in inter–relating them because the attitudes explicitly relate to factual situations, the explanatory measure being concerned with the position "already" in time and the variable to be explained referring to an anticipated future situation when the new child is added to the family.

Rather differently, where information is strictly attitudinal and where the bases and evolution patterns of the attitudes are not set in a logical/temporal framework, the line between two boxes cannot be construed as explanatory, nor is there verifiable sequencing of two or more such boxes. For example a sequence of boxes such as *SUPERNATURAL BELIEFS*, *ANXIETY TOWARDS RETALIATION FROM OTHERS AND FROM THE ENVIRONMENT*, then *PERCEPTION OF THE BEGINNING OF THE PREGNANCY* may quite plausibly be put into this order, but the measurements of association between variables (construed to relate to these concepts) are correlational and non–directional. On occasion, it may be possible to elicit, from research subjects themselves, how they perceive the processes of sequencing and deduction underlying their attitude patterns. These may be sufficiently consistent across subjects to be useful.

7. THE NEED FOR CONTENT ANALYSIS IN DEFINING EMPIRICAL MEASUREMENTS.

In general terms the content analysis of boxes and lines should condition the planning of empirical work such as data collection, providing a clear perception of the purposes of selecting each variable and the use that can be made of it in understanding subjects' reported behaviour. Thus the "content analysis exercise" is not an unfocussed effort to augment or redraft the conceptual model: rather it is part of the systematic construction of an operational instrument for research – with one eye on the general framework in which the instrument is to be used and the other eye on the analysis which will be legitimate for the data, its interpretation and its usefulness.

Given the conceptual model as framework, quantitative empirical research can proceed at a variety of levels. In what might be described as a "top–down" approach, general questionnaire and other instruments can be used to collect a little data on each of a number of boxes comprising the whole model or a sub–model. In a "bottom–up" approach, the

intellectual content of a single box or a linked pair might be considered in greater depth and efforts made, for example, to describe very explicitly how best to measure *ECONOMIC STATUS* or *PERSONAL FEELINGS TOWARDS CURRENT PREGNANCY*. Obviously a mixture of both approaches is needed. Broad data–collection assumes meaningful variables have been agreed and on the other hand the selection of measures of individual concepts is only reducible to practical proportions through consideration of the use to be made of the measures, e.g. economic status is to be measured because it relates to the woman's feelings towards her current pregnancy, and the measures to be used have to be validated and selected for their relevance to this relationship. This question of the scale at which empirical research is carried out affects the answers to the questions posed in the next paragraph.

Let us say that the expert panel of "brainstormers" delegates a subgroup to prepare an empirical research instrument, such as a questionnaire, in order to assess or measure values in the population, of some variables related to the concepts in the conceptual model. Looking at an individual box and its short coded description, the delegates need to ask themselves certain questions:–

(i) do all of the expert panel share a common mental image of what lies behind the short description of the concept?

(ii) does that shared perception lend itself to meaningful observation in the population to be studied?

(iii) can the expert panel be persuaded to agree on a set of feasible measurements which will constitute a non–trivialised and valid profile of the concept?

(iv) can the resulting measures be collected cost–effectively, quickly and reliably?

The work of developing an instrument, such as a questionnaire, is only feasible where satisfactory answers can be given to preliminary questions like these.

The same type of scrutiny needs to be applied to the connecting lines in the conceptual model. Horizontal lines join together concepts which are at the same level within a submodel, and the vertical lines join the successive levels in a presumed causal sequence.

For example in our case–study, two concepts at the same level within a submodel are *PREVIOUS EXPERIENCES WITH PAST PREGNANCIES* and *RELATION WITH HUSBAND, FAMILY MEMBERS AND GROUP.* These entities are individually quite complicated, but pose extra problems when considered together; at certain levels of explanation that joint involvement may be necessary. Beghin et al. (1988) refer to the need to economise on choice of variables and on data collection. It is primarily in relation to the consideration of the vertical lines in the conceptual model diagram that such parsimony is brought about. The vertical linkage between *PREVIOUS EXPERIENCES WITH PAST PREGNANCIES* and *PERSONAL FEELINGS TOWARDS CURRENT PREGNANCY* provides an example. The data collection about the former, the lower item in the heirarchy, should be strictly limited to measures having an a priori plausible relationship with the concept higher in the heirarchy.

The variables to be measured on the "explanatory" (lower) concept still have to be characterised as being primary explanatory measures or as confounders and effect modifiers whose values are used simply to standardise or correct other more important measurements. Demographic variables such as *age at marriage* are often used in this way.

Another aspect which needs to be looked at systematically in setting up complex data collection exercises is the combination of observed values into indices, which are taken to measure variables that are hard to evaluate directly. For example *satisfaction* is difficult to measure consistently across a set of subjects, but can be acceptably scored e.g. in hospital Patient Satisfaction Surveys. Unlike a straightforward variable such as *age at marriage*,

satisfaction is itself a concept that has to be intellectually dissected in the study context, almost as a microcosm of the conceptual modelling process. It is generally inappropriate to create indices by mechanical methods such as principal component, or factor, analysis which produce data–generated linear combinations a priori unlikely to be good measures of the concept sought. In some sense they may reflect internal variability in the input observed values but this may not be the relevant constituent part of the variability of the index, the part that will correlate with other variables attached to other parts of the conceptual model. Levin et al. (unpublished) provides examples of the commonsense creation of indices described as *compliance score* and *risk score* which measure respectively how far the study subjects follow Belgian norms of antenatal care take–up, and the extent to which the subject's previous history means she is at risk of prenatal misadventure. These are commonplace concepts in health programmes, but have to be defined specifically to suit the study.

8. THE RELATIONSHIP OF THE CONCEPTUAL MODEL TO STATISTICAL REASONING

If a direct translation into statistical terms were possible, Fig. 1 might be construed by statisticians to be an example of a "path diagram", to use the terminology of Duncan (1966). But the essence of a path diagram, and the regression–based path analysis which goes with it, is that the nodes of the diagram represent variables which can be, and have been, measured. Furthermore the path analysis methods assume that precisely the variates represented in the diagram are those which are measured and which it is desired to analyse, to report and to subject to further scrutiny. This is not the case with conceptual models of Beghin's type. Rather they aim to depict and structure higher level or more abstract concepts in the neutrally–labelled "boxes". The intermediate stage of constructing and validating an empirical research instrument has been discussed above, and some results of this process quoted from Beghin et al (1988). These measured variables, already intellectually validated, are the input to the stage of statistical model–building, but because of the conceptual model arrive with more explanatory "power" than otherwise.

Many authors, such as Bibby (1977), caution against over-interpreting regression coefficients, correlations or the so-called path coefficients derived in "path analysis". If a set of variables are measured, connecting them arbitrarily in regression equations does not supply evidence of any direct and unambiguous causal link, even if the correlations found are large, and consistent from study to study. In the causal modelling work of da Silveira et al the logical framework is supplied to allow meaningful causal inferences : the data-based, statistical inference about a measure of association is interpretable in causal terms if the association is close and if the measured variables have been agreed by expert consensus to be in a direct cause-effect relationship. It is important to appreciate that this last condition relates specifically to the measured variables, and not in some vaguer way to the broad concepts which may be found in boxes of the conceptual model

9. STATISTICAL ANALYSIS OF DATA FROM SURVEYS PLANNED USING A CONCEPTUAL MODEL

Of course the general procedures of statistical analysis of data are the same with or without a conceptual model, but the existence of the model supplies some guidance as to the sequence of decisions which are always taken in the analysis of a complicated data set (Palloni, 1987). The most important analytical technique in this context will almost inevitably be regression analysis, since this is the primary method used by statisticians to demonstrate and estimate associations between variables in the presence of nuisance variables which may affect or obscure the relationships.

There are several facets of regression analysis which should be borne in mind when the analysis of data from a complicated survey is being started.

(a) UNDERSTANDING THE DATA STRUCTURE

The statistician must be clear how the variables have been measured. For example with a variable taking integer values 1,... 5, is 3 logically between 1 and 5 and is it equidistant between them? The sampling process needs to be made explicit. For example, certain regression results are affected if observations were collected from clusters of individual respondents.

(b) KNOWING THE APPEARANCE OF THE DATA

Even a small proportion of observations, which are in some way extreme, may have much influence on a regression fit. Before beginning the regression modelling process, the individual variables measured should be reviewed using the simple methods of descriptive statistics. Similarly plots and/or cross-tabulations for 2 or 3 variables at a time should be examined. These elementary methods are likely to reveal any "outliers" – individuals which are very different from the main group in respect of a variable or combination of variable values. These may represent errors in data collection or recording, or genuine and self-consistent results which are too different from the others to be meaningfully combined in summaries. A related finding is that there are subgroups of very different individuals in the dataset. Here again a combination analysis may not be sensible. For example, in our case-study, Levin et al. (unpublished), one index intended to separate traditional from modern attitudes produced a breakdown of 74 subjects

with	63 having values spread in the range	60 to 170,
	7 " " " " " "	–70 to –40,
and	4 " " " " " "	–170 to –140.

Clearly the large gaps between these internally relatively homogeneous groups suggest that the groups differ markedly in type, so that a combined analysis may be assuming common structure which is not there.

(c) REMOVING REDUNDANT VARIABLES

Variables which are essentially the same as one another can certainly be expected to waste analysis time in producing duplicate sets of results with the same information content. Wherever these phenomena arise they should prompt an attempt to reduce the data set – and in parallel to reconsider the conceptual model at the corresponding point, if it seems to encourage the inclusion of redundant variables.

Variables which, although logically different, produce rather similar sets of sample values should also be considered carefully. They may confuse analyses in two contrasting ways; making neither seem important when both are, or making both seem important when neither is. Mechanical/statistical techniques of dealing with this situation are less desirable than intellectual efforts to reduce the data set or to think out a new more refined index which truly reflects the important content of the variable set. This should then replace the offending variables in the analysis process.

(d) SAMPLE SIZE AND THE SIGNIFICANCE OF MODEL TERMS

In many cases, and complex structures with conceptual models are likely to be amongst them, there will be numerous candidate "explanatory variables" or "stimulus variables" to associate with a given response. The process of choosing an apparently sensible model of the relationship usually involves selecting certain exploratory variables which have emerged as "statistically significant" in their joint association with the response variable. This process is difficult and unsatisfactory because a small sample size, i.e. few survey respondents, inevitably means few variables are found to be significant. A very large sample size almost always means many more significances will be observed.

If the conceptual modelling process, and that of choosing the variables to be measured, have been closely linked, the logical process ought to be able to say that certain variables belong in a given relational model regardless of their sample significance, thus avoiding the above problem. Of course a confidently postulated relationship may fail to achieve significance even in studies large enough to detect a very small effect. Methodological failings aside, this tends to challenge the model–builders' preconceptions.

(e) CROSS–VALIDATION AND THE HOLD–OUT SAMPLE

With a reasonable sample size, a well–conducted phase of empirical data collection, a good survey instrument and a good conceptual model, it should be the case that the explanatory variables associate quite closely with their corresponding response. This tends to add credence to all parts of the system. However, statistical logic also requires that a good regression relationship, estimated from a sample, should be capable of predicting events in a second sample, i.e. the relationship estimated in one sample, applied to the explanatory variables found in an independent second sample, should predict the responses in the second sample nearly as well as in the estimation sample. If a hold–out sample is used, one data set is divided – usually at random, and usually into a larger and a smaller part. Estimates are based on the larger part – the smaller is the hold–out. Then the prediction capability is assessed in the hold–out sample. In a basic form of cross–validation the data set is halved and estimates from each half are used for prediction in the other half. Prediction performance is the acid test that a conceptual model and data–generating system are both working effectively together.

(f) THE FORM OF THE RELATIONSHIP

In data–collection and modelling situations where there has been no successful effort to explain logically how variates relate to one another statisticians tend to use simple models linear in the explanatory variables e.g.

$$y = a + b_1x_1 + b_2x_2,$$

unless prompted by the data themselves to introduce polynomial terms, e.g.

$$y = a + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{12}x_1x_2 + b_{22}x_2^2,$$

or transformations, e.g.

$$\log y = a + b_1x_1 + b_2x_2.$$

In principle a conceptual model could be used as the basis of more explicit "mathematical modelling", where an accurate form of equation linking the variables is determined by logical consideration of their relationships. Such a model could then be fitted statistically to data, with much improved interpretative potential. A good candidate for such treatment in our case-study would be a set of demographic variables, where a woman's birth history up to the time of interview could be modelled in some detail, having regard to her age at marriage, current age and other factors. This is an example where the case-study sampling procedures need to be considered : by definition all subjects were pregnant or had very recently given birth, so that sampling biases could arise, and other natural variation present in the broader population is excluded from modelling.

10. APPLICATION OF THESE PRINCIPLES

A description of a statistical modelling exercise carried out on data from Studies on Turkish migrant women is reported separately in Levin et al (unpublished). This paper uses data in some cases loosely connected to the model in Fig. 1, but illustrates some of the principles discussed here.

BIBLIOGRAPHY

Beghin, I. (1986). L'approche causale en nutrition. In Lemonnier, D. and Y. Ingelbeeck: La Malnutrition dans les Pays du Tiers-Monde (pp. 615–628). INSERM, Paris. Série colloque n^0 136.

Beghin, I., Cap, M. and Dujardin, B. (1988). A guide to nutritional assessment. WHO, Geneva.

Beghin, I., De Muynck, A., Van der Stuyft, P. & Mentens, H. (1988). Can the causal model approach contribute to the study of the epidemiology and the control of sleeping sickness? Institute of Tropical Medicine, Antwerp. Working Paper n^0 20.

Bibby, J. (1977). The general linear model: a cautionary tale. In C.A. O'Muircheartaigh & C. Payne (Ed.s): The Analysis of Survey Data : Volume 2 – Model Fitting (pp. 35–80). John Wiley & Sons, Chichester.

Duncan, O. D. (1966). Path analysis : sociological examples. Am. J. Sociol., 72, 1–16.

da Silveira, V.C., De Muynck, A., Timmerman, C. and Van der Stuyft, P. (1988). Development and uses of a conceptual model in the study of antenatal services utilisation by migrant women in Belgium. Institute of Tropical Medicine, Antwerp. Working Paper n^0 19.

Kendall, M.G. and O'Muircheartaigh, C.A. (1977). Path Analysis and Model Building. International Statistical Institute, Voorburg. W.F.S. Technical Bulletin, n^0 2.

Palloni, A. (1987). Theory, analytical frameworks and causal approach in the study of mortality at young ages in developing countries. Ann. Soc. belge Méd. Trop., 67 (suppl. 1), p.31–45.

World Fertility Survey (1975). Interviewers' Instructions. International Statistical Institute, Voorburg. W.F.S. Basic Documentation, n^0 6.