Designing and Coding Survey Instruments for Statistical Analysis*

Ntonghanwah Forcheh
Department of Statistics,
University of Botswana
email: Forchehn@mopipi.ub.bw

Abstract
A forgotten part or generally ignored aspect of survey research is the preparation of the research instrument for suitable and efficient data entry and analysis. What ever is available appears to elude most researchers in the social sciences and humanities. The consequence is that these researchers, some of whom are authorities in their fields, frequently fall victim to poorly designed instruments, which can only be used to answer their research questions in the most extraneous manner. After many years of interacting with researchers from several areas of enquiry, and following urging from some of the researchers, I decided to write this paper as a sort of 'where there is no doctor' type of paper. The problems discussed are illustrated using actual questionnaire items from a wide range of instruments, which I have been involved in at one stage or another. The main focus is how to turn questionnaire items into suitable data for data entry and analysis. Other issues discussed include understanding data handling in statistical packages, especially in relation to the traditional classifications of variables that most empirical researchers are familiar with. The paper highlights challenges in designing questionnaire instruments that are relevant in investigating the research hypotheses in social and behavioural enquiries.

Introduction

At the dawn of Statistics as a scientific discipline, an excited enthusiast predicted its future as follows:

> Even on the very threshold of the scientific realization of this valuable adjunct it is clear that a great future awaits our present 'Statistik', and we may reasonably anticipate that the combination of statistical data and analysis will create a science which will excel every other based on mathematics, not even excepting astronomy, mechanics and physics (Zeuner circa 1860, in Pearson 1978).

Within a century of this prediction, Statistics has now penetrated into more disciplines than any other scientific subject. No university degree programme is nowadays complete without a course in at least elementary statistics. In their revolutionary 'Outcome Based Education (OBE)' programme, the South Africa education ministry has stated as one of its goals 'to make all South Africans Statistically literate'. Even self-confessed qualitative researchers still have to depend on statistical methods for sample selection.

Despite this widespread interest in statistical methods and their use in scientific and behavioural inquiries, many researchers continue to stumble as they try to apply the methods in their research endeavours. During the last decade, I have interacted extensively, as consultant and/or adviser, with researchers and consultants from a wide range of backgrounds, notably energy and environmental studies, market

research, HIV/AIDS surveys, agriculture, career and counselling, information and communication, and have identified recurrent problems to most researchers, especially those from the social and behavioural sciences. These problems are identifying appropriate instruments for investigating research hypotheses, preparing such instruments in forms suitable for quick and efficient data entry and analysis, and selecting the appropriate statistical tools for performing the analyses.

The focus of this paper is to provide a systematic and comprehensive approach to preparing quantitative survey instruments when the resulting data are to be captured and analysed using a statistical software package. The outcome of the suggested methods would be a tidy, well-coded questionnaire aimed at avoiding expensive and time-wasting coding after data collection. Even after a questionnaire has been well designed, determining variables from complex questionnaire items can still be a frustrating endeavour. This has been the case in a number of large-scale studies brought to my attention, in which several thousand questionnaires had been captured into the software packages, only to discover that some of the variable definitions were so bad that the data could not be analysed after data entry. The paper therefore, provides illustrated methods of how to turn questionnaire items into variables for purposes of data entry and analysis. Suggestions are also made on how to select variable names that do not follow the usual methods of using fancy and confusing abbreviations, dealing with open-ended questions, linking questionnaire items to research objectives.

A brief discussion of the need to relate survey questions to research objectives is presented in the next section. This is followed in section 3 by a discussion of the main classes of variables and data types, with emphasis on their representation in computer programmes. The different types of questions encountered in surveys in the social sciences, behavioural studies and related disciplines are presented in section 4, and the associated variables that should be derived from these questions are discussed. Questions considered range from the simplest Yes/No-type questions to the complicated grid, combination and open-ended questions.

Throughout, questionnaire items from actual questionnaires are used. The paper should be relevant to any researcher conducting a survey with the view of analysing the resulting data using a statistical package.

Data Representation in Statistical Packages

A benefit of the computer revolution is that standard data analysis packages now include a wide array of data analysis tools which until recently were available only from purposely written programmes. An understanding of the assumptions behind most of the techniques requires more advanced statistical knowledge than is possible from introductory statistics courses offered to most undergraduate students in the social sciences and humanities. Yet with a few clicks of the mouse, pages of results from each technique can be generated, giving the impression that expert knowledge is no longer required, if only one is computer literate. The inappropriate application of these tools to data analysis is becoming the main source of the misuse of statistics.

A common example of this misuse is when techniques designed for the analysis of quantitative data such as comparing means, standard deviation, Pearson correlation coefficient and fitting linear regression are used to analyse pre-coded qualitative data.

For example, in a paper on job satisfaction, an experienced researcher included the following two questions among other questionnaire items:

Q1: How satisfied are you with your job (please tick one)?

1.Very Satisfied	2.Satisfied	3.Neutral	4.Dissatisfied	5.Very Dissatisfied

Q2: Highest Level of Education

1. None	2. Primary	3. Secondary	4. Tertiary

The responses were appropriately coded and labelled using codes:
1='Very satisfied', 2='Satisfied', 3='Neutral', 4='Dissatisfied' and 5= 'Very Dissatisfied'
1 'None', 2 ' Primary', 3 'Secondary', and 4 'Tertiary'
Data for both questions were then entered into the computer as numbers rather than categories.

The researcher wanted to determine if there was any correlation between job satisfaction and educational level, and to compare the level of satisfaction of male respondents with that of female respondents. The Pearson correlation coefficient between job satisfaction and the level of education was found to be -0.333, with a 'two-sided significance' of 0.000 (n=308). The mean and standard deviations of the level of job Satisfaction and the level of education were tabulated as shown in Table 1 below:

Table 1. Comparison of Job Satisfaction and Level of Education of Female and male respondents:

Gender	**Statistic**	**Job Satisfaction**	**Education Level**
Male	Mean	3.25	2.35
	Std. Deviation	1.24	0.75
Female	Mean	3.47	2.01
	Std. Deviation	1.13	0.78
Total	Mean	3.42	2.08
	Std. Deviation	1.15	0.79

The researcher concluded that there was a significant negative correlation between job satisfaction and the level of education, and further that male respondents were slightly more satisfied (mean = 3.25) than females (mean = 3.47). Also, males had attended higher levels of education (mean=2.35) than female respondents (mean = 2.01). Subsequent discussions were based on these and similar results.

The numbers as presented unfortunately do not prove any link between higher education and job satisfaction. To prove such a link, one needs to compare education levels of those satisfied with levels for those not satisfied, using tools such as contingency table analysis, chi-squared tests for association, ordinal regression, etc. The negative correlation coefficient between job satisfaction and level of education is a result of the arbitrary coding system employed and the value is also dependent on

coding scale. The application and interpretation of the mean, standard deviation, linear correlation coefficients and tests based on them such as t-tests, analysis of variance and regression are based on the assumption that the data are measured on a ratio scale rather than nominal or ordinal scale (such as number of years of education as opposed to education level, income as opposed to income category, etc.).

In general, the appropriate tools for analysing data depend on the type of variable/data as well as the objective of the study (or research question). While concepts such as mean age, mean weight, mean number of years of education, have an inherent meaning, concepts such as mean marital status, mean level of satisfaction, mean level of education are virtually meaningless.

Relating Survey Questions to Research Objectives

Relating survey questions to research objectives requires a clear idea of the full meaning of issues to be investigated, the clarification of any ambiguous words or phrases in the stated objectives and the specification of the population to be surveyed. The type and source of information required in order to address the general aims and objectives as well as each of the specific objectives of the study must be properly specified from the onset.

Operationally, it will usually be necessary to break down each main objective into sub-objectives (specific objectives) in which ambiguous terms are clarified. Each specific objective may itself be broken down into further sub-objectives, until the basic concepts and issues to be investigated have been clearly defined. A similar 'step-wise-refinement' approach can then be utilised to determine the information to be gathered. In this regard, the specific objectives could be rephrased in terms of research questions.

Illustration 1 Consider a study whose main objective is *to establish the needs of potential orphans from terminally ill patients in Botswana, with a view to formulating appropriate policy procedures to assist the children.*

Key issues that must be resolved early on include:

- Who is a terminally ill patient and where can they be contacted?
- What questions should the questionnaire contain to ensure that all relevant information is captured and unnecessary information is avoided?
- What age group of children should be included (the usual restriction to under 18 may be inadequate, since some children over 18 may still be in school)?
- How should the questionnaire items be phrased to avoid further distress to the family, who would naturally still harbour hopes that their loved one would recover?

Illustration 2 Consider a study whose main objective is *to investigate the causes of school dropout in Botswana.*

This main objective must be broken down into sub-objectives before any meaningful attempt can be made at designing the relevant questions. Some specific objectives include:

1. To determine the 'causes' of dropout in Primary, junior secondary and senior secondary schools in Botswana.
2. To investigate if the 'causes' of dropout differ between the rural, semi-urban and urban areas.
3. To determine if there are gender-forces at play in school dropout.
4. To suggest recommendations on how to reduce/or eliminate school dropout.

Key terms will also need clarification, for example:

a) What does 'school dropout' actually mean and how would it be measured?

- Does it mean those children of school going age not actually in school?
- Is it to be the difference between the school enrolments at the beginning of the year and the enrolments at the end?
- What about those children who transferred/changed schools during the year, or those who joined in the middle of the school year?

b) Is classification of schools as 'rural', 'semi-urban' and 'urban' clear and unambiguous?

c) How are possible 'gender forces' to be measured? Will differences between male and female dropout rates indicate gender forces? What about sex and or marital status of head of household?

d) What other background factors might be related to causes of dropout that may assist in policy formulation? For example, might the causes of dropout be related to:

- Parents' factors such as their occupation, socio-economic status, culture, religion, marital status, age, income level, etc.?
- Child's factors, such as physical or psychological impairment, age at which the child started school, earlier abuse by adults, etc.?
- School factors such as conflict between school discipline and home discipline, victimisation by peers and/or teachers, dislike of school food, times at which classes begin, etc.?

e) Is the severity of the problem known or is it to be estimated? Does dropout rate vary widely from region to region?

In order to investigate the first two main objectives, the data collection instrument (such as a questionnaire), must differentiate between the different school levels as well as between rural, semi-urban and urban schools. If 'gender forces' are to be measured using differences in male and female dropout rates as well as sex of head of the household, then appropriate variables must be included in the survey instrument to capture this information. Similarly, variables must be included in the instrument to capture information necessary for investigating each of the factors listed in (a) to (e).

Illustration 3 Suppose that your department wishes to set up a Postgraduate degree programme during the next National Development Programme (NDP). As one of the

conditions for the University approving the programme, you must carry out a 'needs assessment'. The primary objective appears to be quite simple: *To carry out a needs assessment of the proposed programme.*

But then: how is *need* to be defined?

a) Does it refer to citizens expressing an interest in joining the programme (as proposed)?

b) Is it the number of students that potential sponsors are prepared to support in a year?

c) Is it the number of job vacancies requiring people with the given qualifications?

d) Is it the existing pool of graduates who meet the entry requirements?

e) Is it the projected enrolments at the undergraduate level?

f) Is it the proportionate number of graduates from sister departments who are joining their own existing programmes?

g) Is it the age-specific national populations?

h) What University and/or Departmental factors (such as content of the proposed programme, previous experience in lower level courses in that department, faculty/university regulations, etc.) may influence interest in the programme?

i) What external factors such as sponsorship, perceived market demand for graduates, perceived prestige of graduating from that department/university might influence demand?

The appropriate definition will be arrived at through consultation with all stakeholders. More than one definition may even be adopted in order to satisfy different stakeholders. The survey population as well as the key questions to include in the survey instrument will depend on which of the above definitions are adopted.

Data Types Vs Variable Types

Data resulting from a survey are generally entered into a statistical package as cases and variables. In a survey, each questionnaire will produce one case if all information collected relate to a single respondent (such as the head of a household), or each questionnaire could produce two or more cases if some of the required information relates to more than one person (such as members within a household). The variables consist of specific responses to each question. Thus one questionnaire item may produce a single variable if it leads to a single response only, or one question could produce two or more variables, if it leads to two or more possible responses.

In statistical analyses, variables are distinguishable in a number of ways depending mainly on the type of values that they take, and sometimes on area of application. The two main classifications are *qualitative* (string) and *quantitative* (numeric) variables.

Qualitative variables are also called categorical variables in the statistics literature, since the underlying values that they take simply divide objects into

different mutually exclusive categories (such as male/female; poor/average/good, etc.). These categories can be stored in the computer as they are (i.e. as strings such as Male, Female, Good, Poor, etc.) or they can be coded into numbers and the numbers entered into the computer instead of their text equivalents. Non-numeric values take up a lot more computer memory than numeric values. Also typing text values such as (Male or Female) is much more time consuming and error prone than typing numbers (such as 1 and 2).

Furthermore, data analysis programmes think of upper and lower case characters as representing different values. Hence the three string values 'male', 'Male', 'MALE' are taken to be different values. As such, the values of qualitative variables are usually given numeric codes (such as 1 for Male and 2 for Female), and these numbers are then entered into the computer in the place of the actual text values. This innovative method of dealing with an otherwise intractable problem has become a source of much confusion in data entry and analysis. Once data have been entered as numbers, analysis tools have no way of telling that these numbers are arbitrary. This is because data analysis programmes 'see' only *data types* and not *variable types*.

The *data type* for a given variable is the way in which the values for that variable are represented in the computer. Typically, the data will be stored either as numbers (*numeric*), letters (*alpha*) or a combination of both (*alphanumeric* or *string*). When numeric codes are used to denote nominal and ordinal categories of a variable, the data type is then numeric, while the variable type is string. Appropriate tools for analysis apply to variable types which the researcher knows, and not the data types which the computer software accepts. The onus is on the researcher to know the type for each variable, and hence use the appropriate statistical tools for data analysis. Failing this, one risks computing meaningless summary measures such as average gender, average age group, average level of satisfaction, and the like, especially when using tools such as regression and analysis of variance, which implicitly compute and utilise these measures in order to provide the model parameters and significance.

Types of Questions and Associated Variables

Synonymous to variable classification is the classification of questions in a questionnaire. This section considers the most common classes of questions and the corresponding variable types that can be derived from such questions. Each type of question has been illustrated by examples from actual surveys. No attempt has been made to improve the phrasing of questions or the codes used for the values of the qualitative variables. The coding systems used for categorical variables is created only by numeric data types, since these are the most efficient data types for both data entry and analysis.

Dichotomous or Yes/No Questions A dichotomous question is one that has only two possible answers. Such questions are also referred to as YES/NO questions. Examples include the following:

A1: Have you registered to vote in the coming election? (DRP, 1999)

Yes ☐ No ☐

A2. District Number ☐☐☐

51 Ngamiland West (Botswana, 1997)
52 Ngamiland East

A3. Sex (of Respondent) (Nkrumah,1995).
1. Male
2. Female

A4. Men must share in house work (IEC, 1997)
__ Agree __ Disagree

A5. What was your final result last academic year?
0 Repeated the year
1 Proceeded to the next year

A6: Health Status (of potential orphan): HIV Positive ☐ (Orphans Form B)

While some questions have a natural dichotomy in their response, such as A1—A3, for some, the dichotomy is only forced on the respondent by the stated options (A4, A5), while for others, there is no ordering at all (A6).

Each dichotomous question produces a *single nominal* (dichotomous) variable at data entry, even when there is a natural ordering in the answers such as in questions A4 and A5. When using numbers to represent the values of a dichotomous variable, it is good practice to use '0' and '1' unless the codes have a particular meaning (as question A2). There are several advantages to this convention during data analysis. For example, the total number of 'Yes' cases can be obtained just by summing the responses. Also, some statistical procedures such as linear regression will give meaningful results only if the binary responses are recoded as 0 and 1 (also called dummy variables).

Multiple-choice Questions with a Single Response Here the respondent is provided with a range of options from which only one option is to be selected. The categories could be nominal or ordinal. For nominal categories, any set of numbers could be used to represent the categories. For ease of data entry, each category should be pre-coded in the question.

B1. Age group: 1) 15—20 2) 21—29 3) 30—39 4) 40—49 5) 50+ (FSS, 1998)
B2. What is the main source of food of your household? (CBPP 1997)

Own production	1	☐
Market purchases	2	
Government food rations	3	
Wages in Kind	4	
Gifts from Relatives	5	
Other (specify)	6	

B3. How do you rate the state of refuse collection in your area? (Gaborone, 1997)
Very good
Good
Satisfactory
Poor
Don't Know

Usually, each multiple-choice question with a single response produces a single nominal or ordinal variable. If the question includes 'Other (specify)' category as in B2, then the specified responses can either be analysed qualitatively, or an additional string variable could be created, and treated as an open-ended question. It should be noted however, that open-ended questions are time consuming to complete, complicated to analyse, and usually contain much higher non-response rates than closed-ended questions.

Question B1 and B3 above lead to one ordinal variable each, while question B2 leads to a nominal variable. Note that the wording of B2 is crucial, if it is to lead to a single variable. It is possible, and indeed likely, that many respondents will have more than one source of food. By requesting just the 'main source', perhaps the researcher wanted to restrict attention only to one source. Had the question been phrased as - 'What are the sources of food for your household?' - then the question would have been a multiple-choice question with several responses, sometimes referred to as 'Check-All-that-Apply' questions. These are discussed in the next section.

Multiple-choice Question with Several Responses A multiple-choice question with several responses is a collection of dichotomous questions under one main question. During data entry, each category produces a dichotomous (Yes/No) variable.

Examples:

C1. Which of the following do you think can promote career guidance activities/services for youths in Botswana? Please put X in front of your response.

(NCSS, 1998)

1. Adequate career resources/materials ----------
2. Trained personnel in career services ----------
3. Upgrading career services/activities in schools ----------
4. Establishing of Career Centres for youth in the regions----------
5. Other (specify)----------
   -

C2. Which of the following languages do you speak at home? (*MPSAQ, 1998*)

1	2	3
English	Setswana	Kalanga

C3. Since 1966, several elections have been held in Botswana in every 5 years. Please indicate whether or not you voted. (DRP, 1999)

	Voted	Didn't Vote
1965		
1969		
1974		
1979		
1984		
1989		
1994		

C4. Identify all the correct ways you know on how to avoid pregnancies (please tick every way you know) (IEC, 1997)

1.	Non-Penetrating Sex	9.	Male Sterilisation
2.	Condoms	10.	Female Sterilisation
3.	Withdrawal	11.	Full Abstinence
4.	IUD	12.	Periodic Abstinence
5.	Injectibles	13.	Use Traditional Doctor
6.	Diaphragm/Foam/Jelly	14.	Morning After Pill
7.	Pills	15.	Home Preparation
8.	Abortion	16.	Have sex with a Virgin

Question C1 from the Career Services study is equivalent to 4 different Yes/No questions on whether each of the specified items 1—4, can promote career guidance, and a fifth open-ended question about which other activity could promote such services. The question thus leads directly to 4 nominal variables:
QC1a 'Adequate Career resources can promote career guidance';
QC1b 'Trained Person in career services can promote career guidance', and similarly, for QC1c and QC1d.

The string variable QC1e 'Other factors that can promote career guidance' should be defined if many respondents select the 'Other (specify)' category.

Question C2 from the Media Profile and Situation Analysis Questionnaire (MPSAQ) leads to 3 dichotomous variables:
QC2a 'Speaks English at home';
QC2b 'Speaks Setswana at home';
QC2c 'Speaks Kalanga at home'.

Question C3 from the University of Botswana 1999 Democracy Project questionnaire leads to 7 nominal 'Voted/Did not vote' variables corresponding to the seven different elections. However, on closer examination this question should really have three categories (1: 'Voted', 2: 'Did not vote' and 3: 'Was not eligible'). The added category is needed to distinguish those people who were eligible and did not vote, from those who were not eligible due to some restrictions, such as age.

Question C4 from the Information, Education and Culture (IEC) study appears to be a collection of 16 Yes/No questions, which could be represented by 16

dichotomous variables. However, the purpose of the question was to assess the knowledge level of respondents about methods of preventing pregnancies rather than measuring how many respondents answered yes/no to a particular category. So ultimately, what is required is an indicator of how knowledgeable the respondents were.

For data entry purposes, the 16 nominal (Yes/No) variables (QC4_1 – QC4_16) should be defined. During data analysis, however, two indicator variables need to be created to measure the knowledge level of the respondents. These new variables could be:

QC4A: 'Number of correct responses' and

QC4B 'Number of incorrect responses'

Without loss of generality, let us suppose that QC4_1 to QC4_11 represent the correct ways, and QC4_12 to QC4_16 represented the incorrect ways of preventing pregnancies. Then the number of correct responses, QC4A would be obtained by counting the number of 'Yes's' among QC4_1 to QC4_11, while QC4B would be obtained by counting the number of 'Yes's' among QC4_12 to Q26_16. This counting would be greatly facilitated if the values of QC4_1 to QC4_16 were pre-coded with 1 used to represent 'Yes' and 0 used to represent 'No' respectively.

Ranking Questions A Ranking question is a multiple-choice type question in which the respondent ranks the given options. The range of values is equal to the number of options. Unlike the usual multiple-choice questions, the answer to each of the categories of a rank question are inter-dependent. Firstly, one needs to see all the options, before deciding on which is the best, and which is the worst. Secondly, tied ranks are usually not allowed, so that once one category has been given a particular rank, no other category is allowed be given the same rank. Hence there should be only 1 tick in each row, and in each column of the table.

D1 Rank the following social activities according to your interest (1=best, 2=2nd best, etc.)

	Rank					
Activity	1	2	3	4	5	6
1. Eating Out						
2. Going to the Movies						
3. Going to Nite clubs						
4. Attending house parties						
5. Watching TV						
6. Recreational sports (Soccer, squash, etc.)						

There are two ways of going about a ranking question. If greater emphasis is on the categories, rather than ranks, then each category will generate an ordinal variable. Hence in question D1, the 6 ordinal variables will be:

QD1_1 '(rank of) Eating out'

QD1_2 'Going to movies'

up-to

QD1_6 'Recreational sports'.

If emphasis is on determining which activities are ranked first, which are ranked second, etc., then each rank will generate a nominal variable with six categories, representing the six activities, that is: 1 'Eating Out'; 2 'going to movies'...'; 6 'Recreational sports'. The six (6) nominal variables are:

QD1a 'Most preferred social activity';
QD1b 'Second most preferred social activity';
up-to
QD1f 'Least Preferred Social Activity'.

Grid Questions Grid Questions are essentially a collection of multiple-choice questions into a single question. Each category in the grid is equivalent to a multiple-choice question (either with a single response or with several responses). Examples are as follows:

E1. Which type of music do you like/dislike hearing on the radio?
0=Hate it; 1=Dislike it; 2=Like it; 3=Like it a lot; 4=Love it!; 9=Do not Know it
(RLS, 1998)

Type of Music	0	1	2	3	4	9
1. R&B						
2. Gospel						
3. Traditional						
4. Reggae						
5. Hip HOP						
6. Jazz						
7. CHOIR						
8. Afro Pop (e.g. Bayette, Mbongeni Ngema, etc.)						
9. De Gong (Thebe etc.)						
10. Kwaito						
11. Kwasa Kwasa						
12. Other (specify)						

Question E1 combines 11 multiple-choice questions, each with a single response into a single question. Each of the 11 questions then leads to a single ordinal variable:
QE1_1 'Extent to which respondent would like to hear *R&B* on the radio'.
up-to
QE1_11 'Extent to which respondent would like to hear *Kwasa Kwasa* on the radio'.

The 12th category (other) is equivalent to asking respondents to list any music they would like/dislike, and then specify to what extent they would like/dislike hearing the music. This can lead potentially to many multiple response questions. In general however, most respondents will leave this option blank and hence the few responses can be analysed qualitatively. Note the inclusion of the last category (Do not know). This ensures that each respondent answers each of the questions.

Some research questions in this particular study required an analysis involving the music that respondents liked and those that they disliked. The required

variables can be computed from QE1_1 to QE1_11. Unlike for rank questions, it is not appropriate to define just 4 variables, each having 12 possible categories corresponding to the different types of music: R&B, Gospel, etc.:

QE1_A 'Music that Hate'
QE1_B 'Music that Dislike'
QE1_C 'Music that Like'
QE1_D 'Music that Like a Lot'

This is because a variable can have only one value per unit of analysis in the database, whereas for each of the variables QE1_A to QE1_D, each respondent can specify more than one answer. Faced with such responses, some researchers have resolved to enter values such as 137, etc. to indicate that the respondent had selected options 1, 3 and 7 for the given variable. Others resort to more fancy/ingenious methods of recoding, but in my experience, none of these approaches is ever adequate beyond frequency analysis of data.

The second example (E2) of a grid question is taken from a 1996 survey of teachers about a proposed curriculum in Population/Family life Education (POP/FLE). Here each of the 10 topics constitutes a multiple-choice question with several (1 to 8) possible responses.

E2. If a Policy is made to include all the topics listed (a-k) below in the POP/FLE curricula circle what problems listed (1-8), you would have in teaching?

1 None
2 Lack of teaching Materials
3 Lack of adequate Knowledge
4 Against Culture
5 Against Religion
6 Opposition from parents
7 Opposition from religious leaders
8 Other (specify)

(POP/FLE, 1996)

a.	Population growth and development	1	2	3	4	5	6	7	8
b.	Economic and Social development	1	2	3	4	5	6	7	8
c.	National Resources and environment	1	2	3	4	5	6	7	8
d.	Marriage, family life and welfare	1	2	3	4	5	6	7	8
e.	Pregnancy and birth control	1	2	3	4	5	6	7	8
f.	Human reproduction	1	2	3	4	5	6	7	8
g.	Social-Cultural factors influencing sexual development and sexual life	1	2	3	4	5	6	7	8
h.	Social responsibility	1	2	3	4	5	6	7	8
i.	Problems and issues relating to sexual conduct i.e. undesired pregnancy, STDs, abortion, HIV/AIDS	1	2	3	4	5	6	7	8
j.	Gender Issues–Traditional and changing social roles and relationships between men and women	1	2	3	4	5	6	7	8
K	Other (specify)	1	2	3	4	5	6	7	8

This grid is equivalent to 11x8 = 88 nominal (Yes/No) variables, corresponding to the 11 questions (a to k) and the eight (8) yes/no responses to each question. The following variables could be defined:

QE2a1_1 'There will be No problem in teaching if Population growth & development is included'

QE2a1_2 'We shall experience lack of teaching materials if Population growth & development is included'

Until:

QE2j_7 'We shall experience opposition from religious leaders if gender issues are included'

QE2k_8 'We shall experience other problems if other specified issues are included'

Any other method of representation will lead to loss in information, or analysis problems. For example, treating each of the question a-k as a multiple-choice question with a single response will not work. This is because each respondent can select more than one reason from lack of teaching materials to 'other'.

During data analysis, different multi-response sets can be generated, depending on the questions of interest. Furthermore, if it turns out that no respondent listed more than 3 problems say, then 4 multi-response sets can be defined for 'No Problem', 1st, 2nd and 3rd problems, each containing 10 categories corresponding to each of the 10 policies. If on the other hand, interest is to see how each problem (say lack of teaching materials) affects the different policies, then 6 multi-response variables corresponding to the 6 problems: a) lack of teaching materials, b) Lack of adequate knowledge, and so on, could be defined. Each of the 6 variables will also contain 10 categories corresponding to each of the 10 policies.

Combination Questions Combination questions arise when more than one type of response is required from each question. Consider question F1 (below) from a survey of Hotel users in Botswana.

F1. **Which hotels do you stay in most often**? (*Under the rank column, record 1 for the first choice, 2 for the second choice and so on).* Then for each hotel ranked, indicate **what was your main reason for choosing this hotel**.

Main Reasons for choosing hotel (Please record the appropriate number next to each hotel)

1=Cost 2=Service 3=Availability 4=Reputation 5=Membership 6=Company choice 7=Other

(BWHU, 1997)

Hotel	Rank	Main reason
Bosele Hotel-Best Western		
Botsalo Travel Inn		
City Lodge		
Gaborone Sun		
Gaborone Travel Inn		
Grand Palm/Sheraton		
Marang		
Morning Star		
Mowana Safari Lodge		
Oasis		
President		
Rileys-Best Western		
Sedia		
Tati-F/Town		
Thapama		
Cresta Lodge		

This question is really a combination question comprising of what looks like a rank question, and what looks like a multiple-choice question with several responses. Note that each respondent would have stayed in anything between 1 and all 16 hotels. He/she will rank only those hotels that they have stayed in, and give corresponding reasons why they chose that hotel. That is why the resulting questions are not exactly rank and multiple-choice questions.

The general approach to organising the resulting data is to define two sets of variables; one set for the information on ranks and the other set for the 'reasons'.

For example, to collect information on the hotel rankings, define 16 rank-type questions corresponding to the 16 hotels, each with ranks 0, 1, 2,..,k, where k is the number of hotels that the respondent has stayed in ($k \leq 16$) and 0 indicates that the respondent has not stayed in that hotel. For instance, you could have:

QF1a1_1 'Rank of Bosele Hotel Best Western'
QF1a1_2 'Rank of Botsalo Travel Inn':
QF1a1_16 'Rank of Cresta Lodge'
QF1a2_1 'Main Reason for choosing Bosele Hotel Best Western'
QF1a2_2 'Main Reason for choosing Botsalo Travel Inn'
up-to
QF1a2_16 'Main Reason for choosing Cresta Lodge'

To organise information on the reasons given, define 16 nominal variables, one for each hotel, and having 8 categories:

1 = Cost; 2 = Service; until 7 = 'Other'; 8 'Combination of reasons'.

For this particular study, it turned out that most people ever stayed only in a few hotels (rarely more than 4). Furthermore, the consultants were interested only in the

top three ranked hotels. Thus an alternative approach was to define 3 nominal variables:
QF1a: 'Top Ranked hotel'
QF1b: 'Second ranked hotel' and
QF1c 'Third ranked hotel'.
These are nominal variables, with possible categories:
1 = 'Bosele Best Western', 2 'Botsalo Travel Inn'..., 16 = 'Cresta Lodge'. In addition, QF1b and QFc will include a 17th category; 0 'Not Applicable' for those who only stay in one or two hotels respectively.

Open-Ended Questions Survey instruments in the so-called qualitative research studies in social science are usually dominated by open-ended questions. Experts in qualitative data analysis can best defend the merits of using open-ended questions to gather information. For large scale surveys or surveys in which the use of statistical packages for data analysis is envisaged, open-ended questions should be avoided. Although great improvements have been made in the ability of statistical packages to handle open-ended questions, any advantages in using open-ended questions may be outweighed by numerous problems from data gathering to data analysis. Typical problems include:

- the time spent to respond to, and to record the response to the questions during data collection,
- the increased likelihood of misunderstanding the question, which may lead to inappropriate responses or non-response,
- the time taken to capture the information into computer, and likely errors,
- the difficulties in organising the responses into meaningful categories and the corresponding costs on data analysis.

Thus as much as possible, open-ended questions should be avoided in large-scale surveys. Flexibility could be added by providing, as the last category of a close-ended quantitative question, the category:
Other (specify).....

The appropriate categories to any qualitative question could be obtained through a pilot study and literature review. Consider the following questionnaire from a *1998 Media Profile and Situation Analysis (MPSAQ)* questionnaire:

G1. Name three types of TV programmes that you watch regularly:

_____

_____

When the data were recorded as three string variables:
QG1a: '1st programme'; QG1b '2nd Programme'; QG1c '3rd programme';
There were 50 different responses for QG1a, 51 responses for QG1b and 40 responses for QG1c. These included duplicates such as 'COMDIES', 'COMEDIES' and 'COMEDY', and similar programmes like 'FILMS' and 'MOVIES'; 'LADUMA', 'MABALENG'; 'SPORTS' and 'SOCCER', etc.

For the purpose of the study, these fine distinctions were unnecessary. It took several hours of analysis, meetings and computation to recode the variables into 10 categories:

1 'Chat Shows'; 2 'Comedies'; 3 'Documentaries'; 4 'Drama';
5 'Music'; 6 'Movies'; 7 'News'; 8 'Soaps'; 9 'Sports'; 10 'Others'.

If a pilot study had been undertaken, a close-ended multiple-choice question with the above options could have been used, and some time and money saved in data capture and coding. In general, it is advisable to always carry out a pilot study, even for small-scale studies. The size of the questionnaire and the diversity of the respondents should guide the size of the pilot study. A large questionnaire or a large-scale study would require more interviews to reveal all the shortcomings. On the other hand, a questionnaire with only a limited number of items would require only a small number of respondents to validate it.

Conclusion

An attempt has been made to address the key issues in the organisation of survey data to facilitate proper data capture and analysis using statistical packages. The choice of the issues covered in the paper is based on the demand from numerous researchers that I have interacted with over the years. The issues are particularly relevant for researchers who wish to collect primary information for research in new areas such as HIV/AIDS, were standard instruments are not yet fully developed. Indeed Forms A, B, C and D currently used in Botswana for collecting information in this area need much revision and standardization. Any researcher wishing to undertake this novel responsibility should benefit from the issues raised and the advice provided in this paper. No doubts that not all areas of concern to survey-researchers have been address. To do so requires an entire book, for which this article can only form a chapter.

The suggestions provided have been tested on many research studies and have been shared with many colleagues and students involved in survey studies. Very positive reviews have also been obtained from researchers on environmental science and biology/ecology involved in conducting and analysing field experimental studies.

Bibliography

Czaja R. and Blair J., (1996), *Designing Surveys. A guide to decisions and procedures*, Pine Forge Press: London.

Freund J. E., Williams F.J and Perles B. M., (1993), *Elementary Business Statistics*, Prentice-Hall: New Jersey.

Stevens S.S., (1951), *Handbook of Experimental Psychology*, John Wiley and sons, Inc.: New York.

Varkevisser C. M., Pathmanathan I. and Brownlee A., (1995), *Designing and Conducting Health Systems Research*, (unpublished training guide) IDRC,Canada

Questionnaires Referred to

Botswana (1997), Monitoring the effects of the CBPP eradication in Ngamiland. Ministry of Agriculture, Botswana

BWHU (1997), Best Western Hotels users survey. Momarketing Options, Botswana

DRP (1999), The 1999 Democracy Research Program Questionnaire, *University of Botswana*.

FSS (1998), A Bachelor of Social Science Project Questionnaire, (personal Communication).

Gaborone (1997), Gaborone City Council Survey of Refuse Collection Services in Gaborone.

IEC, (1997), The 1997 Information Communication and Education Survey, Ministry of Finance, Botswana.

MPSAQ (1998), The 1998 Media Profile and Situation Analysis Questionnaire, Ministry of Finance, Botswana
NCSS, (1998), The 1998. National Careers Services Study, Department of Non-formal Basic Education, Botswana.
POP/FILE (1996), Population and Family Life Education Opinion Survey, (personal communication).
RLS (1998), Radio Listeners Survey. Private radio Consortium, Gaborone.

---

* I wish to acknowledge all those peers and editors who assisted in reviewing this paper.