The overall aims of the book are:

- to assist researchers in mastering the *art of data analysis* and to help students recognize that quantitative analysis is far more than the routine application of statistical tests;
- to identify some of the most common problems that are encountered in the analysis of quantitative social science data;
- to outline ways of detecting whether these problems exist in a given data set;
- to provide a range of ways of dealing with these problems and assistance with deciding which solution is best suited to the situation.

z de Vaus: Analyzing social science data: 50 key problems in data analysis.Sage, 2002.

Part one How to Prepare Data for Analysis

1

How to Code Data

What is the Problem?

Coding is a method of representing categories and values of a variable so that:

- responses are converted to a form suited to statistical analysis;
- data become more manageable by grouping similar responses.

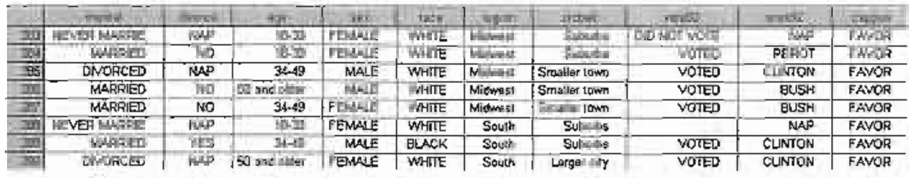
Since coding produces the raw material for data analysis it can fundamentally affect the quality of the analysis. The two central problems of coding are:

- deciding on coding schemes;
- minimizing coding errors.

How to Structure and Store Codes

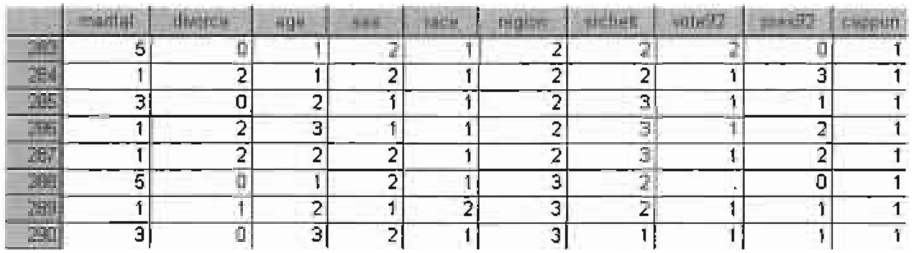
Since coding is designed to facilitate computer-based statistical analysis, the codes have to be stored in such a way that computer programs can access and interpret them. Most statistical analysis programs require coded data to be stored in a particular form which I call a *variable-by-case data grid*. The main characteristics of such a grid are:

- Each column represents a variable.
- Each row represents a case.
- Each cell contains the response (or value) of a particular case to a specific variable.
- Responses are represented as a code: an abbreviated representation of the response for that case on that variable.



	marital	divorce	age	sex	race	region	srcbelt	vote92	pres92	cappun
283	NEVER MARRIE	NAP	18-33	FEMALE	WHITE	Midwest	Suburbs	DID NOT VOTE	NAP	FAVOR
284	MARRIED	NO	18-33	FEMALE	WHITE	Midwest	Suburbs	VOTED	PEROT	FAVOR
285	DIVORCED	NAP	34-49	MALE	WHITE	Midwest	Smaller town	VOTED	CLINTON	FAVOR
286	MARRIED	NO	50 and older	MALE	WHITE	Midwest	Smaller town	VOTED	BUSH	FAVOR
287	MARRIED	NO	34-49	FEMALE	WHITE	Midwest	Smaller town	VOTED	BUSH	FAVOR
288	NEVER MARRIE	NAP	18-33	FEMALE	WHITE	South	Suburbs		NAP	FAVOR
289	MARRIED	YES	34-49	MALE	BLACK	South	Suburbs	VOTED	CLINTON	FAVOR
290	DIVORCED	NAP	50 and older	FEMALE	WHITE	South	Larger city	VOTED	CLINTON	FAVOR

Figure 1.1 *Variable-by-case data grid with labels*



	marital	divorce	age	sex	race	region	srcbelt	vote92	pres92	cappun
283	5	0	1	2	1	2	2	2	0	1
284	1	2	1	2	1	2	2	1	3	1
285	3	0	2	1	1	2	3	1	1	1
286	1	2	3	1	1	2	3	1	2	1
287	1	2	2	2	1	2	3	1	2	1
288	5	0	1	2	1	3	2	.	0	1
289	1	1	2	1	2	3	2	1	1	1
290	3	0	3	2	1	3	1	1	1	1

Figure 1.2 *Variable-by-case data grid with codes*

- Each case must have a 'response' or value for each variable.
- Each case has one and only one value (code) per variable.
- The order of the variables in the file containing coded responses will be the same for every case.
- The same set of codes will be used for all cases on a given variable.

Figure 1.1 provides an example of a variable-by-case data grid in which the actual responses are entered in each cell. The data grid is a small part of the data grid for the US General Social Survey. It indicates actual responses for eight cases (rows) for ten variables. Each cell contains the actual responses for each case. Notice that for case 288 there is a dot rather than an actual value for the variable VOTE92 – this is the 'code' given to a non-response to this question. Figure 1.2 presents the same grid but numeric codes replace the actual answers.

Decisions to Make

Translating answers into numeric codes requires a series of decisions including:

- the type of codes to use – *numeric* or *alphanumeric* (words and letters);
- when to produce a coding system – *before* or *after* collecting data;
- how to code *non-responses*;
- methods of coding *multiple answers* to the one question;

- the method of coding *open-ended* questions – whether to use a pre-established coding scheme or to create one from the responses;
- *how many* codes for the variable – the level of detail to which responses will be coded.

Should Numeric or Alphanumeric Coding be Used?

Numeric coding involves allocating numbers to responses, as illustrated in Figure 1.2. Alphanumeric coding may use letters such as A, B, C, D, E rather than numbers 1, 2, 3, 4, 5. More often, however, alphanumeric coding involves the use of words (e.g. Yes, No; Male, Female; or I voted for X because ...). In nearly all cases where variables are to be used for statistical analysis responses are coded numerically.

Should Codes be Produced Before or After Collecting Data?

When data are collected using a predetermined set of categories, codes can be allocated to these answers before collecting the data. This *precoding* can eliminate the need to manually code later and can improve the accuracy of data entry (see p. 9). Figure 1.3 provides examples of precoded questions.

How Should 'Non-Responses' be Coded?

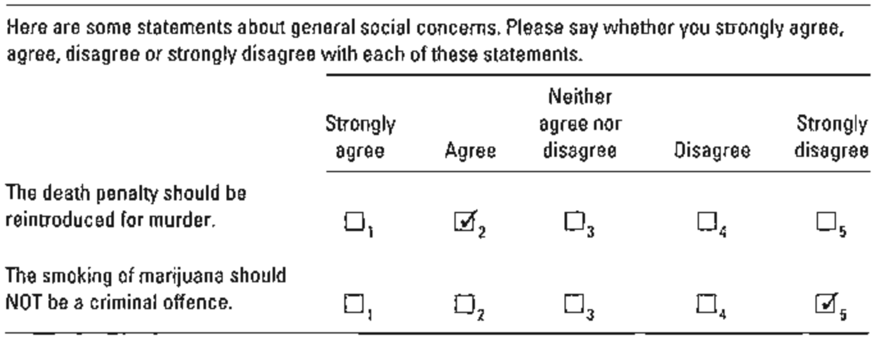
Since every case must have a code for each variable, a system must be developed for coding non-responses to questions. Non-responses may occur because people:

- were not required to answer (question did not apply);
- refused;
- provided an illegible answer;
- responded 'don't know'.

You may wish to distinguish between different types of non-response by allocating separate codes to each type. This provides maximum flexibility during later data analysis.

To avoid confusion, distinctive codes are allocated to non-responses. For variables where only single-digit codes are required, non-responses are often coded 8, 9 or –1. For variables that require two-digit codes (e.g. age), non-responses are given codes such as 98, 99, –1. The important thing with allocating codes for missing or invalid responses is to:

- make them distinctive;
- make them as consistent as possible across questions (e.g. always –1 or always 9 for single-digit variables, always 99 for two-digit variables).



Here are some statements about general social concerns. Please say whether you strongly agree, agree, disagree or strongly disagree with each of these statements.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
The death penalty should be reintroduced for murder.	□1	☑2	□3	□4	□5
The smoking of marijuana should NOT be a criminal offence.	□1	□2	□3	□4	☑5

Figure 1.3 *Examples of precoded questions*

How to Code Multiple Answers to the One Question

Respondents might give several responses to the same question. For example, in response to a question about the anticipated major problem facing the world in 10 years' time a person may identify several possibilities. Problem 2 (see p. 10) discusses methods of coding this type of question.

Approaches to Coding Open-ended Questions

Open-ended questions are not so readily precoded. Open questions fall into four categories:

- closed questions with an open-ended category (e.g. Other, please specify);
- open-ended questions to which there is a defined range of possible responses (e.g. religious group, country of birth, occupation);
- self-coding open questions (e.g. age, number of children);
- open-ended questions to which there is a wide and undefined range of possible responses (e.g. What, in your opinion, is the most important problem facing the world today?).

For open questions of the first two types a set of codes can be developed before collecting data. Self-coding variables are numeric variables, such as income in dollars and age, where the answer is already in numeric form. Answers to these questions are suitable for statistical analysis and do not require further coding. To code the fourth type of open question it is best to look at answers before developing a coding system (see pp. 15–16).

Table 1.1 *Three-level coding scheme for religion*

Level 1: Broad (single-digit codes)	Level 2 Narrow groups (three-digit codes)	Level 3 Detailed classification (four-digit codes)
1 Buddhism	2 Christianity	223 *Orthodox*
2 Christianity	201 Anglican	2231 Albanian Orthodox
3 Hinduism	203 Baptist	2232 Antiochian Orthodox
4 Islam	205 Brethren	2233 Greek Orthodox
5 Judaism	207 Catholic	Autocephalic Greek Orthodox Church of Australia
6 Other religions	211 Churches of Christ	Greek Orthodox (Australian Archdiocese)
7 No religion	213 Jehovah's Witnesses	Greek Orthodox (Old Calendar)
	215 Latter Day Saints	2234 Macedonian Orthodox
	217 Lutheran	2235 Romanian Orthodox
	221 Oriental Christian	The Lord's Army
	223 *Orthodox*	2236 Russian Orthodox
	225 Presbyterian and Reformed	Orthodox Church in America (Australian Mission)
	227 Salvation Army	Russian Orthodox (Ecumenical Patriarchate)
	231 Seventh-day Adventist	Russian Orthodox (Moscow Patriarchate)
	233 Uniting Church	2237 Serbian Orthodox
	240 Pentecostal	2238 Ukrainian Orthodox
	280 Other Protestant	Ukrainian Autocephalic Orthodox Church
	290 Other Christian	2239 Orthodox, not elsewhere classified
		Byelorussian Autocephalic Orthodox Church
		Old Believers (Russian)
		Old Orthodox Church of the Holy Nativity
		Polish Orthodox

Whether to Use Other People's Codes or Develop Your Own

Some excellent coding schemes have been developed for common demographic variables. Where these exist it normally makes sense to use them, since they have been carefully developed and enable comparisons between your results and those from other studies.

Many established classification schemes (e.g. Table 1.1) employ a *multilevel coding scheme* which allows the variable to be coded to different levels of detail. Table 1.1 gives a scheme for classifying religions. At the first, most

Table 1.2 *Examples of coding and classification schemes for core demographic variables*

Variable	URL
Standard Occupational Classification (SOC)(USA)	http://stats.bls.gov/soc/soc_home.htm
Employment status (UN)	http://www.ilo.org/public/english/bureau/stat/class/icse.htm
Industry (UK)	http://www.statistics.gov.uk/nsbase/themes/compendia_reference/Articles/downloads/structur.pdf
Diseases and Related Health Problems (UN)	http://www.who.int/msa/mnh/ems/icd10/icd10.htm
Education (UN)	http://unescostat.unesco.org/en/pub/pub0.htm
Race and Ethnicity (USA)	http://198.137.240.91/textonly/OMB/fedreg/directive_15.html
Countries (Australia)	http://www.abs.gov.au/ausstats/abs@.nsf/StatsLibrary? Open View then select link to ABS classifications and select catalogue 1269.0
Languages (Australia)	http://www.abs.gov.au/ausstats/abs@.nsf/Stats Library? Open View then select link to ABS classifications and select catalogue 1267.0
Religion (Australia)	http://www.abs.gov.au/ausstats/abs@.nsf/StatsLibrary? OpenView then select link to ABS classifications and select catalogue 1266.0
Crime/Offences (Australia)	http://www.abs.gov.au/ausstats/abs@.nsf/StatsLibrary? Open View then select link to ABS classifications and select catalogue 1224.0
Causes of Death (Australia)	http://www.prometheus.com.au/healthwiz/142death.htm

general, level we have the major world religions, along with a general code for other religions and for no religion. The second level distinguishes between the main groupings within a specific religion (e.g. within Christianity we can code different Christian groups). The third level of coding provides most detail. The extract in Table 1.1 makes some very fine-grained distinctions between different types of Orthodox Christianity. Notice the structure of the numerical codes that accompany this classification: the first digit in all the codes indicates the first-level classification (2 = Christian), the second and third digits indicate the second-level classification (223 = Orthodox), while the fourth digit indicates the third-level classification (2237 = Serbian Orthodox).

The decision about the level of detail at which to code depends on various factors, including the likely make-up of the sample, the sample size and the way in which the data will be used. If the sample is likely to consist mainly of Christians we would want to code beyond the first level to enable the distinction to be made between religious affiliations. If everyone belonged to the one group the variable would be of little value in later data analysis (see pp. 48–53). If the sample is fairly small it is unlikely to be worth coding at the third level since these fine distinctions are unlikely to be of much use.

Table 1.3 *First- and second-level codes in an open coding scheme*

First-level codes		Second-level codes	
Codes	Broad type of problem	Codes	**Specific environmental problem**
1	Social	501	Overcrowding
2	Economic	502	Air quality
3	Moral	503	Water quality
4	Military	504	Scarcity of resources
5	**Environmental**	505	Extinction of species
6	Political	506	Greenhouse problems
7	Religious	507	Ozone problems
8	etc.	508	Salinity
		509	Deforestation
		510	etc.

However, this would depend on the nature of the sample. If the sample consisted of migrants from eastern Europe these finer distinctions might be important for the analysis.

There are many excellent coding schemes available. Some of these can be found on the websites listed in Table 1.2.

Fixed coding schemes are inappropriate for some open questions. Suppose that you have asked people what they think will be the major problem facing their country in 10 years' time. Since you do not want to impose your range of possibilities on respondents, you have asked an open-ended question.

To code the responses you should examine the first 50–100 responses (or more if you keep finding new responses). Try classifying these responses into broad groupings. This will require some trial and error. Once these broad headings are developed, you should assign a code to each broad grouping. Then, examining the specific responses, see if there are some distinctions that you should make within each broad grouping. Assign specific codes to these subcategories, but make sure that you retain the broad-level code as the first digit of the code. You might possibly then develop a third-level classification and codes.

Table 1.3 illustrates a set of first-level codes that might emerge from an examination of responses to the major problems question. One type of problem might be broadly classified as 'environmental'. The table illustrates some of the possible subcategories within the environmental grouping. The end product of developing a coding scheme from responses to open-ended questions should have a similar structure to the pre-existing coding schemes.

When developing codes, try to make the categories and codes flexible so that additional codes can be added later as more questionnaires are examined. Use multiple-digit codes so that there are plenty of spare codes if they are needed. For example, allow two digits for specific environmental problems so that you can have up to 99 different codes.

How Detailed Should the Coding be?

There is no simple answer to this question. The fineness of the coding scheme will depend on the:

- size of the sample;
- importance of making particular distinctions;
- way in which the data will be analysed;
- likely distribution of cases within the sample.

However, the general guideline is to code for more rather than less detail. This is particularly true if you are not quite sure how you will analyse the data or how many cases will belong to any one category. The main rationale for detailed coding is that you can always collapse categories at a later stage (see pp. 34–38) but you cannot expand broad categories to reveal finer detail if you have only coded at the broad level.

How to Minimize Coding Error

The best way of reducing coding error is to reduce the number of steps involved in coding and data entry. The fewer the steps, the smaller the chance of error.

How to Reduce Coding Mistakes

A number of procedures can help increase the accuracy of coding:

- Include codes *next to* responses to fixed-choice questions. Data entry operators can then enter data directly from questionnaires (see Figure 1.3).
- Use automated coding programs. For complex coding, such as occupational coding, computer programs are available to allocate codes based on descriptive information.
- Develop written coding schemes and continually update these with guidelines and decisions that are made while coding.
- Use several coders and introduce consistency checks.
- Where electronic data collection methods are used (e.g. Computer Assisted Telephone Interviewing – CATI, web-based surveys, e-mail surveys), they can be programmed to disallow responses outside a specified range of codes and to disallow the wrong type of response code (e.g. alphanumeric rather than numeric). While these safeguards cannot eliminate incorrect 'within-range' responses they can eliminate 'out-of-range' errors.
- Electronic data collection methods automatically code responses and build a database, thus eliminating the need to code at all.

How to Reduce Data Entry Errors

A number of procedures can help improve the accuracy of data entry:

- Automate the data entry by using electronic data collection methods.
- Use a data entry template. Some electronic questionnaires that do not automatically code and create a data file will still automatically generate a data entry form that simplifies manual data entry. (e.g. SPSS Data Entry).
- Use professional data entry personnel. If manual data entry is required, use trained people who are much more likely to enter data accurately than inexperienced people.
- Use double data entry methods. This involves entering data from the same questionnaire twice. Where the second attempt differs in any way from the first attempt, the data entry operator is alerted to the discrepancy.
- Choose data entry software correctly. If a special data entry program such as SPSS Data Entry is unavailable, then more common software can assist with data entry. Avoid entering codes into a text file or word-processing file. A spreadsheet can be useful since the layout of a spreadsheet is the same as the variable-by-case data grid. Even better are database programs. These can be programmed easily to create a data entry form for each person. These forms provide an intuitive method for entering codes and can be programmed to build in data integrity checks.

2

How to Code Questions with Multiple Answers

What is the Problem?

When coding any variable a case must belong to *one and only one category* of the variable. However, some questions allow respondents to provide more than one answer. The problem then is how we code multiple answers to a single question.

What are Multiple-Response Questions?

To identify a solution it is helpful to examine first how the problem arises. There are four main types of question that are widely used in questionnaires and produce multiple responses:

- *Closed question – select as many answers as apply.* This question provides respondents with a single question, but instead of asking them to select only one response allows them to select all that apply (Table 2.1).
- *Closed question – ranking responses.* A different type of multiresponse closed question asks respondents to rank the set of responses (Table 2.2). Rather than selecting some responses, this method requires that each possible answer receives a response.
- *Open questions – limited number of responses.* Open-ended questions that allow a set number of responses (Table 2.3) can produce a large number of possible answers which require particular strategies for coding and analysis.
- *Open questions – unlimited number of responses.* This type of question is similar to the previous type except that there is no limit on the number of responses that can be given (Table 2.4).

The solution to the problem of coding multiple responses to a single question lies in distinguishing between a *question* and a *variable*. While an answer to a single question often only requires a single variable to contain the response code, a single question can produce a number of variables.

Table 2.1 *Closed multiresponse question*

From the list below select the main pressures on your life at the moment (select all that apply).

□ Financial
□ Job insecurity
□ Uncertainty about the future
□ Health
□ Relationship with my partner
□ Relationship with my children
□ Relationships with people at work
□ Housing problems
□ What the future holds for my children
□ Education

Table 2.2 *Closed multiresponse question requiring ranking*

The list below describes various features of jobs. When looking for a job, what are the things you look for most and least? Please rank each of the job features below from most important to least important. Place a 1 in the square next to the most important; a 2 in the square for the second most important; a 3 in the third most important; and a 4 in the square for the job feature that is least important to you. Please place a number in each square and do not use the same number more than once.

□ A good income so that you do not have any worries about money

□ A safe job with no risk of closing down or unemployment

□ Working with people you like

□ Doing an important job which gives you a feeling of accomplishment

Table 2.3 *Open question with a limited number of responses*

Thinking about the future, what do you think will be the three most important problems facing this country in ten years' time?

1. ______________________

2. ______________________

3. ______________________

Where respondents provide several responses to a question, the solution is to create a *set* of variables to 'hold' those responses to the question. There are two approaches to developing these sets of variables:

- multiple-dichotomy method;
- multiple-response method.

The application of these two approaches can be explained in relation to each of the four question types outlined above.

Table 2.4 *Open question with the option of unlimited responses per person*

Thinking about the future, what do you think will be the main problems facing this country in ten years' time. (You can mention more than one problem.)

____

____

How to Code Multiple Responses to a Closed Question

When respondents can 'select as many answers as apply' the multiple answers can be coded using the multiple-dichotomy and multiple-response methods.

Using the Multiple-Dichotomy Method

Using this method, each of the possible responses is treated as a separate variable to which respondents provide either a yes answer (by selecting it) or an implied no answer (by not selecting it). For the life pressures question in Table 2.1, this method results in 10 separate variables (Table 2.5). Respondents are then coded for each of the 10 variables. If they selected all 10 variables they would receive a yes code for each variable. If they selected just two options they would receive a yes code for those two variables and a no code for the remaining eight.

Using the Multiple-Response Method

Instead of creating a separate variable for each response category, this approach involves creating a separate variable to contain each of the responses provided by an individual. Since most people will only select two or three responses to the pressures question and no one will probably select all ten pressures, we do not need to create ten variables. The multiple-response method involves the following steps:

1. Determine the maximum number of responses given by any person (for this example assume that no one selected more than four pressures).
2. Create four variables. Let us call them PRES1, PRES2, PRES3, PRES4.
3. Create 10 categories for each variable. Each category will represent the ten different pressures. Use the same set of categories for each of the four variables. Because each variable will contain all ten pressures this is called the multiple-response method (as opposed to a simple dichotomous yes/no coding).

Table 2.5 *Multiple-dichotomy method for multiple-response questions*

Variable number	Is this a pressure?	No (code = 0)	Yes (code = 1)	Total *N*
1	Financial	380	620	1000
2	Job insecurity	630	370	1000
3	Uncertainty about the future	680	320	1000
4	Health	640	360	1000
5	Relationship with my partner	710	290	1000
6	Relationship with my children	650	350	1000
7	Relationships with people at work	850	150	1000
8	Housing problems	880	120	1000
9	What the future holds for my children	550	450	1000
10	Education	780	220	1000

Table 2.6 *Coding multiple responses for three cases*

CASE	PRES1	PRES2	PRES3	PRES4
A	2	8	–1	–1
B	5	–1	–1	–1
C	2	5	6	10

4. For each case code the responses into the four pressure variables. This can be illustrated with the examples of three cases in Table 2.6. Case A nominated two pressures (job insecurity and housing problems). This person's two responses would be coded into the first two pressure variables (PRES1, PRES2). The remaining two pressure variables (PRES3, PRES4) would be coded –1 to indicate that no information was coded to those variables for this case. Case B nominated only one pressure. This response would be coded into PRES1, with the remaining pressure variables being coded –1. Case C listed four pressures (job insecurity, relationship with partner, relationship with children, and education). These four responses would require the use of all four PRES variables.

Table 2.7 illustrates the coding and distributions of the same 1000 cases as were coded using the multiple-dichotomy method (Table 2.5). In this table these same cases are coded into the four PRES variables using the multiple-response method. Notice that the number of people who indicated a particular pressure is the same as with the multiple-dichotomy method. The selections of a particular pressure are simply represented in a different way using the two methods.

Table 2.7 *Multiple-response method for multiple-response questions*

Code	Is this a pressure?	PRES1	PRES2	PRES3	PRES4
1	Financial	270	160	110	80
2	Job insecurity	120	140	65	45
3	Uncertainty about the future	90	100	85	45
4	Health	90	200	50	20
5	Relationship with my partner	70	70	130	20
6	Relationship with my children	90	80	125	55
7	Relationships with people at work	30	50	35	35
8	Housing problems	20	60	30	10
9	What the future holds for my children	120	65	195	70
10	Education	100	40	45	35
–1	No response	0	35	130	585
	Total *N*	1000	1000	1000	1000

Table 2.8 *Multiple 'dichotomy' coding for ranking questions*

	Variables			
Codes	INCOME	SECURITY	PEOPLE	ACCOMP
Ranked 1st	400	250	200	150
Ranked 2nd	300	250	300	150
Ranked 3rd	200	200	400	200
Ranked 4th	100	300	100	500
N	1000	1000	1000	1000

How to Code Ranking Questions

Rank-ordered responses can be coded using the same general logic as outlined above.

Using the Multiple 'Dichotomy' Method

In Table 2.2 there are four responses to rank, so the multiple 'dichotomy' method will result in four variables – one for each of the responses. The only difference is that instead of a simple yes/no set of responses (i.e. a dichotomous response) the possible ranks become the categories of each variable. Where there are more than two items to be ranked there will be more than two categories used for these variables. Since any job characteristic could receive one of four possible ranks the variable used for that job characteristic will have four categories – one for each possible rank. The resulting variables will indicate, for each job characteristic (e.g. INCOME), how often it was ranked first, how often it was ranked second and so forth. Table 2.8 illustrates the distribution of 1000 cases using this approach. Four variables (INCOME, SECURITY, PEOPLE and ACCOMP) are created. Each has the categories to indicate the number of people who ranked a given variable first, second, third and fourth.

Table 2.9 *Multiple-response coding for ranking questions*

	Variables			
Codes	FIRST ranked characteristic	SECOND ranked characteristic	THIRD ranked characteristic	FOURTH ranked characteristic
Income	400	300	200	100
Job security	250	250	200	300
Nice people at work	200	300	400	100
Feeling of accomplishment	150	150	200	500
N	1000	1000	1000	1000

Using the Response Method

The logic of the multiple-response method can be used for ranking questions. Using this approach, the ranks become the variables and job characteristics become the categories. The resulting variable (e.g. FIRST rank) will indicate the number of times income was ranked first, security was ranked first and so on (see Table 2.9).

How to Code Open Questions with Multiple Answers

Multiple responses to open-ended questions can be handled with the multiple-dichotomy and multiple-response methods.

Using the Multiple-Response Method

This method can be applied to the open question in Table 2.3 using the following steps:

1. Create three variables (PROB1, PROB2 and PROB3).
2. Examine the responses provided by respondents.
3. Create categories for PROB1 for each of the problems listed.

4. Use the same categories for PROB2 and PROB3.
5. For each respondent, code the problems listed into the variables PROB1, PROB2 and PROB3.

Using the Multiple-Dichotomy Method

The multiple dichotomy method can be applied to open questions with a maximum number of responses in the following way:

1. Examine the responses provided to the question.
2. Compile a list of the problems (e.g. environment, overpopulation, no jobs, poverty, crime, low birth rate, moral decline, ...).
3. Where appropriate, combine very similar problems.
4. For each problem in your list create a separate variable with the values 0 and 1 (e.g. ENVIR, POP, NOJOBS, POV, CRIME, BIRTH, MORALS, ...). If a total of 20 different problems were listed across the sample, create 20 different variables.
5. Code each respondent on each of the variables. Where the respondent listed the specific problem, code them as 1. Otherwise code them as 0.

When there is no maximum number of responses to open questions the same basic coding strategies can be used. The only difference is in the way the multiple-response method is implemented. Where respondents are restricted to three responses, only three variables are required to contain these responses. Where there is no limit on the number of responses, you will first need to examine each case and determine the largest number of responses any case has provided. Then create that number of multiple-response variables to contain each response.

Using SPSS

The variables that are used with either the multiple-dichotomy method or the multiple-response method should be created when setting up the initial database. Decisions about the method of dealing with multiple responses should be made at this stage and the variables should be coded accordingly. There is nothing to prevent you using both methods in the same database – it simply requires more coding and data entry.

3

Can the Respondent's Answers be Relied on?

What is the Problem?

Data analysis relies on measurements being both reliable and valid. A *reliable* measure is one for which we can depend on obtaining *consistent* responses. A set of scales is reliable if it gives the same reading each time the same person steps on it (assuming the person does not change weight). A questionnaire item is reliable if it elicits dependable and consistent answers from people. Remember, however, these answers may not be accurate answers. A set of scales that consistently underweighs is still reliable. A question that consistently overestimates happiness is nevertheless reliable (see pp. 25–27).

If we cannot rely on the responses that a questionnaire item elicits then any analysis based on such data will be suspect. If the results we obtain from a sample could just as easily be different if we administered the questionnaire again, how much confidence can we have in any of the findings?

We must, therefore, use reliable items, but this requires a way of evaluating how reliable our measurement instruments are. Would we obtain similar data if the same questions were given to the same people again?

We cannot just use measures because others have found that they are reliable. A measure's reliability can change over time, can vary in different contexts, with different samples and on the method of administering the questions. It is therefore important to assess the reliability of our questions and data.

How to Assess Reliability

A range of methods of evaluating the reliability of measures have been developed. Unfortunately there is no single method that is suitable for all situations.